

# Timbre Saliency, The Attention-Capturing Quality of Timbre

*Song Hui Chon*



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

August 2013

---

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

© 2013 Song Hui Chon



# Abstract

This dissertation proposes a new concept of timbre saliency as the attention-capturing quality of timbre and investigates its effects on the blending of concurrent notes and on the perceptual segregation of voices in counterpoint music. As this is the first effort to consider attentional factors in timbre perception research, a number of listening experiments were needed to define the concept and to establish the field.

The first chapter introduces timbre saliency and connects it to other related research fields. A survey of visual saliency research was particularly helpful in developing experimental methodologies, because research in auditory saliency is still in its infancy.

The second chapter describes two experiments to measure timbre saliency among the set of chosen timbres and discusses the importance of choosing the right experimental paradigm and how it can affect the outcome. This is especially relevant because *saliency* is a function of *context*, which is determined by the experimental setup. The measured saliency seems to be related to the fine structure in harmonic spectrum. These saliency relations became the basis for the following experiments.

The next two chapters examine the effect of timbre saliency in a more realistic setting. First, the perception of blending is analyzed in concurrent unison dyads in terms of timbre saliency. The average blend ratings showed a negative correlation with timbre saliency, confirming the hypothesis that a highly salient timbre would not blend well with

others, although the effect was not as strong as some other factors. Then the scope expands to non-unison intervals in multiple voices in the fourth chapter, where the effect of timbre saliency on the voice recognition in short counterpoint excerpts is studied. The hypothesized systematic effect of timbre saliency was found in neither two- nor three-voice excerpts, although having a distinctive timbre on each voice helped the recognition of the middle voice in three-voice excerpts, which is the most difficult to listen to. The findings from the experiments, as well as a discussion on the general context effects, are summarized in the last chapter.

This research extends traditional timbre research by considering the role of attention in sound and music perception. It provides a bridge between the perception of multi-voice music and auditory scene analysis, and hence has the potential to contribute to research in auditory perception as well as in music perception and cognition.

# Résumé

Cette thèse propose un nouveau concept de saillance de timbre, conçu comme la qualité du timbre qui attire l'attention. Elle étudie les effets de saillance sur le mélange de notes simultanées et sur la séparation perceptive des voix dans la musique contrapuntique. Puisque c'est la première fois que les facteurs attentionnels sont pris en considération dans la recherche sur la perception du timbre, des expériences d'écoute ont été nécessaires pour définir le concept et établir le domaine de recherche autour de lui.

Le premier chapitre introduit la saillance du timbre et la lie à d'autres domaines de recherche connexes. Une revue de la recherche sur la saillance visuelle aide particulièrement à développer les méthodes expérimentales car la recherche sur la saillance auditive est toujours dans son enfance.

Le deuxième chapitre décrit deux expériences qui mesurent la saillance du timbre sur un ensemble de timbres sélectionnés et discute de l'importance de choisir le bon paradigme expérimental et de comment celui-ci pourra affecter le résultat. Cette approche est particulièrement pertinente puisque la *saillance* est fonction du *contexte*, qui est déterminé à son tour par la manipulation expérimentale. La saillance mesurée semble liée à la structure fine du spectre harmonique. Les relations de saillance établies deviennent la base pour les expériences ultérieures.

Le deux chapitres suivants examinent l'effet de la saillance du timbre dans un contexte

plus naturel. D’abord, la perception du mélange est analysée sur des dyades jouées à l’unisson en termes de la saillance des timbres. Les évaluations de mélange montrent une corrélation négative avec la saillance, confirmant l’hypothèse selon laquelle un timbre hautement saillant ne se mélangerait pas bien avec d’autres timbres, bien que cet effet n’était pas si important que d’autres facteurs. Ensuite, dans le quatrième chapitre les intervalles non unisson dans des voix multiples sont étudiés pour évaluer l’effet de la saillance du timbre sur la reconnaissance de voix dans de courts extraits de contrepoint. L’effet hypothétique de la saillance n’est retrouvé ni dans les extraits à deux voix ni dans ceux à trois voix. Ceci étant dit, la présence d’un timbre distinctif sur chaque voix aide la reconnaissance de la voix du milieu dans des extraits à trois voix, cette voix étant la plus difficile à entendre. Le dernier chapitre résume les résultats des expériences et présente une discussion des effets généraux de contexte.

Cette recherche étend la recherche traditionnelle sur le timbre en prenant en considération le rôle de l’attention dans la perception du son et de la musique. Elle fournit un pont entre la perception de la musique à plusieurs voix et l’analyse de scènes auditives et contribuera potentiellement à la recherche sur la perception auditive, ainsi que sur la perception et la cognition musicales.

# Acknowledgments

It has taken a long journey to get here; therefore there are a lot of people for whose support I am deeply grateful.

First of all, I am thankful for my advisor Stephen McAdams. Dr. McAdams is a true scholar with endless zeal for research, as well as support, guidance and discipline for his students. I feel truly blessed to have been one of the advisees under his wing.

There are other professors that I should not fail to mention. Drs. Philippe Depalle and Ichiro Fujinaga helped me refine the idea of timbre saliency when I was first contemplating the topic. The comments by my reviewers of my dissertation, Drs. Al Bregman and Petri Toiviainen, helped make it better and more complete. Dr. Bregman also provided insightful critiques on data analysis, as well as a generous scholarship, both of which were a tremendous help in the course of my graduate years. Dr. Yoshio Takane offered crucial advices on statistical data analysis. Dr. Sean Ferguson co-supervised me on my CIRMMT student award projects to study the effect of timbre saliency on blending and voice perception. Dr. Wieslaw Wosczeke and his student (and my good friend) Doyuen Ko invited me to an interesting research of the perceptual analysis of virtual acoustics, though this topic is not included in this dissertation.

A special thank-you is in order for Tom Beghin, who accepted me as a private student on the fortepiano and clavichord. Thanks to Dr. Beghin, I could enjoy studying rare historic

instruments, which became a haven from research stress. Without these lessons my stay at McGill would not have been as enjoyable.

I am also thankful to the members of the Music Perception & Cognition Lab (MPCL). Bennett Smith deserves the first mention, for providing all his help in implementing computer programs for my experiments. Previous members Dr. Bruno Giordano, now at the University of Glasgow, and Finn Upham, now studying at NYU, helped me greatly with experimental design and behavioural data analysis. David Sears and Meghan Goodchild were my sounding block for music theory questions. Kevin Schwartzbach, my undergrad assistant for the last experiment, was indispensable in helping me think through all complicated details in design. Sven-Amin Lembke provided insights on timbre perception and blending, which is the main focus of his research.

I should also mention CIRMMT for their generous financial support with my student projects for two years. I am also thankful to the CIRMMT staff, Harold Kilianski, Yves Méthot, Julien Boissinot, Jacqui Bednar and Sara Gomez. Hélène Drouin at the Graduate Music Office guided me through not-so-straight administrative requirements throughout my graduate years at McGill.

Many friends, some of whom are already mentioned above, provided support and encouragement during my Ph.D. years here at McGill. I am especially thankful for Dan Steele, Minsoo Ha, Mikyoung Suh, Colleen Nelson, Victoria Miller, Jung Suk Lee, Minkyong Kim, Yoona Jhon, Doyuen Ko, Alix Momperousse, Seok Mo Song and Sandra Duric. I also owe a special thank-you to Anchi Tan Miller who graciously helped with proofreading and editing many manuscripts including parts of this dissertation.

I am forever indebted to Dr. Ik Geun Hwang of Chonbuk National University in Korea. This degree would not have been possible without his thoughtful care of my health for many years. I also thank Dr. ChongHwan Park at Belton Engineering for his generous



financial gifts that helped me throughout these Ph.D. years.

Last but certainly not least, I thank my loving family for their continuous support, both emotionally and financially. Even from afar, they never stopped their encouragement.

Above all, I thank God who gives me strength to do all things. He brought me to McGill, inspired me with the thesis topic, provided means for my projects, and carried me through many difficult times. I dedicate this dissertation to Him and pray that it is a pleasing offering in His sight.

(This page is intentionally left blank.)

## Contribution of Authors

The document is formatted as a monograph dissertation and includes contents from the following conference publications.

- Chapter 2: Chon, S.H., and McAdams, S. (2012). Investigation of Timbre Saliency, the Attention-Capturing Quality of Timbre, *In Proceedings of Acoustics 2012 Hong Kong*.
- Chapter 3: Chon, S.H., and McAdams, S. (2012). Exploring Blending as a Function of Timbre Saliency, *In Proceedings of the 12th International Conference on Music Perception and Cognition, Thessaloniki, Greece*.
- Chapter 4: Chon, S.H., Schwartzbach, K., Smith, B., and McAdams, S. (2013). Effect of Timbre on Voice Recognition in Two-voice Counterpoint Music. *To be presented at the Society for Music Perception and Cognition (SMPC) 2013, Toronto, Canada*.
- Chapter 4: Chon, S.H., Schwartzbach, K., Smith, B., and McAdams, S. (2013). Effect of Timbre on Melody Recognition in Three-voice Counterpoint Music. *Submitted to Stockholm Music Acoustics Conference (SMAC) 2013, Stockholm, Sweden*.

As the person who proposed a new concept, I was responsible for every step involved in designing and carrying out all experiments mentioned in this dissertation, as well as

analyzing collected data and preparing manuscripts for all the publications listed above. My advisor, Stephen McAdams, provided necessary funding, laboratory equipments and space. He also contributed with guidance in experimental design, data analysis and interpretation of the results. I had one research assistant, Kevin Schwartzbach, under my supervision, who worked closely with me on the design and the execution of Experiment 6 in Chapter 4. Bennett Smith, who is the technical manager of the Music Perception and Cognition Lab (MPCL), implemented the graphic user interfaces for all experiments and organized the collected data for analysis.

# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Attention . . . . .	3
1.1.1	Attention in Auditory Perception . . . . .	4
1.2	Review of Saliency Research . . . . .	6
1.2.1	Visual Saliency Studies . . . . .	6
1.2.2	Auditory Saliency Studies . . . . .	11
1.3	Dissertation Structure . . . . .	13
1.3.1	Chapter 2: Definition of Timbre Saliency Space (Experiments I & II)	13
1.3.2	Chapter 3: Perceived Blending as a Function of Timbre Saliency (Experiment III) . . . . .	16
1.3.3	Chapter 4: Effect of Timbre Saliency and Timbre Dissimilarity on Voice Recognition in Counterpoint Music (Experiments IV, V & VI)	17
1.3.4	Conclusion and Future Work (Chapter 5) . . . . .	18
1.4	Academic Relevance . . . . .	18
<b>2</b>	<b>Defining Timbre Saliency</b>	<b>21</b>
2.1	Introduction . . . . .	22
2.2	Experiment I: Measuring Timbre Saliency using an Indirect Method of Tapping	24

---

2.2.1	Methods . . . . .	24
2.2.2	Analysis of Measured Saliency Differences . . . . .	29
2.2.3	Analysis of Behavioural Differences in Tapping Patterns . . . . .	47
2.3	Experiment II: Measuring Timbre Saliency using a Direct Comparison Method	50
2.3.1	Methods . . . . .	52
2.3.2	Analysis . . . . .	55
2.4	Conclusion . . . . .	64
<b>3</b>	<b>Perceived Blend of Unison Dyads</b>	<b>69</b>
3.1	Introduction . . . . .	70
3.2	Experiment III: Blending of Unison Dyads . . . . .	73
3.2.1	Methods . . . . .	73
3.2.2	Results . . . . .	79
3.3	Conclusion . . . . .	90
<b>4</b>	<b>The Role of Timbre Saliency and Timbre Dissimilarity in Voice Recognition in Counterpoint Music</b>	<b>95</b>
4.1	Introduction . . . . .	96
4.2	Experiment IV: Timbre Dissimilarity . . . . .	100
4.2.1	Methods . . . . .	101
4.2.2	Results and Discussion . . . . .	103
4.3	Musical Stimulus Design . . . . .	106
4.3.1	Selection of Musical Excerpts . . . . .	106
4.3.2	Timbre Combinations . . . . .	107
4.3.3	Excerpt Assignment to Timbre Combinations . . . . .	111
4.3.4	Assignment of Same or Modified Comparison Melodies . . . . .	115

---

4.4	Experiment V: Melody Discrimination . . . . .	118
4.4.1	Methods . . . . .	118
4.4.2	Results & Discussion . . . . .	120
4.5	Experiment VI: Voice Recognition in Counterpoint Music . . . . .	122
4.5.1	Methods . . . . .	122
4.5.2	Results . . . . .	125
4.5.3	Discussion . . . . .	138
4.6	General Discussion and Conclusions . . . . .	141
<b>5</b>	<b>Conclusion</b>	<b>149</b>
5.1	Summary . . . . .	149
5.1.1	Proposition of A New Research Topic . . . . .	149
5.1.2	Experiments and Results . . . . .	151
5.2	Discussions . . . . .	157
5.3	Conclusion and Future Work . . . . .	160
<b>A</b>	<b>Stimulus Selection</b>	<b>165</b>
<b>B</b>	<b>Equalization Experiments</b>	<b>169</b>
B.1	Loudness Equalization . . . . .	169
B.1.1	Participants . . . . .	169
B.1.2	Stimuli . . . . .	169
B.1.3	Procedure . . . . .	170
B.1.4	Result . . . . .	171
B.2	Isochronous Rhythm Generation . . . . .	174
B.2.1	Participants . . . . .	174

B.2.2	Stimuli . . . . .	174
B.2.3	Procedure . . . . .	174
B.2.4	Result . . . . .	175
B.3	Melody Loudness Equalization . . . . .	178
B.3.1	Participants . . . . .	179
B.3.2	Stimuli . . . . .	179
B.3.3	Procedure . . . . .	179
B.3.4	Result . . . . .	180
<b>C</b>	<b>Scores of Melodies used for the Voice Recognition Experiment</b>	<b>183</b>
C.1	Two-voice excerpts . . . . .	184
C.2	Three-voice excerpts . . . . .	188
	<b>Bibliography</b>	<b>193</b>



# List of Figures

1.1	An example interleaved melody task (from <a href="#">Bey &amp; McAdams, 2003</a> ) . . . . .	4
1.2	An example of a search display. The target was always oriented 20 degrees clockwise, regardless of the orientation of the distractors. The orientation difference between the target and the distractors was either 20 (A and C) or 90 (B and D) degrees. One distractor was coloured red on half the trials (A and B). All items were coloured black on the remaining half of the trials (C and D). Black items are represented as solid lines; red distractors are represented as segmented lines (from <a href="#">Poiese, Spalek, &amp; Di Lollo, 2008</a> ). . . . .	7
1.3	An example of scenes and associated objects' outlines (from <a href="#">Elazary &amp; Itti, 2008</a> ) . . . . .	9
2.1	Illustration of an isochronous ABAB sequence . . . . .	27
2.2	Distributions of dominance values for all pairs of timbres involving each instrument on the horizontal axis . . . . .	34
2.3	Calculation of proximity values $E_k(i, j)$ from dominance values $D_k(i, j)$ . . . . .	37
2.4	One-dimensional timbre saliency space of 15 timbres . . . . .	41
2.5	Average tap timing according to gender and musicianship. The error bars represent the 95% confidence interval. . . . .	49

2.6	Average number of taps according to gender and musicianship. The error bars represent the 95% confidence interval. . . . .	51
2.7	Screenshot of the graphic user interface for Experiment II . . . . .	54
2.8	A three-dimensional CLASCAL solution . . . . .	60
3.1	Ranges and distributions of saliency measures of the composite sounds . .	75
3.2	Generating composite sounds from the synch process . . . . .	76
3.3	Screenshot of blend rating experiment . . . . .	79
3.4	Distribution of average blend ratings across participants for all pairs of timbres involving each instrument on the horizontal axis. . . . .	81
4.1	Screenshot of the graphic user interface used for the timbre dissimilarity rating experiment . . . . .	102
4.2	Two-dimensional timbre dissimilarity space (CL = Clarinet, EH = English Horn, FH = French Horn, FL = Flute, HA = Harp, HC = Harpsichord, MA = Marimba, OB = Oboe, PF = Piano, TB = Tubular Bells, TN = Trombone, TP = Trumpet, TU = Tuba, VC = Cello, VP = Vibraphone). .	105
4.3	An example of a two-voice excerpt and corresponding comparison melodies	107
4.4	Melody used for loudness equalization across timbres. . . . .	116
4.5	Screenshot of the melody discrimination experiment. . . . .	119
4.6	Screenshot of the voice recognition experiment . . . . .	125
4.7	Results for two-voice excerpts. The error bars show the standard error of the mean. . . . .	126
4.8	Results for three-voice excerpts. The error bars show the standard error of the mean. . . . .	129
4.9	Average performance of two-voice excerpts . . . . .	133

---

4.10 Average performance of three-voice excerpts . . . . .	134
A.1 ADSR estimation . . . . .	166
A.2 A raised cosine decay ramp . . . . .	166
B.1 Screenshot of Pilot 1 . . . . .	170
B.2 Result from Pilot 1 . . . . .	172
B.3 Screenshot of Pilot 2 . . . . .	175
B.4 Median adjustment time as a function of attack time difference . . . . .	178
B.5 Melody used for loudness equalization . . . . .	179
B.6 Result of Pilot 3 with 15 participants . . . . .	181
B.7 Total loudness equalization result of Pilot 3 with 15 participants . . . . .	182

## List of Tables

2.1	Labels, names and durations of 19 stimuli used in Experiment I . . . . .	25
2.2	Mean dominance matrix $\overline{D}$ from 40 dominance matrices, $D_k$ . . . . .	33
2.3	Log likelihood, degrees of freedom, and AIC and BIC values for spatial models	38
2.4	Coordinates for a one-dimensional spatial solution with specificities and two latent classes . . . . .	40
2.5	Distributions of participants in each latent class according to gender, musi- cianship and age . . . . .	42
2.6	Estimated weights in the selected one-dimensional model with specificities for two latent classes . . . . .	42
2.7	Posterior probabilities of class membership . . . . .	44
2.8	Prior distribution of class membership . . . . .	45
2.9	The five timbre descriptors with the highest correlations with the saliency dimension in Figure 2.4. . . . .	45
2.10	Correlation coefficients of popular timbre descriptors with the saliency di- mension. . . . .	46
2.11	Paired-samples t-test result . . . . .	57
2.12	Correlation of $t$ statistics with measures from Experiment I . . . . .	58

2.13	Log likelihood, degrees of freedom, and AIC and BIC values for spatial models, presented in rank order . . . . .	61
2.14	Acoustic correlates of each dimension of the CLASCAL solution . . . . .	61
2.15	Correlation of each dimension of the CLASCAL solution with the saliency dimension from Experiment I . . . . .	62
2.16	Estimated weights in the selected three-dimensional model with specificities for two latent classes from Experiment II . . . . .	63
2.17	Gender, musicianship and age of the five participants in Class 2 . . . . .	63
3.1	One-way ANOVA result of average blend ratings between percussive attack group (HA, HC, MA, PF, TB, VP) and gradual attack group . . . . .	82
3.2	One-way ANOVA result of average blend ratings among three timbre saliency regions . . . . .	83
3.3	Spearman's rank correlation between the 15 saliency values from Experiment I and the 15 IQRs and medians of the average blend ratings . . . . .	84
3.4	Pearson correlations between the average blend and saliency measures of 105 composites . . . . .	85
3.5	Pearson correlations between average blend and timbre descriptors. . . . .	87
4.1	Means and STDs of the ratings in two disjoint sets for ANOVA on SPSS . . . . .	104
4.2	Acoustic correlates of the two-dimensional timbre dissimilarity space. . . . .	104
4.3	Timbre assignments for two-voice excerpts . . . . .	108
4.4	Timbre conditions for three-voice excerpts . . . . .	110
4.5	Timbre assignments for three-voice excerpts . . . . .	110
4.6	Excerpt assignments for two-voice excerpts - Option A . . . . .	112
4.7	Excerpt assignments for two-voice excerpts - Option B . . . . .	112

4.8	Excerpt assignments for three-voice excerpts . . . . .	114
4.9	“Same-fate” cells in two-voice excerpt assignment, example 1 . . . . .	117
4.10	“Same-fate” cells in two-voice excerpt assignment, example 2 . . . . .	117
4.11	Melody discrimination result for melodies from two-voice excerpts, averaged across all participants . . . . .	121
4.12	Melody discrimination result for melodies from three-voice excerpts . . . . .	122
4.13	Summary of Spearman’s rank correlation analysis of each of five voices and the ‘percent correct’ values from the control experiment . . . . .	136
4.14	Summary of paired-sample t-test to study the effect of timbre orders on the average performance in two-voice excerpts . . . . .	137
B.1	Median Loudness Adjustment Values in Decibels . . . . .	173
B.2	Median lag in generating perceptually isochronous sequences of four timbres for training. AF = alto flute, BS = bassoon, CE = celesta, and VN = violin. The timbre specified by the row corresponds to timbre A and that specified by the column to timbre B in ABAB sequences. . . . .	176
B.3	Median lag in generating perceptually isochronous sequences of 15 timbres for testing. CL = clarinet, EH = English horn, FH = French horn, FL = flute, HA = harp, HC = harpsichord, MA = marimba, OB = oboe, PF = piano, TB = tubular bells, TN = trombone, TP = trumpet, TU = tuba, VC = cello, VP = vibraphone. The timbre specified by the row corresponds to timbre A and that specified by the column to timbre B in ABAB sequences.	177
B.4	List of instruments with preliminary loudness equalization and their adjust- ment levels . . . . .	180

# List of Acronyms

2AFC	Two-Alternative Forced-Choice
ADSR	Attack-Decay-Sustain-Release
AF	Alto Flute
ANOVA	Analysis of Variance
ANCOVA	Analysis of Covariance
AP	Absolute Pitch
ASA	Auditory Scene Analysis
BS	Bassoon
CASA	Computational Auditory Scene Analysis
CB	Critical Band
CE	Celesta
CL	Clarinet
DB	Decibels
DF	Degree of Freedom
DV	Dependent Variable
EH	English Horn
ERB	Equivalent Rectangular Bandwidth
ERP	Event-Related brain Potential

FH	French Horn
FFT	Fast Fourier Transform
FL	Flute
HA	Harp
HC	Harpsichord
HL	Hearing Level
IOI	Inter-Onset Interval
IV	Independent Variable
MA	Marimba
MDS	Multidimensional Scaling
MMN	Mismatch Negativity
PF	Piano
STD	Standard Deviation
TB	Tubular Bells
TN	Trombone
TP	Trumpet
TU	Tuba
VC	Cello
VN	Violin
VP	Vibraphone
VSL	Vienna Symphonic Library



# Chapter 1

## Introduction

Timbre is a multidimensional percept. It is what enables us to distinguish a note (for example, C4) played mezzo-forte on a piano from the same note played on a clarinet ([American Standards Association, 1960](#)). The term ‘timbre’ is used to describe the character of sound from a cello in general or the Davidov Stradivarius that Yo-Yo Ma uses. ‘Timbre’ tells us whether the plate I dropped on the floor was broken or not ([McAdams, 1993](#); [Handel, 1995](#)). Sometimes ‘timbre’ gets used in the context of the particular *feel* of the collective sounds from multiple instruments playing together ([Bregman, 1990](#), pp.520–521).

Since the word *timbre* gets used in many different contexts, the research had to start from the very basic case, “is there an underlying scientific explanation about how people perceive the degree of differences among pairs of timbres?” Results from many researchers tend to point to a consensus that (log) attack time and spectral centroid are two of the most important acoustic features related to the perception of timbre dissimilarity in musical instrument tones ([Grey, 1977](#); [Grey & Gordon, 1978](#); [Krumhansl, 1989](#); [McAdams, Winsberg, Donnadieu, De Soete, & Krimphoff, 1995](#); [Lakatos, 2000](#); [Caclin, McAdams, Smith, & Winsberg, 2005](#)). Train sounds, car sounds and air conditioning systems show

other prominent features, such as noise-to-harmonic ratio, loudness, amplitude modulation rate of the temporal envelope and spectral deviation (Susini et al., 2004; Lemaitre, Susini, Winsberg, McAdams, & Letinturier, 2007).

Saliency is defined as “a striking point or feature; a highlight (Saliency, n.d.).” It gets used more often in its adjective form, *salient*. When something is salient, it usually means that the object is more prominent than its neighbours and therefore draws more attention to itself.

As inferred from the definition, saliency is highly related to attention; something striking inadvertently catches one’s attention even when it is against the listener’s intention. The concept of saliency has received much interest in vision research, although not as much research has been done in audition. This is partly because in audition it is difficult to track which auditory stream or object a participant is paying attention to, whereas it is possible to track the direction of gaze in visual saliency.

This dissertation proposes a new concept of *timbre saliency* as the attention-capturing quality of timbre. As it is the first effort to consider attention in timbre perception, a survey of techniques used in saliency research is necessary to set up the ground. This introduction will therefore review some studies on saliency in vision as well as in audition in Section 1.2.

This chapter is organized as follows. A brief discussion on attention is presented in Section 1.1. Section 1.2 offers a literature review of saliency research in vision (Section 1.2.1) as well as in audition (Section 1.2.2). A hypothesis for a saliency model is also presented in Section 1.2.2. The experiments to study timbre saliency are summarized in Section 1.3 as well as the implication of timbre saliency in a perceptual theory of orchestration, which is followed by academic relevance in Section 1.4.

## 1.1 Attention

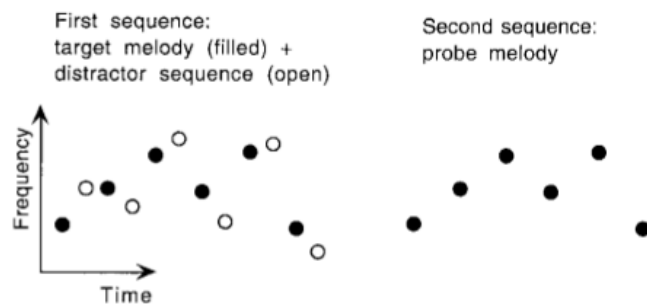
Before starting a discussion on saliency, the topic of attention has to be introduced. Attention is the action of “focusing cognitive processes onto a subset of the information that is currently available” (Luck & Vecera, 2002). There are two main factors affecting how attention flows – stimulus-driven and goal-oriented. Stimulus-driven attention is initiated from salient features in the stimulus itself, hence sometimes called “bottom-up” attention. In this case, the distinct features claim a person’s attention, sometimes even against the individual’s intention. On the other hand, goal-oriented focusing starts from someone’s intentional dispatch of cognitive processes. This is a “top-down process” of purposefully focusing one’s cognitive power on a subset of the given information.

The two attentional directions mentioned above are not independent of each other. In fact, one process affects the other. Consider, for example, that we are looking for a coin that was just dropped on the street, and an ambulance happens to pass us by with a loud siren. It would momentarily catch our attention from focusing on the intended goal of looking for the coin. This is an example of a stimulus-driven process affecting a goal-driven attentional process. On the other hand, there are many studies of change blindness (Levin & Simons, 1997; Simons & Levin, 1998; Levin, Momen, Drivdahl, & Simons, 2000) or change deafness (Vitevitch, 2003; Eramudugolla, Irvine, McAnally, Martin, & Mattingley, 2005), which suggest that people may fail to detect would-be salient changes while focusing on a task. This is an illustration of a top-down process affecting (or rather blocking) the recognition of salient stimulus-driven factors. I suspect that there is a threshold for bottom-up saliency detection (i.e., from the features of stimuli) and that the schema-driven attentional process changes the threshold. When someone’s attention is intentionally focused on one stimulus, the threshold for that specific stimulus would be lower for an easier detection of even smaller

changes. On the other hand, this same process would raise the thresholds of saliency for the detection of other stimuli, making the person more insensitive to changes that could have been salient enough under normal conditions.

### 1.1.1 Attention in Auditory Perception

[Eramudugolla et al. \(2005\)](#) studied the effect of attention on the perception of complex auditory scenes using a collage of various natural sounds. Each object in a scene had a distinctive timbre and spatial location. They showed that with directed attention (i.e., instructing the participants to which object to pay attention) the participants' change detection remained almost perfect even for a complex auditory scene with eight objects. On the other hand, the performance in the undirected-attention case suffered significantly in scenes with more than four objects, which is similar to what [Huron \(1989b\)](#) observed in his study of voice denumerability in monotimbral polyphonic music. While Eramudugolla and his colleagues considered various timbres in the study, they did not analyze the possible effect of timbre in the performance of change detection, so it is not clear if there was any effect of timbre on change detectability.



**Figure 1.1:** An example interleaved melody task (from [Bey & McAdams, 2003](#))

Although attention is one of the major factors determining the perception of auditory scenes, it is not a necessary condition. [Bey \(1999\)](#) showed this by studying the effect

of timbre on stream segregation using an interleaved melody task (shown in Figure 1.1). Participants listened to interleaved melodies of target and distractor sequences and were asked whether the following probe melody (the second set of black dots) was the same as the presented target (the first set of black dots in Figure 1.1). Four synthesized instrument timbres were used (bassoon, guitar, trombone and vibraphone sounds) that were chosen from a previous timbre similarity study by [McAdams et al. \(1995\)](#). The melody recognition in general was better as the timbral distance between the interleaved melodies increased. Although there was no significant main effect of the target timbre found, an asymmetry was observed for trombone-vibraphone and guitar-vibraphone combinations. The correct response rates were lower when the distractor was the vibraphone timbre and the target was either trombone or guitar. Other timbre pairs showed fairly symmetric response rates. Bey could not explain the cause of this asymmetry and hypothesized that there might be an order effect, based on an observation by [Dowling, Lung, and Herrbold \(1987\)](#) that it was easier to recognize a target melody interleaved with distractor notes in the same pitch range when a melody was played by odd notes of the composite sequences.

The asymmetry in Bey’s data could have been caused by the intrinsic saliency of the vibraphone timbre affecting the recognition of target melodies in trombone or guitar timbre even at the highest allocation of schema-driven attention (i.e., when the participants intentionally focused their attention on the target melody stream). The (top-down) schema-driven process is focused on forming a stream of the melody sequence, and this in turn minimizes the effect of the distractor sequence by attenuating the unattended stream as [Botte, Drake, Brochard, and McAdams \(1997\)](#) suggested. But somehow, the intrinsic saliency of the vibraphone timbre grabbed the listeners’ attention away from the other timbre, thus causing an attenuation of the now unattended target timbre.

I propose to call this unique attention-capturing quality of timbre that led to the asym-

metry described above *timbre saliency*. My conjecture is that each instrument timbre has a different degree of saliency so that some timbres tend to grab listeners' involuntary attention more easily. I have not found any previous research to prove or disprove this hypothesis yet, because auditory saliency in general has not been studied very much and timbre saliency has not been studied at all. A survey of visual and auditory saliency studies is presented in Section 1.2 to review the techniques used and results reported.

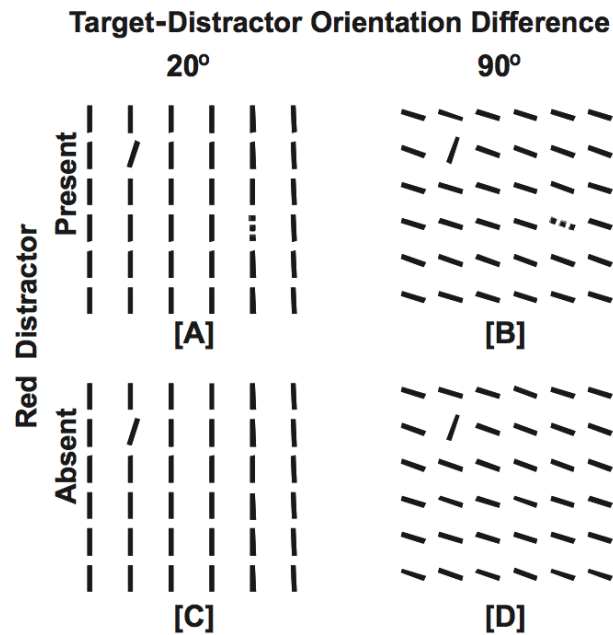
## 1.2 Review of Saliency Research

### 1.2.1 Visual Saliency Studies

Saliency is the quality by which something is prominent, striking and attention-grabbing. We perceive something as salient when the object is different and stands out from its surroundings. As there has been more research on visual saliency than on auditory saliency, this section reviews selected studies of perceptual saliency in vision and the techniques used to study it.

In vision, an object is defined by features such as form, colour and texture. The saliency of an object, therefore, must also be a function of these features, among which form and colour are the simplest to study. This is why they are often used as independent variables in many visual saliency studies. Theeuwes (1992) studied saliency of form and colour by asking participants to search for a target, which was always a green circle, with or without form distractors (i.e., green squares surrounding the green circle) or colour distractors (i.e., red circles or squares). Based on the response time data, he found that colour was more salient than form in his experimental conditions.

Poiese et al. (2008) studied the attentional effect of visual distractors using different colours and orientations. An example display is shown in Figure 1.2. There were always



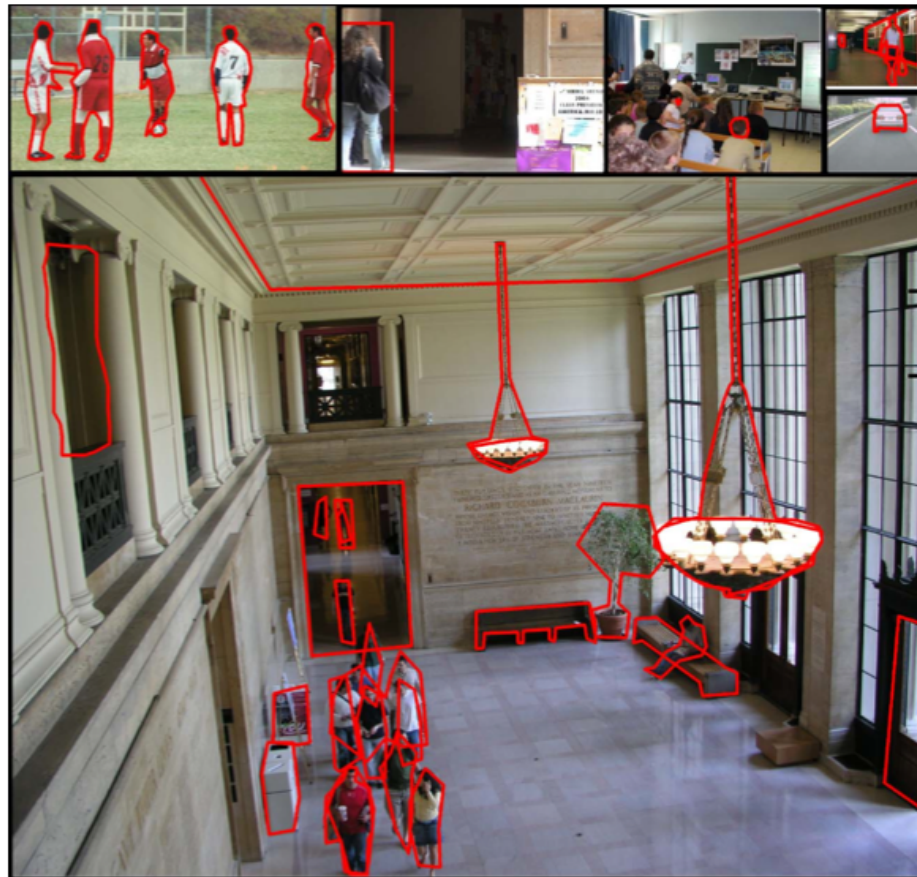
**Figure 1.2:** An example of a search display. The target was always oriented 20 degrees clockwise, regardless of the orientation of the distractors. The orientation difference between the target and the distractors was either 20 (A and C) or 90 (B and D) degrees. One distractor was coloured red on half the trials (A and B). All items were coloured black on the remaining half of the trials (C and D). Black items are represented as solid lines; red distractors are represented as segmented lines (from [Poiese et al., 2008](#)).

two salient stimuli amongst others, the target and the salient distractor. The target was different in its orientation from the other stimuli (by either 20 or 90 degrees). The salient distractor was of the same orientation with the other distractors but always in red, while all the other stimuli (including the target) were in black, making the red distractor highly salient. The task was to indicate, using specific keyboard keys, whether or not the target was present in the given display, as fast and accurately as possible. The results showed that “the salient-distractor effect is influenced by target-distractor similarity”: when the target and the distractors were more different (i.e., by 90 degrees), this difference was apparently more prominent than the target-distractor colour difference. This study is different from [Theeuwes \(1992\)](#) in that they explored different levels of saliency in form (i.e., orientation) and how it might interact with the fixed saliency of colour difference.

A salient object tends to draw attention. So does an interesting object. Then, might interestingness and saliency be correlated? [Elazary and Itti \(2008\)](#) studied interestingness of stimuli in visual scenes using the images annotated by humans. Annotation refers to the manual tracing of the outlines of the objects in an image, in this case using a computer-based tracing tool. The authors defined interesting objects as those annotated by the participants. The images were collected and annotated by anonymous participants via the World Wide Web, with no restrictions in terms of task or time. Figure [1.3](#) shows an example from the paper, where participants’ annotations are shown in red lines around scene objects. They compared the annotated objects to the result from a computational model of a visual saliency map by [Itti, Koch, and Niebur \(1998\)](#) and concluded that the saliency map provides a good approximation of what humans find interesting in visual scenes.

How long does the effect of saliency last? Does it last a short time or the entire duration of a stimulus? These are the questions [Donk and Zoest \(2008\)](#) asked in their research.





**Figure 1.3:** An example of scenes and associated objects' outlines (from [Elazary & Itti, 2008](#))

Their stimuli were similar to those used by [Poiese et al. \(2008\)](#), except the distractor was in the same colour as the target, but in a different orientation from other stimuli as well as the target. Participants were required to make eye movements as quickly and accurately as possible to the target, and the eye movements were recorded. Data showed that the probability of eye movement towards the correct target decreased as the latency of eye movement increased. This relationship between latency and probability means that it took longer to figure out which was the target when there was less difference in orientation between the target and the distractor. The saliency of the target (controlled with the angle difference from the other stimuli's orientation) had an effect on performance up to about 340 ms; after that time, there was no performance difference observed. The number of 340 ms was noticed from the eye tracking data but its possible significance was not discussed in the paper. The authors concluded, "this result implies that after some time had elapsed, relative salience differences between conditions did not affect performance." ([Donk & Zoest, 2008](#), p. 735)

[Schubö \(2009\)](#) studied the effect of irrelevant salient distractors on capturing attention by varying both form and colour of the target and the distractor. She used event-related brain potentials (ERPs) and response latencies to determine the effect of the salient distractor. Her result was supportive of [Theeuwes \(1992\)](#)'s observation that colour was more salient than form. Response times were slower when the task was to recognize the direction of the line element in the form target while a salient colour distractor was present than when recognizing the direction of the line element in the colour target while ignoring a salient form distractor. A slow response means more mental processing was required to ignore the salient distractor. Therefore, the more salient the distractor, the slower the response. ERP results revealed that the processing patterns for target-only cases were different from those for target-and-distractor cases regardless of the saliency and type (colour or form) of

the distractor. The author concluded that “bottom-up signals may be modulated by task constraints already at an early processing level” (Schubö, 2009, p. 242). This might be an illustration of schema-driven attention subduing stimulus-driven attention.

In summary, in the studies mentioned above, saliency was studied in terms of its noticeability using some sort of quantification of the similarity of the target and the distractor as well as the reaction time, which also reflects the similarity of the two. Saliency and ‘interestingness’ are correlated, although the effect of saliency may be rather short, even with the help of attention (Donk & Zoest, 2008), which adds to the complexity of the problem of studying saliency.

### 1.2.2 Auditory Saliency Studies

In his seminal book *Auditory scene analysis*, Bregman (1990) defines two types of stream segregation: primitive-integration and schema-driven selection. Primitive integration is based on the features of the stimuli, whereas a schema-driven segregation is more dependent on the listener’s focus of attention.

Let us imagine a live concert scene with Brahms’s second piano concerto, and the third movement has just started. I am enjoying the cello solo opening the movement, which somehow manages to stand out from the rest of violas, cellos and basses accompanying it. Then the melody goes over to the violins, and I have a hard time tuning into the cello that played the solo any more. After a little while the orchestra hushes and the piano starts the solo.

In the scenario above, the solo cello, the violins and the piano are examples of auditory objects causing primitive stream segregation. Auditory saliency must be a function of multiple variables such as loudness and pitch, as we tend to have a hard time ignoring a loud or high-pitched sound. The louder a sound is, the more salient it is; high-pitched

sounds are usually more salient, which may be why melodies are often carried by the top voice.

While loudness and pitch are two important factors in auditory saliency, there must also be a timbral factor. I further hypothesize that there is something inherent in each instrument timbre that is related to its ability to grab the listener’s attention. The asymmetry in data observed by [Bey \(1999\)](#) in Section 1.1.1 could be a good example. Timbre is also used in auditory scene analysis research to obtain a ‘saliency map’ that visualizes the time and frequency location of a salient object in an auditory scene ([Kayser, Petkov, Lippert, & Logothetis, 2005](#); [Kalinli & Narayanan, 2009](#)).

While our attention often gets directed to the most salient object or stream, the brain senses much more than that. [Paavilainen, Arajarvi, and Takegata \(2007\)](#) showed, using mismatch negativity (MMN), that the brain is aware of nonsalient features of auditory stimuli. MMN is observed in ERP patterns when a deviant stimulus is detected within otherwise consistent stimulus patterns. MMN is useful for studying preattentive detection because it is obtained even when the deviance goes unnoticed (and is therefore unattended) by participants.

Recently, the auditory attention team at the Telluride Neuromorphic Workshop reported that they could use EEG signals to correctly predict which auditory object a listener was paying attention to ([Slaney, Lalor, et al., 2012](#)). This is exciting news, although we will have to interpret it with a caution until another team verifies the result.

Reflecting the result by [Paavilainen et al. \(2007\)](#), I hypothesize that there are a few different levels of timbre saliency. Some timbres such as violas are usually very difficult to focus on, even with the knowledge and an intentional dispatch of top-down attention. They may form a “sub-average” saliency group. They rarely cause a primitive integration based on timbre. There also could be some in the “average” saliency group, where each timbre

is usually not very salient, but can become salient with the help of top-down attention. The timbres in this group may be salient enough to cause primitive integration, but they would become more salient with schema-driven segregation. Then there must be a group of highly salient timbres, which tend to catch listeners' attention by itself, like the vibraphone timbre that [Bey \(1999\)](#) observed. They would be so salient as to cause primitive integration without any schema-driven segregation.

Saliency is a degree of uniqueness of the object with respect to its surroundings. Visual saliency was defined by the saliencies of form and colour. I suppose timbre saliency could be defined in a similar manner – by some of the major acoustic correlates of timbre, such as attack time, spectral centroid, spectral flux and spectral irregularity. My hypothesis, therefore, is that it would be possible to define timbre saliency in terms of the known acoustic correlates. The experiments to study timbre saliency in this dissertation are summarized in [Section 1.3](#).

## 1.3 Dissertation Structure

### 1.3.1 Chapter 2: Definition of Timbre Saliency Space (Experiments I & II)

With the hypothesis of timbre saliency being a function of acoustic correlates of timbre, the first task is to define the timbre saliency space, where the distance between two timbres would correspond to the timbre saliency difference. Two experiments were carried out for this purpose.

The first experiment employed an indirect measurement protocol through a tapping experiment using ABAB perceptually isochronous sequences, presented in [Section 2.2](#). A direct comparison method was used in Experiment II ([Section 2.3](#)) to remove the unforeseen confound with rhythmic saliency in the task of tapping in the first experiment. In both

experiments, timbres A and B were perceptually equalized in terms of pitch, loudness and duration to minimize their impact on the perceived timbre.

A set of instrument sounds was selected from those that have been considered in many timbre perception studies ([Grey, 1977](#); [Grey & Gordon, 1978](#); [Krumhansl, 1989](#); [McAdams et al., 1995](#); [Lakatos, 2000](#); [Caclin et al., 2005](#)). The timbres had to be short enough to form a comfortable tapping rhythm, so an artificial offset was imposed on the recorded samples of natural instrument sounds to make them short. Attack patterns, which are known to be related to timbre perception, were left intact. The recorded instrument sounds were taken from the samples in the Vienna Symphonic Library ([Vienna Symphonic Library GmbH, 2011](#)). The stimulus selection, equalization and creation of perceptually isochronous ABAB sequences are presented in detail in Appendices [A](#), [B.1](#), and [B.2](#), respectively.

## Experiment I: Timbre Saliency using Indirect Measurements

In the first experiment, participants were asked to listen to each ABAB sequence and tap to the timbre that sounded like a strong beat to them. If timbre A happens to be much more salient than the other, then participants would tap to A more frequently. If A and B are about the same saliency, then they would be tapped to with an almost equal probability.

From this tapping ‘choice’ data, we acquired ‘dominance’ measures, which reflect how often one timbre would be chosen as strong beat across all the other timbres with which they could be combined in a sequence. ‘Equal dominance’ can be interpreted as ‘high similarity’ and ‘extreme (in)dominance’ as ‘high dissimilarity’ in terms of dominance patterns. Using these rules, the ‘dominance’ data could be transformed into ‘dissimilarity’ data, to which then a multi-dimensional scaling (MDS) algorithm such as CLASCAL ([Winsberg & De Soete, 1993](#)) could be applied to obtain the timbre saliency space. The best CLASCAL solution turned out to be in one dimension with two latent classes and specificities. The

acoustic correlate of the saliency dimension was obtained using the acoustic features calculated from the Timbre Toolbox (Peeters et al., 2011). This timbre saliency dimension became the basis for the next experiments.

### Experiment II: Timbre Saliency using Direct Comparison

One unexpected confound with rhythmic saliency was found in terms of the tapping task for the first experiment. The task was to tap to a timbre acting like a strong beat in a perceptually isochronous ABAB sequence, which is essentially in a duple rhythm. Duple rhythms are popular in music, especially in rock music, where a strong beat would in general be associated with a bass drum (with more low-frequency energy) whereas a weak beat with a snare (with more high-frequency energy). As this custom has been widely practiced for the entire history of rock music, the general public learned through repeated exposure to have an implicit expectation of “lower-sounding” beat to be a strong beat. This was revealed during the post-experiment interviews with participants and verified in the data analysis.

To minimize this unwanted relationship between a strong beat and low frequency energy, a second experiment was performed using a direct comparison Two-Alternative Forced Choice (2AFC) protocol in Section 2.3. In this experiment, participants listened to timbre A presented three times, followed by timbre B presented three times, and determined which one would “stand out more or grab their attention more.” It was not possible to repeat the stimuli more times, to go back to a previous trial or to respond that the two timbres were of equal saliency.

Participants’ choice data turned out to have a significant order effect, which did not exist in data from the first experiment. Since the purpose of the second experiment was to compare the result with that from the first experiment, the same process of data transfor-

mation was performed (from choice to dominance to dissimilarity), ignoring the order effect. The CLASCAL result on this new data turned out to be a three-dimensional solution with two latent classes and specificities. Only two dimensions showed mild to moderate correlations with the saliency dimension from Experiment I. It might suggest that the contexts used for the two experiments may not be exactly compatible. The acoustic correlates of the three dimensions also seemed to have problems in interpretation. These complications may have originated from the strong order effect that we ignored. Hence, the results from this experiment were not used for any of the following experiments. But it shows the importance of the experimental design, as the two approaches to the same problem resulted in what seems to be inconsistent results.

### 1.3.2 Chapter 3: Perceived Blending as a Function of Timbre Saliency (Experiment III)

The definition of saliency requires the object to stand out with respect to its neighbours, implying little blending between the object and its surroundings. With the timbre saliency measured in Experiment I (Section 2.2), we studied the degree of perceived blending of two concurrent unison sounds as a function of timbre saliency. Following the definition of saliency, we hypothesized that a sound with a high timbre saliency level would not blend well with other sounds. On the other side of the spectrum, a not-so-salient sound would be hypothesized to blend better with other sounds.

Stimuli for this experiment were the composites, or sums, of every pair of sounds from the set of stimuli used in Experiment I. The analysis of acoustic characteristics of the individual as well as the composite sounds were carried out using the Timbre Toolbox (Peeters et al., 2011). Data from a listening experiment were analyzed in terms of the acoustic parameters as well as the positions in the timbre saliency space to find the relationship



between the perceived degree of blending and these acoustic features. Our hypothesis of blend being negatively correlated with timbre saliency was verified. Previous reports in the literature that a sound with a longer attack time or a lower spectral centroid showed better blending on average were also confirmed with our data (Sandell, 1995; Tardieu & McAdams, 2012).

### 1.3.3 Chapter 4: Effect of Timbre Saliency and Timbre Dissimilarity on Voice Recognition in Counterpoint Music (Experiments IV, V & VI)

The last experiment expanded the scope from isolated notes to a more realistic case of music with multiple voices. As Huron (1989b) observed, we know that the inner voices are more difficult to follow than outer voices in multi-voice music. Given that, is there any way to enhance the recognizability of a voice by using a highly salient timbre on it?

Two- and three-voice counterpoint excerpts were chosen from the *Trio Sonatas for Organ*, BWV 525 – 530 by Bach (1730). These excerpts were instrumented with a set of timbres that were determined by considering their positions in timbre saliency and timbre dissimilarity spaces. There was also a mono-timbre version of each excerpt to contrast the participants' performance between mono-timbre and multi-timbre stimuli. Stimuli were generated from the Logic program (Apple Computer, 2012) using the sound sample database from the Vienna Symphonic Library (Vienna Symphonic Library GmbH, 2011).

An embedded melody recognition experiment was carried out using the same/different paradigm as in Bey and McAdams (2003). Participants listened to a multi-part excerpt followed by a monophonic melody and had to choose whether the monophonic melody was the same as or different from one of the embedded melodies in the excerpt. The obtained data were analyzed in terms of timbre conditions (i.e., timbre saliency condition and timbre dissimilarity condition), as well as the embedded melody position.

### 1.3.4 Conclusion and Future Work (Chapter 5)

This chapter summarizes all the experiments presented in this dissertation and recounts the lessons learned. It also presents ideas for future experiments and possible implications for other fields of research.

## 1.4 Academic Relevance

Even with over a century of history since [Helmholtz \(1877\)](#), research in timbre perception still has room for more developments. The first stage of timbre research was to scientifically understand the relationship of dissimilarity between two timbres ([Grey, 1977](#); [Grey & Gordon, 1978](#); [Krumhansl, 1989](#); [McAdams et al., 1995](#); [Lakatos, 2000](#); [Caclin et al., 2005](#)). Next was to understand the blending of two sounds with respect to dissimilarity ([Kendall & Carterette, 1991](#); [Iverson & Krumhansl, 1993](#); [Sandell, 1995](#); [Tardieu & McAdams, 2012](#)). As a new line of timbre research, the topic of timbre saliency is proposed in this dissertation.

Timbre is one of the auditory features affecting stream segregation in auditory scene analysis ([Bregman, 1990](#); [Kayser et al., 2005](#); [Kalinli & Narayanan, 2009](#)). This dissertation, however, is the first to consider the attention-capturing ability of timbre rather than timbre's effect on the organization of an auditory scene. The hypothesis is that each timbre has intrinsic timbre saliency, which may be related to traditional roles of some instruments as a melody carrier and others in a more supporting position.

Listening experiments were designed in order to define timbre saliency and to study its effect in blending and voice recognition. As saliency must be a function of context, which affects the extrinsic saliency (the degree and nature of what is being measured), there must be more than one way to measure timbre saliency. A novel methodology using tapping was employed to define timbre saliency, the result of which was not affected by the order

of presentation of sound stimuli that was observed in early dissimilarity experiments using the direct comparison method (Grey, 1977; Grey & Gordon, 1978).

This dissertation is the first step towards understanding the attention-related character of timbre that may shed light on some much-used examples in orchestration treatises. When developed sufficiently, timbre saliency may provide a novel tool for applications such as computer-aided orchestration programs or design of effective sound alarms.

By considering the attention-capturing quality of timbre, the problem of how listeners perceive different voices in music becomes a special case of the stream organization problem in auditory scene analysis. As musical scenes have a harmonic coherence that does not exist in general auditory scenes, this study of timbre saliency may bring about an interesting finding that could also be useful in auditory scene analysis research.

(This page is intentionally left blank.)

## Chapter 2

# Defining Timbre Saliency

### Abstract

Timbre saliency is a new concept that we propose here to refer to the attention-capturing quality of timbre. For example, [Bey \(1999\)](#) observed in her doctoral dissertation that certain timbres tended to catch the participants' involuntary attention in spite of their intention. This illustrates that there must be something intrinsic about an instrument timbre that reflects its level of saliency. If each timbre indeed has a different level of saliency, then when two timbres with equal pitch, loudness and duration are presented together, the more salient one will draw more attention to itself. Saliency is a difficult subject to deal with, and the research on the effect of saliency in music has been almost nonexistent.

This Chapter presents the first step towards understanding timbre saliency, which is to define the timbre saliency space, in which the distance between a pair of timbres corresponds to the difference in timbre saliency. We present two experiments to measure timbre saliency and discuss different data obtained as well as the importance of choosing the right experimental paradigm.

## 2.1 Introduction

Timbre is a multi-faceted percept, which enables us to distinguish two sounds from different instruments with the same pitch, loudness and duration ([American Standards Association, 1960](#); [Plomp, 1970](#)). Due to its multidimensional nature, timbre has not received as much interest from researchers as pitch or loudness until recently, and a great deal of initial effort was focused on finding acoustic features that best explain how people perceive timbral differences ([Grey, 1977](#); [Grey & Gordon, 1978](#); [Krumhansl, 1989](#); [Iverson & Krumhansl, 1993](#); [McAdams et al., 1995](#); [Lakatos, 2000](#); [Caclin et al., 2005](#)). This research led us to understand the perception of timbre dissimilarities, often presented in two- or three-dimensional *timbre dissimilarity spaces*, where the distance between a pair of timbres is proportional to the perceived degree of dissimilarity between them. In many of these studies, *(log) attack time* and *spectral centroid* turned out to be two of the most successful acoustic correlates that explain timbre dissimilarity spaces.

While these timbre dissimilarity studies provide scientific explanations of perceived dissimilarities of isolated notes, they do not explain interactions that occur when two or more timbres are combined in a more realistic setting. In music, notes are combined concurrently (at the same time on various pitches) and/or sequentially (at different times), and somehow people can focus on one particular stream in the mixture of streams. Can we then apply the principles from *auditory scene analysis* to understand how people perceive the general musical scene, such as picking one particular stream of sounds to focus on, although sometimes that focus is lost by a highly salient event?

An average person deals with ever-changing visual and auditory scenes everyday without any problem. Most people do not even think about *how* they (or rather, their brains) organize the vast amount of information coming in from various senses and figure out what

is more important than the others. As discussed in Chapter 1, the study of *scene analysis* has advanced a lot more in vision studies than auditory studies, mainly because the analysis is easier in vision.

In audition, the progress has been much slower. There are currently only two models of auditory saliency (Kayser et al., 2005; Kalinli & Narayanan, 2009). The algorithm by Kayser et al. (2005) is fairly similar to visual saliency models, which Kalinli and Narayanan (2009) extended by explicitly adding pitch considerations. Recently, Slaney, Agus, Liu, Kaya, and Elhilali (2012) simulated a simplified *cocktail party effect* and compared five *machine listening* strategies, although none were comparable to human performance. Slaney, Lalor, et al. (2012) reported that EEG signals could predict which auditory object a listener was paying attention to, although they did not specify if it was top-down or bottom-up or a mixture of the two that was reflected in the EEG result.

While timbre has been considered one of the elements that affect auditory grouping (Bregman, 1990), there has not been any research on timbre’s attention-capturing ability. In this Chapter we propose the concept of *timbre saliency*, the attention-capturing quality of timbre. Our hypothesis is that each instrument timbre has a unique level of saliency and that it might provide a scientific explanation of some well-known combination examples in orchestration treatises. It may also explain why certain instruments have been favoured as a melody carrier, whereas some others have been in a more supporting role.

Two experiments were conducted to define timbre saliency. One used an indirect measurement employing a tapping technique, and the other used the traditional direct comparison applying a two-alternative forced-choice (2AFC) method. As we will see in later Sections, the result of the two experiments are quite different. Moreover, data from one experiment shows a significant order effect, whereas the other does not. Pros and cons of the two experiments will be also discussed.

## 2.2 Experiment I: Measuring Timbre Saliency using an Indirect Method of Tapping

### 2.2.1 Methods

#### Participants

Participants ( $N = 41$ ) were recruited from a classified advertisement on the McGill University website. Their hearing was tested before the experiment ([International Organization for Standardization, Geneva, 2004](#); [Martin & Champlin, 2000](#)), and one participant was rejected because her hearing did not meet the requirement of being able to hear within 20 dB HL at frequencies 250, 500, 1000, 2000, 4000 and 8000 Hz. There were 20 males and 20 females, and each group contained an equal number of musicians and nonmusicians based on self reports. Musicians are defined as those with four or more years of training (instruments, voice, composition, recording, etc.) during which time they spent at least five hours per week on their training (lessons, practices, etc.). Their ages ranged from 19 to 40, with a median of 22 years. All participants received monetary compensation.

#### Stimuli

Nineteen orchestral instrument sounds were chosen from the Vienna Symphonic Library (VSL) ([Vienna Symphonic Library GmbH, 2011](#)) listed in Table 2.1. Specific sound files were selected considering their relative positions in a three-dimensional timbre space, the dimensions of which are spectral centroid, attack time and spectral flux, based on the timbre dissimilarity space by [McAdams et al. \(1995\)](#). Spectral centroid, attack time and spectral flux were among the 87 acoustic features in four types (temporal energy envelopes, frequency information based on short-term Fourier transform, frequency information based



**Table 2.1:** Labels, names and durations of 19 stimuli used in Experiment I

Set	Abbrev.	Instrument	Attack Dur. (ms)	Effective Dur. (ms)	Physical Dur. (ms)
Test Set	CL	Clarinet	92.9	200.5	333.0
	EH	English Horn	62.2	200.5	298.0
	FH	French Horn	55.3	200.4	307.5
	FL	Flute	54.9	200.3	290.0
	HA	Harp	40.7	200.5	271.0
	HC	Harpsichord	30.4	200.2	259.0
	MA	Marimba	45.1	200.0	289.0
	OB	Oboe	70.1	200.4	302.7
	PF	Piano	33.2	200.2	264.6
	TB	Tubular Bells	28.3	200.4	257.0
	TN	Trombone	57.8	200.3	301.0
	TP	Trumpet	59.9	200.3	303.0
	TU	Tuba	41.6	200.5	279.3
	VC	Violoncello	75.1	200.5	334.3
	VP	Vibraphone	28.5	200.5	260.3
Train Set	AF	Alto Flute	98.2	200.3	341.0
	BS	Bassoon	86.6	200.5	323.0
	CE	Celesta	31.6	200.3	263.0
	VN	Violin	95.3	200.4	360.3

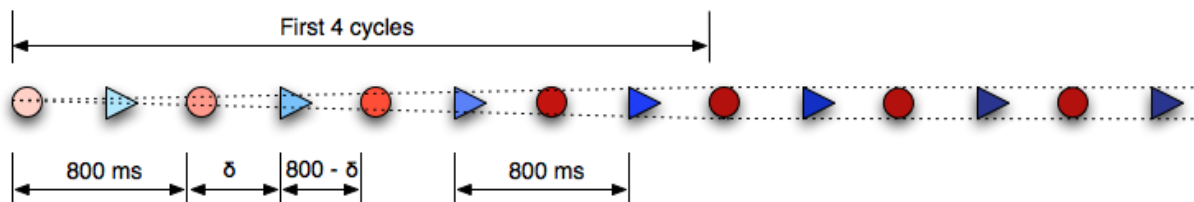
on an auditory filter model, and sinusoidal harmonic distributions) calculated by the Timbre Toolbox (Peeters et al., 2011) on the sound files. Default VSL files are in stereo format, but we decided to use only the left-channel information and send them to both ears of headphones in order to minimize any possible unintended stereo effects.

These instrument sounds all had a fundamental frequency of C4 (261.6 Hz). Small pitch shifts were necessary on some of the sound files using AudioSculpt v.3.0.9 (IRCAM, n.d.) so that their median fundamental frequencies would be 261.6 Hz. These micro pitch shifts did not create any audible artifacts. The effective duration, one of the acoustic features calculated from the Timbre Toolbox, of all sounds was matched to about 200 ms by imposing a 50-ms raised cosine decay ramp somewhere around 250 - 330 ms from the beginning of a sound file. The effective duration is a measure corresponding to the perceived duration of a signal, which is defined to be the time during which the energy envelope of the signal is above a given threshold level (40% of maximum). The beginning of each sound file was kept intact because the attack time has been reported to be one of the main acoustic factors of timbre perception (Berger, 1964; Saldanha & Corso, 1964; McAdams et al., 1995; Lakatos, 2000; Caclin et al., 2005). The stimulus selection process is explained in detail in Appendix A. The sounds also had to be equalized in terms of loudness because saliency would change with loudness. Loudness equalization was carried out in a pilot study with 17 participants. The results are presented in detail in Appendix B.1.

After equalizing pitch, loudness and duration of isolated instrument sounds, the next step was to generate perceptually isochronous ABAB sequences. This was necessary because each sound had a different perceptual starting point depending on the attack pattern, which would result in a perceptually non-isochronous long-short sequence if two segments were combined into one sequence with physically equal inter-onset-intervals (IOIs). Figure 2.1 shows an example. The circles represent one timbre and the triangles another. These

isochronous ABAB sequences were generated by interleaving two uni-timbre isochronous sequences (AAAA and BBBB) in another pilot study (see Appendix B.2 for details). The uni-timbre sequences all have 800 ms IOIs, which corresponds to the time differences between adjacent circles or adjacent triangles, as specified in Figure 2.1. In the pilot study, participants were asked to adjust the IOI between the two sequences ( $\delta$  in Figure 2.1) to achieve a regular rhythm with a tempo twice as fast as the original sequences. Interpersonal variances were observed in the data, from which median values of  $\delta$  were determined. These median values for each pair of timbres presented in Tables B.2 and B.3 in Appendix B.1 were then used in the main experiment.

A linear ramp was imposed to fade in over the first four cycles of each ABAB sequence (shown in the gradual darkening of colours) in Figure 2.1. This fade-in was important to avoid an unwanted saliency boost on the first beat, because musicians are taught to treat the first beat they hear as the strong beat.



**Figure 2.1:** Illustration of an isochronous ABAB sequence

## Procedure

The experiment was carried out in a sound-attenuated booth ([Industrial Acoustics Company](#), model 1203) in the Music Perception and Cognition Laboratory of the Schulich School of Music of McGill University. Mono signals were presented to both ears of the participants via Sennheiser HD 280 headphones after amplification by a Grace m904 stereo

monitor controller, the average level was set to 59 dBA as measured with a Brüel & Kjær type 2250 sound level meter coupled with a Brüel & Kjær type 4153 artificial ear.

After an audiogram test to make sure their hearing meets our minimum criteria of being able to hear within 20 dB HL at frequencies 250, 500, 1000, 2000, 4000 and 8000 Hz, participants were presented with an instruction sheet that describes what was expected from them. They were instructed to listen to each sequence carefully before starting to tap to what they perceived as strong beats. They were told that they had to tap 10 times to one of the two timbres in each sequence before moving to the next sequence, which meant that they could stop after a few initial taps to one timbre and change their mind to tap 10 times to the other timbre. For each sequence, both the downbeat timbre and the relative timing of each tap to the physical onset of the corresponding timbre were recorded. Nonmusician participants who were unfamiliar with the concept of strong beats (or downbeats) were instructed to tap to the timbre that seemed “more important” in a given sequence.

Each session started with a training block with 6 sequences for participants to become familiar with the experiment. These sequences were constructed with the four instrument timbres – Alto flute (AF), Bassoon (BS), Celesta (CE) and Violin (VN) – which are not included in the sequences for the actual experiment. Two participants did not feel comfortable enough after going through the training block once, in which case the same block was repeated one time. These data were not analyzed.

After the training block, the actual experiment was presented in two blocks. Each block had 105 sequences (i.e., the total number of possibilities of choosing 2 out of 15). The two blocks were identical except that each sequence started with the other timbre. In other words, if Block 1 had a sequence of PF-VP (piano followed by vibraphone), Block 2 would have the opposite sequence of VP-PF (vibraphone followed by piano). The purpose was to make sure to present all possible cases to examine the possible effect of the timbre of the

first sound. Sequences in each block were presented in random order to counter possible order effects.

Participants had a chance to take a break after the first block, but most chose to proceed directly to the next block. The entire experiment took between an hour and 20 minutes to two hours including time for filling out questionnaires, debriefing and feedback.

### 2.2.2 Analysis of Measured Saliency Differences

#### Participants' Choices of Downbeats

For each participant, a 15-by-15 choice matrix was created where each cell would have either 0 or 1, according to the participant's responses. In other words, for a participant  $k$ , the choice matrix  $C_k$  (for  $k = 1, 2, \dots, 40$ ) is defined as follows.

$$C_k(i, j) = \begin{cases} \text{choice of participant } k \text{ for a sequence starting with timbre } i \text{ followed by timbre } j \\ 1, \text{ if participant } k \text{ tapped to timbre } i \text{ in the sequence with timbre } j \\ 0, \text{ otherwise} \end{cases} \quad (2.1)$$

where  $1 \leq i \neq j \leq 15$ .

The diagonal entries in  $C_k$  are all zeros because same-timbre sequences were not included in the experiment. Also, if the  $(i, j)$  pair was presented in the first block, the opposite sequence, the  $(j, i)$  pair, was presented in the second block; no two sequences with the same two timbres were presented in the same block. For example, if  $C(\text{HC}, \text{CL}) = 0$ , meaning that the participant tapped to the timbre corresponding to CL given a sequence of timbre HC followed by timbre CL, and the corresponding cell entry in the other block  $C(\text{CL}, \text{HC})$  is 1, meaning that the participant tapped to the timbre corresponding to CL

given a sequence of timbre CL followed by timbre HC, then timbre CL was judged to be more salient than timbre HC when these two were combined, regardless of the presentation order.

Let's consider another example. If  $C(\text{FL}, \text{EH}) = 1$  and  $C(\text{EH}, \text{FL}) = 1$ , indicating that the participant tapped once to each of the two timbres, the timbres EH and FL were probably of similar saliency level to this participant.

### Effect of Presentation Order

The choice matrices were summed over all participants ( $N = 40$ ) and the upper and lower triangular matrices were compared to examine the order effect. A paired-sample t-test verified that there was no significant difference between them,  $t(104) = 0.768, p = .444$ . As the 95% confidence interval is from  $-1.0708$  to  $2.4232$ , which is relatively small and includes the zero difference points, we can interpret the results as indicating that the participants' average decisions were probably consistent in the two blocks. It may also indicate that participants were not affected by the timbre of the first beat of a given sequence. This can also be seen from the fact that the first beat timbres were chosen as downbeats in 49% of cases across all participants, which is not statistically different from the chance level of 0.5.

A further t-test reveals that there was a slight difference in how often participants chose the first beat timbre in two blocks. The dependent variable (DV) was the average probability of each participant choosing the first beat timbre as the strong beat in each block. The first beat timbres were chosen as strong beats in the first block 48.6% of the time,  $t(39) = -1.763, p = .086$ , which is slightly less often than 49.5% in the second block,  $t(39) = -0.533, p = .597$ . As these numbers suggest that the first beat timbre was not favoured over the other beat timbre in the given task.

As there was lack of significant difference to suggest an effect of order, data were aver-

aged across the two presentation orders. A two-way Analysis of Variance (ANOVA) on these average choice rates with independent variables (IVs) of gender and musicianship showed that there were no further differences according to gender,  $F(1, 416) = 0.018, p = .893$ , or musicianship,  $F(1, 416) = 0.449, p = .503$ . No significant interaction effect was observed between gender and musicianship,  $F(1, 416) = 0.001, p = .979$ . To summarize, there was no presentation order found in the data and it was universal across different gender and musicianship groups.

### Forming Dominance Matrices from Choice Matrices

Since the upper and lower triangular matrices of the choice matrices  $C_k$  were shown to be statistically equivalent in the previous section, the two parts in each of 40 matrices were combined to form 40 dominance matrices by applying the following formula:

$$\begin{aligned}
 D_k(i, j) &= \text{dominance perceived by participant } k \\
 &= \frac{C_k(i, j) + (1 - C_k(j, i))}{2} \\
 &= \begin{cases} 1, & \text{if } C_k(i, j) = 1 \text{ and } C_k(j, i) = 0 \\ 0.5, & \text{if } C_k(i, j) = C_k(j, i) \\ 0, & \text{if } C_k(i, j) = 0 \text{ and } C_k(j, i) = 1 \end{cases}
 \end{aligned} \tag{2.2}$$

where  $k$  ( $k = 1, 2, \dots, 40$ ) and  $1 \leq i \neq j \leq 15$ .

What Eq. 2.2 means is as follows.  $D_k(i, j) = 1$  means total dominance of timbre  $i$  over timbre  $j$ , regardless of the combined order. The opposite case of  $D_k(i, j) = 0$  indicates total dominance of timbre  $j$  over timbre  $i$ , regardless of the combined order.  $D_k(i, j) = 0.5$

occurs when  $C_k(i, j)$  and  $C_k(j, i)$  have the same values, which implies that neither timbre  $i$  nor timbre  $j$  was dominant over the other. This case also illustrates that the choice of the timbre to tap to was essentially random due to the lack of a dominance relationship between these timbres.

Note that the matrices  $D_k$  are *not* symmetric. The entry  $D_k(i, j)$  is in fact the complement of  $D_k(j, i)$  in the sense that the sum of the two corresponding off-diagonal entries is always 1 (where  $i \neq j$ ).

### Mean Dominance as Average Tapping Probability

A mean dominance matrix,  $\overline{D}$ , was computed across all participants according to Eq. 2.3. This matrix gives us an average probability of someone tapping to timbre  $i$  (which is the *row* index) when it was presented in combination with timbre  $j$  (which is the *column* index). For example, consider the following grand-average dominance matrix  $\overline{D}$  in Table 2.2.

$$\begin{aligned}\overline{D}(i, j) &= \text{dominance of timbre } i \text{ to timbre } j, \text{ averaged across participants} \\ &= \frac{1}{40} \sum_{k=1}^{40} D_k(i, j)\end{aligned}\tag{2.3}$$

where  $k$  ( $k = 1, 2, \dots, 40$ ) and  $1 \leq i \neq j \leq 15$ .

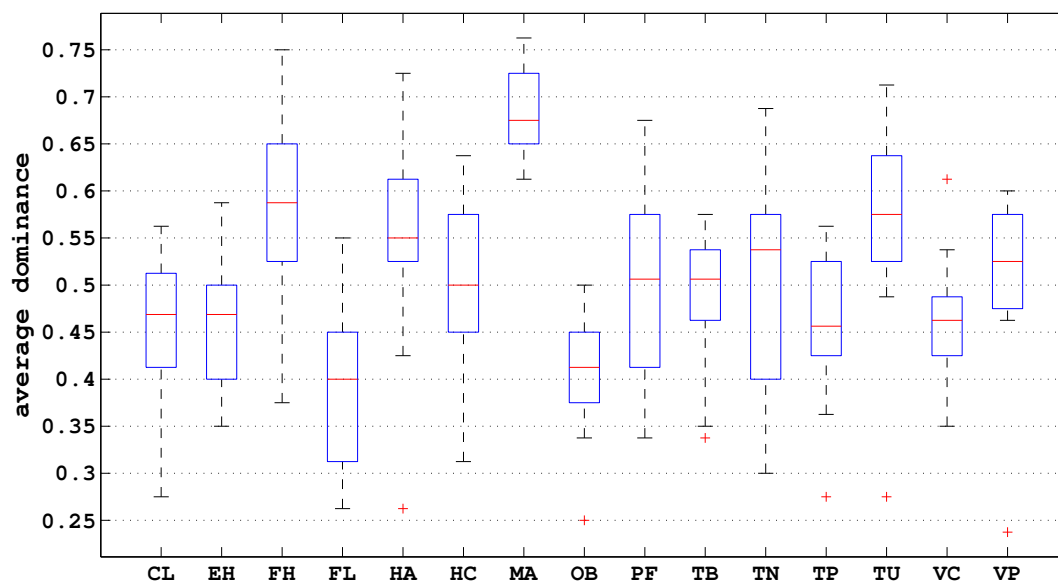
Since an equal dominance would be 0.5 in this table, any entry that is higher than 0.5 means that the instrument on the corresponding row is more dominant than the instrument on the column when they are combined in a sequence.

Figure 2.2 shows the distribution of dominance values of 15 instruments in grand average dominance matrix  $\overline{D}$ , when each timbre on the abscissa is combined with 14 other timbres. This box plot was obtained from each row of the matrix  $\overline{D}$ , as  $\overline{D}(i, j)$  is the average dominance of timbre  $i$  to timbre  $j$ . So the entries on the  $i$ -th row of  $\overline{D}(i, :)$  reflect the



Table 2.2: Mean dominance matrix  $\bar{D}$  from 40 dominance matrices,  $D_k$ 

	$j$															
	CL	EH	FH	FL	HA	HC	MA	OB	PF	TB	TN	TP	TU	VC	VP	
$i$	CL	0.5000	0.4500	0.4125	0.5250	0.2750	0.4875	0.3125	0.5625	0.4875	0.5125	0.4250	0.5375	0.3500	0.5125	0.4250
	EH	0.5500	0.5000	0.4125	0.5875	0.3875	0.3625	0.3500	0.5000	0.4000	0.5375	0.4375	0.4750	0.4625	0.5000	0.4750
	FH	0.5875	0.5875	0.5000	0.7375	0.4750	0.5875	0.3750	0.7500	0.6250	0.5625	0.6500	0.7250	0.3750	0.6125	0.5250
	FL	0.4750	0.4125	0.2625	0.5000	0.3500	0.4000	0.2625	0.5500	0.3250	0.4250	0.3125	0.4500	0.2875	0.5375	0.4000
	HA	0.7250	0.6125	0.5250	0.6500	0.5000	0.6125	0.2625	0.5500	0.6500	0.4250	0.6000	0.5500	0.4750	0.5375	0.5250
	HC	0.5125	0.6375	0.4125	0.6000	0.3875	0.5000	0.3125	0.6125	0.4625	0.4875	0.5250	0.5750	0.4500	0.5500	0.4750
	MA	0.6875	0.6500	0.6250	0.7375	0.7375	0.6875	0.5000	0.6625	0.6625	0.6625	0.7000	0.6125	0.7250	0.6375	0.7625
	OB	0.4375	0.5000	0.2500	0.4500	0.4500	0.3875	0.3375	0.5000	0.4250	0.4750	0.3750	0.4375	0.3750	0.3875	0.4000
	PF	0.5125	0.6000	0.3750	0.6750	0.3500	0.5375	0.3375	0.5750	0.5000	0.5000	0.4375	0.5750	0.4625	0.5250	0.4125
	TB	0.4875	0.4625	0.4375	0.5750	0.5750	0.5125	0.3375	0.5250	0.5000	0.5000	0.3500	0.4750	0.5125	0.5750	0.5375
	TN	0.5750	0.5625	0.3500	0.6875	0.4000	0.4750	0.3000	0.6250	0.5625	0.6500	0.5000	0.5375	0.4000	0.5375	0.4750
	TP	0.4625	0.5250	0.2750	0.5500	0.4500	0.4250	0.3875	0.5625	0.4250	0.5250	0.4625	0.5000	0.3625	0.4625	0.4500
	TU	0.6500	0.5375	0.6250	0.7125	0.5250	0.5500	0.2750	0.6250	0.5375	0.4875	0.6000	0.6375	0.5000	0.6500	0.5250
	VC	0.4875	0.5000	0.3875	0.4625	0.4625	0.4500	0.3625	0.6125	0.4750	0.4250	0.4625	0.5375	0.3500	0.5000	0.4500
	VP	0.5750	0.5250	0.4750	0.6000	0.4750	0.5250	0.2375	0.6000	0.5875	0.4625	0.5250	0.5500	0.4750	0.5500	0.5000



**Figure 2.2:** Distributions of dominance values for all pairs of timbres involving each instrument on the horizontal axis

average dominance of the timbre  $i$  across all possible combinations. For example, the first row of  $\overline{D}$  has 14 numbers corresponding to the average dominance of CL when it is combined with other timbres, which are shown by a box and whiskers in the first column with the corresponding label ‘CL’ on the abscissa in Figure 2.2. Each box shows values from the 25th to the 75th percentiles and the horizontal (red) bar in each box is the median value. The whiskers outside a box designate the range of values within  $\pm 2.7\delta$ , where  $\delta$  is the standard deviation, and the outliers are noted with (red) crosses outside of these whiskers.

Of importance is the size of each box as well as the location of the median bars. Some timbres, for example PF and TN, have bigger boxes than do others (e.g., MA or VC). A bigger box indicates the dominance of that particular timbre changed a lot depending on the other timbre that it was combined with. A smaller box means the opposite, that the timbre’s dominance value did not depend much on the other timbre. According to Figure 2.2, VC’s dominance was least affected and PF most affected by other timbres.

One more thing to notice is the location of MA. Its median value is much higher than other medians. Also, the size of the box for MA is one of the smallest ones (second only to VC and possibly OB), which means participants tended to tap to MA over other instruments fairly consistently, regardless of what other instrument it was combined with.

Out of all timbre descriptors from the Timbre Toolbox (Peeters et al., 2011), *harmonic skew* turned out to be the best correlate of the median values of  $\overline{D}$ , explaining 62% of the variance in data,  $r(13) = .7875, p = .0005$ . Tristimulus statistics, introduced by Pollard and Jansson (1982), give energy ratios of harmonics in the spectrum, as defined by Eq. 2.4 (from Peeters et al., 2011) – Tristimulus Band 1 is the ratio of the fundamental energy to the total energy of the first  $H$  harmonics, Tristimulus Band 2 is the ratio of the energy of harmonics 2, 3 and 4 to the total energy of the first  $H$  harmonics, and Tristimulus Band 3 is the ratio of the energy of harmonics 5 and above to the total energy of the first  $H$

harmonics. In the Timbre Toolbox, the default value of  $H$  is set to 20.

$$\begin{aligned} T1(\tau) &= \frac{a_1(\tau)}{\sum_{h=1}^H a_h(\tau)}, \\ T2(\tau) &= \frac{a_2(\tau) + a_3(\tau) + a_4(\tau)}{\sum_{h=1}^H a_h(\tau)}, \\ T3(\tau) &= \frac{\sum_{h=5}^H a_h(\tau)}{\sum_{h=1}^H a_h(\tau)}, \end{aligned} \tag{2.4}$$

The median values of dominance showed a high positive correlation with Tristimulus Band 1,  $r(13) = .6565, p = .0079$ , and negative correlations with Tristimulus Bands 2 and 3,  $r(13) = -.3432, p = .2105$ , and  $r(13) = -.7572, p = .0011$ , respectively. What these correlation coefficients illustrate is that on average people tapped to sounds with more energy on the fundamental rather than on the higher harmonics. This is evident from the directions of the correlations: Tristimulus Band 1 is positively correlated with the median dominance values, indicating that the more energy in the fundamental the more often it was chosen as downbeat; Tristimulus Bands 2 and 3 show negative correlations, which signifies that timbres with less energy in higher harmonics were tapped to more often. This is in line with some participants' post-experiment comments that they tended to tap to "lower-sounding" instruments, i.e., instruments with more energy in the lower frequency range of the spectrum.

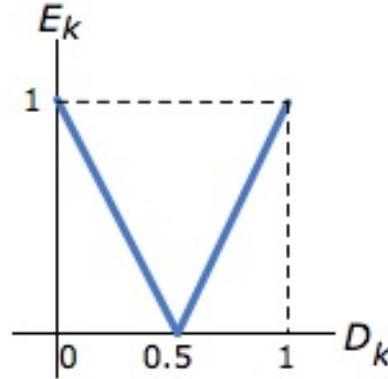
### Obtaining Proximity Matrices from Dominance Matrices

A dominance relationship between a pair of stimuli can be considered to reflect a type of proximity between them. For example, if a pair of objects have a highly polarized dominance relationship, it would mean that these objects are highly non-proximal in the given context. On the other hand, if none of the two objects is significantly more dominant

than the other, it implies that they are fairly proximal to each other and hence equally dominant. Note that we are not making any assumptions on a possible implication of perceptual dissimilarity on dominance here. Our suggested function of proximity in terms of dominance can be summarized using the following formula:

$$\begin{aligned} E_k(i, j) &= \text{distance (inverse proximity) between timbres } i \text{ and } j \\ &= |2 * (D_k(i, j) - 0.5)| \end{aligned} \quad (2.5)$$

where  $k$  ( $k = 1, 2, \dots, 40$ ),  $1 \leq i \neq j \leq 15$  and  $0 \leq D_k \leq 1$ .



**Figure 2.3:** Calculation of proximity values  $E_k(i, j)$  from dominance values  $D_k(i, j)$

Figure 2.3 is a graphic representation of Eq. 2.5. Forty dominance matrices were converted into 40 proximity matrices using Eq. 2.5 to model the perceptual space underlying these matrices.

### Saliency Map of 15 Instrument Timbres

Multidimensional scaling (MDS) (Shepard, 1962a, 1962b; Kruskal, 1964a, 1964b) is an algorithm that computes underlying low-dimensional spaces from a collection of proximity data. It helps understand relationships among stimuli through visualization in space where

the distance between a pair of stimuli is monotonically related to some measure of proximity between them.

CLASCAL (Winsberg & De Soete, 1993) is an extended MDS algorithm that estimates not only the positions of stimuli along the dimensions in a common space, but also the number of latent classes of participants, specificity values that describe each stimulus’s unique characteristic that cannot be explained by the common dimensions, a set of weights per latent class on the dimensions and specificities, and the probability of each participant’s membership in each latent class. We decided to use CLASCAL rather than the general MDS algorithm, since we were interested in finding different groups of listeners according to their saliency judgments from the latent class information.

**Table 2.3:** Log likelihood, degrees of freedom, and AIC and BIC values for spatial models

Rank	# Dim.	Specificities	# Class.	logL	df	AIC	BIC
1	1	Yes	2	−2802	4167	5670	<b>5880</b>
2	1	Yes	3	−2799	4164	5669	5898
3	1	Yes	1	−2825	4170	5709	5899
4	1	Yes	4	−2798	4161	5674	5921
5	1	Yes	5	−2798	4158	5679	5946
6	1	Yes	6	−2798	4155	5685	5971
7	2	Yes	2	−2786	4152	5667	5972
8	2	Yes	1	−2809	4157	5705	5978

We ran CLASCAL on 40 individual Distance matrices. Table 2.3 lists the log likelihood, degrees of freedom, and Akaike’s information criterion (AIC) (Akaike, 1977) and Bayesian information criterion (BIC) values for spatial models with the eight least BIC values. Both AIC and BIC are model selection criteria among a finite set of models, which describe the

relative quality of a statistical model to fit the given data. The difference is that while AIC is proportional to the degrees of freedom of the model minus the log likelihood, BIC takes into account the sample size in addition to the number of degrees of freedom of the model as a penalty. With multiple estimated models to compare, the best solution is usually the one with the lowest AIC and BIC numbers, as these numbers tend to increase with the unexplained variance in the dependent variable. When the minimum AIC and the minimum BIC do not coincide (as in Table 2.3), one can choose to use either AIC or BIC to determine the best model. We decided to use BIC numbers for our analysis.

The best model without specificities had the BIC value of 6176, which is much larger than those in Table 2.3, which is why all the models listed above have specificities. The best model (with the lowest BIC) is the one-dimensional solution with two latent classes, although it is followed by two close contenders of one-dimensional solutions with one and three latent classes.

CLASCALMC, which compares nested models using Monte-Carlo tests, confirms that the best model is indeed the one-dimensional solution with specificities and two latent classes, the positions and the specificities of all timbres of which are listed in Table 2.4. The one-dimensional solution is shown in Figure 2.4. Note that even though the “map” is presented in two dimensions, what is important is the y-axis positions of the data points. This two-dimensional presentation is to help distinguish points that are very close in value around the zero points of the saliency scale (VC, HC, FL, TU, VP, OB, HA). Note that the positive or the negative values have no psychological meaning; the value distribution by CLASCAL is centered on zero.

The saliency map in Figure 2.4 seems to have three groups of saliency levels – a “high” saliency group of positive values (PF, TB, TP, MA), a “middle” saliency group around zero (VC, HC, FL, TU, VP, OB, HA), and a “low” saliency group of negative values (EH, TN,

**Table 2.4:** Coordinates for a one-dimensional spatial solution with specificities and two latent classes

Timbre	Saliency Value	Saliency Region	Specificities
MA	0.473	High	0
TP	0.223		0.087
TB	0.124		0.213
PF	0.069		0.133
HA	0.014	Middle	0.213
OB	0.009		0.160
VP	0.008		0.233
TU	−0.016		0.232
FL	−0.017		0.204
HC	−0.026		0.284
VC	−0.031		0.220
FH	−0.148	Low	0.175
CL	−0.165		0.208
TN	−0.218		0.071
EH	−0.301		0.157



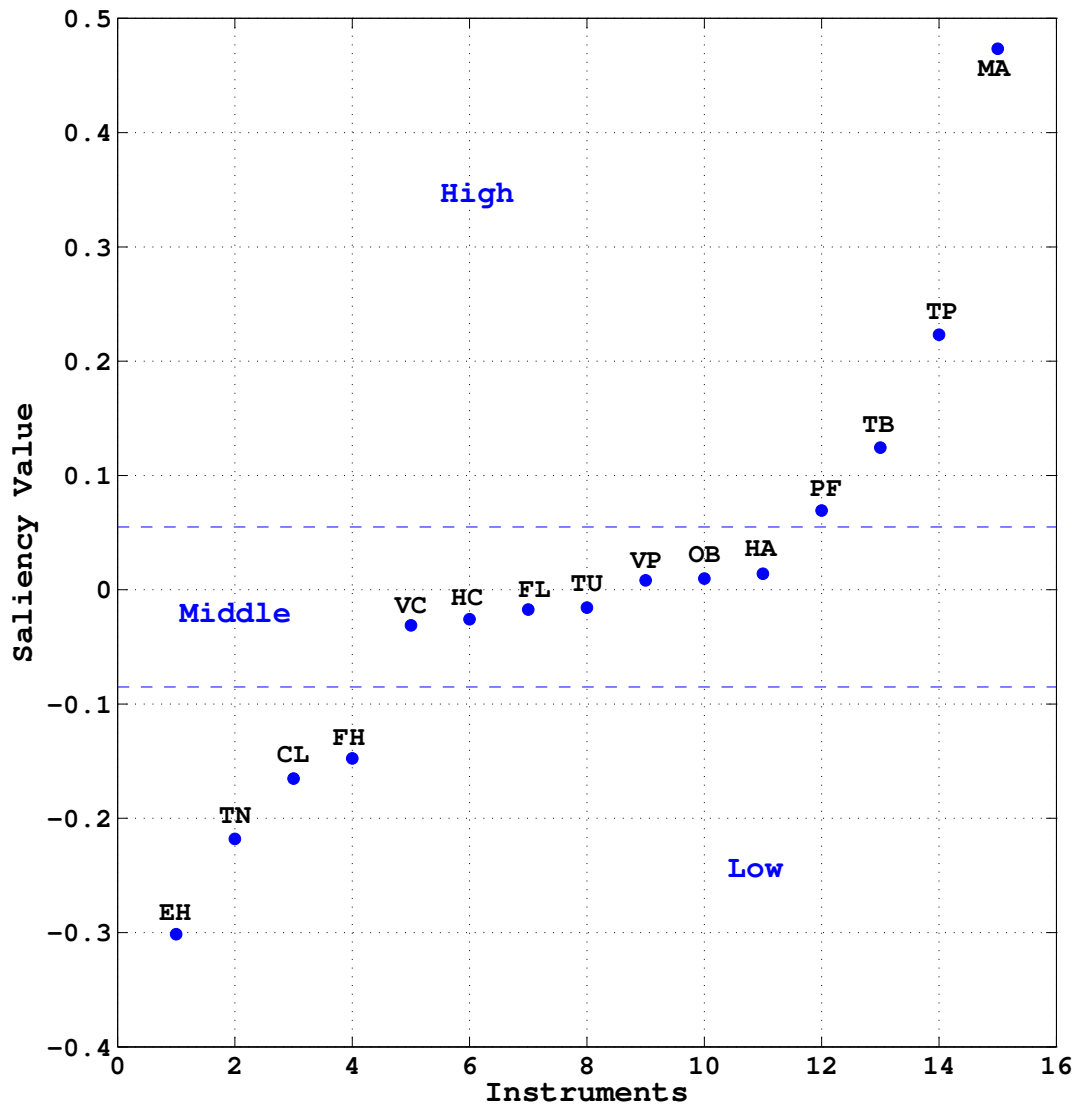


Figure 2.4: One-dimensional timbre saliency space of 15 timbres

CL, FH). This map means, for example, that when MA was presented in an isochronous sequence, participants tapped to MA more often than any other paired timbre, whereas they tended to tap to the other timbre when EH was one of the timbres in the sequence. Those in the “middle” saliency level group are more or less interchangeable and therefore the tappings were almost at a chance level when two of them were combined in an isochronous sequence.

**Table 2.5:** Distributions of participants in each latent class according to gender, musicianship and age

Criterion		Class 1	Class 2
Gender	Male	11	9
	Female	13	7
Musicianship	Musician	12	8
	Nonmusician	12	8
Age		19 to 35 years with a median of 21.5	19 to 40 years a median of 22

**Table 2.6:** Estimated weights in the selected one-dimensional model with specificities for two latent classes

Class	Dim 1	Specif
1	1.34	1.18
2	0.66	0.82

As Table 2.5 shows, the two latent classes turned out to have little correlations with gender, musicianship, or age. The difference seems to purely come from how an individual weighs the saliency dimension and the specificities in judging the relative saliencies,

according to Table 2.6; the participants belonging to Class 1 tend to put more weights on the saliency dimension in Figure 2.4 and the specificities, whereas those in Class 2 seem to apply less weights on both of them. In fact, what seems to really distinguishes the two classes is the weights on the saliency dimension; much greater for Class 1 than for Class 2 (in fact twice as much!).

The posterior probabilities of class membership of each participant is summarized in Table 2.7. Each participant is assigned to the class of which the posterior probability is the greatest. As we observe from Table 2.7, the class membership is pretty clear in general with a number close to 1 for one of the classes. Note that there are five participants with rather ambiguous memberships – numbers 7, 9, 14, 22 and 38. These listeners are ambiguous in the sense that a part of their data appears to belong to one class and the rest to the other class. For example, the participant 22 looks like he/she could belong to either class with almost equal probability. To evaluate these participants' latent class memberships, the prior distribution of class membership information in Table 2.8 needs to be compared against the posterior probabilities in Table 2.7. What Table 2.8 shows is the rough probability of a listener belonging to each class according to the null model; more than 60% of listeners would belong to Class 1 and the rest to Class 2. If there is someone with the posterior probabilities close to these prior distributions, it means that it would be impossible to find the latent class for that person. These numbers were estimated before the actual analysis of each participant's data. For the ambiguous participants 7, 9, 14, 22 and 38, we compare their posterior probabilities with the prior distributions and find that all participants' numbers are different from the prior distribution and that all five participants belong to Class 2 as their posterior probability of Class 2 is greater than that of Class 1.

**Table 2.7:** Posterior probabilities of class membership

Participant No.	Class 1	Class 2	Participant No.	Class 1	Class 2
1	0.00	1.00	21	0.10	0.90
2	0.00	1.00	22	0.46	0.54
3	0.95	0.05	23	0.00	1.00
4	0.90	0.10	24	1.00	0.00
5	1.00	0.00	25	0.01	0.99
6	0.86	0.14	26	0.11	0.89
7	0.31	0.69	27	0.90	0.10
8	0.99	0.01	28	1.00	0.00
9	0.22	0.78	29	1.00	0.01
10	0.04	0.96	30	0.99	0.01
11	0.99	0.01	31	1.00	0.00
12	0.93	0.07	32	0.00	1.00
13	1.00	0.00	33	0.87	0.13
14	0.24	0.76	34	0.98	0.02
15	0.02	0.98	35	1.00	0.00
16	0.91	0.09	36	1.00	0.00
17	0.85	0.15	37	0.01	0.99
18	1.00	0.01	38	0.21	0.79
19	0.93	0.07	39	0.94	0.06
20	0.00	1.00	40	1.00	0.00

**Table 2.8:** Prior distribution of class membership

Class	Prior Distribution
1	0.62
2	0.38

### Acoustic Correlates of the Saliency Dimension

To find the acoustic correlate of the saliency dimension in Figure 2.4, Pearson correlation coefficients were calculated between the saliency dimension in Table 2.4 and each of the 87 timbre descriptors calculated from the Timbre Toolbox (Peeters et al., 2011). The best correlate turned out to be the odd-even harmonic ratio,  $r(13) = .7080, p < .005$ , which alone explains 50% of the variance in the data. Table 2.9 lists the five timbre descriptors with the highest correlation coefficients.

**Table 2.9:** The five timbre descriptors with the highest correlations with the saliency dimension in Figure 2.4.

Descriptor	$r(13)$	$p$
Odd-even harmonic ratio	.71	$< .005$
Harmonic spectral decrease	-.68	$< .005$
Spectral amplitude rolloff	.54	$< .05$
Harmonic kurtosis	.53	$< .05$
Tristimulus Band 2	-.50	.057

The descriptors in Table 2.9 mostly describe harmonic details, which may indicate that the participants made saliency judgments based on harmonic details of each pair of timbres. There are a few timbre descriptors that have been confirmed in previous timbre

dissimilarity studies (Grey, 1977; Grey & Gordon, 1978; Krumhansl, 1989; Iverson & Krumhansl, 1993; McAdams et al., 1995; Lakatos, 2000; Caclin et al., 2005) – spectral centroid, attack time, log attack time and spectral variation. Apparently none of these well-known timbre dissimilarity descriptors is included in Table 2.9. The correlations of these descriptors with the saliency dimension are listed in Table 2.10 for comparison.

**Table 2.10:** Correlation coefficients of popular timbre descriptors with the saliency dimension.

Type	Descriptor	$r(13)$	$p$
Spectral	Harmonic spectral centroid	−.28	.31
Spectral	Power spectral centroid	.25	.37
Spectral	Magnitude spectral centroid	.24	.39
Spectrotemporal	Harmonic spectral variation	.18	.52
Spectrotemporal	Power spectral variation	.15	.59
Spectrotemporal	Magnitude spectral variation	.11	.71
Temporal	Log attack time	−.11	.71
Temporal	Attack time	−.09	.75

The correlation coefficients in Table 2.10 are not quite as high as those in Table 2.9, implying that participants may have used different timbre characteristics and properties for the saliency judgment from those for the dissimilarity judgment in the aforementioned timbre dissimilarity studies. Also noticeable are the different signs of the correlations with harmonic and power spectral centroids. These may result from different patterns in distributions of harmonic and inharmonic energy. Combining the results from both Tables 2.9 and 2.10, it appears that saliency judgments were made more on harmonic energy distributions of sounds rather than temporal information (such as attack time, log attack

time and spectral variations).

### 2.2.3 Analysis of Behavioural Differences in Tapping Patterns

#### Importance of the First Beat Timbre

Musicians are taught to consider the first beat as a strong beat. This places an unwanted emphasis on the first beat and the timbre of it in our experimental task. To minimize this additional saliency of the first beat timbre, we applied a linear ramp at the beginning of each sequence so that the first beat would be much quieter and hence its additional unwanted importance would be limited. The question then becomes whether it was worth the effort.

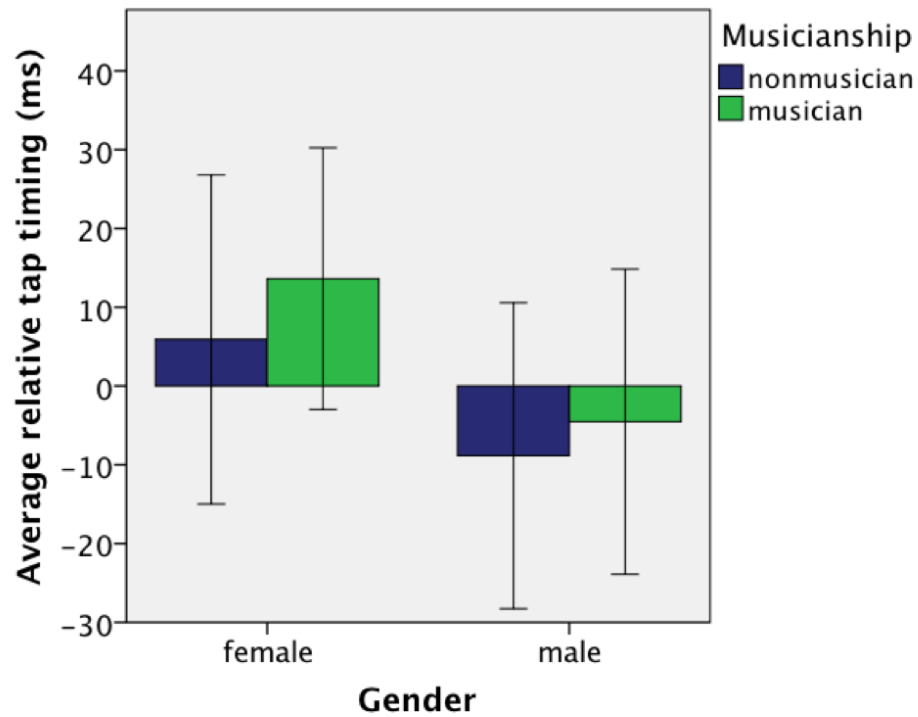
Since the experiment was done in two blocks, we first ran the paired-sample t-test to see if the participants' behaviours changed over time. The dependent variable was the average of each participant's tap timing (in milliseconds) relative to the start of each sequence played per block. The result showed the tapping patterns did not differ significantly in two blocks,  $t(39) = -0.869, p = .390$ . In fact, the first beat timbre was chosen as the strong beat for 49% of time in all trials, which is not statistically different from a chance level of 50%. These both indicate that there was no effect of the first beat timbre and it was consistent in both blocks, verifying that the fade-in worked as we intended.

#### Tap Timing

Since the participants' behaviour did not change in two blocks as we saw in the last section, all tapping timing data relative to the physical onset of each corresponding sound were averaged across both blocks per participant to calculate average tap timing values for further analysis. These values reflect the tapping patterns of each participant. The hypothesis was

that musicians might be tapping earlier than nonmusicians with their training to anticipate the upcoming beats, regardless of gender. One-way ANOVAs were carried out with the dependent variable of the average tap timing (in milliseconds) relative to the start of each sequence averaged across all trials for each participant. Gender and musicianship were the IVs. The result reveals that males tend to anticipate (which means to tap before the onset of a sound) on average, which was marginally significant,  $F(1, 38) = 3.942, p = .054$ , whereas there was no musicianship effect in tap timing,  $F(1, 38) = 0.482, p = .492$ , which was against our expectations. Figure 2.5 shows the average values of tap timings relative to the physical onset of each corresponding sound played in a sequence according to gender (on the horizontal axis), and musicianship [nonmusicians in blue (or dark gray in monochrome prints) bars and musicians in green (or light gray) bars]. The values on the y-axis mean relative tap timing (in milliseconds) to the start of each sound played. First, notice the big error bars, suggesting quite a large range of interpersonal differences in tap timing. These variances seem consistent across gender or musicianship. Second, we can clearly see that males tended to tap in advance (*anticipatory*) with negative average timings, and females tended to tap afterwards (*reactive*). A three-way Analysis of Covariance (ANCOVA) on the average tap timing across all trials per participant with IVs of musicianship and gender and a covariate of age shows that the gender effect is not as significant as in the one-way ANOVA,  $F(1, 35) = 2.893, p = .098$ . Neither musicianship nor age had a significant effect:  $F(1, 35) = 0.193, p = .663$ , for the main effect of musicianship,  $F(1, 35) = 0.660, p = .422$  for the main effect of age, and  $F(1, 35) = 0.002, p = .969$  for the interaction of musicianship and gender.





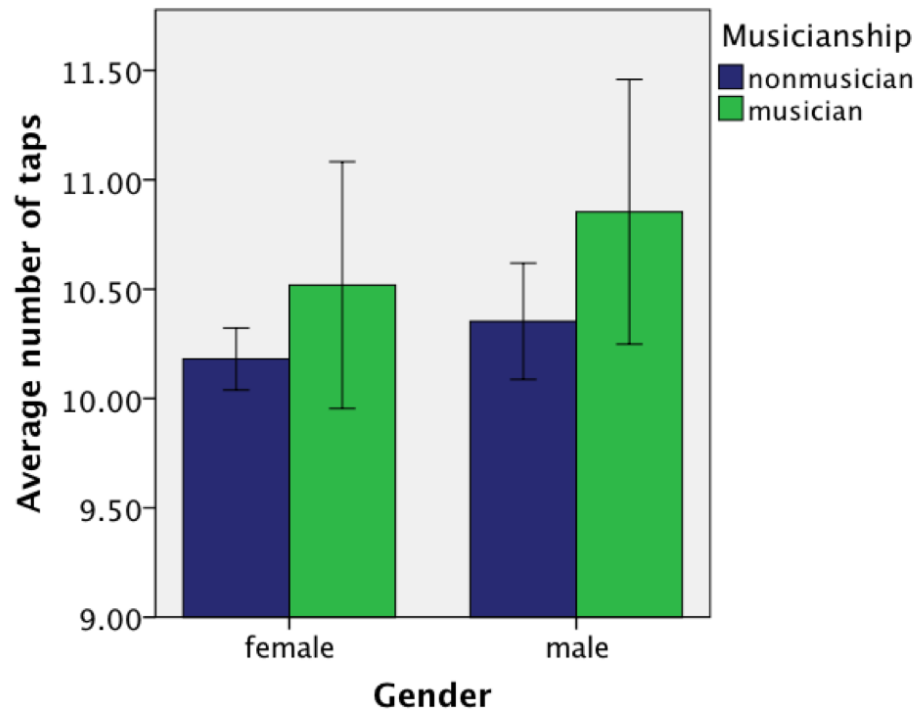
**Figure 2.5:** Average tap timing according to gender and musicianship. The error bars represent the 95% confidence interval.

### Total Number of Taps

Was there any difference according to gender or musicianship in how often a participant changed his/her mind on which timbre to tap to? For this, a two-way ANOVA was carried out on the average number of taps per participant (averaged across all stimuli) with IVs of gender and musicianship. Age was not included as a covariate as it was not significantly correlated with tap timing in previous analyses. The result shows that on average musicians tended to tap more times than nonmusicians,  $F(1, 36) = 4.645, p = .038$ , which indicates that musicians were more prone to change their mind. This might mean that musicians were more careful and attentive about the task, hence more often changed their mind after a few initial taps. There was no significant gender difference in the average number of taps,  $F(1, 36) = 1.693, p = .201$ , or gender-musicianship interaction,  $F(1, 36) = 0.173, p = .680$ , as Figure 2.6 illustrates. Musicians changed their minds even after five initial taps significantly more often than nonmusicians,  $F(1, 38) = 4.342, p = .044$ , according to a one-way ANOVA on the average number of taps per participant with musicianship as IV.

## 2.3 Experiment II: Measuring Timbre Saliency using a Direct Comparison Method

The task of tapping to strong beats in a perceptually isochronous sequence, used for the first experiment in Section 2.2, might have had an unforeseen confound with rhythmic saliency in duple rhythms. Duple rhythms are often used in rock music, where strong beats usually coincide with the bass drum, which gives more low-frequency energy to the music, whereas weak beats tend to have snare sounds, which add high-frequency energy. Hence, participants may have had an implicit expectation of more low-frequency energy for a strong beat timbre in a given sequence, as some participants' post-experiment comments



**Figure 2.6:** Average number of taps according to gender and musicianship. The error bars represent the 95% confidence interval.

revealed in Section 2.2. The saliency dimension obtained from this experiment was negatively correlated with spectral centroids in power and harmonic spectrum, which means that participants judged a timbre with more low-frequency (harmonic) energy to be more salient.

Another experiment was performed to evaluate timbre saliency without this confound of rhythmic saliency. This time, a direct comparison technique was employed using the two-alternative forced-choice (2AFC) paradigm.

### 2.3.1 Methods

#### Participants

Participants ( $N = 60$ ) were recruited from a classified advertisement on the McGill University website. There were equal numbers of males and females and of musicians and nonmusicians based on self reports. Musicians are defined as those with four or more years of musical training (instruments, voice, composition, recording, etc.) during which time they spent at least five hours per week on their training (including lessons and practices). All participants passed the preliminary hearing test before the experiment to make sure that they were able to hear within 20 dB HL at frequencies 250, 500, 1000, 2000, 4000 and 8000 Hz ([International Organization for Standardization, Geneva, 2004](#); [Martin & Champin, 2000](#)). Their ages ranged from 19 to 44 with a median of 23 years. They were paid \$5 upon completion of the experiment, which took about 30 minutes including the time for filling out a questionnaire and debriefing.

### Stimuli

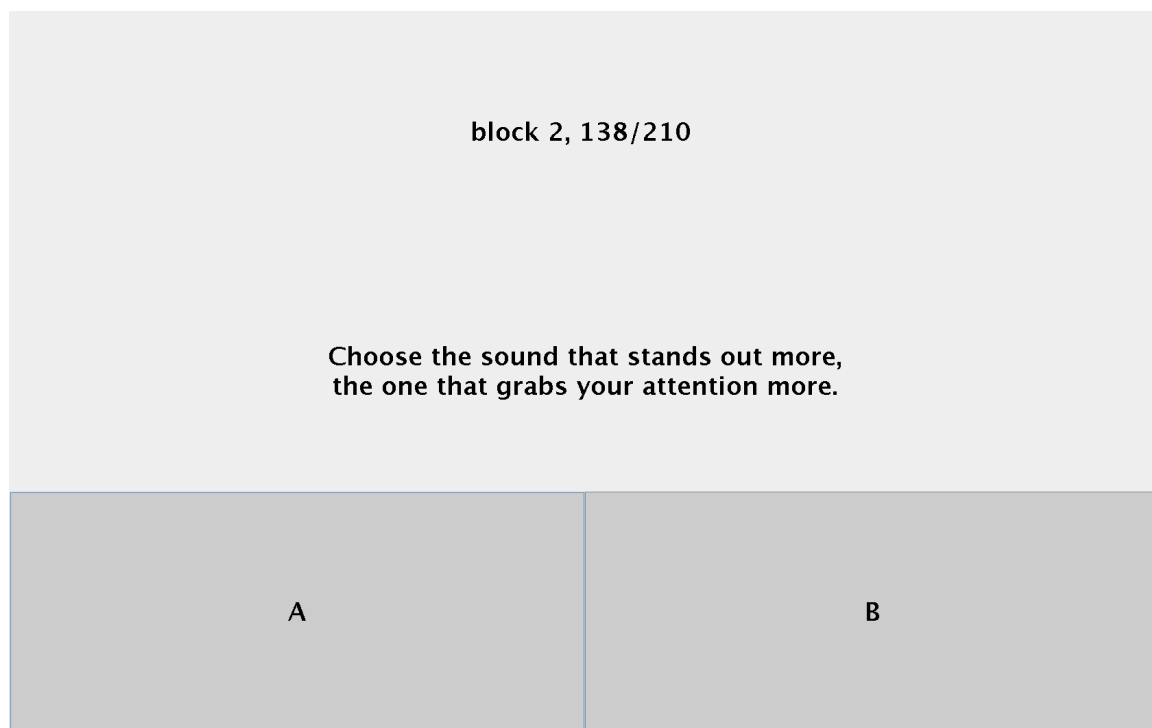
The same set of 19 instrumental sounds was used that formed isochronous sequences in Experiment I (Section 2.2). They are listed in Table 2.1. These sounds were equalized in terms of pitch (C4 or 261.6 Hz), effective duration (200 ms) and loudness. The stimulus selection and loudness equalization pilot are presented in detail in Appendices A and B.1, respectively.

### Procedure

The experiment was carried out in a sound-attenuated booth ([Industrial Acoustics Company](#), model 1203) in the Music Perception and Cognition Laboratory of the Schulich School of Music of McGill University. Mono sound signals, to further eliminate any undesired cue for separation that might affect the saliency judgment, were presented to both ears of the participants via Sennheiser HD 280 headphones. The average level was set to 60 dBA by a Brüel & Kjær type 2250 sound level meter coupled with a Brüel & Kjær type 4153 artificial ear.

There was a training block with six pairs of timbres before the actual experiment, so as to help participants familiarize themselves with the task. These trials used the training timbres in Table 2.1, that are different from those used for test trials. After the training block, the experimenter, who was present in the booth during the training block, provided any further clarifications if required. Participants performed the task alone after the experimenter left the booth.

The experiment used the graphic user interface in Figure 2.7, developed in PsiExp ([Smith, 1995](#)) on an Apple PowerPC G5. For each trial, participants listened to one timbre presented three times followed by the other timbre presented three times and decided which



**Figure 2.7:** Screenshot of the graphic user interface for Experiment II

of the two was more salient than the other. When they pressed the button corresponding to the more salient stimulus, it would automatically bring the next trial. It was not possible to repeat the current stimuli or to go back to an earlier trial. For each pair of particular instruments, both orders were considered for the experiment (e.g., one trial with PF for A and VC for B, and another trial with VC for A and PF for B) to study a possible order effect.

### **2.3.2 Analysis**

#### **Participants' Choices of More Salient Timbres**

For each participant, a 15-by-15 choice matrix was created, in the same way as for Experiment I in Section 2.2.2. Each cell in a choice matrix would have either 0 or 1, indicating the participant's choice of a more salient timbre for a given pair. The definition of a choice matrix,  $C_k$ , for a participant  $k$ , is the same as Equation 2.1.

First, group averages were calculated from the individual judgment data for each of the 210 stimulus pairs across the 15 listeners within each combination of gender and musicianship categories. These average choice values were analyzed as the DV in ANOVA. Gender and musicianship were the IVs. A two-way ANOVA reveals that there were no effects of gender or musicianship on the choice of a more salient timbre,  $F(1, 836) < 1$ , for the main effect of gender;  $F < 1$ , for the main effect of musicianship;  $F(1, 836) = 1.103, p = .294$ , for the interaction of gender and musicianship. This finding confirms the result from Experiment I that neither gender nor musicianship had a significant impact on saliency judgments and the variance in the data comes from saliency differences in the stimuli.

### Effect of Presentation Order

As there were no significant main effects of gender or musicianship found in the data, the individual choice matrices were all averaged across 60 participants and the resulting upper and lower triangular matrices were examined for the order effect. Two column vectors of length 105 were created, one ( $V_1$ ) from the entries of the lower triangular portion of the average choice matrix  $C$ , and the other ( $V_2$ ) from the entries of the upper triangular portion. The second vector was in fact calculated from subtracting the upper triangular cell values from one, because if timbre  $i$  was more salient than timbre  $j$  regardless of the presentation order, the value at  $C(i, j)$  should be close to 1 while the entry at  $C(j, i)$  should be close to 0. The relative orders of the two vectors were kept the same so that if  $V_1(k) = C(i, j)$ , then  $V_2(k) = 1 - C(i, j)$ . A paired-sample  $t$ -test was carried out on these two vectors to find a significant difference,  $t(104) = -6.067, p < .000001$ . This effect of order did not exist in Experiment I.

A further investigation reveals that about two thirds of the stimuli had the order effect. Table 2.11 shows the result of paired-samples  $t$ -tests for each of 15 instruments. For example, the  $t$ -statistic and the corresponding  $p$ -value on the first row (CL) was calculated from comparing the average choice of CL as the salient timbre when it was presented first and the average choice when it was presented second. Since there are 14 other instruments to compare with CL, these comparison vectors of CL are of length 14, which results in 13 degree of freedom.

Note that both EH (positioned the lowest on the saliency dimension) and MA (positioned the highest) exhibit significant effects of order, which makes us suspect that this order effect may be little related to the saliency dimension from Experiment I. The  $t$  statistics in Table 2.11 can be considered to reflect the susceptibility to the order effect, as the



Table 2.11: Paired-samples t-test result

Instrument	$t(13)$	$p$
CL	-3.121	.0009*
EH	-2.185	.003*
FH	-0.732	.533
FL	-2.060	.010*
HA	-1.365	.246
HC	-2.867	.001*
MA	-2.816	.001*
OB	-2.225	.002*
PF	-2.258	.015*
TB	-2.463	.006*
TN	-1.050	.767
TP	-1.095	.677
TU	-1.096	.677
VC	-2.559	.0003*
VP	-0.845	.779

\* indicates significant cases of order effects with  $p < .05$

nine instruments with the smallest  $t$  statistics showed the effect of order. To figure out what might be the explanation for the  $t$  statistics and hence the susceptibility to order effects, Pearson correlation analysis was performed on the  $t$  statistics in Table 2.11 with the timbre saliency coordinates and specificity values in Table 2.4 and the median dominance values (red bars within boxes in Figure 2.2) from Experiment I. None of these measures turned out to have a significant correlation with the  $t$  statistics, as shown in Table 2.12.

**Table 2.12:** Correlation of  $t$  statistics with measures from Experiment I

Measure	$r(13)$	$p$
Saliency coordinates	-.18	.53
Specificities	-.06	.84
Median dominance	.21	.46

As the measures from Experiment I failed to explain the  $t$  statistics in Table 2.11, correlations were calculated on the 87 timbre descriptors from the Timbre Toolbox (Peeters et al., 2011). The best descriptor turned out to be the *Harmonic Spectral Deviation* with  $r(13) = -.7360, p = .0018$ , which is a measure of how different the harmonic amplitudes are from the estimated smoothed amplitude envelope (Krimphoff, McAdams, & Winsberg, 1994):

$$HDEV(\tau) = \frac{1}{H} \sum_{h=1}^H (a_h(\tau) - SE(f_h, \tau)) \quad (2.6)$$

where  $SE(f_h, \tau)$  denotes the global (smoothed) spectral envelope at harmonic frequency  $f_h$  and time  $\tau$ . The spectral envelope is a rough estimate using a smoothing window of three adjacent partials:

$$SE(f_h, \tau) = \frac{1}{3} (a_{h-1}(\tau) + a_h(\tau) + a_{h+1}(\tau)) \quad (2.7)$$

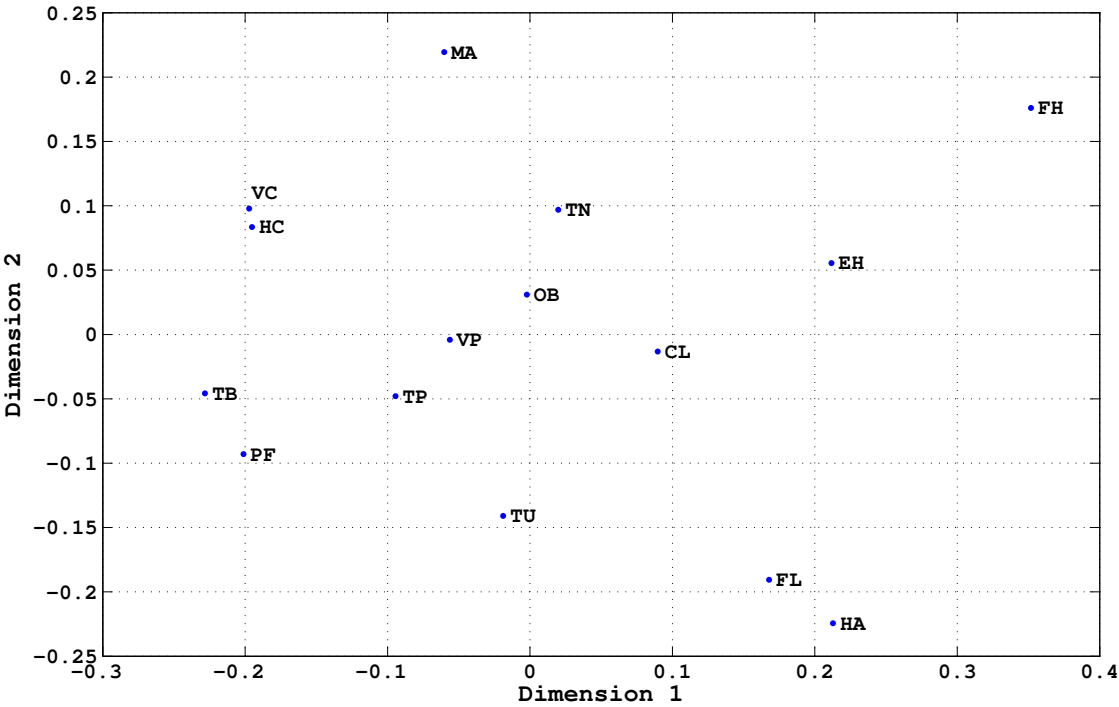
for  $1 < h < H$ .  $H$  is the number of harmonic peaks the Timbre Toolbox estimates.

The negative direction of correlation between harmonic spectral deviation and the  $t$  statistics suggests that the bigger the harmonic spectral deviation value, the smaller the  $t$  statistic would be, corresponding to a higher chance to show the order effect. A large harmonic spectral deviation results from irregular (or not so smooth) harmonic amplitudes (e.g., CL), which in turn is correlated with a higher likelihood of showing order effects.

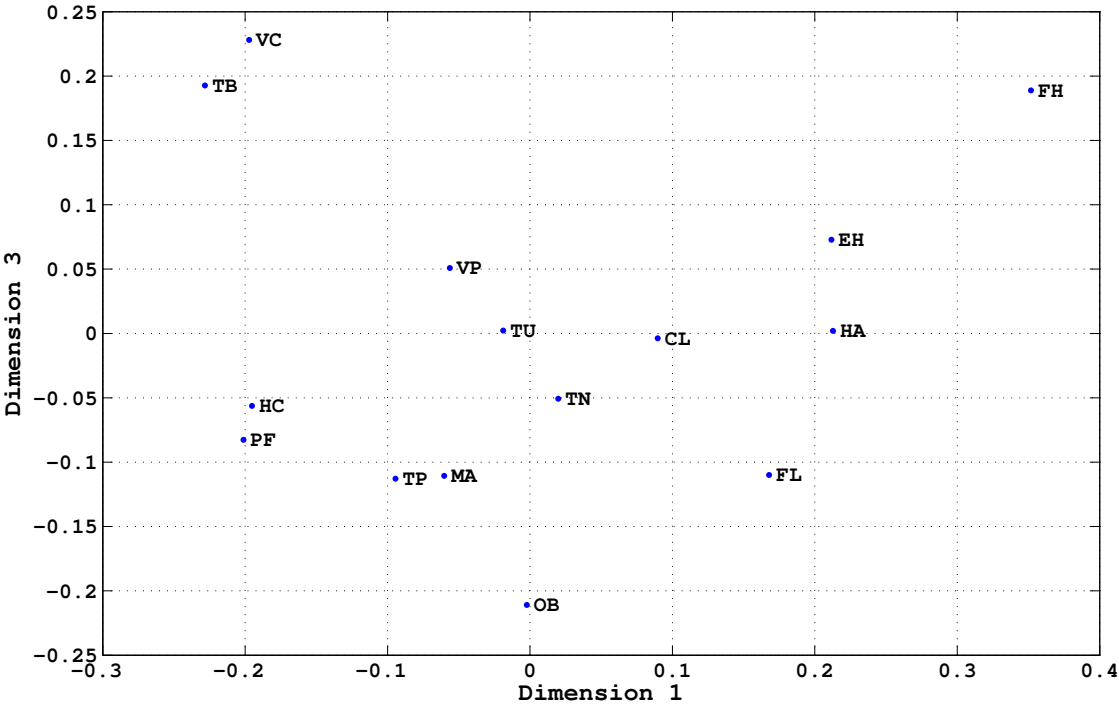
### Multidimensional Scaling (MDS) Analysis

In spite of significant effects in presentation order of timbre pairs, average data across orders were used for multidimensional scaling (MDS), which cannot deal with distance asymmetries. Following the same sequence of analysis in Experiment I, the 60 individual choice matrices were converted into 60 dominance matrices using Equation 2.2, which in turn were converted into 60 dissimilarity matrices using Equation 2.5. These dissimilarity matrices were analyzed with CLASCAL (Winsberg & De Soete, 1993), resulting in a solution with three dimensions, specificities and two latent classes. Table 2.13 compares the eight best spatial models in terms of the log likelihood, degrees of freedom, AIC and BIC values. CLASCALMC, comparisons of nested models using Monte-Carlo tests, confirmed that the three dimensional space with specificities and two latent classes was indeed the best model.

The three-dimensional CLASCAL space is presented in Figure 2.8. Unlike the one-dimensional solution from Experiment I, this three-dimensional space is a bit difficult to understand with respect to the stimulus coordinates. Acoustic correlates were calculated for each dimension of this three-dimensional space using the timbre descriptors from the Timbre Toolbox (Peeters et al., 2011), listed in Table 2.14. Notice that the correlations are moderate, which might have come from ignoring the significant order effects. This time it seems that participants tended to focus more on the timing information as the acoustic



(a) Dimensions 1 and 2



(b) Dimensions 1 and 3

Figure 2.8: A three-dimensional CLASCAL solution

**Table 2.13:** Log likelihood, degrees of freedom, and AIC and BIC values for spatial models, presented in rank order

Rank	# Dim.	Specificities	# Class.	logL	df	AIC	BIC
1	3	Yes	2	-2571	6237	5269	<b>5694</b>
2	3	Yes	3	-2565	6232	5265	5724
3	3	Yes	4	-2559	6227	5263	5756
4	2	Yes	2	-2669	6252	5433	5757
5	4	No	2	-2614	6238	5352	5770
6	2	Yes	3	-2659	6248	5421	5772
7	5	No	2	-2653	6223	5259	5779
8	3	Yes	5	-2554	6222	5264	5790

correlates of two out of the three dimensions belong to the Time descriptor group. The attack slope is defined to be the weighted average temporal slope of the energy during the attack segment, where the weights form a Gaussian distribution to emphasize slope values in the middle of the attack segment. Therefore the two may not be perfectly correlated to each other. The signal duration is the time duration between the estimated beginning and the ending of the signal, thus includes the attack duration.

**Table 2.14:** Acoustic correlates of each dimension of the CLASCAL solution

Dimension	Descriptor (Group)	$r(13)$	$p$
1	Spread (ERB Gammatone)	-.54	.04
2	Attack Slope (Time)	-.53	.04
3	Signal Duration (Time)	-.62	.01
3	Attack time (Time)	.58	.02

**Table 2.15:** Correlation of each dimension of the CLASCAL solution with the saliency dimension from Experiment I

Dimension	$r(13)$	$p$
1	-.47	.0775
2	.05	.8580
3	-.31	.2602

Table 2.15 shows the correlations between each dimension of the three-dimensional CLASCAL space and the saliency dimension obtained from Experiment I. The correlations are moderate at best here, which may also reflect the fact that this space was obtained from data ignoring the significant order effect. The negative correlations indicate that, in comparison with the measured saliency values from Experiment I, the smaller the coordinates are in this new space, the more salient the stimulus is. From the acoustic correlates in Table 2.14, we can deduce that participants judged a timbre to be salient when it had a wide spectral spread (dimension 1) and a shorter attack time and a longer signal duration (dimension 3), regardless of the attack slopes (dimension 2). This is a quite unexpected finding. First, the implication of the first dimension is that a brighter sound was considered to be more salient, which is the opposite of the interpretation of the saliency dimension obtained from Experiment I (Section 2.2).

Next, judging from the little correlation coefficient, the second dimension can be considered virtually independent of the saliency dimension from Experiment I. The saliency dimension from Experiment I also showed almost no correlation with attack slope,  $r(13) = .08, p = .79$ , which is the acoustic correlate of dimension 2 in Table 2.14. This makes the interpretation of dimension 2 difficult.

Table 2.16 lists the estimated weights in the three-dimensional CLASCAL solution for

each of the two latent classes. This time the class membership seems to come from whether a participant relies more on the saliency space in Figure 2.8 or on the specificities for saliency judgment. It is interesting to see that these class membership criteria are more polarized than those from Experiment I in Table 2.6.

It turns out that the majority of participants belong to Class 1; only five participants are assigned to Class 2. Table 2.17 lists the gender, musicianship and age of these five participants. We see roughly equal numbers of males and females while only one of them is a musician. As there are only five samples, we cannot say if the musicianship had any significant effect on the latent class membership at this time. Age seems to be insignificant either, although without certainty.

**Table 2.16:** Estimated weights in the selected three-dimensional model with specificities for two latent classes from Experiment II

Class	Dim 1	Dim 2	Dim 3	Specif
1	2.00	1.61	2.00	0.0025
2	0.00	0.39	0.00	1.9975

**Table 2.17:** Gender, musicianship and age of the five participants in Class 2

Participant No.	Gender	Musicianship	Age
13	Female	Nonmusician	40
27	Male	Nonmusician	29
38	Female	Musician	29
39	Male	Nonmusician	21
42	Female	Nonmusician	20

## 2.4 Conclusion

Timbre saliency, which is the attention-capturing quality of timbre, has been proposed and measured in two different experiments. The first experiment used a tapping technique, which is an implicit measure based on the hypothesis that a more salient timbre would capture listeners' attention and therefore be chosen as the strong beat more often. The second experiment employed a direct comparison method with the 2AFC paradigm. Surprisingly, the results from the two experiments were quite different from each other.

In the first experiment, data from 40 participants showed no effect of gender, musicianship or age on the choice of strong beats, meaning that the variety of responses came from each pair of timbres combined. No effect of presentation order was observed either. The choice data were transformed into dominance data, which in turn were transformed into proximity data, which were then processed with CLASCAL, resulting in a one-dimensional solution for the 15 instrument sounds with two latent classes and specificities. The latent classes were not according to gender, musicianship or age, but according to how participants weighed the saliency dimension and specificities in their judgments.

The saliency dimension seems to have three regions – low, middle and high. One half of the stimuli were positioned in the middle region around the zero point, which may imply that they are more-or-less equally salient (although some are positioned lower or higher than others) and therefore a strong beat in the sequence with a pair of such timbres might have been chosen almost randomly. The odd-even harmonic ratio showed the highest correlation with the saliency dimension among the descriptors out of the Timbre Toolbox. From the correlation analysis, it appears that participants used descriptors related to harmonic energy distributions rather than temporal information to decide on a strong beat. It was also interesting to see that none of the well-known timbre descriptors for the dissimilarity



perception (spectral centroid, attack time, spectral flux or spectral variation) showed a high correlation with the saliency dimension. This may suggest that listeners used different spectral properties for saliency judgments from those for dissimilarity judgments.

The mean dominance, which is an average probability of one timbre being chosen as the strong beat regardless of the other timbre it is combined with, was also analyzed in terms of the distributions and the medians. The distributions varied quite a bit depending on the instrument, indicating that some timbre's dominance value did not change very much as a function of the other timbre with which it was combined, whereas some other timbre's dominance value was affected much by different timbres. The acoustic correlate of the median dominance values turned out to be the harmonic skew, also confirming that participants chose strong beats based on comparisons of spectral properties.

One possible complication with the task of tapping in Experiment I that we did not foresee was that the perceptual ABAB sequences are essentially in duple rhythms, which are often used in various types of music including rock, jazz and Southeast Asian music. In these cases, strong beats tend to be accompanied with a bass drum adding more low-frequency energy to the sound, whereas weak beats played with a snare, which adds more high-frequency energy. As some participants mentioned in their post-experiment interviews, these implicit schema, obtained from many years of exposure to such use of a duple rhythm, highly affected the choice of a strong beat. This confound with the "rhythmic saliency" is also reflected in the correlation that the saliency dimension is negatively correlated with the harmonic spectral centroid.

As an effort to separate this confound, we carried out another experiment to measure timbre saliency. This time a direct comparison technique was employed using the 2AFC paradigm. Even though it succeeded in removing the rhythmic saliency, this approach resulted in many problems. First, a strong effect of presentation order was found in nine

out of 15 instruments, which was not found in the first experiment. This could be due to arbitrary contexts with respect to different pairs of timbres to compare. Participants were asked to decide which of the two given timbres “grabbed their attention better,” without any specific context that ran consistently through the entire experiment. This might have required participants to imagine a context for comparison, which may have varied from trial to trial. It is also interesting to observe that some timbres were judged to be more salient when they were presented first, whereas others were favoured when they were presented later, for some unknown reasons.

Following the same process of data transformations, the 60 individual choice data were converted into dominance data, which were then transformed into dissimilarity data ignoring the order effect. These dissimilarity data were then analyzed with CLASCAL to result in a solution of three-dimensions with specificities and two latent classes. Two dimensions in this new three-dimensional CLASCAL space were only moderately correlated with the saliency dimension from Experiment I, which made us suspect the applicability of this new space. The acoustic correlates were obtained for the three dimensions, and the interpretations of the acoustic correlates were problematic. These all add to the suspicion that this space may not be actually meaningful.

Since the result from Experiment II has more problems than Experiment I, future studies will be built on the findings from Experiment I. The important lesson learned from Experiment II is the importance of the experimental paradigm used to examine a certain hypothesis. We started with the same hypothesis for both experiments, but using two different paradigms resulted in two different findings. Also, saliency depends on the context, and it seems that the two experiments set two different contexts to measure timbre saliency, the result of which may not be entirely compatible.

This is the first step towards understanding timbre saliency, which might be one of the

factors underlying well-known examples in orchestration treatises such as combinations of certain instruments for a timbral effect or the use of an instrument as a melody carrier or another as to support the “body”. The work is by no means complete. Saliency must be a function of context – a red dot surrounded by green dots is highly salient, but when it is surrounded by red dots it is no longer salient. The red dot might still be salient if there are many red triangles, although not in the same way as with green dots. By similar logic, what we measured in these experiments was the effect of timbre saliency in the context of each experiment. A third experiment might have yielded a different result, especially if it employed another paradigm. Accepting the fact that the saliency dimension from Experiment I will not be perfect in all contexts, we still move forward with this research knowing that it does reflect saliency in some specific context.

The next step will consider a more realistic scenario with concurrent sounds. When two sounds happen at the same time, they tend to be blended to a certain degree, although the degree of blending depends on the particular sounds. How does timbre saliency play a role in blending? It will be answered by a study of the perceived degree of blending in terms of timbre saliency. We will consider the composite sounds by combining a pair from the same 15 timbres. The results will be compared with previous findings in the blending perception literature.

(This page is intentionally left blank.)

## Chapter 3

# Perceived Blend of Unison Dyads

### Abstract

There are orchestration handbooks with much-used examples of what instrument combinations work and what others do not, although their underlying principles have not yet been scientifically studied. We conjecture that timbre saliency might be a factor that could explain why these examples work.

A rating experiment was conducted with 60 people. Stimuli were composite sounds made of two concurrent unison timbres varying in degree of timbral and saliency differences. Participants were asked to rate each composite sound's degree of blend on a continuous scale between "very blended" and "not blended".

The average blend across all participants turned out to be mildly correlated with the sum, minimum and maximum of saliency values of two individual timbres. This means that a highly salient sound will not blend well. The best acoustic correlate to describe the average blend is the minimum attack time of the two individual timbres, which alone explains 57% of the variance. It means the longer the minimum attack time is, the better

the blend. This confirms previous observations that a sound with a longer attack tends to blend better. Previous findings that sounds with lower spectral centroids are likely to blend better were also confirmed.

### 3.1 Introduction

Orchestration is the art of combining various instruments to realize a sound image that the composer wants to portray. Berlioz claimed that “... this art [composition] can no more be taught than the writing of beautiful melodies or beautiful chord progressions or original or powerful rhythmic patterns.” (p. 6 of [Berlioz, 1855](#)). What can be taught, according to Berlioz, is how to combine instruments for a certain effect using examples by great composers. This might explain why the study of orchestration is often based on descriptions of various instruments, as well as good examples to mimic and bad examples to avoid by well-known composers.

It is rather strange to find that there have not been many perceptual studies of orchestration, especially from the timbral point of view, because timbre is an important element in orchestration. [Kendall and Carterette \(1993\)](#) studied the blend of five wind instruments in six musical contexts. Participants were asked to rate the blend of each combination of two instrument tones in terms of multiplicity, with one end marked “one” and the other end “two” denoting the number of independent sound objects perceived. They found that blend was inversely correlated with the correct identification rate of the individual timbres in each combination. They also reported that the blend ratings showed a high correlation with both “nasality” and “brilliance / richness,” which were two acoustic correlates of the two-dimensional timbre similarity space found from their earlier work ([Kendall & Carterette, 1991](#)). They also observed that the blend of highly nasal sounds was poor; in

fact, the oboe consistently showed the worst blend.

Sandell (1995) used 15 timbres from Grey (1977) to study blend and the most significant acoustic factors determining it. He found that blend worsened as each of the following four values increased: composite centroid (which is the sum of the centroids of two timbres), centroid difference, attack asynchrony and offset asynchrony. The centroid measures reflect in fact that a bright sound does not blend well with other sounds, however bright or not the other sound may be. Attack and offset asynchronies are known to be important for segregation in musical ensembles (pp. 490 – 491 of Bregman, 1990), that the soloists “should not be synchronous with those of the rest of the ensemble” for “maximum distinctiveness.” This is based on the fact that the sounds with similar onset and offset patterns tend to get grouped together into the same auditory stream in auditory scene analysis (Darwin, 1981).

Sandell’s observations above are consistent with findings in timbre dissimilarity perception studies (Grey, 1977; Grey & Gordon, 1978; Krumhansl, 1989; Iverson & Krumhansl, 1993; McAdams et al., 1995; Lakatos, 2000; Caclin et al., 2005) where attack time and spectral centroid were two of the most significant acoustic correlates of timbre perception.

Sandell (1995) also noticed that in some cases the individual timbre’s spectral centroid mattered more than composite centroid or centroid difference. For example, a timbre with high spectral centroid would not blend well with other sounds even when the centroid difference is small. He also noted that this is consistent with the way singers sing in a choir – they “darken” their tones and they must have known from their experience (rather than science) that darker tones blend better (Goodwin, 1989). This is also supported by Bregman (1990, p. 521), who wrote “Higher partials would fuse better into a complex tone if their amplitudes were less than those of the lower partials,” meaning that if two timbres have the same partials, the one with more energy in the lower spectrum (i.e. lower spectral centroid) will blend better than the one with more energy in the higher spectrum.

More recently, [Tardieu and McAdams \(2012\)](#) considered the perception of dyads of impulsive and sustained instrument sounds, in contrast to the earlier two studies considering only the sustained instruments. This therefore can be thought of as a generalization of the findings by [Kendall and Carterette \(1993\)](#) and [Sandell \(1995\)](#). [Tardieu and McAdams \(2012\)](#) reported that “longer attack times and lower spectral centroids increased blend.” They pointed out that since onset asynchrony contributes to streaming of simultaneous sounds ([Darwin, 1981](#)), a longer attack time may make it difficult to identify the onset, therefore resulting in less streaming and more blending. They also separated the perceived degree of blend from the emergent timbre and found that the overall blend was determined by the impulsive sound more than the sustained, whereas the emergent timbre was more influenced by the sustained sound.

Timbre saliency is defined to be the attention-capturing quality of timbre. In [Chapter 2](#), we measured saliency differences of pairs of 15 timbres using a tapping technique. The result from CLASCAL ([Winsberg & De Soete, 1993](#)) was a single timbre saliency dimension with two latent classes and specificities. The timbre saliency dimension seems to have three regions – lower, middle and higher. The timbres in the lower region would almost never be more salient than those in other two regions. On the opposite side, the timbres in the higher region would be more salient than others most times. As the saliency dimension was obtained from a comparison of a sequence of isolated notes, how can it explain the blend when these timbres are combined concurrently? Would a highly salient timbre not blend well with other sounds since the definition of saliency requires little blend with its surroundings? Could timbre saliency be an important factor determining the perceived blend like attack time and spectral centroid? We aim to answer these questions in this chapter.

This chapter presents a rating experiment to study the perceived degrees of blend in



concurrent unison notes in terms of timbre saliency measures as well as other acoustic features from the Timbre Toolbox (Peeters et al., 2011). Stimuli are composites of 15 isolated instrument notes on the pitch C4, which were used in Chapter 2. As the composite sounds are concurrent unison blends of those isolated notes, it must be a situation where the maximum degree of blend could happen between two different instruments. This experiment is expected to verify the previous findings in the blending literature (Kendall & Carterette, 1993; Sandell, 1995; Tardieu & McAdams, 2012) and to examine the implication of the timbre saliency dimension that was obtained from Experiment I in Chapter 2. The hypothesis is that a highly salient timbre will not blend well with other sounds, regardless of how salient the other sounds may be.

## 3.2 Experiment III: Blending of Unison Dyads

### 3.2.1 Methods

#### Participants

Participants ( $N = 60$ ) were recruited from a classified advertisement on the McGill University website. There were equal numbers of males and females, and musicians and non-musicians, based on self reports. Musicians are defined as those with four or more years of musical training (instruments, voice, composition, recording, etc.), during which time they spent at least five hours per week on their training (including lessons and practice). All participants passed the preliminary hearing test before the experiment (International Organization for Standardization, Geneva, 2004; Martin & Champlin, 2000) to make sure that they were able to hear within 20 dB HL at frequencies 250, 500, 1000, 2000, 4000 and 8000 Hz. Their ages ranged from 19 to 44 with a median of 23 years. They were paid \$10 upon completion of the experiment, which took about 45 minutes including the time for

filling out a background questionnaire and debriefing.

## Stimuli

The stimuli were composite sounds generated from 15 Western orchestral instrument sounds from the Vienna Symphonic Library ([Vienna Symphonic Library GmbH, 2011](#)), all of which were equalized in terms of pitch (C4), loudness and effective duration to minimize the impact of these parameters in the experimental task. The 15 instruments are clarinet (CL), English horn (EH), French horn (FH), flute (FL), harp (HA), harpsichord (HC), marimba (MA), oboe (OB), piano (PF), tubular bells (TB), trombone (TN), trumpet (TP), tuba (TU), violoncello (VC) and vibraphone (VP). There were 105 composite sounds generated for the experiment by combining all pairs of the 15 above-mentioned instrument sounds. Four more instrument sounds – AF (alto flute), BS (bassoon), CE (celesta) and VN (violin) – were used to create composite sounds for the training session. Sound selection and equalization processes are described in detail in Appendices A and B.1, respectively. Figure 3.1 shows the ranges and distributions of saliency measures (sum, difference, minimum and maximum) of the 105 composite sounds. The horizontal axes show the range of each measure and the vertical axes the number of samples that belong to each bin (or subrange) of the horizontal axis.

The 15 timbres’ individual saliency values are listed in Table 2.4 in a decreasing order. There seem to be three timbre saliency regions, as indicated in Table 2.4. The “middle” region timbres are all crowded around zero, suggesting that they might be pretty much equivalent to one another in saliency. But the timbres in the lower or upper regions have distinctive values within each region that seem to indicate an internal hierarchy within the groups. These timbres are not as interchangeable as those in the middle region.

Composite sounds were created by matching perceptual onsets, rather than physical

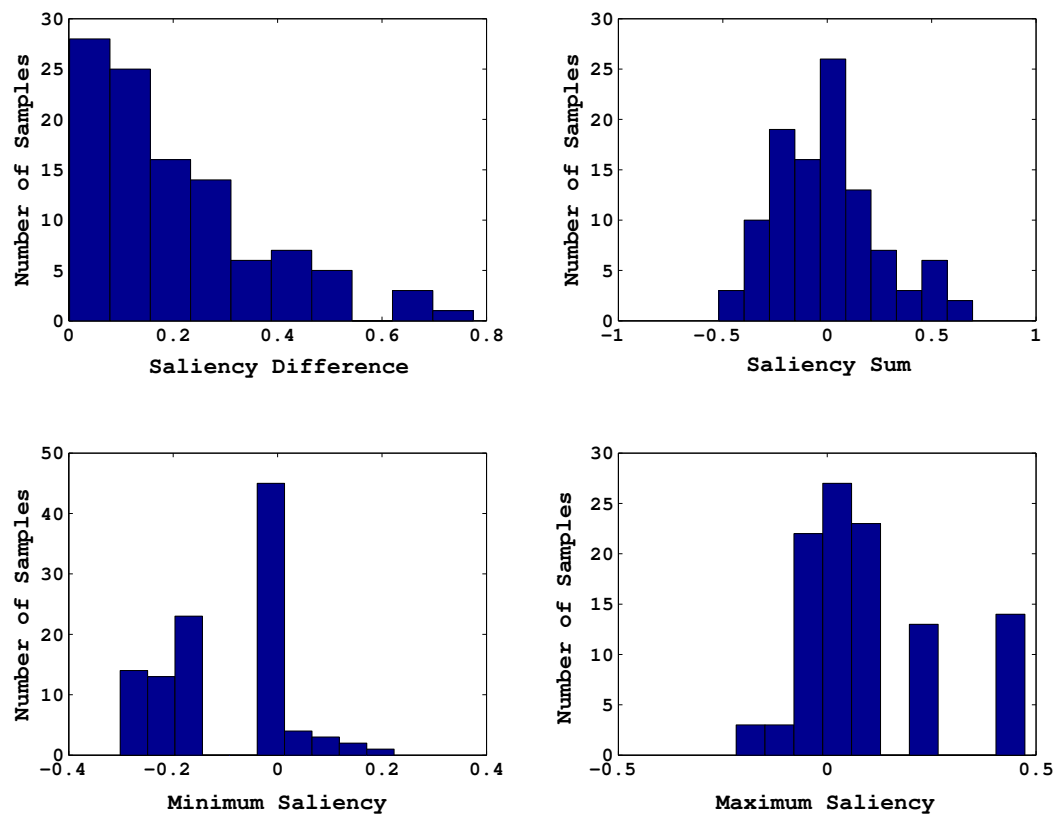
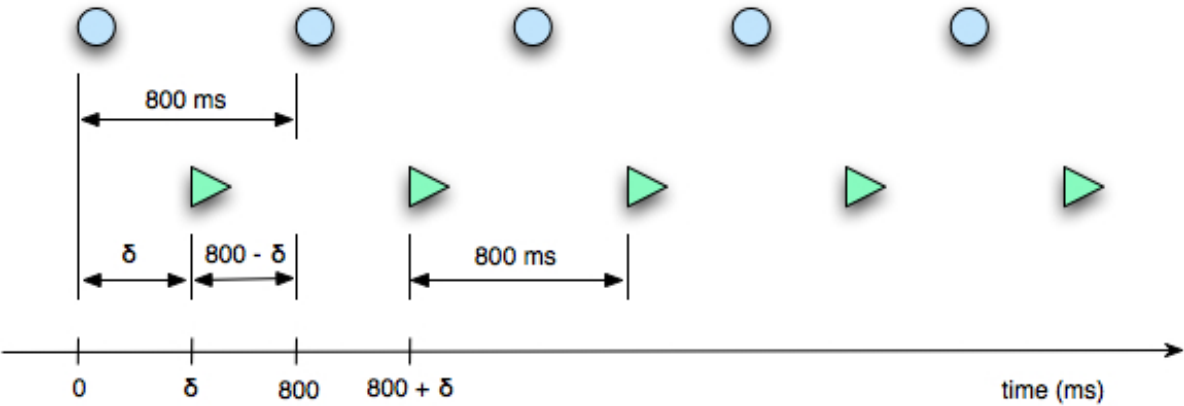
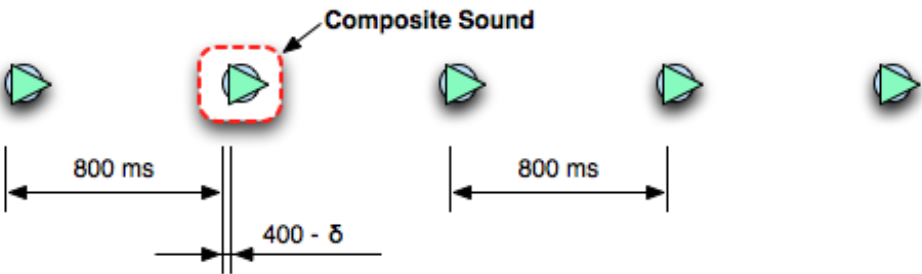


Figure 3.1: Ranges and distributions of saliency measures of the composite sounds



(a) Combining two isochronous sequences into one perceptually isochronous ABAB sequence



(b) Generation of a composite sound

**Figure 3.2:** Generating composite sounds from the synch process

onsets (as in [Tardieu & McAdams, 2012](#)), of the two individual sounds based on the isochronous sequence generation pilot experiment (see [Appendix B.2](#) for details). The underlying logic here is that if a delay  $\delta$  was used to combine two mono-timbral sequences AA and BB into a perceptually isochronous ABAB sequence, both of which have 800 ms IOI, then the delay of  $400 - \delta$  should match the perceptual onsets of A's and B's on top of each other, as depicted by [Figure 3.2](#). First, [Figure 3.2a](#) shows how an ABAB sequence was generated using a delay  $\delta$ , which is the time between the physical onset of the first sound of the AA sequence and the physical onset of the first sound of the BB sequence.

Now, let us consider a special case where the timbres A and B are the same. Then  $\delta$  must be exactly 400 ms to combine the two sequences into one sequence, the rhythm of which would be twice as fast. When A and B are different timbres (which was the case in [Appendix B.2](#)), the delay  $\delta$  deviates from 400, depending on the attack patterns of the pairs of timbres combined together. When we combine the sequences AA and BB with the delay  $400 - \delta$ , they are perceptually on top of each other, creating a new mono-timbral sequence of CC where C is the new composite timbre of A and B, which is pictured in [Figure 3.2b](#). We listened to composite sounds from both the physical and perceptual onset matchings and verified that the perceptual onset-matched composite sounds indeed have a higher degree of blend due to more synchronous onsets.

## Procedure

The experiment was carried out in a sound-attenuated booth ([Industrial Acoustics Company](#), model 1203) in the Music Perception and Cognition Laboratory of the Schulich School of Music of McGill University. Mono sound signals, to further eliminate any undesired cue for separation, were amplified by a Grace m904 stereo monitor controller and the average level was 60 dBA by a Brüel & Kjær type 2250 sound level meter coupled with a Brüel &

Kjær type 4153 artificial ear.

Prior to the experiment, every participant went through a screening audiogram to make sure their hearing met our minimum criteria ([International Organization for Standardization, Geneva, 2004](#); [Martin & Champlin, 2000](#)). After the audiogram, participants were presented with an instruction sheet that described what was expected of them. The experiment had three parts. First, there was a familiarization block in which participants listened to all composite sounds to be used in the experiment one after another in one sequence. This was to make sure that they became aware of the wide range of degrees of blend in the composite sounds, so that they could have an idea of extreme cases for the rating interface. The second was a training block with six composite sounds made of four training timbres (AF, BS, CE, and VN) so that participants could get used to the task and the graphic user interface in Figure 3.3, developed in PsiExp ([Smith, 1995](#)) on an Apple PowerPC G5. Participants could ask the experimenter questions after the training block. The experimenter was present in the sound booth up to this point and left after making sure participants understood the task and there were no more questions. Then participants performed the third part with 105 test trials in random order.

For each trial, participants listened once to a composite of two simultaneous, unison instrumental sounds varying in degree of timbral difference and saliency difference and were asked to rate the degree of blend on a continuous scale by placing a marker between “not blended” and “very blended” as shown in Figure 3.3. Participants submitted their ratings by clicking on the “ok” button, at which point the position of the marker was converted into a number between 0 and 1 before proceeding to the next composite sound. On this scale, 0 was mapped to the “not blended” end and 1 to the “very blended” end. No repeat was allowed nor could participants go back to earlier sounds and change the rating once submitted.

21/105

How much do these two sounds blend into one?

not blended

very blended

ok

**Figure 3.3:** Screenshot of blend rating experiment

### 3.2.2 Results

#### Biographical Factors

First, we wanted to see if there was any systematic variation in the data related to gender, musicianship or age of the participants. For this, a repeated-measures analysis of variance (ANOVA) was carried out on blend rating data from 60 participants in SPSS. Dependent variables (DVs) were each participant's blend rating on each of 105 composite sounds. Gender and musicianship were between-subjects independent variables (IVs) and age was a covariate. Stimuli were the repeated factor. Due to the limitation in SPSS that allows only up to 99 repeated measures in an ANOVA, the analysis had to be done in two disjoint sets. The first set considered the blend rating data for composite sounds 1 through 53 and the second set for composites 54 through 105. Neither set showed an effect of gender,  $F(1, 556) = 0.312, p = .579$ ;  $F(1, 55) = 1.311, p = .257$ , for the two re-

spective sets, musicianship,  $F(1, 55) = 0.633, p = .430$ ;  $F(1, 55) = 0.519, p = .475$ , or age,  $F(1, 55) = 0.961, p = .331$ ;  $F(1, 55) = 0.039, p = .844$ , on blend ratings.

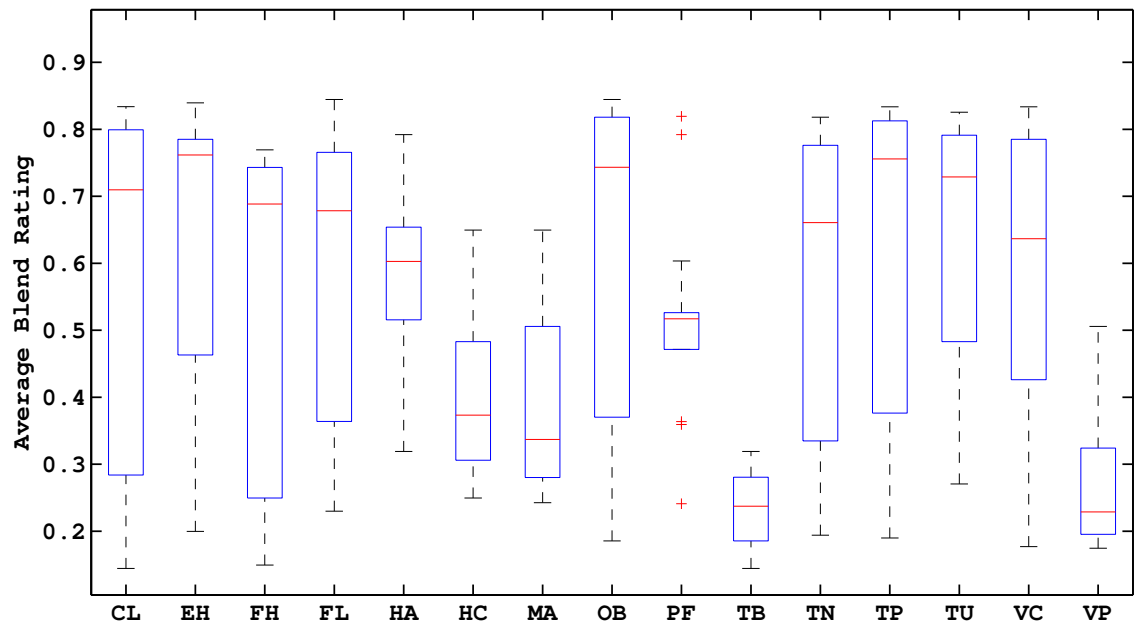
This means that the only factor affecting the variance in responses was the instrument combination. As it turns out, stimuli did not meet Mauchly's sphericity condition,  $\chi^2(1377) = 2694.235, p < .000001, \epsilon = 0.182$ ;  $\chi^2(1325) = 2402.409, p < .000001, \epsilon = 0.266$ , so the main and interaction effects of stimuli were corrected with the Greenhouse-Geisser epsilon. The main effect of stimuli was significant,  $F(9.484, 521.619) = 4.728, p < .00001$  for the first set;  $F(13.561, 745.862) = 3.087, p < .0005$  for the second set, but none of the interactions were significant. The blend rating data for each instrument pair were averaged across all participants for further analysis.

### Average Blend Ratings

As there was no difference coming from gender, musicianship or age of the participants, the next step is to see the distribution of blend ratings averaged across all participants. Do they look homogeneous across all composite sounds, or are there any variable yet distinctive patterns in small groups of them? What could be the distinguishing factor between composites with higher average blend ratings from those with lower average blend ratings? This section aims to answer these questions using the analysis of average blend judgments.

Figure 3.4 shows the distribution of blend ratings averaged over participants for all pairs involving each instrument on the horizontal axis. For example, the box anchored at the 'CL' column on the horizontal axis is the distribution and range of blend ratings of all composite sounds with CL in it, averaged across all participants. The red line in a box represents the median. The edges of a box show the upper and lower limits of the interquartile range (IQR). The whiskers outside a box designate the range of values within





**Figure 3.4:** Distribution of average blend ratings across participants for all pairs of timbres involving each instrument on the horizontal axis.

$\pm 2.7\delta$ , where  $\delta$  is the standard deviation, and the outliers are noted with red crosses outside of these whiskers.

A longer box indicates that the average blend rating varies more over all composites that include the corresponding timbre, which in turn signifies a bigger effect of ‘the other’ timbre in combination. On the other hand, a smaller box means a more limited range of blend ratings, implying that ‘this’ timbre has a more significant impact on the perceived blending of the composites.

As previous research reported the correlation between attack patterns and the perceived blend (Sandell, 1995; Tardieu & McAdams, 2012), we grouped the 15 timbres into two attack types of ‘longer’ and ‘shorter’ and tried to find a possible systematic difference in the average blend ratings between the two attack types. Is this attack type-based difference significant in box sizes (IQRs)? Would the medians show the same difference in terms of the attack types?

**Table 3.1:** One-way ANOVA result of average blend ratings between percussive attack group (HA, HC, MA, PF, TB, VP) and gradual attack group

Measures	$F(1, 13)$	$p$
IQR	60.207	< .0001
Median	38.142	< .0001

To answer these questions, a closer examination using one-way ANOVA was performed, which is summarized in Table 3.1. DVs are IQR and median values for each column of the average blend ratings shown in Figure 3.4. IV is the group designation: either the group with percussive attacks (HA, HC, MA, PF, TB, VP) or the rest of timbres with gradual attacks. The average IQR showed a significant difference between the two groups: the average IQR of the percussive attack group was 0.137, which was less than half of

the average IQR of the other group (0.414). Lower variability (i.e., smaller IQR) of the percussive attack group suggests that the ratings did not vary much because these timbres tended to dominate the blend, whereas greater variability of the gradual attack group implies that the degree of blend depended more on the other timbre. The median values in Figure 3.4 were also significantly different: percussive attack group median = 0.383, other group = 0.707. These lower medians suggest that the overall perceived degree of blend was poor for composites that involved sounds with percussive attacks. To summarize, the timbres with percussive attacks seem to have dominated perceived blend, regardless of the other timbre with which it was combined. The overall perceived blend was rather poor in comparison with the composites with the timbres with two gradual attacks. This confirms the previous findings by Tardieu and McAdams (2012) that sounds with longer attacks tend to blend better.

**Table 3.2:** One-way ANOVA result of average blend ratings among three timbre saliency regions

Measures	$F(2, 12)$	$p$
IQR	3.416	.067
Median	1.837	.201

A similar ANOVA was carried out on the average blend ratings of 105 composite sounds in terms of three timbre saliency regions to see if there is any systematic difference in blend ratings in terms of the saliency values. The result is presented in Table 3.2. The only difference that approached significance was observed for IQR. The average IQRs were 0.44, 0.28 and 0.20 for the timbres in the low, middle and high saliency regions, respectively. A Bonferroni test reveals that the IQRs of the low and high saliency regions were almost significantly different ( $p = .078$ ) and nothing else. This suggests that composites with more

salient timbres had less variance in blend ratings, possibly because these timbres dominated the other timbres in combination. However, the medians did not show any significant effect of timbre saliency region, which suggests that the analysis in terms of three saliency regions may not be effective due to the loss of information by collapsing 15 saliency values into three categories.

To examine a possible monotonic nonlinear relationship between the blend ratings and the 15 saliency values, Spearman’s rank correlations were calculated between the saliency values and the IQR and median values. The result is summarized in Table 3.3.

**Table 3.3:** Spearman’s rank correlation between the 15 saliency values from Experiment I and the 15 IQRs and medians of the average blend ratings

Measures	$\rho(13)$	$p$
IQR	-0.51	.054
Median	-0.38	.165

The result in Table 3.3 is not too different from that in Table 3.2. The range of average blend ratings were negatively correlated with the saliency values from Experiment I, suggesting that a salient timbre would not blend well and at the same time dominate the perceived blend with others. The median values of average blend ratings also showed a mild negative correlation with saliency, confirming that high saliency is correlated with less blending.

### Timbre Saliency and Audio Descriptors of Timbre

In Section 2.2, a tapping technique was used on perceptually isochronous ABAB sequences to measure timbre saliency on 15 instrument timbres, the pitch, loudness and effective duration of which were all equalized. The hypothesis was that a more salient timbre

would catch listeners' attention more easily therefore be chosen as the strong beat more often. Data from 40 participants, balanced in gender and musicianship, showed no effect of gender, musicianship, or age on the choice of strong beat timbre. The choice data were transformed into proximity measures, which were analyzed with an MDS algorithm, CLASCAL (Winsberg & De Soete, 1993) The result was a one-dimensional solution with 2 latent classes and specificities. The exact saliency values are listed in Table 2.4.

As saliency requires little blend with neighbouring objects, we hypothesized the perceived blend would be inversely related to saliency. Hence the average blend ratings were analyzed in terms of four saliency measures: absolute difference, sum, minimum and maximum of saliency values of the two individual timbres in a given composite sound. To investigate a possible linear relationship between these saliency measures and the average blend ratings, Pearson correlation coefficients were calculated.

**Table 3.4:** Pearson correlations between the average blend and saliency measures of 105 composites

Saliency Measures	$r(103)$	$p$
Sum	-.34	< .0005
Maximum	-.30	.002
Minimum	-.26	< .008
Difference	-.12	.222

Table 3.4 reports weak negative correlations observed between the average blend and saliency measures. Absolute values of saliency differences on the last row show a non-significant negative correlation. The other statistics are all significant, although saliency measures account for 11.6% of the variance in blend ratings at best. The best correlate among them is the saliency sum, followed by the maximum then the minimum of the two.

These mean that for a better blend it is important to have both components be less salient. It also implies that a highly salient sound will not blend well, as we expected, although the degree of correlation is weak. In addition, it is the individual component's saliency level and the sum of saliencies of all sounds combined that contribute to the overall degree of perceived blend, rather than the difference in saliency between individual sounds.

One possible reason behind this weak correlation could be that saliency is a function of context and the context within which the timbre saliency was measured may not be highly related to the blending context. If this is indeed true, then it might suggest that saliency judgments in one context may not be easily transferrable to another context. It could also imply that the importance of timbre saliency may be *modulated* with the task. Timbre saliency may be the most prominent feature in another context, even though it may not be so in this current context.

The weak correlations between timbre saliency and the average blend ratings may be explained by the loss of information in the MDS analysis, especially in comparison with the raw dominance scores. However, a visual comparison of the dominance distributions in Figure 2.2 and the average blend ratings in Figure 3.4 seems to suggest that these two graphs may not have much in common. Two vectors were created from the average blend ratings and the average dominance values for a correlation analysis. Note that there were 105 blend ratings and 210 dominance judgments, so the blend rating of the same timbre pair  $i$  and  $j$  were used twice to compare with the dominance judgments in both orders. The Pearson correlation of the two vectors were negligible ( $r < 10^{-16}$ ), leading us to conclude that the two vectors are independent of each other. What then can provide a satisfactory explanation of the blend ratings as neither timbre saliency statistics nor the average dominance values do?

Pearson correlations were calculated between the average blend ratings and four statis-

**Table 3.5:** Pearson correlations between average blend and timbre descriptors.

Descriptor Group	Descriptor Statistic	$r(103)$	$p$
Time	minimum temporal centroid	.66	< .0001
ERB FFT	sum of spectral variations	−.62	< .0001
ERB FFT	maximum spectral variation	−.62	< .0001
ERB FFT	difference of spectral variations	−.60	< .0001
Harmonic Spectrum	sum of noisinesses	−.60	< .0001
Harmonic Spectrum	maximum noisiness	−.60	< .0001
Harmonic Spectrum	difference of noisinesses	−.59	< .0001
Magnitude Spectrum	maximum spectral centroid	−.46	< .0001
Magnitude Spectrum	sum of spectral centroids	−.44	< .0001
Magnitude Spectrum	difference of spectral centroids	−.35	< .0005
Magnitude Spectrum	minimum spectral centroid	−.22	.0214
Time	minimum log attack time	.42	< .0001
Time	sum of log attack times	.30	< .005
Time	maximum log attack time	.19	.0524
Time	difference of log attack times	−.01	.9210
Time	minimum attack time	.25	.0115
Time	sum of attack times	.22	.0237
Time	maximum attack time	.17	.0808
Time	difference of attack times	.08	.4414

tics (sum, difference, minimum and maximum) of each timbre descriptor from the Timbre Toolbox (Peeters et al., 2011). Table 3.5 lists the seven best descriptors with the highest correlation values, as well as the correlation coefficients of the measures of spectral centroids, log attack times, and attack times. The best acoustic correlate to describe the average blend turned out to be the minimum temporal centroid of the two individual timbres,  $r(103) = .66, p < .0001$ . Temporal centroid is the time position corresponding to the centre of gravity of the energy envelope, which is useful in distinguishing percussive sounds from sustained sounds (Peeters, McAdams, & Herrera, 2000). Since the correlation coefficient is positive, it means that a better blend is achieved with two sounds both of which have larger temporal centroid, which essentially refers to non-percussive sounds. This is in agreement with previous reports by Tardieu and McAdams (2012), as well as what we have seen in Section 3.2.2 that the timbres with more gradual attacks were on average judged to have better blending. The minimum temporal centroid alone explained 44% of the variance in blend data.

Three descriptors from the ERB FFT group also showed high correlations, although in a negative direction. ERB stands for *Equivalent Rectangular Bandwidth*, and it is a model of the *critical band* (CB) that Moore and Glasberg (1983) proposed to approximate peripheral auditory filtering. There are two types of ERB implementations, one uses the Fast Fourier Transform (FFT) and the other Gammatone filter banks (Patterson et al., 1992). Spectral variation reflects the amount of variation of the spectrum over time (Krimphoff et al., 1994). Three statistics (sum, difference and maximum) of spectral variation from the ERB FFT model show a fairly high correlation with the average blend ratings. Negative correlation coefficients suggest that the smaller the descriptor values were, the better the observed blend. In other words, a higher blend was associated with smaller values of the sum, difference and minimum of the spectral variation of the two underlying components,



suggesting that a better blend was achieved when both sounds had little spectral variation. Since timbres with percussive attacks would have a number of inharmonic partials in the beginning of sounds, which would die down quickly, they would have larger spectral variation than timbres with gradual attacks. Hence, what the spectral variations measures imply for blend ratings seems to agree with what minimum attack and minimum temporal centroid suggest in terms of the effect of the attack patterns on the spectral content.

The other three best correlations come from the measures in the harmonic spectrum. Noisiness is the ratio of the noise energy to the harmonic energy, where noise energy is defined as the difference between the total energy and the harmonic energy. A high noisiness value suggests that the signal is highly inharmonic. As the sum, maximum and difference of noisiness values are all moderately correlated with the average blend ratings in Table 3.5, they basically tell the same story that the spectral variation measures do: the smaller the noisiness and spectral variation values of individual sounds are, the better blend they would create. To summarize, a sound would contribute to a better blend if it is highly harmonic and non-percussive with little spectral variation.

Spectral centroid has been reported as a main contributor to the perception of blend. It is the spectral centre of gravity and has been shown to be important in timbre perception (Grey, 1977; Grey & Gordon, 1978; Krumhansl, 1989; McAdams et al., 1995; Lakatos, 2000; Caclin et al., 2005). The middle four rows of Table 3.5 show the Pearson correlation results between the average blend rating and measures of spectral centroids (difference, sum, minimum and maximum) in the magnitude spectrum.

The best correlation is observed with the maximum of the two spectral centroids. The mild negative correlation ( $r(103) = -.46$ ) indicates that a composite of two sounds tends to achieve a better blend when the maximum spectral centroid of the two is low. All four measures of spectral centroid relations exhibit significant negative correlations with the

average blend rating, which indicates that composites of two sounds blend better when both sounds have low spectral centroid and the two spectral centroids are close to each other in frequency. This supports previous findings by [Sandell \(1995\)](#) and [Tardieu and McAdams \(2012\)](#) that sounds with lower spectral centroids are likely to achieve better blending.

Another descriptor that has been previously related to the perceived blend is attack time. The four statistics of attack times and log attack times are listed on the low eight rows of [Table 3.5](#). The minimum log attack time seems to be the best descriptor for the average perceived blend. These statistics suggest that a timbre with a longer (log) attack time will blend better with other sounds, which is in line with what the high correlation with the minimum temporal centroid implies.

### 3.3 Conclusion

Orchestration is the art of combining timbres. As a first step towards understanding scientific principles behind orchestration, the perception of blending of two concurrent unison timbres was studied.

A rating experiment was carried out with 60 participants to judge the degree of blend of composite sounds formed by combining pairs of timbres from 15 instrument sounds used for the timbre saliency experiment in [Chapter 2](#). The composites were created by adding two individual sound files together with aligning the perceptual onsets rather than physical onsets, making it blend better than the composites created by aligning the physical onsets. No participant mentioned being able to hear double attacks in the composites with percussive instruments, although they would rate them to be less blended. This confirms that the perceptual onset matching contributed to a better degree of blending even with

percussive timbres. The rating data showed no significant effect of gender, musicianship or age.

The blend ratings were then averaged across all participants for further study. Composites containing timbres with percussive attacks tended to be judged as less blended than those with gradual attacks. The percussive-attack timbres tended to have less variance in the rating data, suggesting these timbres dominated the perceived blend.

The average blend ratings were analyzed with measures (sum, absolute difference, minimum and maximum) of timbre saliency values as well as timbre descriptors from the Timbre Toolbox (Peeters et al., 2011). Mild negative correlations were observed with timbre saliency measures, which confirmed our hypothesis that timbre saliency would have a negative relationship with perceived blend. However, the low degree of correlation with timbre saliency measures was disappointing as the best one (sum of two saliency values) explained only 11.6% of the variance in the data. As saliency is context-dependent, this low correlation might be caused by the fact that the tapping context within which the timbre saliency was measured may not be highly related to the blending context of this experiment. More studies on perceptual saliency and context transferability are required to understand the exact reason behind this low correlation.

Mild correlations between the average blend and timbre saliency measures suggest that there may be other more significant factors that affect the blend of two unison instrument sounds, such as the attack time and spectral centroids, two of which were reported in the blend perception literature. Among the acoustic descriptors obtained from the Timbre Toolbox, the best acoustic correlate of the average blend ratings turned out to be the minimum temporal centroid, the time position of the centre of gravity in the energy envelope, explaining almost half of the variance in data. Taken together with the mild correlations between the average perceived blend and the (log) attack time measures, these

results confirm previous reports that sounds with slower attacks tend to achieve a better blending ([Sandell, 1995](#); [Tardieu & McAdams, 2012](#)). This effect is probably related to the role of onsets in auditory grouping ([Darwin, 1981](#); [Bregman, 1990](#)) in music: the grouping of objects starting around the same time contributes to a better perceived blend. The high correlation with the minimum temporal centroid reflects the fact that non-percussive sounds tend to blend better, confirming previous reports by [Tardieu and McAdams \(2012\)](#).

Three measures of spectral variation in the ERB FFT spectrum, as well as the maximum spectral variation in magnitude spectrum also showed a high correlation with the average blend ratings, though the direction was negative this time.

Spectral centroid, the centre of gravity of the spectrum, was reported to be one of the acoustic correlates of average blend ([Sandell, 1995](#); [Tardieu & McAdams, 2012](#)). Among the spectral centroids in various spectra (magnitude, power, harmonic, ERB FFT and ERB Gammatone), the centroid of magnitude spectrum showed the best correlation with the average blend ratings in our data, but the degree of correlation was moderate at best. The negative direction of correlations indicate that a bright sound does not blend well with other sounds, as previously reported by [Sandell \(1995\)](#) and [Tardieu and McAdams \(2012\)](#).

To summarize the result from analyses, the average perceived blend seems to be determined by the attack and spectral patterns of the underlying components. It is also interesting to see these temporal effects in the composite sounds created by matching perceptual onsets rather than physical ones. This importance might be related to the grouping rules by onsets in auditory scene analysis ([Darwin, 1981](#); [Bregman, 1990](#)), even though most composites did not have any perceivable onset delays. Perhaps these attack timing differences, too small to cause different auditory groupings, are still prominent enough to perceptually signify that these concurrent events may not be from one single source.

Spectral brightness of underlying sounds is important as “darker” sounds tended to

achieve better blending. More important than spectral centroid seems to be spectral variation, which was one of the dimensions of the timbre dissimilarity space reported by [McAdams et al. \(1995\)](#). This might be related to the musical practice that a new string quartet group will first work on matching their vibratos before anything else to achieve the maximum degree of blending as a group. It is interesting to see that all three acoustic correlates of the dimensions of the timbre dissimilarity space by [McAdams et al. \(1995\)](#) show mild to high correlations with the average blend ratings. Perhaps this implies that we are using the same acoustic features to judge the degree of blend and the degree of dissimilarity.

The fact that the average blend ratings were more strongly correlated with acoustic features from the Timbre Toolbox rather than the saliency statistics may be due to the loss of information in the complex transforms on the tapping data to yield the saliency values using MDS. Also, if both timbre saliency and the perceived blend are both correlated with some acoustic descriptor, the correlation between timbre saliency and the average blend ratings may exhibit a weaker correlation, as both are functions of a parameter (the common acoustic descriptor) rather than parameters themselves.

In this experiment, we used concurrent unison dyads, which maximized the perceived degree of blending in the composites. Combining instruments in non-unison contexts may lead to a lesser degree of perceived blend due to pitch differences. In the next step, the stimulus range will be expanded to consider various pitches in a more musical setting to study the effect of timbre saliency on voice perception using a melody recognition paradigm used by [Bey and McAdams \(2003\)](#) and [Gregory \(1990\)](#).

(This page is intentionally left blank.)

## Chapter 4

# The Role of Timbre Saliency and Timbre Dissimilarity in Voice Recognition in Counterpoint Music

### Abstract

Timbre saliency refers to the attention-capturing quality of timbre. Can we make one musical line stand out of multiple concurrent lines using a highly salient timbre on the line? This is the question we ask in this chapter using a voice recognition task in counterpoint music.

Two- and three-voice excerpts were generated using instrument timbres that were chosen following specific conditions of timbre saliency and timbre dissimilarity. A listening experiment was carried out with 36 musicians without absolute pitch. A significant effect of timbre dissimilarity was found on the recognition of low voice in two-voice excerpts. In the other two- and three-voice cases, timbre saliency and timbre dissimilarity conditions

did not appear to have systematic effects on the average recognition rate as we hypothesized. This could be due to the variability in the excerpts used for certain conditions, or more fundamentally, because the context effect of each voice position might have been much bigger than the effects of timbre conditions we were trying to measure. A further discussion is presented on possible context effects.

## 4.1 Introduction

So far, we have seen the definition of timbre saliency (Chapter 2) and its impact on the perceived blending of two isolated concurrent notes with the same pitch, loudness and duration (Chapter 3). The next step would be an investigation of the role of timbre saliency in a more musically realistic scenario. The question is, “would timbre saliency affect how listeners hear different lines in music?”

To answer this question, we decided to employ a melody recognition task for the experiment. Melody-based experimental techniques have been used in various musical and auditory research. Melody recognition has been examined in short-term memory (Dowling, 1973, 1978) as well as in both short-term and long-term memory (Dowling & Fujitani, 1971). Using atonal melodies and distorted folk tunes, Dowling and Fujitani (1971) found the effect of melodic contour to be independent of scale. Later Dowling (1978) confirmed the independence of contour and scale still hold for tonal melodies, suggesting that the functions of melodic contour and mode are separate. Deutsch (1972) found that listeners could not recognize a well-known melody when every note in the melody was played in randomized octaves that would not preserve the original melodic contour. Continuing in the same direction, Massaro, Kallman, and Kelly (1980) showed that melodic contour, pitch height and tone chroma are all important in melody recognition, but listeners could



generalize across octaves if tone chroma and melodic contour were preserved.

Melody recognition of interleaved melodies has been used in stream segregation studies based on pitch (Dowling, 1973), pitch and time (Dowling et al., 1987; Bey & McAdams, 2003) or timbre differences (Iverson, 1995; Bey & McAdams, 2003). Dowling (1973) used mono-timbre interleaved melodies and confirmed the effect of pitch separation on auditory stream formation. Concerning the effect of familiarity on melody recognition, it has been reported that if a listener is familiar with one of the melodies in a polyphonic texture, it affects their ability to recognize melodies by either enhancing or inhibiting it depending on whether or not the familiar melody is the target melody. Dowling et al. (1987) proposed the concept of “expectancy windows” in pitch and time and hypothesized that a listener has a series of expectancy windows “through which expected events (target notes) could be clearly perceived (p. 643).” The result showed that any changes to a melody might be more easily detected if it falls within these expectancy windows, particularly with regards to time. Bey and McAdams (2003) studied recognition of interleaved melodies as a function of differences in average pitch, timbre dissimilarity, and time interval. The pitch experiment confirmed previous findings in auditory streaming (Bregman & Campbell, 1971; Van Noorden, 1977; Anstis & Saida, 1985): a greater pitch difference led to a better recognition rate. The result of the timbre experiment was in agreement with earlier work by Iverson (1995): having two highly dissimilar timbres on target and distractor melodies helped the comparison of the target and a subsequent probe melody. The time interval between the target and probe melodies did not have a significant effect on recognition performance.

At the other end of the spectrum, using concurrent melodies, Huron (1989b) had participants listen to mono-timbre polyphonic music and count the number of voices they heard. It was observed that in general musicians were capable of correctly identifying the number of voices although the performance degraded as the number of voices increased, especially

beyond three. [Gregory \(1990\)](#) investigated whether listeners can simultaneously perceive two or more melodic lines in polyphonic music or whether they tend to attend to only one line. He found that melodies that had simultaneous note onsets in the same pitch range in a related key tended to be easier to perceive concurrently if they were distinguished by timbre differences. Although this result suggests that listeners can attend to more than one musical line at a time, it might need to be interpreted with caution since the voices in musical excerpts were not controlled carefully and some excerpts might have been too well-known for the study (such as the one from Mozart's Don Giovanni).

Melody analyses in scores also shed light on the perception of melodies in counterpoint music. [Huron and Fantini \(1989\)](#) examined the voice entries in 75 fugues by J.S. Bach and found that there is a significant reluctance to have an inner voice enter in 5-voice textures, in contrast to no such reluctance in three- or four-voice textures. The authors hypothesized that Bach endeavoured to minimize perceptual confusion in his polyphonic works as the textural density increased. [Huron \(1989a\)](#) also analyzed 105 polyphonic keyboard works by J.S. Bach to report that Bach avoided voice crossings, which could lead to perceptual confusion of concurrent streams.

Counterpoint music has multiple musical voices (or lines) “communicating” with one another. They follow certain stylistic rules, but the biggest difference between counterpoint music and *other* harmonic music is that in counterpoint music *all* the lines are musically equally important, whereas in other harmonic music there is usually one line, a.k.a. “*the melody*,” that is more important than the others, which are in a more of an accompanying role.

As we aimed to expand the study of the effect of timbre saliency in a more musically realistic setting, the method of melody recognition in counterpoint music was deemed to be appropriate. There are two or more musical lines with virtually equal musical importance.

It is impossible for listeners to be able to attend to every note of every voice, therefore they would tend to focus on whatever voice that catches their attention. Hence, if we can control the timbre saliency of the voices in music, listeners' tendency to tune to a specific voice must reflect the voice's saliency. But since it is difficult for us to figure out what voice each listener is hearing out at a given moment, we decided to make it a comparison task based on melody recognition. If a listener happened to focus more on the high voice in multi-voice excerpt, for example, and was asked to identify if the following monophonic comparison melody in the high voice range was the same as or different from the original line in the polyphonic excerpt, he or she would be more likely to answer correctly than someone who happened to focus on the low voice. Therefore performance in this task should covary with voice saliency.

The basic questions we are asking with this experiment are quite simple. What is the effect of timbre saliency on concurrent voice perception? How does the instrument timbre's saliency affect the ability to hear out a given voice as a function of its relative register in the polyphony? For example, it has been known that the entries of inner voices are more difficult to detect than those of outer voices in polyphonic music ([Huron, 1989b](#)). Then can we enhance the detection of an inner voice by applying a salient instrument timbre on it?

As we saw in Chapter 2, the context in which saliencies are measured affect the outcome, because the context influences the extrinsic saliency of the unique intrinsic saliency of the object being measured. It could be why the measures of timbre saliencies showed only mild correlations with the perceived blend of concurrent unison dyads in Chapter 3. Since more than one single pitch (C4) is considered in this experiment, we are not sure about the extent to which the timbre saliency from Chapter 2 will still hold. The fact that instrument timbres often do change depending on the register (e.g., clarinet) adds another concern.

Still, we were confident that we would be able to obtain the effect of timbre saliency in this experiment using multi-timbre combinations, especially considering the recognition performance in the mono-timbre conditions as a baseline. Our hypothesis was that there would be significant effects of timbre saliency in melody recognition that a highly salient timbre on a voice would indeed make it easier to listen to. In case of multiple voices in salient timbres, we hypothesized that the listeners' attention would be equally divided among those voices. As [Iverson \(1995\)](#); [Bey and McAdams \(2003\)](#) observed, we also expected to see the improvement in recognition performance as distance between timbre dissimilarity of the target voice and other voices would increase.

Since this was a very complex experiment, we had to run two experiments for preparation. One was to study the dissimilarity of the timbres that were used in our saliency experiment. It is presented in [Section 4.2](#). The other was a melody comparison experiment to make sure that the changes on a voice were easy enough to hear out in isolation, described in [Section 4.4](#). The design of musical stimuli, which took place before the melody comparison experiment, is explained in detail in [Section 4.3](#). [Section 4.5](#) discusses the main experiment, then finally a general discussion and conclusions are presented in [Section 4.6](#).

## 4.2 Experiment IV: Timbre Dissimilarity

A timbre dissimilarity space is needed to provide a solid ground for the design of the voice recognition experiment to build a bridge between timbre saliency, a new concept, and timbre dissimilarity, which has been studied for the past 35 years. A short experiment was implemented using PsiExp ([Smith, 1995](#)) as in previous experiments, to obtain the dissimilarity space of the timbres used in earlier experiments. It is a classic timbre dissimilarity experiment that has been repeated many times over the years ([Grey, 1977](#); [Grey & Gordon,](#)

1978; [Krumhansl, 1989](#); [McAdams et al., 1995](#); [Lakatos, 2000](#); [Caclin et al., 2005](#)).

### 4.2.1 Methods

#### Participants

Twenty-one participants were recruited from a classified advertisement on the McGill University website. Data from one participant were excluded because he failed to identify any of the identical timbre pairs. The remaining 20 participants included 10 males and 10 females, and 10 musicians and 10 non-musicians, based on self reports, aged from 19 to 39 with a median age of 26.5 years. Everyone passed an audiometric test of their hearing to verify that they were able to hear within 20 dB HL at frequencies 250, 500, 1000, 2000, 4000 and 8000 Hz ([International Organization for Standardization, Geneva, 2004](#); [Martin & Champlin, 2000](#)). The experiment took about 30 minutes and the participants were compensated \$5 upon completion.

#### Stimuli

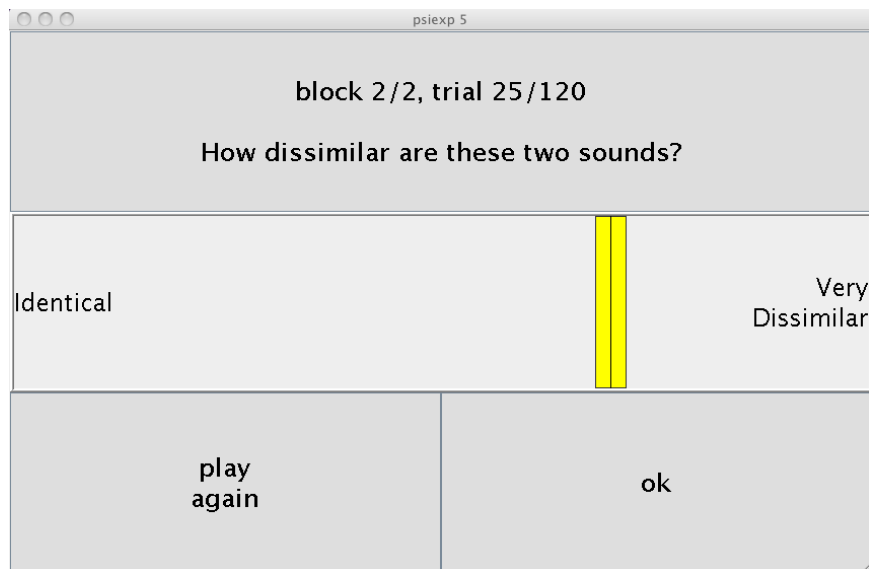
The same set of 19 isolated instrument sounds used in two previous experiments described in Chapters 2 and 3 were employed here. These sounds were all equalized in terms of pitch (C4), loudness (from a pilot study, presented in Appendix B.1) and effective duration, which is a acoustic descriptor calculated from the Timbre Toolbox ([Peeters et al., 2011](#)).

#### Procedure

The experiment was carried out in a sound-attenuated booth ([Industrial Acoustics Company](#), model 1203) in the Music Perception and Cognition Laboratory of the Schulich School of Music of McGill University. The sound signals were all prepared in mono and were pre-

sented to both ears via Sennheiser HD 280 headphones after amplification with a Grace m904 stereo monitor controller. The average level was set to 60 dBA by a Brüel & Kjær type 2250 sound level meter coupled with a Brüel & Kjær type 4153 artificial ear.

Participants first heard all 19 sounds presented one after another in random order before the training phase began. The purpose of this initial exposure was to allow the listeners to determine the overall range of dissimilarity among the sounds they would soon rate. A training phase with 10 trials followed with pairs composed of four training timbres. Four out of 10 training trials used same-timbre pairs to verify that participants understood the task. During the training session, the experimenter was present in the booth with the participant. Once the training session was over and the experimenter answered questions, the participant was left alone to start the testing phase, which consisted of 105 trials with different timbre pairs and 15 trials with identical timbre pairs.



**Figure 4.1:** Screenshot of the graphic user interface used for the timbre dissimilarity rating experiment

Figure 4.1 shows a screenshot of the graphic user interface used for this experiment.

For each trial, participants heard a pair of instrument sounds in sequence, with a silence of 500 ms between them. They could hear the same pair as many times as they desired using the “play again” button. Participants indicated their ratings by placing the marker at an appropriate spot on the continuous scale between “Identical” and “Very Dissimilar” in the graphic user interface shown in Figure 4.1. When they submitted their answer by clicking on the “ok” button, the program would automatically proceed to the next trial and play a new pair of sounds. The same task was used for both the training and the testing phases. There was no break because the entire experiment took less than 30 minutes long. After completing the test phase, participants were asked to fill out a questionnaire with information on their background. They were paid and debriefed and signed the receipt.

#### 4.2.2 Results and Discussion

The continuous scale in Figure 4.1 was coded between 0 and 1, which correspond to the “Identical” and “Very Dissimilar” extremes, respectively. Repeated-measures Analysis of Variance (ANOVA) was run on the dissimilarity rating data from 20 participants in SPSS. The dependent variable (DV) was the numeric dissimilarity rating data on each pair of sounds. Gender and musicianship were between-subject factors and age a covariate. The repeated factor was the timbre pairs. Two disjoint sets of repeated-measures ANOVA had to be carried out because SPSS limits the number of repetitions to 99 in repeated-measures ANOVA. The first set considered the dissimilarity rating data for pairs 1 through 60 and the second set for pairs 61 through 120. The means and STDs of the two sets are listed in Table 4.1. Both the mean and the STD are slightly larger with the second set, although the difference is negligible. No significant effect of gender,  $F(1, 15) = 1.096, p = .312$  for the first set,  $F(1, 15) < 1$  for the second set; no significant effect of musicianship,  $F(1, 15) < 1$  for the first set,  $F(1, 15) < 1$  for the second set; no significant effect of age,  $F(1, 15) < 1$

for the first set,  $F(1, 15) < 1$  for the second set, was found.

**Table 4.1:** Means and STDs of the ratings in two disjoint sets for ANOVA on SPSS

Set Designation	Mean	STD
First set	0.53	0.33
Second set	0.57	0.36

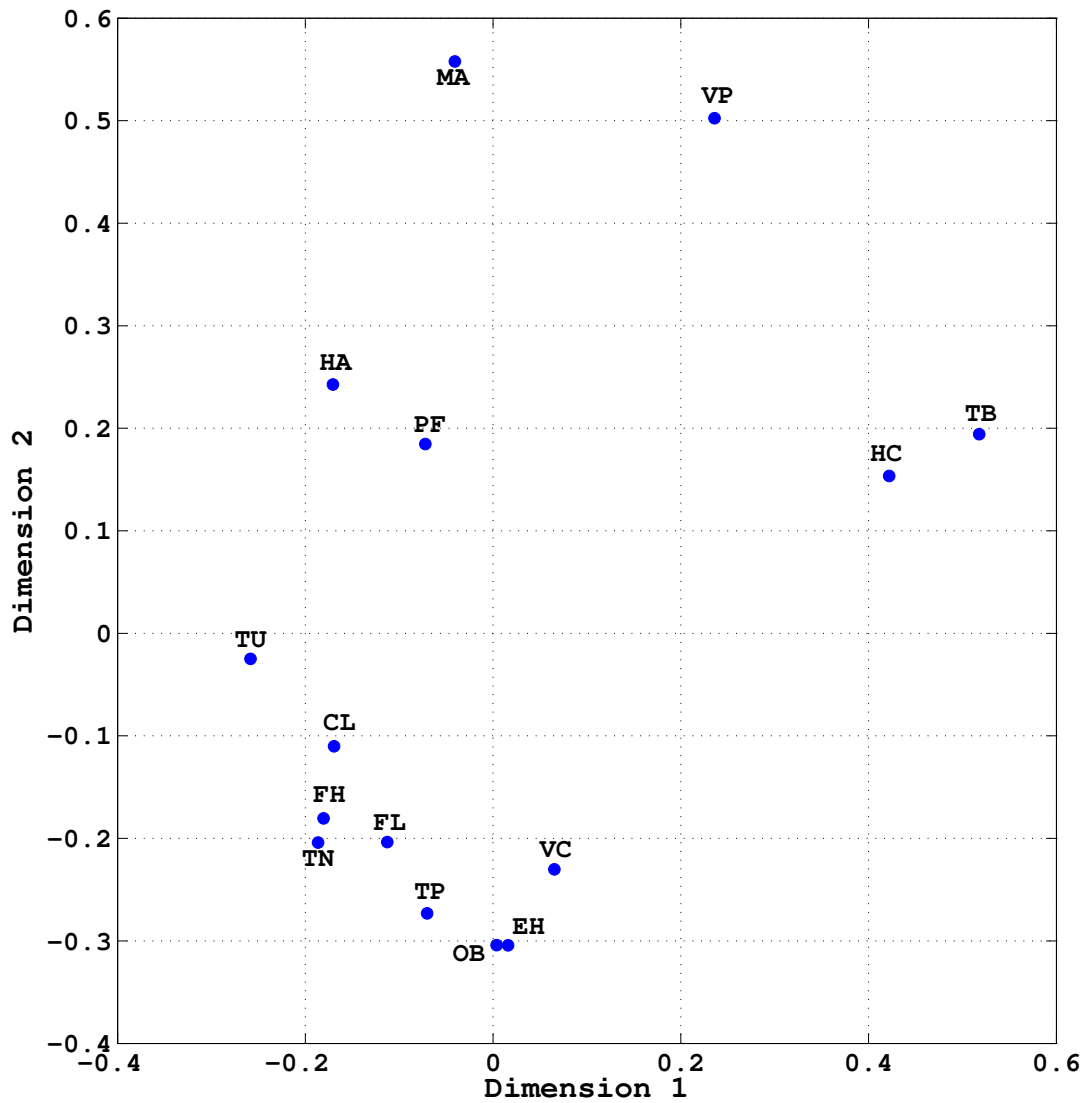
Dissimilarity judgments from 20 individuals were formed into 20 individual lower triangular matrices, which were analyzed by CLASCAL (Winsberg & De Soete, 1993) to obtain the dissimilarity space. The best solution turned out to have two dimensions with specificities and 5 latent classes of participants (Figure 4.2).

**Table 4.2:** Acoustic correlates of the two-dimensional timbre dissimilarity space.

Dimension	Acoustic Feature	Feature Group	$r(13)$	$p$
1	Spectral Centroid	ERB FFT	.86	$< .0001$
2	Temporal Centroid	Time	-.68	$< .01$

Note that the percussive instruments are all located above the  $y = 0$  line. This suggests that the second dimension may be related to attack time. Correlations were calculated between each of the two dimensions and the acoustic features were computed by the Timbre Toolbox (Peeters et al., 2011). As Table 4.2 shows, the first dimension is highly correlated with spectral centroid in the ERB-FFT spectrum,  $r(13) = .8552, p < .0001$ , and has moderate correlations with spectral centroid measures in other domains (such as linear amplitude, power and harmonic spectra). The second dimension is moderately correlated with the temporal centroid,  $r(13) = -.6778, p = .0055$ . This is mostly in agreement with previous studies in timbre dissimilarity showing that attack time and spectral centroid are





**Figure 4.2:** Two-dimensional timbre dissimilarity space (CL = Clarinet, EH = English Horn, FH = French Horn, FL = Flute, HA = Harp, HC = Harpsichord, MA = Marimba, OB = Oboe, PF = Piano, TB = Tubular Bells, TN = Trombone, TP = Trumpet, TU = Tuba, VC = Cello, VP = Vibraphone).

two of the most important acoustic features (Grey, 1977; Grey & Gordon, 1978; Krumhansl, 1989; McAdams et al., 1995; Lakatos, 2000; Caclin et al., 2005).

The two-dimensional timbre dissimilarity space in Figure 4.2 will provide a basis in the design of stimuli for Experiments V and VI, in selecting a subset out of the 15 timbres we have been studying so far. This is necessary because it is not feasible to study all 15 timbres' effects on voice recognition, and therefore we need to select timbres that best represent the experimental conditions. This timbre dissimilarity space will also be essential in data analysis as the dissimilarity distance is one of the main parameters for Experiment VI.

## 4.3 Musical Stimulus Design

### 4.3.1 Selection of Musical Excerpts

A number of excerpts and their comparison melodies are needed for this voice recognition experiment. We selected excerpts from J.S. Bach's *Trio Sonatas for Organ*, BWV 525 – 530 (Bach, 1730), because the music was already clearly written for three voices (right hand, left hand and pedal) and it is relatively unknown in comparison with some other two- and three-voice pieces (such as the *Inventions and Sinfonias*).

In the trio sonatas, there are many parts that had only two voices (i.e., no melody on the pedal). Some other parts have all three voices but with slow note transitions on the pedal voice. We selected two-voice excerpts from those parts, removing the slow pedal voice if necessary. For three-voice excerpts, we looked for the parts with all three voices clearly active with equivalent note densities. Any excerpts with voice crossings were avoided. We also did some editing in some excerpts, such as transposing the melodies to a new key (often to accommodate the playing ranges of selected instruments), changing the pitch of a

note (often by an octave) to avoid voice crossing, or breaking a longer note into two shorter notes to keep the note density constant.

An example is shown in Figure 4.3. The first two measures show the two-voice excerpt and the last two measures the corresponding modified melodies. These modified melodies shown in Figure 4.3 are different on both voices, but in the actual experiment they will never be heard together. All two- and three-voice excerpts and their comparison melodies are presented in Appendix C.1 and C.2, respectively.



**Figure 4.3:** An example of a two-voice excerpt and corresponding comparison melodies

For each trial in the experiment, one excerpt was played to the participant in all voices, followed by a comparison melody (only one voice, which may be the same as or different from the original melody), and the listener had to make a forced-choice response concerning whether the comparison melody was the same as or different from the corresponding voice in the excerpt he or she just heard.

#### 4.3.2 Timbre Combinations

For Experiments V and VI, a subset of instruments was chosen that would best represent the timbre dissimilarity and timbre saliency conditions from the two spaces in Figures 4.2 and 2.4, respectively, which were found from earlier experiments.

For two-voice excerpts, we have two timbre dissimilarity conditions (*close* and *far*) and three timbre saliency conditions (*both higher*, *both lower*, or *opposite* saliencies). Combining

these conditions yields a complete  $2 \times 3$  design that specifies all possible two-voice timbre combinations according to timbre saliency and timbre dissimilarity conditions. As each combination can result in two orderings of timbre assignment to two voices, the number of possible *ordered* combinations is 12, which is not too large to implement in an experiment. We therefore decided to consider all possible cases. Table 4.3 lists the timbre combinations selected for two-voice excerpts. Each instrument in each pair appears once in the high voice and once in the low voice. Notice, for example, in the “Both Higher Saliency” column and “Close in Timbre Space” row that Piano (PF) and Harp (HA) appear once in each voice. This enables us to thoroughly examine the effect of timbre saliency on each of the voice positions (high or low). One may notice that certain instruments appear more often than others in Table 4.3. This could not be avoided with the positions of the timbres in the timbre saliency dimension in Figure 2.4 as well as the timbre dissimilarity space in Figure 4.2. We also did not expect it to affect the result of the experiment, since each timbre combination would be tested with multiple excerpts and melodies.

**Table 4.3:** Timbre assignments for two-voice excerpts

		Both Higher Saliency		Both Lower Saliency		Opposite Saliency	
Close in Timbre Space	High	PF	HA	CL	TN	TP	EH
	Low	HA	PF	TN	CL	EH	TP
Far in Timbre Space	High	MA	TP	CL	HC	EH	MA
	Low	TP	MA	HC	CL	MA	EH

The number of possible combinations for three-voice conditions turns out to be much larger than 12. In the timbre dissimilarity space, all three timbres can be *close* together, all *far apart*, or *two timbres are close and one timbre is far away from them*. On the timbre

saliency dimension, the three timbres can be all in *one* saliency region (high, middle or low), or two timbres are in the same saliency region and the other from a different region. As each of these cases also has three possible orderings in assigning timbres to three voices, the resulting number of complete three-voice timbre combinations becomes too large to be considered for this experiment. We therefore decided to focus on a subset in which two timbres are similar and the other one is different (i.e., two are close to each other and the third one is far from these two in timbre *dissimilarity* space), and one is a highly salient timbre and two others are of lower saliency. This  $3 \times 3$  design, which combines three timbre dissimilarity conditions with three timbre saliency conditions, results in nine three-voice conditions (Table 4.4).

D1, D2 and D3 represents the three *dissimilarity* conditions according to the assignments of three timbres to three voices. Among the three timbres, T1, T2 and T3, *T3* is always the “far” timbre and is highlighted in (blue) italics. Similarly, S1, S2 and S3 represent the three *saliency* conditions. The “**H**igh” saliency timbre of the three timbres is highlighted in bold (red) fonts. For example, the D1S1 column in Table 4.4 shows that in this condition high and middle voices have timbres that are of low saliency and close in dissimilarity space. In the same condition, the low voice has the salient timbre that is far from the other two timbres in timbre space. In a similar manner, the D1S2 condition shows that the highly salient timbre is now on the middle voice and the far timbre is on the low voice. This factorial combination of saliency and dissimilarity allows us to test their separate contributions to voice recognition, as well as their potential interaction.

Even though there are nine three-voice conditions, it turns out that only four sets of timbre assignments are required – {D1S1, D2S2, D3S3}, {D1S2}, {D1S3, D2S3, D3S2}, and {D2S1, D3S1}, as shown with four colours and four types of fonts in columns in Table 4.5. These combinations were chosen considering not only the relative positions in timbre

**Table 4.4:** Timbre conditions for three-voice excerpts

	D1S1	D1S2	D1S3	D2S1	D2S2	D2S3	D3S1	D3S2	D3S3
High	T1L	T1L	T1 <b>H</b>	T2L	T2L	T2 <b>H</b>	<i>T3L</i>	<i>T3L</i>	<i>T3H</i>
Middle	T2L	T2 <b>H</b>	T2L	<i>T3L</i>	<i>T3H</i>	<i>T3L</i>	T1L	T1 <b>H</b>	T1L
Low	<i>T3H</i>	<i>T3L</i>	<i>T3L</i>	T1 <b>H</b>	T1L	T1L	T2 <b>H</b>	T2L	T2L

dissimilarity and timbre saliency spaces, but also the instrument ranges, because some instruments cannot play higher notes in the top voice and others lower notes in the bottom voice. For example, the gray columns D1S1, D2S2, D3S3 in Table 4.5 use the timbre set of CL, TN, and MA. CL and TN are not highly salient and are located close together in the dissimilarity space. MA is the most salient timbre and positioned far away from both CL and TN. Different requirements in conditions D1S1, D2S2 and D3S3 are met by assigning MA to each of three voices.

**Table 4.5:** Timbre assignments for three-voice excerpts

	D1S1	<u>D1S2</u>	<b>D1S3</b>	<i>D2S1</i>	D2S2	<b>D2S3</b>	<i>D3S1</i>	<b>D3S2</b>	D3S3
High	T1L	<u>T1L</u>	<b>T1H</b>	<i>T2L</i>	T2L	<b>T2H</b>	<i>T3L</i>	<b>T3L</b>	T3H
Middle	T2L	<u>T2H</u>	<b>T2L</b>	<i>T3L</i>	T3H	<b>T3L</b>	<i>T1L</i>	<b>T1H</b>	T1L
Low	T3H	<u>T3L</u>	<b>T3L</b>	<i>T1H</i>	T1L	<b>T1L</b>	<i>T2H</i>	<b>T2L</b>	T2L
T1	CL	<u>EH</u>	<b>TP</b>	<i>MA</i>	TN	<b>TN</b>	<i>VP</i>	<b>TP</b>	CL
T2	TN	<u>TP</u>	<b>TN</b>	<i>VP</i>	CL	<b>TP</b>	<i>MA</i>	<b>TN</b>	TN
T3	MA	<u>HC</u>	<b>HC</b>	<i>CL</i>	MA	<b>HC</b>	<i>CL</i>	<b>HC</b>	MA

In addition to the cases presented in Tables 4.3 and 4.5, we need to test the same-timbre condition in both two- and three-voice conditions, to determine baseline performance in the absence of timbre differences for comparison. We decided to use PF for this, not only

because the piano has a sufficient range for all excerpts, but because it is probably the most widely-used instrument in music in general. It was pointed out that PF has a highly percussive onset, which would help voice segregation based on onset asynchronies. But we expected this effect to be negligible in comparison with the tonal fusion based on the same timbre on all voices.

In searching for the right timbre combinations for the conditions specified in Tables 4.3 and 4.5, we had to make some compromises by using some of the instruments with medium saliency. More specifically, Harpsichord (HC) was used in place of some “lower saliency” instruments and Harp (HA) in place of higher saliency instruments. This was the best we could do with the two given spaces (Figures 4.2 and 2.4), especially where nine out of fifteen timbres were all located together in the lower left corner of the timbre dissimilarity space (Fig. 4.2).

### 4.3.3 Excerpt Assignment to Timbre Combinations

#### Two-Voice Excerpts

In a two-voice block, 12 excerpts in multi-timbre and two excerpts in mono-timbre instrumentation were tested. Each mono-timbre two-voice excerpt appeared twice in each block to test the high and low voices. In a three-voice block, nine excerpts in multi-timbre and one excerpt in a single timbre were tested. The mono-timbre three-voice excerpt appeared three times in each block to test the high, middle and low voices. This arrangement was counterbalanced across participants.

In order to counterbalance the possible effect of a given excerpt on the participants’ performance, each excerpt was assigned to multiple timbre conditions. Additionally, each timbre condition was tested with multiple excerpts. For the sake of proper counterbalanc-

ing, it would be ideal to encode each voice of all excerpts in every assigned timbre. But this would require too large a number of stimuli, so we decided to use the assignment strategy shown in Tables 4.6 and 4.7.

**Table 4.6:** Excerpt assignments for two-voice excerpts - Option A

	Close-Higher		Close-Lower		Close-Opp.		Far-Higher		Far-Lower		Far-Opp.	
	PF	HA	CL	TN	TP	EH	MA	TP	CL	HC	EH	MA
	HA	PF	TN	CL	EH	TP	TP	MA	HC	CL	MA	EH
H	1	3	5	7	9	11	12	2	10	4	8	6
L	2	4	6	8	10	12	1	11	3	9	5	7
H	6	10	12	2	4	8	3	5	11	7	9	1
L	11	5	3	9	7	1	8	4	6	2	12	10

**Table 4.7:** Excerpt assignments for two-voice excerpts - Option B

	Close-Higher		Close-Lower		Close-Opp.		Far-Higher		Far-Lower		Far-Opp.	
	PF	HA	CL	TN	TP	EH	MA	TP	CL	HC	EH	MA
	HA	PF	TN	CL	EH	TP	TP	MA	HC	CL	MA	EH
H	4	2	8	6	12	10	1	11	9	3	7	5
L	3	1	7	5	11	9	2	12	4	10	6	8
H	5	11	9	3	1	7	4	8	2	6	10	12
L	10	6	2	12	8	4	5	3	7	11	1	9

Table 4.6 shows the assignment of timbres to excerpts, as well as which voice for a given excerpt is to be tested. Each column represents a given timbre assignment, and each row



represents which voice was used for comparison. This table is built in such a way that for a given row, excerpts 1 through 12 each occur only once. Also, all the entries are unique in each column. Numbers 1 - 12 appear in each of the four  $2 \times 6$  blocks (“close block” and “far block”) only once. The upper two rows are for one block of testing and the lower two for another separate block. Note that each excerpt is arbitrarily assigned to a number from 1 to 12 for an arbitrary assignment of timbre sets to excerpts.

What Table 4.6 shows is what timbres will be used in encoding a particular excerpt. For example, according to the cell highlighted in gray, Excerpt 1 is played with PF (high) and HA (low) and is tested on the high voice (same or modified, which is to be decided pseudo-randomly at the time of the experiment, which will be discussed later). The same combination (PF-HA) is used on Excerpt 2 for testing the low voice. This design ensures that for a given listener, every excerpt is tested twice on the high voice and twice on the low voice but in different timbre combinations every time. Additionally, it ensures that for a given listener, every timbre condition is tested with different excerpts. For further counterbalancing, the opposite case of what is shown in Table 4.6 (“1” in HA-PF in gray-highlighted cell in Table 4.7) would be presented to another participant for proper counterbalancing.

Table 4.7 was constructed by rotating elements in the  $2 \times 2$  submatrices of Table 4.6 by 180 degrees. This is to make sure to counterbalance the voice (high or low) tested for each excerpt. For example, Excerpt 1 (in the gray cell) was played by PF-HA and tested on the high voice in Table 4.6. The corresponding gray cell in Table 4.7 indicates that Excerpt 1, played by HA-PF, would be used to test the low voice. In this way we can test the effect of PF on voice position (high or low) in polyphonic melody recognition.

Having two complementary tables means that half of the participants are tested with the stimuli in Table 4.6 and the other half with Table 4.7.

## Three-Voice Excerpts

Table 4.8: Excerpt assignments for three-voice excerpts

	D1S1	<u>D1S2</u>	<b>D1S3</b>	<i>D2S1</i>	D2S2	<b>D2S3</b>	<i>D3S1</i>	<b>D3S2</b>	D3S3
T1	CL	<u>EH</u>	<b>TP</b>	<i>MA</i>	TN	<b>TN</b>	<i>VP</i>	<b>TP</b>	CL
T2	TN	<u>TP</u>	<b>TN</b>	<i>VP</i>	CL	<b>TP</b>	<i>MA</i>	<b>TN</b>	TN
T3	MA	<u>HC</u>	<b>HC</b>	<i>CL</i>	MA	<b>HC</b>	<i>CL</i>	<b>HC</b>	MA
High	①	<u>4</u>	<b>7</b>	<i>6</i>	8	③	<i>9</i>	②	5
Middle	2	<u>5</u>	<b>8</b>	<i>4</i>	9	①	<i>7</i>	③	6
Low	3	<u>6</u>	<b>9</b>	<i>5</i>	7	②	<i>8</i>	①	4
High	6	<u>3</u>	<b>2</b>	<i>8</i>	4	<b>9</b>	①	<b>5</b>	7
Middle	9	<u>8</u>	<b>4</b>	<i>7</i>	①	<b>5</b>	<i>2</i>	<b>6</b>	3
Low	5	<u>7</u>	①	<i>3</i>	2	<b>6</b>	<i>4</i>	<b>9</b>	8

A similar table for three-voice assignments is shown in Table 4.8. Four colours and four font types are used to designate each of four sets of timbres. This table is constructed in a similar way to the two-voice tables; each number appears only once for each row and for each column, as well as for each of the  $3 \times 3$  matrices. For example, Excerpt 1 (designated with ① in the cell highlighted in purple) is to be coded for 6 trials – only once per row (meaning never tested in the same voice twice). The top three rows make up for one testing block and the bottom three rows another. Also, no excerpt is repeated in the same timbre set (for example the gray columns have entries 1 - 9 only once), except the lavender columns in D2S3 and D3S2. The excerpt numbers on the top three rows in these columns are the same (see the ②'s and ③'s blue cells), and so are those on the bottom three rows. To resolve this conflict, we would need a new set of timbre combinations for one of these two conditions, which however turned out to be not feasible according to the locations of

timbres in timbre saliency and timbre dissimilarity spaces. Therefore we decided to accept the design as it is and proceed.

### Organization of Sound Files

Each of 12 two-voice excerpts resulted in five versions. In one version both voices were played by Piano and the other four versions had different timbre pairs (see Tab. 4.3). In the versions with two different timbres, both timbres appeared in both voices on different excerpts (e.g., Harp high, Piano low on one and Piano high and Harp low on another). The total number of two-voice stimuli was  $12 \times (1 + 4 \times 2) = 108$ . The individual voices from each two-voice excerpt and their corresponding modified melodies were also stored in separate sound files for a total of  $108 \times 4 = 432$ .

For three-voice stimuli, there were nine excerpts in six multi-timbre combinations and one Piano-only version, resulting in 63 sound files ( $= 9 \times (1 + 3 \times 2)$ ). As each voice in these files would have two versions of “same” and “modified”, there were 378 ( $= 63 \times 6$ ) individual melody files. The two- and three-voice excerpts are shown in Appendix C.1 and C.2, respectively.

The stimuli were roughly equalized in terms of loudness in the following manner. Fifteen versions of the melody in Figure 4.4 were created using 15 timbres. A loudness-equalization pilot study was carried out using these files. Fifteen participants took part. The median adjustments in dB were calculated and used to offset the levels in creating sound files in Logic (Apple Computer, 2012). The details are presented in Appendix B.3.

#### 4.3.4 Assignment of Same or Modified Comparison Melodies

The original intention for comparison melody assignment was to leave it random, i.e. to flip a coin for every trial. However, since we have additional conditions for counterbalancing



**Figure 4.4:** Melody used for loudness equalization across timbres.

the number of “same” trials and the number of “modified” trials per voice and per timbre set, this further restricts the degree of randomness in comparison melody assignments.

For example, let’s reconsider the two-voice assignment option A in Table 4.6. If we flip a coin for Excerpt 1 in the top-left gray cell and got a “same” (designated by ‘s’), it determines the assignments for three more trials, all in gray cells, as shown on Table 4.9.

This is because of our counterbalancing rules. Since the first high voice trial in the PF-HA timbre set was tested with the “same” melody, the next high voice trial in the same timbre (using a different excerpt) will have to be in the “modified” melody. This in turn determines that the trial with excerpt 6 in the other block (in the first block in this case) has to be tested with the “same” melody, which decides that the next high voice trial in MA-EH uses the “modified” melody for test. Since Excerpt 1 is what we started with, this iteration ends here. Note that the gray cells in the top block (top two rows of Table 4.9) both have “same” trials and those in the bottom block (bottom two rows) both have “modified” trials. We will soon see that this pattern holds in other cases too.

Table 4.10 shows the next set of excerpts affected by the decision on the first high-voice trial in HA-PF timbre set (with excerpt 3). Again, the lavender cells in the top row all have “same” melodies while those in the bottom row have “modified” melodies. Combining Tables 4.9 and 4.10 together, all the high-voice trials now have comparison melody assignments. In the same fashion, the melody assignments for the other cases can be figured out.

However, these “same-fate” cells caused a serious problem in our block design. Remem-

**Table 4.9:** “Same-fate” cells in two-voice excerpt assignment, example 1

	Close-Higher		Close-Lower		Close-Opp.		Far-Higher		Far-Lower		Far-Opp.	
	PF	HA	CL	TN	TP	EH	MA	TP	CL	HC	EH	MA
	HA	PF	TN	CL	EH	TP	TP	MA	HC	CL	MA	EH
H	<b>1s</b>	3	5	7	9	11	12	2	10	4	8	<b>6s</b>
L	2	4	6	8	10	12	1	11	3	9	5	7
H	<b>6m</b>	10	12	2	4	8	3	5	11	7	9	<b>1m</b>
L	11	5	3	9	7	1	8	4	6	2	12	10

**Table 4.10:** “Same-fate” cells in two-voice excerpt assignment, example 2

	Close-Higher		Close-Lower		Close-Opp.		Far-Higher		Far-Lower		Far-Opp.	
	PF	HA	CL	TN	TP	EH	MA	TP	CL	HC	EH	MA
	HA	PF	TN	CL	EH	TP	TP	MA	HC	CL	MA	EH
H	1	<b>3s</b>	<b>5s</b>	<b>7s</b>	<b>9s</b>	<b>11s</b>	<b>12s</b>	<b>2s</b>	<b>10s</b>	<b>4s</b>	<b>8s</b>	6
L	2	4	6	8	10	12	1	11	3	9	5	7
H	6	<b>10m</b>	<b>12m</b>	<b>2m</b>	<b>4m</b>	<b>8m</b>	<b>3m</b>	<b>5m</b>	<b>11m</b>	<b>7m</b>	<b>9m</b>	1
L	11	5	3	9	7	1	8	4	6	2	12	10

ber that the top two rows of Table 4.10 are supposed to form a block. If the random coin tosses for the top-left light blue cell in Table 4.9 and the top-left lavender cell in Table 4.10 turned out to use the “same” melody, it means *all* the high voice trials in block 1 of two-voice excerpts will use the “same” comparison melodies! A good listener might be able to find the pattern after a couple of trials and just answer using the pattern without really paying attention to later trials. As a solution to this problem, we decided to mix the same/modified assignment results based on the “same-fate” phenomenon of all cells together and divide them into two blocks of roughly the same number of “same” and “modified” trials.

## 4.4 Experiment V: Melody Discrimination

After all modified melodies were created, we wanted to verify how easy or difficult it is to detect changes in pairs of melodies in isolation. This was necessary before carrying out the main experiment, because if participants cannot hear changes in corresponding melodies in isolation, they will not be able to hear out changes on one voice in a mixture with other voice(s).

### 4.4.1 Methods

#### Participants

Twenty musicians (10 males) were recruited. None had absolute pitch based on self reports. Their ages ranged from 18 to 37, with a median of 24 years. They had various music education backgrounds, from five years of training to over 20 years. They all successfully passed hearing tests prior to the experiment to make sure that they could hear at least 20 dB HL at frequencies of 250, 500, 1000, 2000, 4000 and 8000 Hz ([International Organization for Standardization, Geneva, 2004](#); [Martin & Champlin, 2000](#)). Participants were compensated

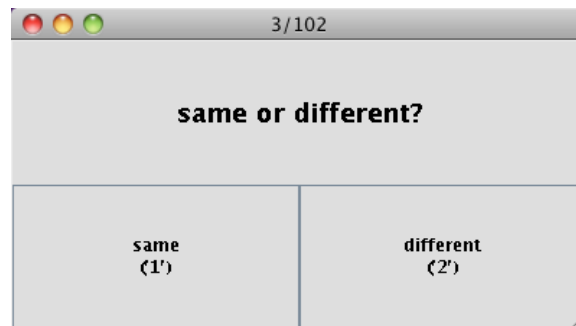
\$10 upon completion of their session.

### Stimuli

The stimuli were ordered pairs of “original” and “modified” melodies in all two- and three-voice excerpts: original-original, original-modified, modified-original, and modified-modified. There were 96 melody pairs from two-voice excerpts and 108 from three-voice excerpts, hence 204 pairs in total. These were all presented to the participants in random order.

### Procedure

The experiment was carried out in a sound-attenuated booth ([Industrial Acoustics Company](#), model 1203) in the Music Technology Area of the Schulich School of Music of McGill University. The mono sound signals were amplified with a Grace m904 stereo monitor controller and presented to both ears of the participants via Sennheiser HD 280 headphones. The level was set to an average of 68 dBA as measured with a Brüel & Kjær type 2250 sound level meter coupled with a Brüel & Kjær type 4153 artificial ear.



**Figure 4.5:** Screenshot of the melody discrimination experiment.

After completing the audiogram, participants were presented with an instruction sheet that described what was expected of them. There was no training session because the task was simple and straightforward. The 204 pairs of stimuli were randomized and divided into

two blocks of 102 trials each. For each trial, one melody of a pair was randomly selected to play first, then a 500 ms pause followed before the other melody started to play. There were no repeats. Participants indicated their answers by clicking on either the “same” or the “different” button in the graphic user interface shown in Figure 4.5, which was developed in PsiExp (Smith, 1995) on an Apple PowerPC G5. As soon as one of the buttons was clicked, the answer was recorded and the program would automatically move to the next trial. There was a chance for participants to take a break after finishing the first block. The entire experiment took about an hour including 45 to 50 minutes for the two blocks and time for questionnaires and collecting feedback.

#### 4.4.2 Results & Discussion

There was quite a large variability in the participants’ average performances, ranging from 69% to 92% correct, with a median of 84%. These numbers are significantly higher than a chance level of 50%, hence we moved forward with data analysis in terms of specific melody pairs. The results are summarized in Tables 4.11 and 4.12 for melodies from two- and three-voice excerpts, respectively.

We had initially hoped to have the percent correct rate to be at least 75% for every pair, but there were three melodies in the two tables with percent correct values less than 0.75 (in bold red), two from two-voice low voice pairs and one from three-voice high voice pairs. These values were less than 0.75, but still higher than 0.7, so we decided to use these melodies for the main experiment without any further modifications. This discrimination performance will be included as an analysis variable for Experiment VI.



**Table 4.11:** Melody discrimination result for melodies from two-voice excerpts, averaged across all participants

Excerpt No.	Proportion Correct	
	High Voice	Low Voice
1	0.83	0.76
2	0.88	0.79
3	0.79	0.89
4	0.75	0.76
5	0.86	0.83
6	0.81	0.84
7	0.83	0.78
8	0.75	0.85
9	0.84	0.81
10	0.80	0.73
11	0.88	0.93
12	0.78	0.71

**Table 4.12:** Melody discrimination result for melodies from three-voice excerpts

Excerpt No.	Proportion Correct		
	High Voice	Middle Voice	Low Voice
1	0.78	0.81	0.79
2	0.88	0.81	0.83
3	0.83	0.83	0.88
4	0.78	0.84	0.85
5	0.81	0.84	0.78
6	<b>0.73</b>	0.75	0.88
7	0.79	0.83	0.84
8	0.86	0.84	0.79
9	0.88	0.85	0.91

## 4.5 Experiment VI: Voice Recognition in Counterpoint Music

The main experiment studied the role of timbre dissimilarity and saliency in voice recognition in counterpoint music, using the results from Experiments IV and V.

### 4.5.1 Methods

#### Participants

Thirty six musicians without absolute pitch in two groups, recruited from a classified advertisement on the McGill University website, took part in the experiment. According to self reports, all participants had at least eight years of musical training on an instrument, voice or in sound recording or composition, during which time they spent at least five hours per week on their training including lessons and practice. The participants' ages ranged

from 18 to 37, with a median of 24 years. There were equal numbers of males and females. Nineteen of them identified themselves as “professional” musicians and the rest as “amateurs”. In terms of their listening habits, 15 claimed to be “harmony-listeners” and 21 to be “melody-listeners.”

All participants passed a preliminary hearing test before the experiment to make sure that they were able to hear within 20 dB HL at frequencies 250, 500, 1000, 2000, 4000 and 8000 Hz ([International Organization for Standardization, Geneva, 2004](#); [Martin & Champlin, 2000](#)). They were paid \$10 upon completion of the experiment, which took about an hour including the time for filling out a questionnaire and debriefing.

### Stimuli

The multi-voice excerpts, chosen from J.S. Bach’s *Trio Sonatas for Organ*, BWV 525 – 530 ([Bach, 1730](#)) were used for this experiment, as well as individual original voices in those excerpts and the comparison melodies (Appendices [C.1](#) and [C.2](#)). Experiment V had verified the listeners’ ability to detect the changes in corresponding pairs of original and comparison melodies in isolation (Section [4.4](#)). The musical stimulus design is explained in detail in Section [4.3](#). The stimuli were roughly equalized in loudness, as described in Appendix [B.3](#).

### Procedure

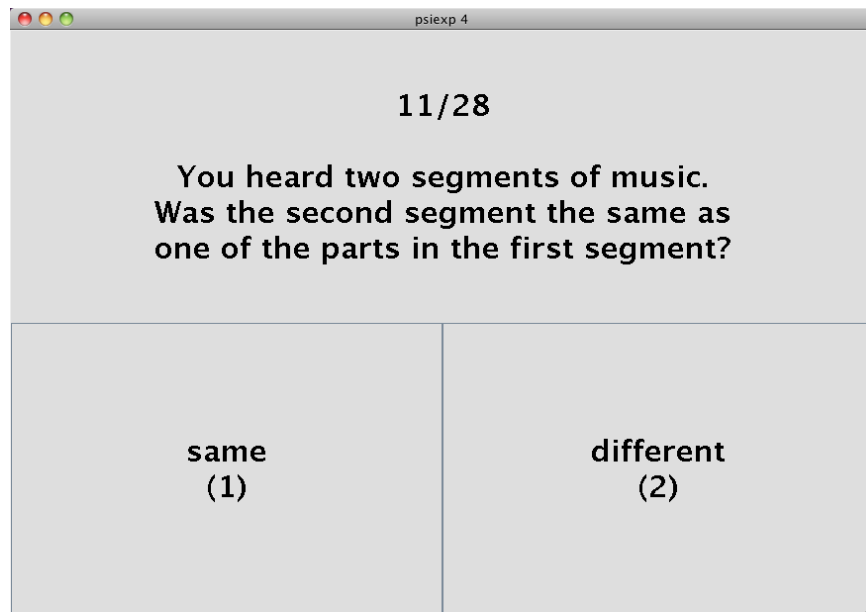
The experiment was carried out in a sound-attenuated booth ([Industrial Acoustics Company](#), model 1203) in the Music Perception and Cognition Laboratory of the Schulich School of Music of McGill University. The sound signals were all prepared in mono to further eliminate any undesired cue for separation, which were in turn amplified by a Grace m904 stereo monitor controller and presented to both ears of the participants via Sennheiser HD 280

headphones. The average level was set to 64.5 dBA as measured by a Brüel & Kjær type 2250 sound level meter coupled with a Brüel & Kjær type 4153 artificial ear.

Participants were first presented with an instruction sheet. If they had a question on the instructions, they could ask the experimenter before the training session began. The first block was a training session with two-voice excerpts, which were in AF, BS and CE timbres, during which the experimenter was in the booth with the participant. After answering any more questions after the first training session, the experimenter left the booth and the participant carried out the two two-voice testing blocks. Participants could take a break between the two testing blocks, but most of them did not.

When participants reached the end of the second block, they saw a message on the screen "Step outside and go see the experimenter". This was to ensure that everyone took a break between the two-voice and three-voice blocks because the task demanded a great deal of attentional focus. Afterwards, the experimenter and the participant went back in to the booth to do the three-voice training session. The participant then completed the two three-voice testing blocks alone.

The participants used the graphic user interface in Figure 4.6 to provide their answers, which was developed in PsiExp (Smith, 1995) on an Apple PowerPC G5. For each trial, a multi-voice excerpt would play first, followed by a one-second silence before a monophonic melody would play. The monophonic melody could be the same as or different from one of the voices in the preceding excerpt. The buttons were disabled until one second after the monophonic melody finished playing. There was no option to repeat the stimuli. When the participant submitted his or her answer by either clicking the appropriate button using a mouse or pressing the key "1" or "2" on the keyboard (for "same" or "different", respectively), the program would automatically proceed to the next trial and start playing the next stimulus after a pause of 200 ms.



**Figure 4.6:** Screenshot of the voice recognition experiment

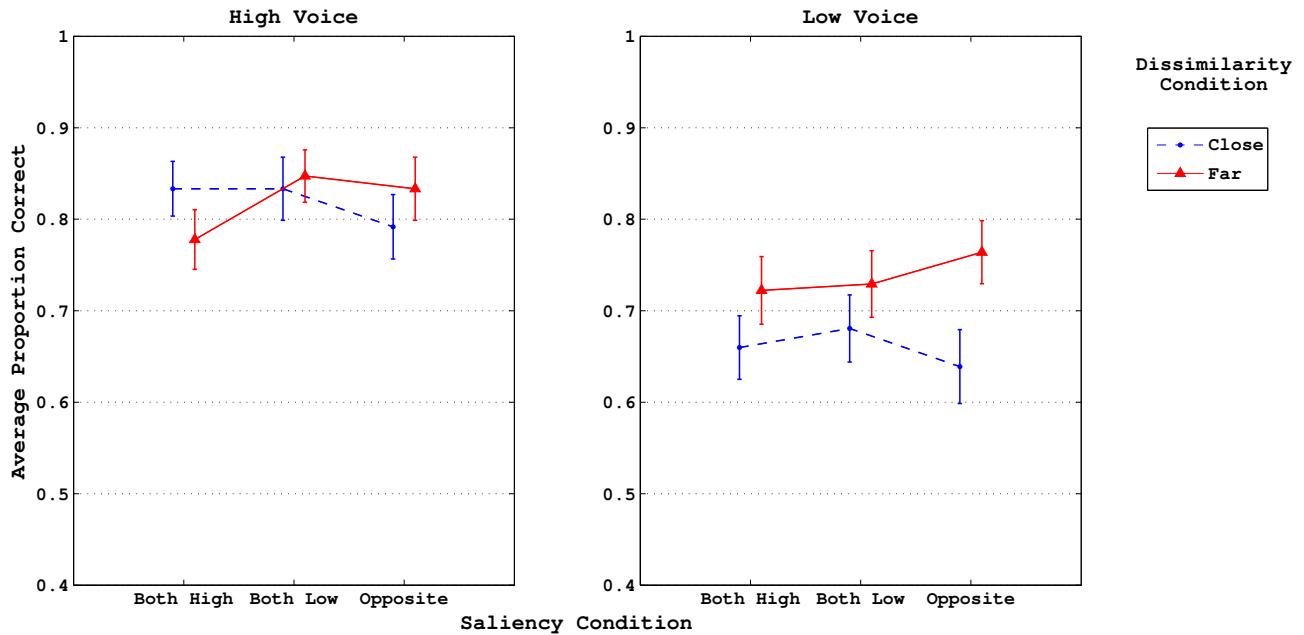
#### 4.5.2 Results

First, each participant's average performance was calculated. These per-participant averages ranged from 58.9% to 92.9% for two-voice cases and from 51.7% to 83.3% for three-voice cases. The average performance across participants was 75.6% for two-voice and 65.9% for three-voice cases, suggesting that the recognition of individual voices were much more difficult with three voices as one might expect.

For each participant, an average performance score was calculated across all two- and three-voice trials. A three-way ANOVA on these data showed that there were no significant differences observed between groups defined on the basis of gender  $F(1, 4528) = 0.462, p = .497$ , self-identification as either “professional” or “amateur”,  $F(1, 4528) = 2.309, p = .129$ , or self-identification as either “melody-listener” or “harmony-listener”,  $F(1, 4528) = 0.557, p = .455$ .

### Average Performance Per Condition

The main goal of this experiment was to examine the change recognition performance in terms of timbre conditions based on timbre saliency and timbre dissimilarity. For this purpose, we computed the average of recognition rates of all the melodies used per voice per condition and compared those average values. The results are presented in Figures 4.7 and 4.8 for two- and three-voice conditions, respectively. The horizontal axis shows the saliency conditions and each line represents the dissimilarity conditions.



**Figure 4.7:** Results for two-voice excerpts. The error bars show the standard error of the mean.

In Figure 4.7, the left graph shows the participants' performance in the high voice and the right in the low voice. The blue dashed lines with dots represent the conditions where the two timbres are located close in timbre dissimilarity space, whereas the red lines with triangles signify the conditions where the two timbres are far in timbre space. Each of

the points on lines is the mean performance per condition and the vertical lines show the standard error of the mean. Note that some error bars overlap. This suggests that the average performance within one condition still varies a bit, possibly due to the individual participant's aptitude. It could also reflect various degrees of how easily the excerpts and their melodies can be heard out.

Focusing on the mean values (blue dots and red triangles), we see that the means of the high voice on the left are distinctly higher than those of the low voice on the right. This could imply a significant effect of voice position on average recognition performance. On the high voice alone, the blue dots and red triangles are all around the same area, suggesting that there be little effect of timbre dissimilarity condition. On the other hand, the red triangles of the low voice are positioned higher than the blue dots, which indicates that there may be an effect of the dissimilarity condition on low voice recognition performance. If we find mid-points of blue dots and red triangles corresponding to each of the timbre saliency conditions, they seem to be all around the same points on both the high and the low voices. This seems to hint that timbre saliency probably did not affect the recognition performance of either voice.

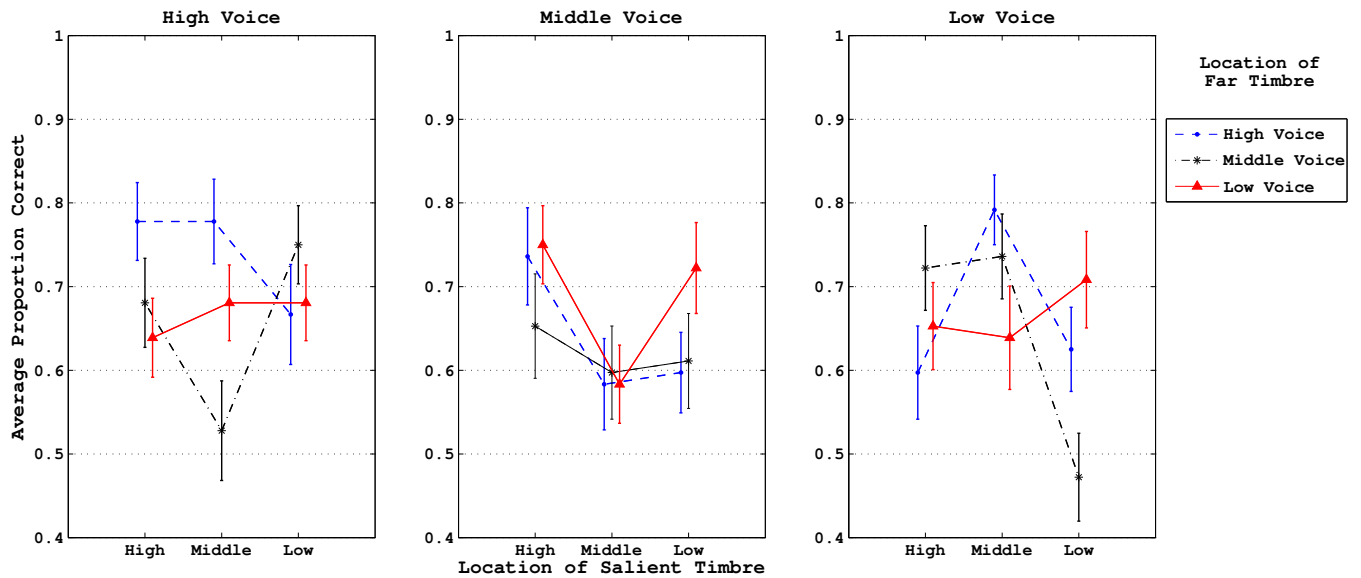
A three-way repeated measures ANOVA was performed on the average recognition rate per condition as DV. The voice type (high or low), dissimilarity condition (both close or far in dissimilarity space) and saliency condition (both high saliency, both low saliency, or opposite saliency) were within-subject factors. The main effect of voice type was significant as we guessed from Figure 4.7,  $F(1, 35) = 32.500, p < .00001$ , that the performance varied systematically with the voice type. The dissimilarity condition was also significant,  $F(1, 35) = 4.608, p = .039$ . The saliency condition did not meet Mauchly's Sphericity condition, hence Huynh-Feldt correction was used as  $\epsilon$  was greater than 0.75. The saliency condition showed neither a significant main effect,  $F(1.732, 60.608) = 0.782, \{\chi -$

$square(2) = 7.814, p = .020\}$ ,  $\epsilon = 0.866, \eta_p^2 = .022, p = .445$ , nor significant interaction effects:  $F(2, 70) = 0.135, p = .874$  for the interaction with voice,  $F(2, 70) = 1.652, p = .199$  for the interaction with dissimilarity, and  $F(2, 70) = 0.383, p = .683$  for the interaction with voice and dissimilarity. On the other hand, there was a significant interaction effect of voice and dissimilarity,  $F(1, 34) = 5.623, p = .023$ , indicating that the effect of dissimilarity is dependent on the voice. This voice-dependent effect of timbre dissimilarity is analyzed further in the following two-way ANOVAs.

A two-way ANOVA on the high voice showed no effect of timbre saliency,  $F(2, 70) = 0.640, p = 0.531$ , or timbre dissimilarity,  $F(1, 35) = 0.000, p = 1.000$ , as we expected from the left graph of Figure 4.7. The interaction of timbre saliency and timbre dissimilarity was not significant either,  $F(2, 70) = 1.358, p = .264$ . This lack of any significant effect on the high voice seems to mean that, in the given experiment, any additional boost from dissimilarity and saliency conditions did not provide any additional benefit to the high voice in two-voice excerpts, which is the most salient to recognize in the chosen musical form. On the low voice, a significant effect of timbre dissimilarity,  $F(1, 35) = 9.072, p = .005$ , was found, but not of timbre saliency,  $F(2, 70) = 0.104, p = .901$ , or of their interaction,  $F(2, 70) = 0.681, p = .510$ . This is in agreement with what we conjectured from the right graph of Figure 4.7; there is a clear offset between the two lines, confirming a significant main effect of dissimilarity. It illustrates that having highly dissimilar timbres on two voices helped the recognition of the low voice. It may be from the fact that two dissimilar timbres would tend to segregate more in auditory scene analysis (Iverson, 1995; Bey & McAdams, 2003), which would make it easier to listen to the low voice melody.

Figure 4.8 presents the effects of timbre saliency and timbre dissimilarity conditions on the average recognition rate in three-voice excerpts. Again, the timbre saliency conditions are specified on the horizontal axis and each of the timbre dissimilarity conditions are shown





**Figure 4.8:** Results for three-voice excerpts. The error bars show the standard error of the mean.

by different lines. The error bars in this figure are larger than those in the two-voice case in Figure 4.7. Also the range of the average performance per condition varies much more than that in the two-voice condition.

Considering the mean values (blue dots, black stars and red triangles) only, we see that they form a v-shape, although the angles seem to vary in the high and the low voices. The three v-shaped lines in the middle voice appear to maintain the same direction, which suggests that the timbre saliency condition may play an important role in the recognition of the middle voices. The fact that the lines keep a similar shape in the middle voice graph but not in other two voices implies a possible main effect of voice type or an interaction between timbre saliency and voice type.

A three-way repeated measures ANOVA was run in the same way as with the two-voice case. The only significant effects were interactions: between voice type and saliency,

$F(4, 140) = 3.859, p = .005$ , and between voice type, saliency, and dissimilarity,  $F(8, 280) = 3.142, p = .002$ . None of the other effects were significant: the main effects of voice type,  $F(2, 70) = 1.084, p = .344$ , timbre dissimilarity,  $F(2, 70) = 1.658, p = .198$ , and timbre saliency,  $F(2, 70) = 2.297, p = .108$ , the interaction between voice type and dissimilarity,  $F(4, 140) = 1.051, p = .384$ , and the interaction between saliency and dissimilarity,  $F(4, 140) = 1.911, p = .112$ .

The significant voice type-saliency interaction means the effect of saliency condition differs per voice. This may imply that the innate ‘voice prominence’ from this musical structure may have a bigger impact on melody recognition than the timbre saliency condition. The significant three-way interaction of voice type, dissimilarity and saliency indicates that the two-way interaction effect between dissimilarity and saliency interaction differs depending on the voice type. This is in agreement with the fact that in Figure 4.8, the dissimilarity-saliency interaction (i.e., the angles of v-shape lines) seems to be higher for high and low voices, but not for the middle voice.

As in the two-voice case, two-way ANOVAs were performed to study the effect of timbre dissimilarity and timbre saliency for each voice type. On the high voice, the interaction effect was significant,  $F(4, 140) = 3.123, p = .017$ , but not the main effects of timbre dissimilarity,  $F(2, 70) = 2.283, p = .110$ , or of timbre saliency,  $F(2, 70) = 0.912, p = .406$ . The significant interaction can be seen from three non-parallel lines in the high-voice graph of Figure 4.8. Note that the performance of eight conditions (i.e., the locations of eight points) are in the range between 0.63 and 0.78, and there is one outlier with the performance around 0.53 (a black star). This happens in the condition where the far timbre is on the high voice (and the salient timbre on the middle voice), which should have helped the recognition of the high voice melody according to our hypothesis. Perhaps this low performance may come from the musical excerpts used in this condition and somehow the low voice melodies

were a lot more salient (average performance of 73% correct) than the top melodies.

On the middle voice, the main effect of timbre saliency turned out to be significant,  $F(2, 70) = 4.692, p = .012$ , but not timbre dissimilarity,  $F(2, 70) = 1.041, p = .358$ , nor their interaction  $F(4, 140) = 0.709, p = .587$ . The three v-shaped lines in the middle voice graphs of Figure 4.8 have similar shapes (hence no significant interaction effect) and locations (hence no significant main effect of dissimilarity). What is strange is that the performance on the middle voice was at its worst when the salient timbre was on the middle voice. This can be observed in all three dissimilarity conditions, probably suggesting that the effect of a salient timbre was minimal on the middle voice. It is also hard to understand that the recognition performance on the middle voice (black dash dotted line connecting stars) was worst when the far timbre was assigned to the middle voice. In summary, this graph (three lines) seems to suggest that there are no effects of dissimilarity or saliency on the middle voice as we had hypothesized.

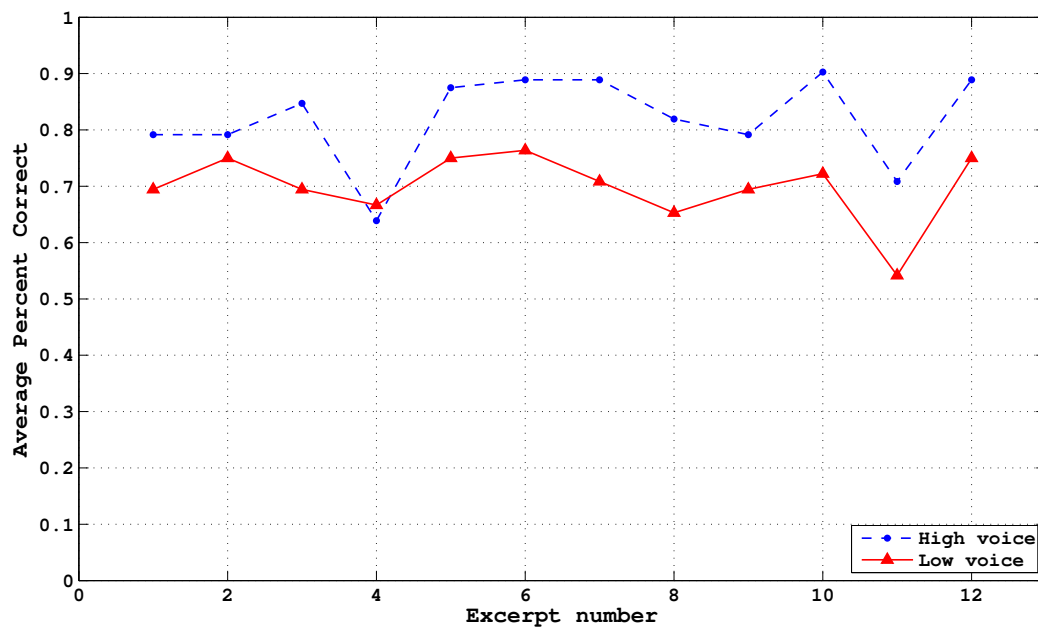
A two-way ANOVA on the low voice showed two significant effects: the main effect of timbre saliency,  $F(2, 70) = 3.661, p = .031$ , and its interaction with timbre dissimilarity,  $F(4, 140) = 4.562, p = .002$ . The main effect of timbre dissimilarity was not significant,  $F(2, 70) = 0.284, p = .754$ . The v-shapes face different directions, reflecting the significant interaction effect. However, it is strange to see that the lowest performance was when the salient timbre was on the low voice. Having the salient timbre on the low voice was expected to help the recognition performance, but apparently it did not. A close look reveals that performance is not too bad when the salient timbre is on the low voice and the far timbre is on the high or low voice. But somehow having a far timbre on the middle voice hindered the recognition of the low voice melody so much that the performance actually fell below 50%. This might be coming from the saliency differences inherent in stimuli; somehow the low voice melodies were not salient at all and participants' attention was drawn to the

salient high-voice melodies in the given condition.

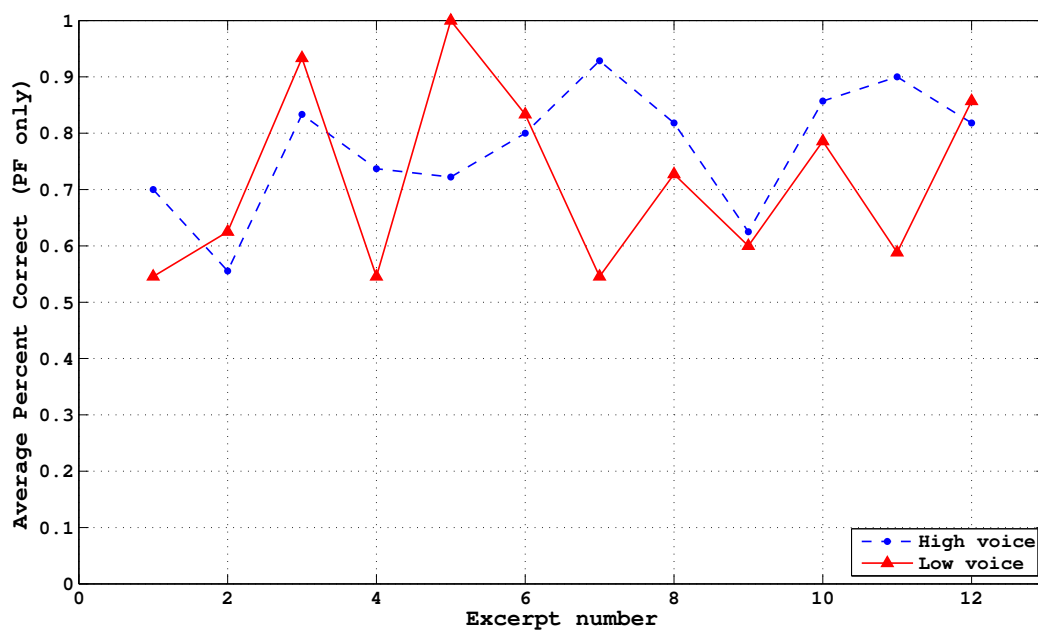
Overall, it is quite disappointing to see that the recognition was not at the highest (with an exception of the high voice) when a voice had both the salient and the far timbre, which had been hypothesized to maximize the effects of these conditions on the recognition task. For example, the high voice graph on the left of Figure 4.8 reaches the maximum performance at the left blue dot, when the salient and far timbre happened to be on the high voice, but this is not the case in the other two graphs. The black star in the middle of the dash dotted line, which was hypothesized to be the highest point, of the middle voice graph is located much lower than the actual highest point (a red triangle). In fact, it is puzzling to see the low performance on the middle voice when it was played with the salient timbre. We began to wonder if the middle voice melodies used for this condition happened to be too difficult. To study this, we decided to study the average recognition performance for each excerpt, which is presented in the next section.

### Average Performance Per Excerpt

A number of excerpts were used in this experiment and the average percent correct performance across all participants is shown in Figures 4.9 and 4.10 for two- and three-voice cases, respectively. As we can see from these figures, there is quite a bit of variability across the excerpts used. This might be due to the fact that some excerpts are more difficult to remember than others. At first glance, the multi-timbre average curves (Figures 4.9a and 4.10a) look a bit different from the mono-timbre ones (Figures 4.9b and 4.10b). A paired-sample t-test on these average performances show that these seeming differences are not significant:  $t(11) = 1.312, p = .216$  for the high voice and  $t(11) = -0.392, p = .702$  for the low voice in two-voice excerpts;  $t(8) = -0.212, p = .838$  for the high voice,  $t(8) = 0.928, p = .380$  for the low voice in three-voice excerpts. A marginally significant difference was found on the

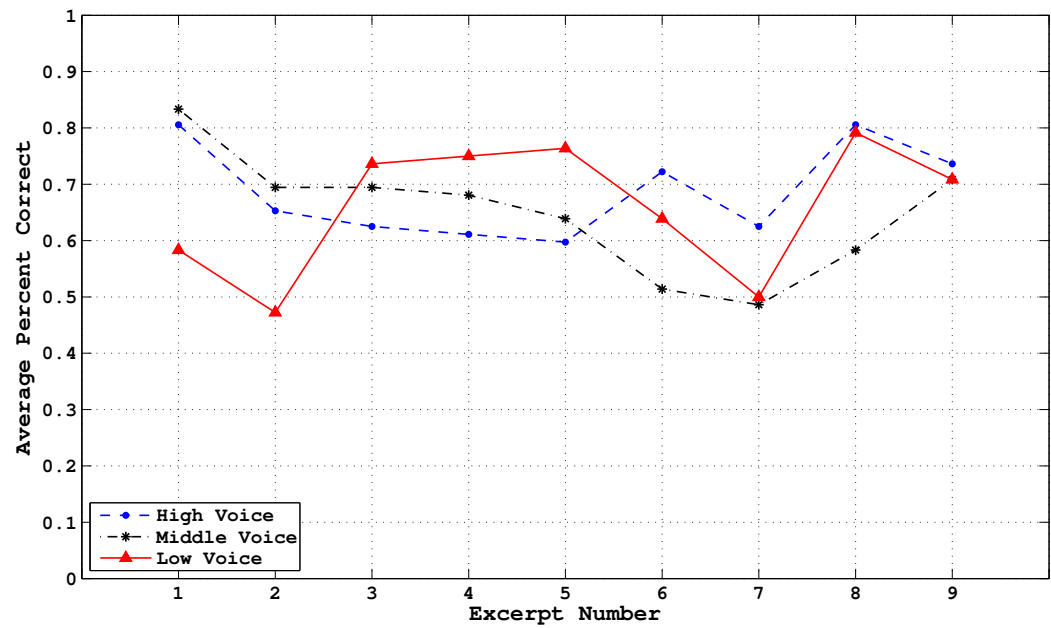


(a) Two-voice multi-timbre cases

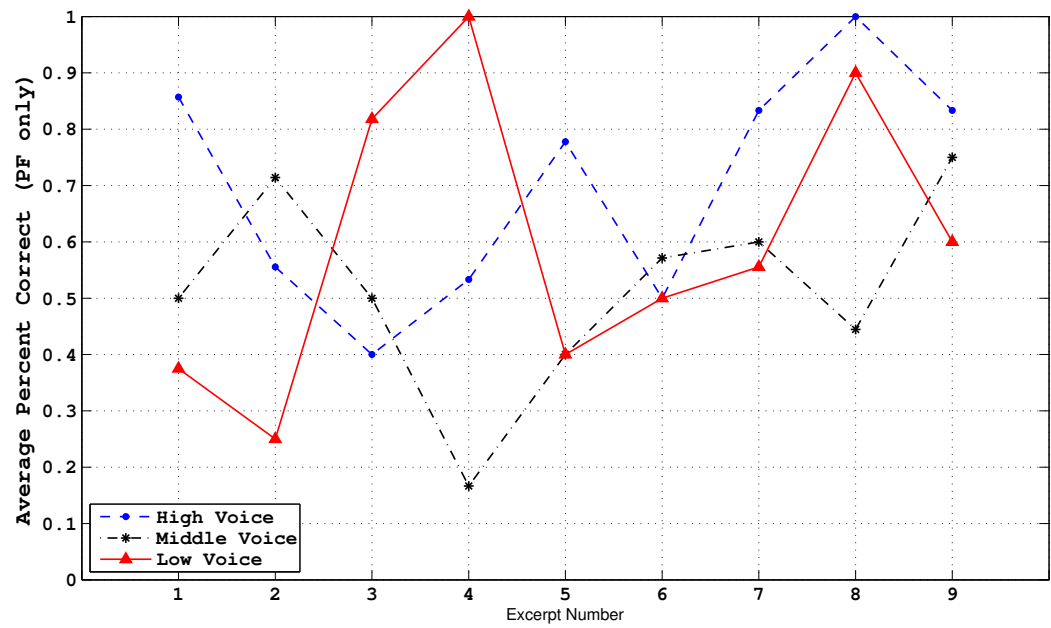


(b) Two-voice mono-timbre cases

**Figure 4.9:** Average performance of two-voice excerpts



(a) Three-voice multi-timbre cases



(b) Three-voice mono-timbre cases

Figure 4.10: Average performance of three-voice excerpts

middle voice in three-voice excerpts,  $t(8) = 1.887, p = .096$ , where the average recognition rate of the middle voice in multi-timbre condition was 0.65 ( $STD = 0.11$ ) whereas that in mono-timbre condition was 0.52 ( $STD = 0.18$ ). This may suggest that having a distinctive timbre on the middle voice helps its recognition, which would usually be the most difficult to listen to according to the musical structure.

An independent-samples t-test on the average recognition rate across voices in two-voice excerpts and that in three-voice excerpts shows a significant difference,  $t(148) = 4.324, p = .000014$ . DVs are the average recognition performance of each two- and three-voice melody in multi-timbre conditions. The average recognition rate in two-voice excerpts was 76%, which is 9 percent higher than the average in three-voice excerpts of 67%. This probably reflects the higher complexity in the task with three-voice excerpts. A further analysis shows that the difference in average recognition rates of two- and three-voice excerpts comes from the recognition of high voice. The average recognition of the high voice in two-voice excerpts was 82%, which was significantly different from that in three-voice excerpts of 69%,  $t(64) = 4.190, p = .000087$ . This difference was unique to high voice recognitions. The low voice recognition rate in two-voice excerpts was not statistically different from the middle or the low voice performance in three-voice excerpts on average:  $t(64) = 1.518, p = .134$  for the middle voice;  $t(64) = 1.110, p = .271$  for the low voice. These results suggest that only the high voice recognition significantly suffers on average when the stimulus texture increases from two to three voices.

Since the average performance per excerpt varied quite a bit, we came to wonder if this is related to how easily the changes in corresponding voices could be heard out in Experiment V (Section 4.4). Hence, the average recognition rate per voice in each excerpt was correlated with the proportion correct values from Tables 4.11 and 4.12. Spearman's rank correlation was used to investigate a possible monotonic but nonlinear relationship.

Table 4.13 provides the analysis result that there was almost no correlation, except for the case of the low voice in three-voice excerpts. This lack of correlation could reflect the fact that the current experimental task is too complex to be successfully predicted by the control experiment result.

**Table 4.13:** Summary of Spearman’s rank correlation analysis of each of five voices and the ‘percent correct’ values from the control experiment

Voice		Correlation coefficient $\rho$
two-voice	High	0.065 ( $df = 10$ )
	Low	−0.333 ( $df = 10$ )
three-voice	High	−0.000 ( $df = 7$ )
	Middle	0.153 ( $df = 7$ )
	Low	0.587 ( $p = 0.096, df = 7$ )

### Effect of Timbre Orders

The experimental design of two-voice conditions was full factorial in the sense that we considered the effect of one timbre on both voices in a combination with another specific timbre (e.g., PF-HA and HA-PF). This was not possible for the three-voice conditions, so in this section we analyzed the two-voice results only.

The questions that prompted this analysis were “is there a performance difference on a specific voice depending on which timbre was on it? For example, did the voice position of PF affect the recognition performance when it was combined with HA?” To answer these questions, paired-sample t-tests were performed on the average performance of each voice (high or low) with each timbre on the voice (e.g., PF or HA) for each of six timbre conditions.



For example, the numbers in the ‘High Voice’ column on the first row of Table 4.14 show the result of a paired-sample t-test on each participants’ average recognition rate of the high voice in the PF timbre in Condition 1 with that of the high voice in the HA timbre, which were the corrected using the Holm-Bonferroni method. As the numbers show, there were no significant recognition difference that came from the position of PF and HA timbres on either voice. The three significant cases from the paired-sample t-test turned out to be still significant after the Holm-Bonferroni method, which are specified with the  $p$ -values in bold fonts: both high and low voices in Condition 2 and the low voice in Condition 5.

**Table 4.14:** Summary of paired-sample t-test to study the effect of timbre orders on the average performance in two-voice excerpts

Condition No.	Dissimilarity Condition	Saliency Condition	High Voice		Low Voice	
			$t(35)$	$p$	$t(35)$	$p$
1	Close	Both Higher	−0.941	.353	−0.533	.597
2	Close	Both Lower	−3.618	<b>.001</b>	3.090	<b>.004</b>
3	Close	Opposite	0.422	.676	−0.349	.729
4	Far	Both Higher	−0.466	.644	0.000	1.000
5	Far	Both Lower	−0.941	.353	−2.142	<b>.039</b>
6	Far	Opposite	1.673	.103	−0.403	.689

Let’s look further into what excerpts these three conditions use. In Condition 2, on the high voice Excerpts 5, 8, 9 and 12 in CL-TN are compared with Excerpts 2, 3, 6 and 7 in TN-CL, the latter (TN-CL) of which show the better performance. On the low voice in Condition 2, Excerpts 2, 3, 6 and 7 in CL-TN are compared with Excerpts 5, 8, 9 and 12 in TN-CL, and the first have the higher performance. On the low voice in Condition 5, Excerpts 3, 4, 6 and 7 in CL-HC are contrasted with Excerpts 2, 9, 10 and 11 in HC-CL,

and the latter performed better.

This is quite mysterious, especially if we contrast the result with the saliency dimension in Figure 2.4. According to our hypothesis that a more salient timbre would make the voice easier to listen to, the voice in CL should have better performance when it is combined with TN. Instead, in Condition 2 on both voices, the performance was better when TN was on the voice in question. The same phenomenon is observed on the low voice in Condition 5 that the recognition performance was higher when the low voice was in CL rather than HC, when HC is more salient than CL on the saliency dimension.

The inverse relationships found in Condition 2 against the hypothesis provide arguments against a systematic effect of the voice position of the two timbres, since the timbres remain the same but the results are inverted. If we consider the average performance in Condition 2, it was higher when CL was on the target voice

What is surprising and unexpected is that in the three significant conditions, the performance is better when a less salient timbre is on the voice in question. As there is no scientific explanation of what might be the cause, we conjecture that it could be due to certain unforeseen (and still unidentifiable) interactions between timbre and melodies in the excerpts used. For a complete analysis, another experiment seems necessary to test the reverse set of current timbre-excerpt combinations such as testing the high voice of Excerpts 5, 8, 9, 12 in TN-CL and contrast their performances.

#### 4.5.3 Discussion

In this experiment, we studied the effect of timbre saliency and timbre dissimilarity on the voice recognition in counterpoint music. Twelve two-voice and nine three-voice excerpts were chosen. Comparison melodies were composed by changing the pitches of two notes in each voice in these excerpts that would result in a different melodic contour. The stimuli

were created in different timbre combinations that were selected to best represent differences in timbre saliency and dissimilarity.

Considering previous work in auditory streaming, which demonstrated that greater timbre dissimilarity led to a higher recognition of interleaved melodies (Iverson, 1995; Bey & McAdams, 2003), as well as our measurement of timbre saliency (Chapter 2), we hypothesized that a highly dissimilar or a highly salient timbre would enhance a voice's prominence in a multi-voice texture. We were also confident of our choice of counterpoint music excerpts, where each voice has about equal musical importance, unlike some of stimuli that Gregory (1990) used that have a more classical melody-and-accompaniment type of structure.

The result from 36 musicians was mostly unexpected and against our hypothesis. The low voice in two-voice excerpts was the only case with a significant effect of timbre dissimilarity: having highly dissimilar timbres contributed to a higher recognition performance of the low voice. This agrees with reports by Iverson (1995) and Bey and McAdams (2003). But no effects of timbre saliency were found in the two-voice excerpts. This lack of effects of timbre saliency may come from the fact that Experiment VI employed concurrent melodies, unlike the aforementioned studies that used interleaved melodies. The only melody recognition study with concurrent melodies that we are aware of is the work by Gregory (1990), but it still does not provide a fair comparison because the timbre conditions were not tightly controlled in that study, as well as because the excerpts had a more traditional melody-and-accompaniment type of structure rather than counterpoint where all voices have about equal importance.

Middle and low voices in three-voice excerpts showed a significant effect of saliency, although not the way we expected: the average performance was poorer when the salient timbre was located on the target voice. This certainly does not support our hypothesis.

There could be many reasons for this lack of support, as there must be more than one way (i.e., our experimental design) to test our hypothesis. One difficulty with the experimental paradigm for Experiment VI came from the fact that it made use of a short-term memory, which is different from an average listening situation where listeners would listen to longer excerpts rather than many short excerpts (without having to remember the melodies in them). They also may have developed a listening strategy for this experiment, although none of our listeners articulated any particular strategy that they were consciously using.

In searching for an answer to this unexpected pattern, we looked at the average recognition performance for each of the excerpts used. It turned out that there was a large difference in per-excerpt performance, which could have come from various degrees of memorability that affected the recognition performance. This variance in per-excerpt performance could also have contributed to differences in per-condition performance.

As there was no significant difference in average recognition of each excerpt-voice according to the timbre conditions (multi-timbre vs. mono-timbre), with the only exception of the middle voice in three-voice excerpts; the lack of effects of timbre saliency may actually indicate a greater voice prominence within the given musical structure than whatever timbral effects we expected. After all, we had not studied the intrinsic saliency of each voice in the two- and three-voice counterpoint structure. This could be an example of the extrinsic saliency (that comes from the experimental setup) affecting the measure of intrinsic saliency to the degree that almost annuls our hypothesis.

However, the fact that the average recognition of the middle voice in three-voice excerpts was almost significantly higher in the multi-timbre condition in comparison with the mono-timbre condition does speak for the case of timbral effects. The middle voice, which is the most difficult to listen to in three-voice music, became easier to recognize with the use of a timbre different from those on other voices. Unfortunately for us, this effect seems too

mild to be reflected and measured systematically in the current experimental setup.

The big variance in the recognition performance also make us hesitate in drawing firm conclusions based on the analysis. In the per-condition performance, all the means and the respective confidence intervals overlapped without exception. Hence, the analyses based on mean values lose their effectiveness when we consider the large variance.

It was disappointing not to see the strong effects of timbre saliency and timbre dissimilarity we expected. What we learned instead was another incidence of a context effect, which was possibly a lot stronger than our planned timbral effects in this experiment. To clarify unanswered questions, another experiment using the untested portion of the current stimuli seems to be in order, which will provide data that can complement this experiment so that we can apprehend the big picture.

## 4.6 General Discussion and Conclusions

To examine the effect of timbre saliency and timbre dissimilarity in a more realistic music listening, a voice recognition experiment was carried out as a natural extension of the previous study of the perception of blend in concurrent unison dyads (Chapter 3). As a mild negative relationship between timbre saliency and the perceived blend was observed in the concurrent unison dyads, we hypothesized a similar finding that a highly salient timbre would show little blend with other voices in the musical texture therefore be heard out more distinctively. Considering the effect of timbre dissimilarity for the first time in this dissertation, we also expected to confirm previous findings in the auditory streaming literature (Iverson, 1995; Bey & McAdams, 2003) that a highly dissimilar timbre on a voice would help recognize that voice more easily compared to other voices in multipart music.

To realize this big experiment, two small experiments had to be performed to set up the

basis. Experiment IV (Section 4.2) studied the dissimilarity perception of the timbres that have been studied in the saliency experiments, the result of which assisted in the stimulus design of Experiments V and VI (Sections 4.4 and 4.5, respectively). In Experiment V, the changes in corresponding voices that were chosen for Experiment VI were verified in terms of how easily they could be heard in isolation. All pairs had at least 71% recognition rate, suggesting these changes were quite obvious when only one voice was presented. However, when we combined these individual voices into a two- or three-voice texture, the change recognition became more difficult.

For Experiments V and VI, excerpts were selected from *Trio Sonatas for Organ*, BWV 525 – 530 by Bach (1730) and modified melodies composed by changing the pitch of two notes for each voice of the excerpts. The choice of counterpoint music as stimuli seemed natural as we wanted to have a musical form where all voices have almost equal importance unlike some stimuli that Gregory (1990) used that had a more melody-and-accompaniment type of relationship. Sound stimuli were generated using the selected excerpts in different instrument combinations, which were determined from timbre saliency and timbre dissimilarity conditions (Section 4.3).

The low voice in two-voice excerpts showed a significant effect of timbre dissimilarity condition that having two highly dissimilar timbres boosted the recognition of the low voice. This is in agreement with previous reports in the auditory streaming literature that timbre dissimilarity increased the recognition of the target melody in interleaved melodies. The high voice, however, did not show any effect of timbre dissimilarity or timbre saliency, probably because it is already the most prominent voice in the chosen musical structure. This ‘voice prominence’ was probably a lot more salient than any possible additional benefits from timbre saliency and dissimilarity conditions. The recognition performance on the high voice is about 85% on average, which implies that there is still room for improvement.

However, the fact that neither timbre saliency nor timbre dissimilarity conditions affected the recognition may suggest that 85% is about the best performance using the short-term memory task in the context of two or three competing melodies.

In three-voice excerpts, the high voice also did not show any main effect of timbre saliency and timbre dissimilarity condition either, probably due to the same reason concerning the high voice in two-voice excerpts. There was a significant interaction effect observed, however, suggesting that the effect of timbre dissimilarity varied with timbre saliency (and vice versa). Middle and low voices showed a significant effect of timbre saliency condition, but not in the same direction as our hypothesis. In fact, the average recognition performance was lowest when the salient timbre was located on the target voice. This was completely unexpected, and we are still puzzled with no apparent explanation. Perhaps the excerpts happened to have very difficult middle and low voice melodies that were used in these conditions?

So we decided to look into the per-excerpt average performance, hoping that it would shed light that could explain the aforementioned observations on middle and high voices. The average performance varied over a wide range, suggesting the variability in each excerpt's (and each voice's) memorability. As one can easily expect, the average recognition in two-voice excerpts was significantly better than in three-voice excerpts. What was interesting, however, was that the performance of "lower parts" (i.e., the low voice in two-voice excerpts and the middle and low voices in three-voice excerpts) were not statistically different. The performance difference resulted from the recognition of the high voices! It seems the high voice in two-voice excerpts, which is the most salient of all five voices that we considered, loses quite a bit of its saliency when there is one more voice in the lower part. It will be interesting to compare this result with a four-voice recognition experiment in the future, to see if this decreased recognition of the high voice and the unchanged recognition

of the lower voices still hold.

When each excerpt's average recognition performance in the multi-timbre conditions was contrasted with that in the mono-timbre condition, the only almost significant difference was observed on the middle voice in three-voice excerpts. The recognition performance was much higher on average (by 13%) in the multi-timbre condition. This suggests that the middle voice, which has the least 'voice prominence' in the chosen musical structure, benefited from having a different timbre from the other voices, which agrees with previous literature on timbral effects on auditory streaming.

However, the fact that this additional benefit did not make any significant difference in average performance per timbre condition led us to think about context effects again. As we hypothesized that there is an intrinsic saliency for each object and extrinsic saliency for each context in which the object's saliency is being measured, the limit in our experiment might have been that we did not consider the inherent prominence of each voice position in the musical form that was selected for the experiment. In the current setup, the 'voice prominence' might have been a lot more salient than our timbre conditions, where the extrinsic saliency might have overpowered the intrinsic saliency.

More questions arose when we analyzed the effect of timbre orders in two-voice excerpts. Unlike the three-voice excerpts, the two-voice conditions involved a full factorial design, because the number of possible combinations was small enough. As one timbre (for example, PF) is tested on both the high voice and the low voice in combination with one specific timbre (e.g., HA), is there any systematic performance difference according to the specific timbre's position within the given combination? It turned out that these orders did not make any difference for most conditions. However, in the "close in timbre dissimilarity space with both lower saliency values" condition, the recognition was better for both high and low voices when *the less salient* timbre was on the voice in question. The same pattern was



observed on the low voice in the “far in timbre dissimilarity space with both lower saliency values” condition. These were completely opposite of our hypothesis. It could be that the specific excerpts and voices used in these conditions were unusually difficult to recognize. It could also reflect some unexpected interaction between timbre and melodies in the excerpts used. Whatever the real cause might be, we are still puzzled. Perhaps another experiment using the untested portion of the same stimuli is in order so that contrasting the results from the two experiments could provide us a big picture with more complete explanations.

Reflecting on the complexity of Experiment VI, we wonder if we should have started with a simpler experiment. Perhaps it would help to carry out a new experiment with simplified conditions to verify the effect of timbre saliency and timbre dissimilarity, where the stimuli have only two conditions – a “high” condition with a highly salient and dissimilar (i.e., far in dissimilarity space) timbre and a “low” condition with a not so salient and similar timbre. This should be able to clearly contrast the performance in each condition to examine the effect of timbre saliency and timbre dissimilarity. We can also execute this same experiment again with the set of stimuli that were not tested currently. Because each three-voice excerpt in a particular timbre combination was tested with only one voice, we can make use of the untested voices and run the same analysis on the new data to contrast with the current findings.

Another idea is to conduct an experiment also utilizing top-down attention instead of the current melody recognition paradigm, which depends on bottom-up attention and short-term memory. Imagine a short cue, an isolated note in a certain pitch and timbre, is played before a polyphonic excerpt is played. What happens to the recognition rate? Do listeners tend to get drawn more towards the voice close to the pitch of the cue? Or to the voice that has the same timbre? This may bring us to an interesting interaction of top-down and bottom-up attention together.

Also, more fundamentally, the relationship between timbre saliency and timbre dissimilarity needs to be examined. In the design of experiments in this chapter, we started under the assumption that timbre saliency and timbre dissimilarity would be at least somewhat related to each other and that there would not be any negative interaction between them. But the result from Experiment VI seems to suggest that in some cases auditory streams based on timbre saliency and timbre dissimilarity differences might have been competing for listeners' attention, which in turn may have contributed to the result which was not supportive of our hypothesis.

What then is the difference between saliency and dissimilarity? Consider timbres that are close on the saliency dimension in Figure 2.4 and far in the timbre dissimilarity space in Figure 4.2, for example, VC and HC. Apparently they are highly dissimilar, according to the large distance between them in the timbre dissimilarity space, but they are positioned right next to each other on the saliency dimension, suggesting they are almost equally salient. Another pair of timbres, EH and TP, illustrates an opposite scenario. They are quite similar in terms of the dissimilarity, but they are located on the opposite ends of the saliency dimension. These examples might suggest that the relationship between timbre saliency and dissimilarity is not a straight forward one where one can explain the other.

Perhaps the relationship between timbre saliency and dissimilarity could be considered analogous to that between hue (or brightness) and dissimilarity in colour perception. If there are a pair of red and green of similar hue levels, they may be balanced in terms of visual saliency even though they are highly dissimilar. But if a subdued red is placed next to a neon pink, the neon pink will likely draw more attention to it even though the two colours are not highly dissimilar in terms of their underlying RGB values. Reflecting these examples to our question of the relationship between timbre saliency and dissimilarity also confirms that it would not be a simple one.

There is one more detail to consider to understand the possible relationship between timbre saliency and dissimilarity. The MDS results would be different from the current ones, if there was one new sound considered for our saliency and dissimilarity judgment experiments, as the relative proximity and dissimilarity judgments would be changed with the additional stimulus. This means that the two MDS spaces we obtained from the tapping and the dissimilarity experiments need to be explained within the context of the stimuli used. Hence, the relative distances between a pair of stimuli carry more importance than the absolute positions of them in a MDS space. This is easy to apply in the timbre dissimilarity space. However, it is not clear how to interpret the distance between two stimuli on the timbre saliency dimension. For instance, the distance between EH and TN, the two timbres at the lower end of the timbre saliency dimension, is greater than the distance between PF and TB, which are at the upper end of the timbre saliency dimension. What does it mean in terms of the timbre saliency? Does it mean that the EH-TN pair has a greater saliency difference than the PF-TB pair even if the latter belongs to the group of more salient timbres than EH and TN do? A further investigation is required to answer these questions. Perhaps, after studying these questions, we might have a new insight to bring to understanding the results in this chapter.

One thing that we learned from carrying out this complex and complicated experiment is that counterpoint music is such a sophisticated art that could not be sufficiently analyzed with our model. Saliency is a function of context, and our measure of timbre saliency might not have been effective in the context of melody recognition in counterpoint music, especially when each voice position's prominence is unknown. As this was our first attempt to explain the perception of music in terms of timbre saliency, any findings are important. However disappointing or puzzling the findings were, these will lead to a new journey with more questions to answer, which will eventually help us understand what catches our

attention in music, which was the starting point of this dissertation.

## Chapter 5

# Conclusion

This chapter summarizes the content of this dissertation, including analysis results and findings. Also presented are a discussion on the relevance of this work and future directions in this line of research.

### 5.1 Summary

#### 5.1.1 Proposition of A New Research Topic

This dissertation presents the novel research topic of timbre saliency (Chapter 1) and a number of experiments which were carried out to establish the concept (Chapters 2, 3, and 4). Timbre saliency refers to the attention-capturing quality of timbre. Saliency means the quality of standing out with respect to its neighbouring objects, often catching listeners' involuntary attention. Auditory saliency is still little understood in comparison with visual saliency where there is a luxury of being able to track someone's visual attention patterns with their eye movements.

As with visual saliency, auditory saliency must be a function of context. There must

be two parts of saliency – *intrinsic* saliency, that is, the unique degree of saliency of an object and *relative* (or extrinsic) saliency. It is the relative saliency that gets affected by the context, although there is no way of knowing exactly how much of what is being measured is the intrinsic saliency and how much is the relative one, because each experiment sets up a certain context. This is probably why in visual saliency research some reported the form to be more salient than colour (Poiese et al., 2008), whereas others found colour to be more salient (Theeuwes, 1992). Their conclusions are all valid, because the contexts were all different. Hence what they measured probably consisted of different portions of intrinsic and relative saliencies, which makes it even harder to study saliency.

Apart from the context, auditory saliency must be a function of loudness and pitch. We have all had the experience that a loud or high-pitched sound is difficult to ignore. But what happens if we control loudness and pitch? Is there something that makes a particular timbre more efficient than others in catching listeners' attention, even when they have the same pitch, duration and loudness? Are there any different levels of saliency based on timbre?

Imagine listening to an orchestra. When there is a trumpet playing with other instruments, it stands out even when it is not too loud. A French horn, which shares a good portion of the playing pitch range with the trumpet, is quite the opposite. French horns are not as noticeable in general, which is probably why they are commonly used in blending with other instruments. Here, the French horn and the trumpet may denote opposite sides of intrinsic saliency. But when a trumpet is combined with a trombone, something interesting happens with the blended timbre and the trumpet is still recognizable but it is not as prominent as the trumpet without the trombone. This can be an example of the “context” of blending to affect the relative level of saliency.

This led me to wonder if every instrument has a varying degree of ability to catch the

listeners' attention, which I refer to as timbre saliency. If this is indeed true, then timbre saliency might have been an implicit factor in the orchestration practices in the history of Western music, which affected popular instrument combinations. When further developed, timbre saliency may provide a novel tool for composing music and sound in applications such as computer-aided composition programs and sound alarm design.

### 5.1.2 Experiments and Results

As a part of this dissertation work, I have conducted the following experiments to lay the ground for this new field. Below is a brief summary of each experiment.

#### Experiment I: Timbre saliency using indirect measurements (Section 2.2)

Saliency differences between 15 instrument timbres were measured using a tapping technique. Listeners were instructed to tap to what sounded like the strong beat in ABAB perceptually isochronous sequences. The basic idea was that the more salient timbre would capture listeners' attention and be chosen more often as the strong beat. Stimuli were generated with 15 orchestral instrument samples from the Vienna Symphonic Library ([Vienna Symphonic Library GmbH, 2011](#)), and further equalized in pitch, loudness and duration (Appendix B.1) before being organized into perceptually isochronous ABAB sequences (Appendix B.2).

Data from 40 participants yielded a one-dimensional CLASCAL ([Winsberg & De Soete, 1993](#)) solution with two latent classes and specificities. The saliency dimension seems to have three regions – low, middle and high. Seven out of 15 timbres fell into the “middle” saliency region. This probably means that these timbres in the middle region are more or less equally salient to each other, although they are still more salient than the timbres in the low saliency region and less salient than those in the high saliency region. It appears

that this equal saliency within a group is unique to this middle group, as the timbres are located at different points along the saliency dimension in the other two groups.

Latent class structure showed no relationship with gender, musicianship or age, which is in agreement with an earlier report (McAdams et al., 1995). Testing audio descriptors from the Timbre Toolbox (Peeters et al., 2011), the odd-even harmonic energy ratio explains 51% of the variance along this dimension. This seems to suggest that saliency comparisons between timbres depend more on spectral envelope jaggedness. The result from this experiment became the basis for the following experiments.

## Experiment II: Timbre saliency using direct comparisons (Section 2.3)

Some participants' post-experiment comments in Experiment I suggested there might be an unexpected effect of rhythmic saliency confounded in the obtained data, since the perceptually isochronous ABAB sequences in Experiment I formed a duple rhythm, where the strong beats tend to be associated with lower or darker sounds and the weak beats with higher or brighter sounds, as is found in many musical cultures, including rock and jazz. We then decided to eliminate this rhythmic saliency by asking a straightforward instruction of "Choose the sound that stands out more, the one that grabs your attention more." The same sounds were used from Experiment I.

The result was different in that an effect of order was observed in more than a half of the 15 timbres, which was not found in the previous experiment. The same series of analyses was carried out ignoring these order effects, the result of which was a three-dimensional timbre saliency space with two latent classes and specificities. Two out of three dimensions showed only mild correlations with the timbre saliency dimension from Experiment I. Furthermore, there were problems with interpreting the acoustic correlates of these new dimensions, which made us less confident of this new timbre saliency space.



As the findings in analysis of this experiment did not seem to make full sense, the result was not reflected in designs of the following experiments. However, this illustrates the importance of employing the right experimental paradigm, which could affect the outcome as we can see here, especially in the context-sensitive cases such as measuring timbre saliency.

### **Experiment III: Exploring instrument timbre blending as a function of timbre saliency (Chapter 3)**

For an object to be perceptually salient, it should be more prominent than its neighbours, which requires little blend between the object and its surroundings. Therefore, the timbre saliency measure from Experiment I (Section 2.2) should have an inverse relationship with the perceived degree of blending that a highly salient timbre will not blend well with others. For this experiment, composite sounds were created from combining pairs of sounds from the 15 instrument sounds used in Experiments I and II. Participants were asked to judge the degree of perceived blending of these composite sounds on a continuous scale from “not blended at all” to “very blended.”

The mean blending ratings from 40 people showed statistically significant negative correlations with the sum, the minimum and the maximum of the two individual timbre saliency values. This means that a sound would blend better when it is not very salient, which goes in the direction of our hypothesis, but the correlations were moderate at best, implying that there may be other more important factors in explaining the perceived blend ratings. Using the acoustic descriptors from the Timbre Toolbox (Peeters et al., 2011), the best acoustic correlate turned out to be the minimum temporal centroid of the two underlying timbres, supporting previous reports by Tardieu and McAdams (2012) that non-percussive sounds would achieve better blending. This seems to suggest that even though timbre

saliency may influence the perceived blending, the outcome will be affected more by temporal centroids with performance choices such as controlling the attack pattern of a sound or the compositional choices in instrument combinations. Other previous findings in the literature were also confirmed that sounds with lower spectral centroids are likely to blend better by [Sandell \(1995\)](#) and [Tardieu and McAdams \(2012\)](#).

#### **Experiment IV: Perception of timbre dissimilarity (Section 4.2)**

Timbre dissimilarity is the first topic studied in the timbre perception research and there is a general consensus on the scientific factors that best explain the perception of timbre dissimilarity. While designing Experiment VI, we realized that a timbre dissimilarity space obtained from the sounds used in Experiments I and II would provide a firm basis to complement the timbre saliency dimension. A classic timbre dissimilarity experiment was carried out with the stimuli from Experiments I and II. Twenty participants judged the dissimilarity of pairs of timbres, the result of which turned out to be a two-dimensional space with five latent classes and specificities.

The acoustic correlates of the two dimensions of this timbre dissimilarity space turned out to be spectral centroid and temporal centroid, which is mostly confirmative of previous reports in the literature. The constellations of 15 timbres in the dissimilarity space was essential in the stimulus design of Experiment VI.

#### **Experiment V: Melody discrimination (Section 4.4)**

Two- and three-voice counterpoint excerpts were chosen for Experiment VI, the task of which was to be a comparison of a voice in the multipart excerpt followed by a single voice melody. The second melody could be exactly the same as one of the voices in the excerpt or slightly different with two notes changed in pitch. Since this task was quite complex, it

was necessary to make sure that the changes in pair of corresponding melodies were easy enough to hear out in isolation. Experiment V was conducted to verify this.

All pairs of original and modified one-voice melodies in two- and three-voice excerpts were presented in random order to 20 musicians, whose task was to report whether the given pair of melodies were the ‘same’ or ‘different.’ All four possible combinations were examined: original-original, original-modified, modified-original and modified-modified. Most melody pairs showed the correct discrimination rate above 75% threshold that we determined before the experiment, with the exception of three melody pairs. The worst performance was 71% correct, but we decided to go ahead and use these melodies in Experiment VI without any further modifications.

### **Experiment VI: Effect of timbre saliency and timbre dissimilarity on voice recognition in counterpoint music (Section 4.5)**

This was the last experiment as part of my dissertation work. As [Huron \(1989b\)](#) reported, the entries of outer voices are easier to notice than those of inner voices. Given that, can we enhance a voice’s recognizability by applying a highly salient timbre on it and not so salient timbres on other voices? That is the main question we were asking with this experiment. We decided to study two- and three-voice cases using *Trio Sonatas for Organ* BWV 525 – 530 by [Bach \(1730\)](#). This music was already written for three voices (right hand, left hand, and pedals) and is not very well known, which was important in order to avoid an undesirable familiarity effect. The experiment was a melody recognition task in which participants listened to an excerpt (with two or three voices) followed by a monophonic melody and judged whether the monophonic melody was the same or not as one of the voices in the excerpt.

In preparation for this experiment, we additionally carried out a loudness equalization

pilot using an eight-note melody (Appendix B.3) to obtain a set of equalization coefficients to apply in stimulus generation, a timbre dissimilarity experiment (Section 4.2) and a melody discrimination experiment to study how easily the difference could be recognized between an original (that Bach wrote) and the comparison melodies (Section 4.4).

Stimuli for the voice recognition experiment were generated according to conditions of timbre saliency and timbre dissimilarity. The hypothesis was that a highly salient timbre would enhance the recognition of the corresponding voice. The voice with the most dissimilar timbre was also expected to be easier to hear out of the multi-voice texture.

The result showed that the recognition in two-voice texture was higher on average than in three-voice, although the only difference came from the performance on the high voice. This result suggests that having one extra voice in the texture (the middle voice) takes the attention away from the most salient voice (the high voice) and not from the already sub-salient one (the low voice).

In studying the effect of timbre conditions, the only significant effect of timbre dissimilarity was observed from the average recognition of low voice in two-voice excerpts. Two highly dissimilar timbres helped the recognition of the low voice, as we hypothesized. This is in agreement with the previous findings in the literature on the effect of timbre dissimilarity on auditory streaming and melody recognition.

To our disappointment, no other effects of timbre conditions we hypothesized were found in the data. This lack of timbral effects could be caused by possibly unusual difficulty of memorability in specific excerpts and melodies used in certain conditions, or by an unforeseen interaction between excerpts (and melodies) and timbre combinations. Whatever the real reason is, we have not been able to identify it.

Even though the result lacked significant effects that we hoped to find, this experiment made us think again about possible context effects. For example, no main effects of timbre

dissimilarity and timbre saliency were found on the high voice in either two- and three-voice excerpts. This is apparently because the high voice is already the most salient in the given musical format so that no additional impact from timbre conditions affected the recognition performance. This is an example of the relative (or extrinsic) saliency in the experimental structure affecting (and hindering to a certain degree) the effect of conditions on intrinsic saliency. More experiments will be needed to clarify the unanswered questions.

## 5.2 Discussions

Timbre has been considered as one of the factors affecting auditory grouping in auditory scene analysis ([Bregman, 1990](#); [Kayser et al., 2005](#); [Kalinli & Narayanan, 2009](#)). Timbre research is relatively young in comparison to the research in pitch, due to its multidimensionality. This dissertation is the first in considering the attentional factor in timbre perception. By taking attention into account in the study of timbre, the problem of melody perception in multi-part music becomes a special case of the auditory scene analysis problem.

A series of experiments were performed to define timbre saliency and to study its effect on the perception of blending and segregation of concurrent sounds. First, two experiments were carried out to define timbre saliency. Even though the goal was the same for both experiments, the results were not totally compatible, to our surprise. This discrepancy illustrates the importance of choosing the right experimental paradigm to operationalize an idea. The degree of incompatibility might have been more evident here than usual, because we were measuring differences in saliency, which is a function of context. After all, many visual saliency studies reported either form or colour to be more salient than the other depending on the experimental context ([Theeuwes, 1992](#); [Poiese et al., 2008](#); [Schubö,](#)

2009), so the incompatibility we found is not an exception in saliency research.

With the timbre saliency dimension obtained from Experiment I, more experiments followed to analyze the effect of timbre saliency on the perceived blending in unison dyads, as well as the effect of timbre saliency on the perceived segregation in the form of the voice recognition in counterpoint music.

The scientific study of the perceived blend is surprisingly rare and all reports seem to confirm the findings from earlier works. [Kendall and Carterette \(1993\)](#); [Sandell \(1995\)](#); [Tardieu and McAdams \(2012\)](#) all found that spectral centroid is an important acoustic descriptor for the perceived blending. A timbre with a high spectral centroid would not blend well with others. [Sandell \(1995\)](#) and [Tardieu and McAdams \(2012\)](#) also reported the importance of attack patterns in determining the blend, relating it to auditory streaming ([Darwin, 1981](#); [Bregman, 1990](#)) that a longer attack of a sound may make it more difficult to identify the onset, which results in less streaming and more blending.

Perceived blend may need to be studied based on the time-varying spectrum of the blended mixture rather than individual components' acoustic features. There may be a good reason that most listeners do not discern the individual sounds in a composite sound. However, this ecological and holistic approach will be very difficult to realize, as it will have to account for a wide range of differences in blend, such as those that come from different instruments, different pitches and different pieces of music. So far we have not come across any effort using this approach.

Since a salient object tends to stand out with respect to its surroundings, implying little blending between the object and its neighbours, we hypothesized that the perceived blend will have a negative correlation with timbre saliency. The result did confirm the hypothesized negative correlation between timbre saliency and the average blend ratings, but the degree of correlation was only at a mild level. This suggests that the saliency

dimension is not an effective descriptor like spectral centroid and attack time that were reported in literature. This was perhaps the first hint that combined sounds may be much more complex than what the one-dimensional measure of timbre saliency could account for.

We still hoped to be able to extend and apply the concept of timbre saliency to the voice perception in counterpoint music. It may have been a big leap of faith as we have not studied the effect of timbre saliency in different pitches or registers, or the saliency of the voice positions within the counterpoint structure. Looking back, this information would have been beneficial in the design and the analysis of the last experiment.

In the last experiment, we considered the effect of timbre dissimilarity for the first time. Timbre dissimilarity has been studied for a long time, and we expected it to help study the effect of timbre saliency on the voice perception in counterpoint music. We also hoped to relate timbre saliency, a new concept, to timbre dissimilarity so that we can find how to apply the findings in timbre dissimilarity to the study of timbre saliency.

To our disappointment, most of the hypothesized effects were not found from the data. In our attempts to come up with an answer for this lack of expected effects, we remembered the matter of context effect again. Are the experimental contexts in Experiment I and VI compatible? If so, what could be the cause of this lack of effects? Could it be the result of extrinsic saliency (context effect) taking over whatever effect timbre conditions resulted in on intrinsic saliency?

Consider the recognition performance on the high voice. The high voice is the most prominent in the given context of counterpoint music. No effects of timbre conditions on the high voice in both two- and three-voice excerpts were probably because any additional boost from timbre conditions was too small to make a difference to this already significant ‘voice prominence.’ The fact that the effect of timbre dissimilarity and timbre saliency varied depending on the voice position also points to an effect of ‘voice prominence,’ which

is apparently what we did not account for in the design.

As [Huron \(1989b\)](#) observed, the entries of inner voices are more difficult to identify than those of outer voices, which also reflects different degrees of ‘voice prominence.’ Even though the data did not support our hypothesis that a salient timbre on a voice would enhance its recognizability, we did see a virtually significant improvement in recognition performance on the middle voice in the multi-timbre conditions from the mono-timbre conditions. This seems to reflect the additional boost from timbre conditions, although it may not have been as systematic as we expected.

The goal of the last experiment was to study the effects of timbre saliency and timbre dissimilarity in voice perception in counterpoint music, but we came out with more questions unanswered. A series of experiments are necessary to clarify these effects as well as the context effect. This may be a natural way of progression in the beginning of a new research field. I believe when I have answers for these questions, then I may be able to explain what makes one part of music catch listeners’ attention, which prompted the proposal of a new concept, timbre saliency.

### 5.3 Conclusion and Future Work

In this dissertation a new concept of timbre saliency has been proposed and a number of experiments have been carried out to define it and examine its roles in the context of blending. I am glad to have opened a new area of research that could be related to many existing research questions, even though the research in this dissertation is by no means complete. There are a few ideas for future work in the field.

As it was pointed out in [Section 5.2](#), the task of the first experiment had an unexpected confound with rhythmic saliency in duple rhythms, which surely affected the result. In



an attempt to remove this confound in measuring timbre saliency, another experiment was carried out with direct comparison. The data this time showed a significant effect of order, which contributed to a result that was problematic in the interpretation, which therefore was not used in any further studies.

As yet another attempt to separate timbre saliency from the aforementioned confound, a similar experiment with same sounds can be done, using a ternary rhythm. A ternary rhythm is often called a waltz rhythm, which is a continuation of *strong-weak-weak* beats. Even though it will not be completely free from the implication of spectral centroids on the choice of downbeat timbre, it will still provide insights especially in contrast to the result with that from Experiment I. The design might need some extra care in terms of counterbalancing and the number of total trials as the stimuli will now become a set of ABCABC perceptually isochronous sequences.

One of the questions that I have been curious about but was not able to clarify yet is the relationship between timbre saliency and timbre dissimilarity. The definition of saliency requires an object to stand out with respect to its surroundings, implying dissimilarity between the object and its neighbours. What then might be the relationship between timbre saliency and timbre dissimilarity? In Chapter 3, we reported that the blending of two unison timbres is negatively correlated with the timbre saliency of individual timbres. What can be said about the perception of blending in terms of timbre dissimilarity? Will similar sounds blend better? Kendall and Carterette (1993) reported that the mean blending judgment was negatively correlated with mean identification rate, suggesting that an instrument that achieves a better blending tends to have a weaker identifiability. Can we confirm their reports with our data? A new analysis of our existing data will help answer these questions. More fundamentally, how is timbre saliency different from timbre dissimilarity? In the last experiment, we just assumed that timbre dissimilarity would not have any negative

interaction with timbre saliency and used both of them in design. Was our assumption true? Another experiment may need to be performed to compare and relate timbre saliency and timbre dissimilarity, especially in the context of the voice perception in counterpoint music.

A study of individual consistency in timbre saliency responses might also be interesting. A small group of people participated in the experiment to measure timbre saliency by tapping (Chapter 2) as well as in the experiment to study the effect of timbre saliency on voice perception in multipart counterpoint music (Chapter 4). We are all aware of large interpersonal variability in auditory and music perception data including those from my first experiment. But is there a personal consistency? Does a person's timbre saliency judgment predict anything in their performance in melody recognition in multipart counterpoint music? For example, if someone happened to think the harpsichord is more salient than the trumpet, did he/she recognize more easily melodies in the harpsichord timbre? Are there different groups showing different behaviours? This may lead us to an interesting finding in the heightened timbre saliency due to personal preference or training.

I also feel that a study of timbre changes according to the pitch or register will be beneficial. Much timbre research has focused on the range of perceived timbres with a fixed pitch. While this has been useful for many studies, we also know that many instruments have different distinctive timbres according to the register, as many orchestration textbooks point out (Berlioz, 1855; Adler, 2002). This study may be much more difficult than it sounds, since it will have to consider different playing ranges of different instruments. With such information, it may be possible to better predict voice perception in counterpoint music.

In relation to other research fields, timbre saliency may be useful in showing brain responses to salient auditory stimuli. There are only two computational models of auditory saliency to date (Kayser et al., 2005; Kalinli & Narayanan, 2009), because it is difficult to

figure out what listeners are attending to without the luxury of eye tracking in the visual saliency research. Very recently (July 2012) the auditory attention team at the Telluride Neuromorphic Engineering Workshop reported that they successfully showed that "it is possible with EEG signals to correctly determine to which auditory stream a participant is attending." (Slaney, Lalor, et al., 2012) This was the first report of auditory attention reflected in EEG measurements of which the researcher is aware. Their experiment used two streams of a male narrator, where the two streams had no coherence between them. This may markedly change the result if we want to use Western tonal music as stimuli where there are governing temporal and harmonic rules among different voices in a piece. Can we still figure out what stream in the music the person is paying attention to? Furthermore, would it be possible to predict which stream a particular person will attend to, given his/her timbre saliency data? This will require a careful design of the experiment. But if we can successfully correlate different brain responses with timbre saliency, this will open a new and exciting approach to studying auditory saliency and attention in general. Furthermore, this research may help facilitate a better quality of human life with its application to a better design of hearing aids and sound alarm systems.

Timbre saliency is a new concept that just started its life in research. It can be applicable in research in other fields such as auditory attention and auditory scene analysis, as well as the design of auditory signals such as alarms, car horns and even hearing aids. As we have seen in this dissertation, the timbre saliency dimension was not as successful as other acoustic features in explaining the blending of two concurrent unison notes (Chapter 3) or in counterpoint music (Chapter 4). In fact, after the last experiment we ended up with more unanswered questions than we began with. The biggest question is the matter of context effects. If I learned one important lesson from these experiments, it is the fact that we cannot foresee the outcome regardless of however carefully designed an experiment is.

This is exciting in a sense that it leads to more experiments and more findings and the cycle goes on. With a series of experiments described above, as well as many others that we cannot imagine right now, I believe the timbre saliency model can be improved and able to contribute to research in other fields.

# Appendix A

## Stimulus Selection

Many previous timbre studies have used synthesized tones based on analyses of acoustic instruments for the ease of control and analysis (Grey, 1977; Grey & Gordon, 1978; Krumhansl, 1989; McAdams et al., 1995; Caclin et al., 2005). We decided to use recorded samples of real instrument sounds from the Vienna Symphonic Library (VSL) (Vienna Symphonic Library GmbH, 2011). With the decision of fixing the pitch to *C*4 (= 261.6 Hz), we studied the acoustic features computed from the Timbre Toolbox (Peeters et al., 2011) on multiple instruments' recordings to find the best stimulus set.

There are many instruments that play *C*4, among which we narrowed the candidates to Western classical instruments that appeared in previous timbre studies. Fifteen instruments were chosen for experimental tests – Clarinet (CL), English Horn (EH), French Horn (FH), Flute (FL), Harp (HA), Harpsichord (HC), Marimba (MA), Oboe (OB), Piano (PF), Tubular Bells (TB), Trombone (TN), Trumpet (TP), Tuba (TU), Cello (VC), and Vibraphone (VP). Four more were chosen for practice trials – Alto Flute (AF), Bassoon (BN), Celesta (CE), and Violin (VN).

Most of these samples are at least a few seconds long, so all of them had to be shortened

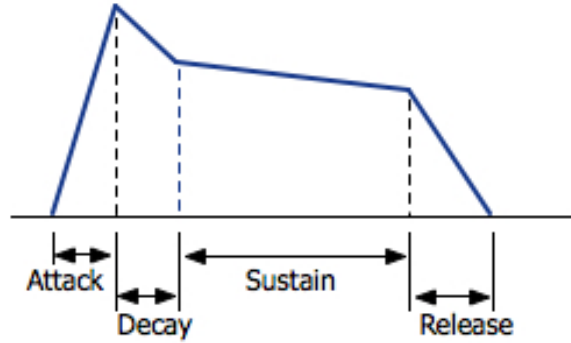


Figure A.1: ADSR estimation

for the experiments. We decided to keep the attack part intact, because the attack is known to be one of the key features in timbre perception studies in the literature (Berger, 1964; Saldanha & Corso, 1964; McAdams et al., 1995; Lakatos, 2000; Caclin et al., 2005), and therefore we wanted to see how important attack is in the perception of timbre saliency. The natural attacks of the 19 chosen instruments vary quite a bit; according to the estimated attack time from the Timbre Toolbox, the shortest was 28.3 ms (TB) and the longest 98.2 ms (AF). The mean was 57.2 ms and the median 55.3 ms (which was from the French Horn). The attack time is estimated from energy fluctuations (Peeters, 2004) as a part of the Attack-Decay-Sustain-Release (ADSR) signal approximation, which is depicted in Figure A.1.

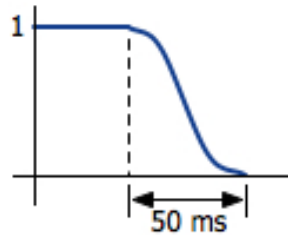


Figure A.2: A raised cosine decay ramp

Since we were following the traditional timbre studies approach by equalizing the other perceptual variables of pitch, loudness and duration, we decided to equalize the duration by

keeping the effective duration, one of the parameters calculated from the Timbre Toolbox, around 200 ms. A raised cosine decay ramp of 50 ms duration (shown in Figure A.2) was applied at appropriate points of an instrument sound signal to achieve the desired effective duration. This meant cutting sound files in different physical lengths (257 – 360.3 ms with a median of 298 ms). Table 2.1 lists attack duration, effective duration and physical duration of each instrument sound.

(This page is intentionally left blank.)



## Appendix B

# Equalization Experiments

Traditional timbre studies ([Grey, 1977](#); [Grey & Gordon, 1978](#); [Krumhansl, 1989](#); [McAdams et al., 1995](#); [Lakatos, 2000](#)) used instrument tones that were equalized in pitch, loudness, and duration to minimize their impact on timbre perception. Since loudness also affects the perceived saliency, we needed to equalize the loudness of the stimuli.

### B.1 Loudness Equalization

#### B.1.1 Participants

Seventeen participants (12 males) were recruited from the Schulich School of Music of McGill University. All reported normal hearing. There was no monetary compensation for participation.

#### B.1.2 Stimuli

Nineteen orchestral instrument sounds from the Vienna Symphonic Library ([Vienna Symphonic Library GmbH, 2011](#)) were prepared with equalized pitch ( $C4$ ). Table [2.1](#) lists their

abbreviations and various measurements of durations. Refer to Appendix A for details on stimulus selection.

Only the left channel information was used for this experiment to prevent any undesired stereo effects. These mono signals were then duplicated to the right channel at the time of the experiment. These segments were stored in a 16-bit PCM wav format with a 44.1kHz sampling rate.

An initial loudness equalization was attempted on the effective-duration-equalized sounds for long-term loudness levels at 90 phons using Brian Moore’s time-varying loudness program (<http://hearing.psychol.cam.ac.uk/Demos/TVL.zip>, as used in Glasberg & Moore, 2002). The result still had a wide range of loudness variations; therefore it was decided to do a perceptual study for further equalization.

### B.1.3 Procedure

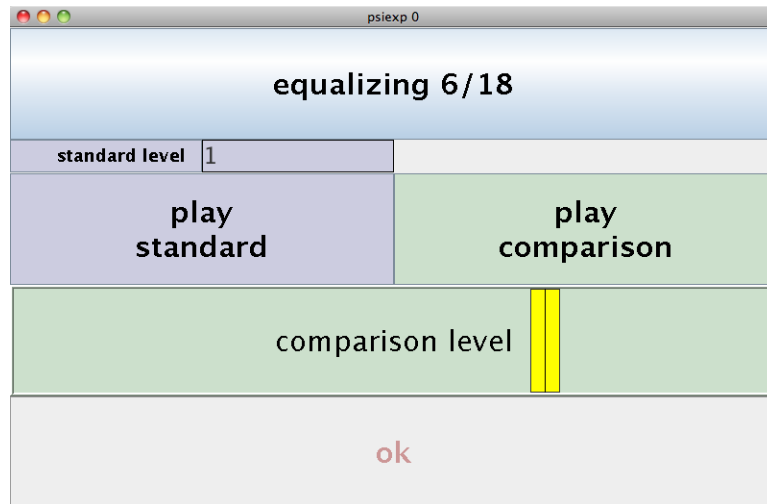


Figure B.1: Screenshot of Pilot 1

The study took place in a sound-attenuated booth (Industrial Acoustics Company, model 1203) in the Music Technology Area of the Schulich School of Music of McGill

University. Mono signals were presented to both ears of the participants via Sennheiser HD 280 headphones after amplification by a Grace m904 stereo monitor controller, the average sound pressure level was set to 56 dBA as measured with a Brüel & Kjær type 2250 sound level meter coupled with a Brüel & Kjær type 4153 artificial ear.

The graphic user interface (shown in Figure B.1) was developed in PsiExp (Smith, 1995) on an Apple PowerPC G5. Participants were asked to match the loudness of each of 18 tones using a scroll bar at the bottom of the screen to that of the “standard tone”, which was the violin sound because it seemed the quietest after Moore’s TVL adjustment. The last position of the scroll bar before the “OK” button was clicked was recorded for each stimulus per participant for analysis.

#### B.1.4 Result

There was quite a bit of diversity in the data. For each of the 19 sounds, the median adjustment value was taken for the next step. Figure B.2 shows the adjustment values of 19 sounds from 17 participants. Each participant’s data have been connected with dotted lines, which do not have any additional meaning. The median adjustments are connected by the solid black line, which also does not have any additional meaning. Although the dotted lines may appear to suggest that some listeners may have consistently rated most sounds lower than the group median and others consistently higher, there was no such participant according to the data. Mean adjustment error around the medians was calculated for each participant to quantify this. The average error values of 17 participants range from  $-3.8$  dB to  $2.5$  dB, with the grand mean of  $-0.07$  dB, which is quite close to  $0$  dB. A single-sample t-test confirms that participants were, on average, close to the median values in their loudness adjustments,  $t(16) = -0.187, p = 0.854$ . The 19 median loudness adjustment values in Figure B.2 are listed in Table B.1.

There is one concern with loudness equalization, that it may affect the perceived saliency of each stimulus. Saliency must be a function of many things, including pitch and loudness, hence adjusting the loudness of a stimulus might result in a different perceived saliency. The adjustment values in Table B.1 look as though the instruments with percussive attacks may have been attenuated more, which may have weakened their saliencies, which in turn may have had an undesired impact on the main experiments. Although we are aware of this possible distortion that might result from this equalization, we decided to use the loudness-equalized stimuli, as their unequalized levels were set somewhat arbitrarily from the recording process for VSL. Note, however, that MC and HC, which were included in the main experiment, had the greatest reduction in level (3 – 4 dB attenuation), but received high and medium saliency values in Experiment I.

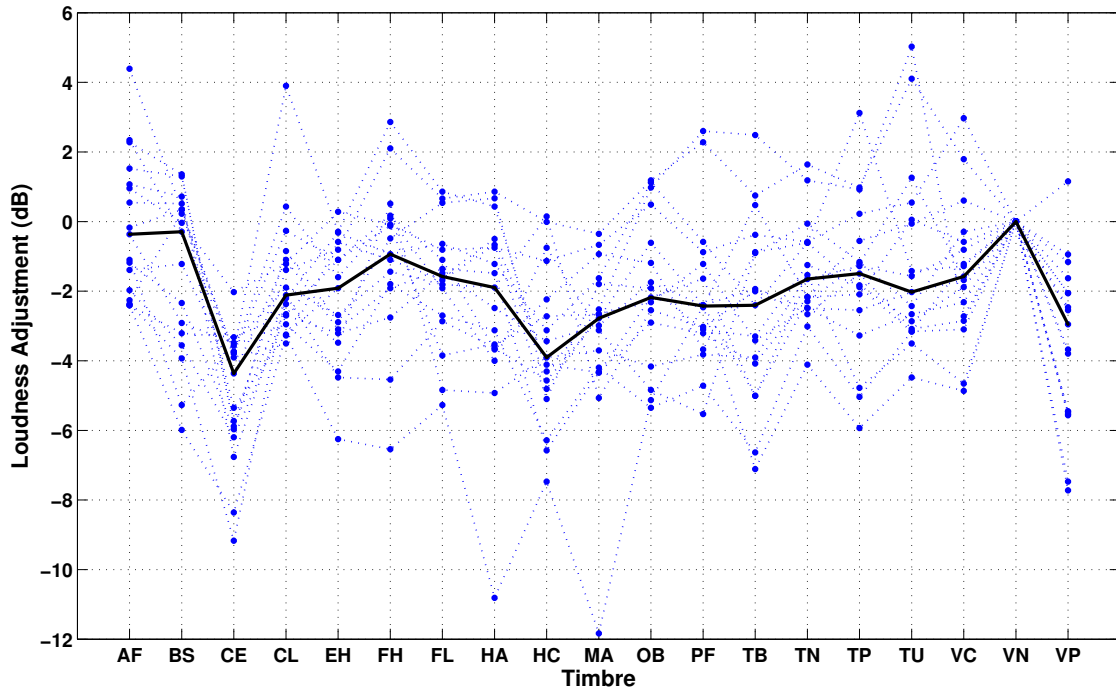


Figure B.2: Result from Pilot 1

**Table B.1:** Median Loudness Adjustment Values in Decibels

Set	Abbrev.	Loudness Adjustment (dB)
Test Set	CL	−2.1
	EH	−1.9
	FH	−0.9
	FL	−1.6
	HA	−1.9
	HC	−3.9
	MA	−2.8
	OB	−2.2
	PF	−2.4
	TB	−2.4
	TN	−1.6
	TP	−1.5
	TU	−2.0
	VC	−1.6
	VP	−3.0
Train Set	AF	−0.4
	BS	−0.3
	CE	−4.4
	VN	0

## B.2 Isochronous Rhythm Generation

### B.2.1 Participants

Seventeen participants from the Schulich School of Music of McGill University took part in this experiment. All reported normal hearing. There was no monetary compensation for participation. One participant's data were excluded from analysis because they were clear outliers; whereas most participants' adjustments were around 400 ms, her timing adjustments included extreme values such as 114 or 689 ms, which could not result in an isochronous rhythm. Data from the other 16 participants (12 males) were analyzed.

### B.2.2 Stimuli

The 19 instrument sounds from the loudness equalization experiment in Appendix B.1 (see Table 2.1 for the list) were scaled according to the loudness adjustment result in Table B.1. This new set of sounds was used in generating mono-timbre isochronous sequences with 800 ms IOI (inter-onset interval). One hundred eleven pairs of two sounds were created, since there are 105 ways to choose 2 out of 15 (for testing) and 6 ways to choose 2 out of 4 (for training) and no cross-group combinations to be tested in the Experiment I.

### B.2.3 Procedure

The task was to combine two mono-timbral isochronous sequences to create another isochronous sequence, the rhythm of which is twice as fast. The equipment setup was the same as in Pilot 1 in section B.1. A graphic user-interface in PsyExp (Smith, 1995) was used, as shown in Figure B.3. There were two buttons ("hear sound 1" and "hear sound 2") to turn on or off each of the two sounds. In the beginning of a new trial, the two timbres would alternate every 400 ms, which may sound like a gradual-abrupt or abrupt-gradual rhythm,

depending on their attack patterns. It was expected that listeners would shorten the IOI between abrupt and gradual sounds and lengthen the IOI between gradual and abrupt sounds to equalize the perceived IOIs. In the middle of the screen was a scroll bar that participants could move to control the time lag between the two sequences. Some found it useful to turn off one of the sounds occasionally when they felt they needed to remember the original rhythm. When a participant felt the resulting rhythm was regular and twice as fast as the original, he/she would click on the “ok” button to move to the next pair of sounds. The value of the time lag for each timbre pair was recorded.

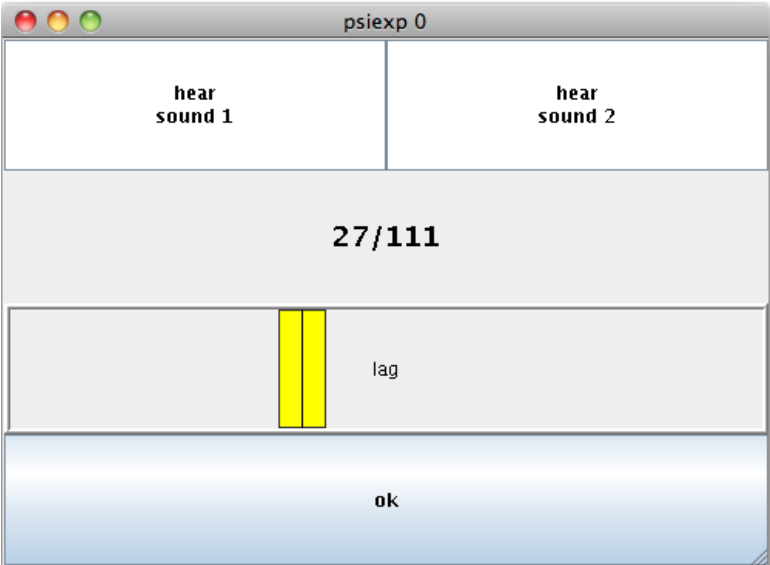


Figure B.3: Screenshot of Pilot 2

B.2.4 Result

A large interpersonal difference was also observed in the data. The median values were obtained for every pair of combinations, the result of which are shown in Tables B.2 and B.3. In both tables, the diagonal entries are empty because the cases of same timbre pairs were not considered. The value of the cell  $(i, j)$ , on the  $i$ -th row and  $j$ -th column, shows

**Table B.2:** Median lag in generating perceptually isochronous sequences of four timbres for training. AF = alto flute, BS = bassoon, CE = celesta, and VN = violin. The timbre specified by the row corresponds to timbre A and that specified by the column to timbre B in ABAB sequences.

	AF	BS	CE	VN
AF	—	409	433	405
BS	391	—	422	400
CE	367	378	—	362
VN	396	401	439	—

the median value of the time lag in milliseconds when the first timbre is the one specified by the row and the second by the column. For example, the value of 391 on the second row and the first column of the Table B.2 means to start the second timbre, AF, 391 ms after the start of the first timbre, BS, to form an isochronous sequence. Note that the opposite case of BS following AF (on the cell (1, 2)) shows a different time lag of 409, which is the 800's complement of 391. For all entries in both tables B.2 and B.3, the values in the cells  $(i, j)$  and  $(j, i)$  should always sum to 800, which is the original time lag in mono-timbral sequences. Some entries actually sum up to 801, due to the rounding error in calculation. These median values were used to generate two-timbre isochronous sequences for Experiment I.

A brief analysis reveals that the median adjustment values (the values on the Table B.3 minus 400) are highly correlated with the attack time difference of the two stimuli. The attack time of each sound was calculated from the ADSR (attack-decay-sustain-release) estimation values using the Timbre Toolbox (Peeters et al., 2011). Next, the difference of attack times for each pair was computed. The median delta values were obtained by subtracting 400 (a half of 800 ms, which means an ideal case of same-timbre pairs) from



**Table B.3:** Median lag in generating perceptually isochronous sequences of 15 timbres for testing. CL = clarinet, EH = English horn, FH = French horn, FL = flute, HA = harp, HC = harpsichord, MA = marimba, OB = oboe, PF = piano, TB = tubular bells, TN = trombone, TP = trumpet, TU = tuba, VC = cello, VP = vibraphone. The timbre specified by the row corresponds to timbre A and that specified by the column to timbre B in ABAB sequences.

	CL	EH	FH	FL	HA	HC	MA	OB	PF	TB	TN	TP	TU	VC	VP
CL	—	439	436	448	446	454	452	435	463	451	436	432	456	424	442
EH	362	—	396	399	411	417	416	388	408	416	404	399	411	394	420
FH	365	405	—	405	410	421	413	395	417	410	411	404	413	407	416
FL	352	401	395	—	399	407	409	389	419	408	394	400	413	379	417
HA	354	390	391	401	—	398	406	389	402	400	390	391	401	382	402
HC	347	384	379	394	403	—	402	380	399	400	490	388	395	388	396
MA	349	385	387	392	398	399	—	389	404	393	399	378	391	383	400
OB	366	412	405	412	412	420	412	—	420	406	412	408	410	400	415
PF	338	393	383	382	398	402	396	381	—	400	392	389	403	375	400
TB	350	384	391	392	400	400	408	394	400	—	393	388	400	381	399
TN	364	396	390	406	410	411	401	389	409	408	—	399	411	379	412
TP	368	402	397	401	409	413	423	393	411	412	412	—	411	386	407
TU	344	390	387	388	400	406	409	391	398	400	390	390	—	377	400
VC	377	406	394	421	419	413	418	400	425	420	422	414	424	—	402
VP	359	381	384	383	399	404	400	386	401	402	388	394	401	398	—

the values on the tables B.2 and B.3. The correlation coefficient between the median delta values of non-diagonal entries and the attack time difference is 0.878 at the  $p = .01$  significance level, reflecting the fact that participants were adjusting the time according to the attack time difference. The Figure B.4 clearly shows this trend.

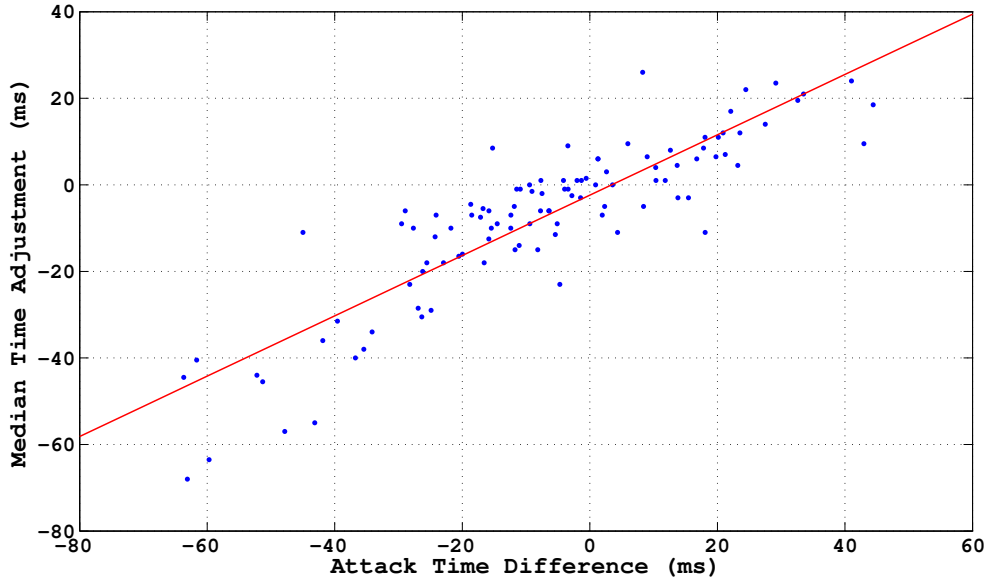


Figure B.4: Median adjustment time as a function of attack time difference

### B.3 Melody Loudness Equalization

An instrument's timbre might change according to its pitch or register, hence also affecting the timbre saliency of a melody. It would therefore be complete if we could equalize the loudness of every single note on every single instrument that we would be using for Experiment VI, but in reality it is quite impossible. As an anecdote, we decided to equalize the loudness of all instruments using a melody that is in the middle of the range of all melodies used for the experiment, which turned out to be the middle voice of a training

melody in three-voice excerpts, shown in Figure B.5.



**Figure B.5:** Melody used for loudness equalization

### B.3.1 Participants

Fifteen participants (8 males), all of whom reported normal hearing, from Schulich School of Music of McGill University participated in the study. There was no monetary compensation in exchange for participation.

### B.3.2 Stimuli

The melody in Figure B.5 was generated in each of 19 instrument timbres (AF, BS, CE, CL, EH, FL, FH, HA, HC, MA, OB, PF, TB, TN, TP, TU, VN, VC, and VP) using VSL instruments in Logic ([Apple Computer, 2012](#)). They were in stereo format. Initially, there were quite big loudness differences of the 19 files, due to various reasons such as instrumental characteristics and different recording conditions. I did a rough gain control on several files so that they would not be shockingly louder than the other files, which are specified in Table B.4. These modified files were used for the equalization experiment.

### B.3.3 Procedure

The pilot study took place in a sound-attenuated booth ([Industrial Acoustics Company, model 1203](#)) in the Music Technology area of the Schulich School of Music of McGill University. The experimental setup was exactly the same as in the previous two experiments

**Table B.4:** List of instruments with preliminary loudness equalization and their adjustment levels

Instrument	Loudness Adjustment (dB)
FH	−9
FL	−6
HC	−18
OB	−9
PF	−9
TB	−7
VC	−15
VN	−9
VP	−9

in this Appendix – participants heard the sounds using Sennheiser HD 280 headphones, and the average level was set to 65 dBA as measured with a Brüel & Kjær type 2250 sound level meter coupled with a Brüel & Kjær type 4153 artificial ear.

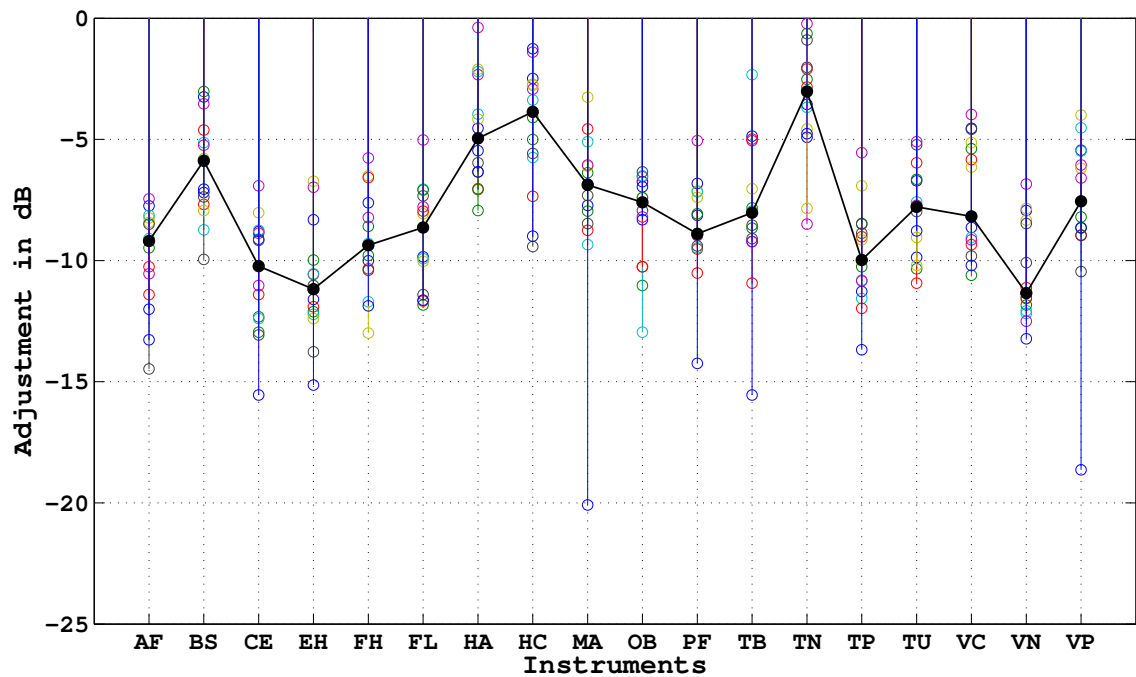
The same graphic user interface that was used for the previous loudness equalization pilot in section B.1 was used again for this experiment. Participants clicked on two buttons (“play standard” and “play comparison”) to hear the standard sound and the comparison sound, as captured in Figure B.3. The standard sound always stayed the same – it was the melody in CL because it was the quietest. The comparison sounds were the other 18 files.

### B.3.4 Result

There were quite large inter-personal differences in the data. The data and median values are shown in Figure B.6. The medians are specified in filled black dots, which are connected

with a solid black line for ease of visualizing the pattern.

The string and wind instruments showed less variance (e.g., TU and VC), whereas struck or plucked instruments showed much larger variation across listeners. The biggest range is observed with MA, TB and TP, all of which have a long decay, which affects the overall perceived loudness of those sounds. CL is not shown because it was the standard sound, hence there was no adjustment.



**Figure B.6:** Result of Pilot 3 with 15 participants

Combining the experimental results in Figure B.6 with the preliminary loudness equalization in Table B.4, we obtain the total loudness adjustment of the 19 files shown in Figure B.7 to be used for Experiment VI in Chapter 4.

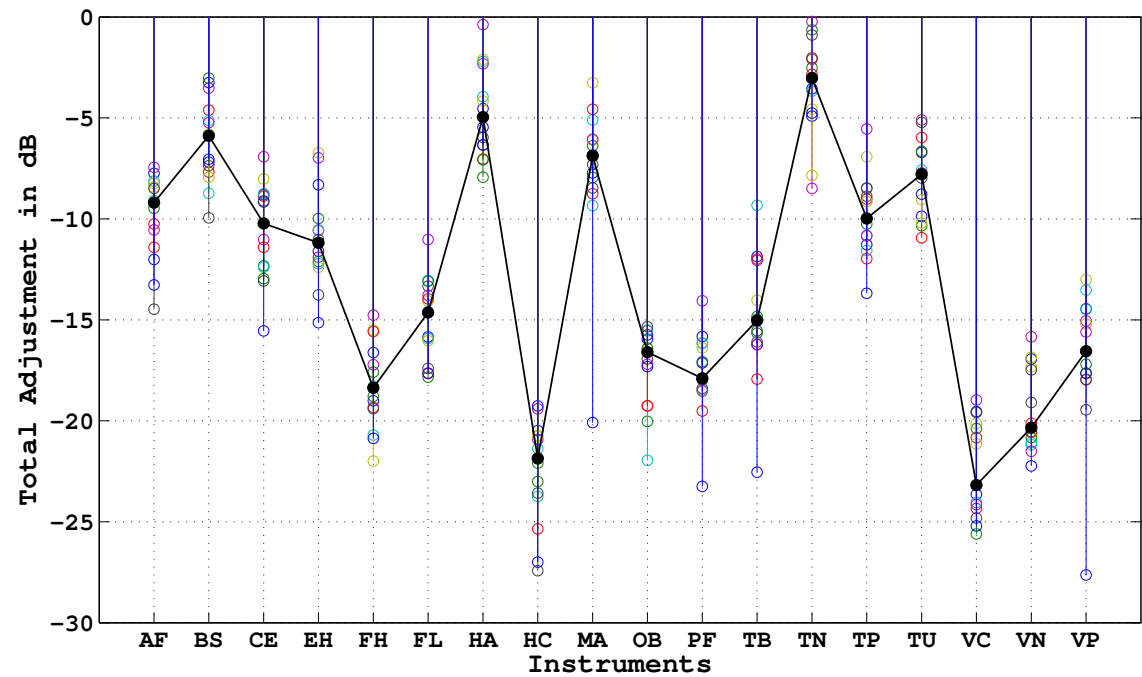


Figure B.7: Total loudness equalization result of Pilot 3 with 15 participants

## Appendix C

# Scores of Melodies used for the Voice Recognition Experiment

Here are melodies used for the experiment to study voice recognition in multipart counterpoint music, as presented in Chapter 4. The “original” melodies were chosen from J.S. Bach’s *Trio Sonatas for Organ*, BWV 525 – 530 (Bach, 1730). Sometimes modifications on Bach’s music were necessary to create *our* original melodies, such as breaking up a long note or to transpose the key to accommodate an instrument’s playing range.

On each stave, on the left side of a double bar is an original melody and on the right is the corresponding “modified” melody that we composed by changing two notes of the original melody. In the composition of a “modified” melody, we adhered to the following rules for most cases:

- There is no change of rhythm between the original and the corresponding modified melodies.
- The two changed notes cannot be on the first or the last note, which would be easier

to notice than other notes.

- The two changed notes will change the resulting melodic contour.
- The two changed notes will still make sense harmonically and stylistically, in order to avoid giving a possible unintended cue with a different harmony or style.
- For a high-voice melody, the pitches of the changed notes will not be higher than the highest pitch in the original melody. They can be lower than the lowest pitch in the original melody.
- For a low-voice melody, the pitches of the changed notes will not be lower than the lowest pitch in the original melody. They can be higher than the highest pitch in the original melody.
- For a middle-voice melody, there are no restrictions in terms of pitch height.
- Change notes on a more important beat position (e.g., beats 1 or 3 rather than 2 or 4 in a 4/4 meter) to maximize salience of the changes.
- Change notes on a more important metric positions (e.g., first or third 16th-note rather than second or fourth in beat 3) to maximize salience of the changes.

## C.1 Two-voice excerpts

The 12 two-voice excerpts used for the voice recognition experiment are presented below. The sonata numbers refer to Bach's trio sonata numbers. The measure numbers specify from which measures the melodies were taken. The page numbers correspond to those in the Bärenreiter edition ([Bach, 1730](#)). The melodies used for training are distinguished with "Tr." preceding the excerpt number.



## Two Voice Excerpts

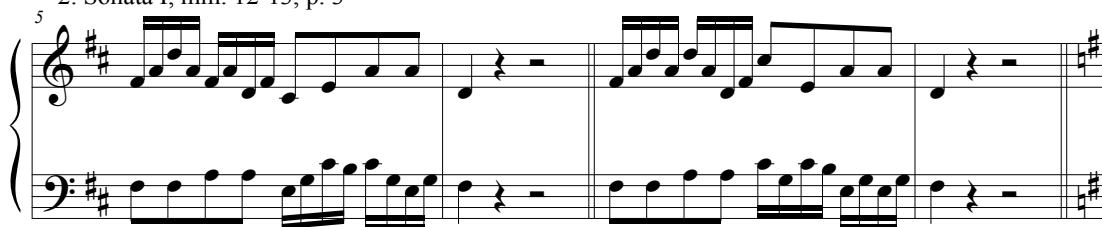
Excerpt

Comparison

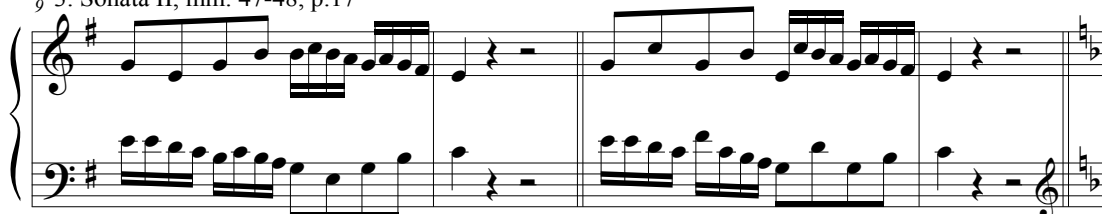
1. Sonata I, mm. 5-6, p.2



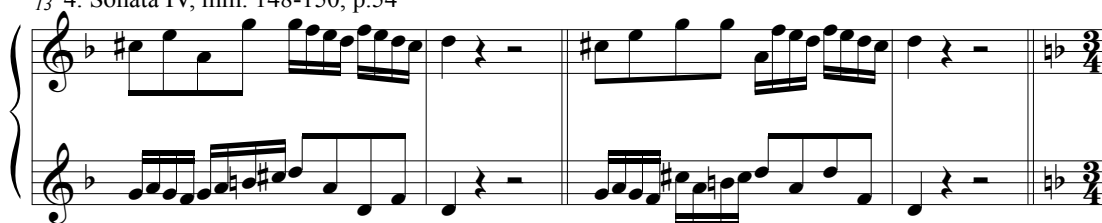
2. Sonata I, mm. 12-13, p. 3



3. Sonata II, mm. 47-48, p.17



4. Sonata IV, mm. 148-150, p.54



2	Excerpt	Two Voice Excerpts	Comparison
17	5. Sonata V, mm. 111-112, p.63		
21	6. Sonata V, mm. 120-121, p. 63		
25	7. Sonata II, mm. 133-134, p. 26		
31	8. Sonata VI, mm. 1-2, p. 76		
37	9. Sonata IV, mm. 76-77, p. 78		

Excerpt	Two Voice Excerpts	Comparison	3
41 10. Sonata VI, mm. 78-79, p. 78			
47 11. Sonata VI, m. 82, p. 78			
51 12. 1. Sonata VI, mm. 137-138, p.81			
55 Tr. 1. Sonata VI, mm. 102-104, p. 79			
59 Tr. 2. Sonata V, mm. 9-10, p. 71			

C.2 Three-voice excerpts

Following is the nine three-voice excerpts used for the voice recognition experiment.

Three Voice Excerpts

Original

Modified

1. Sonata III, mm.65-66



2. Sonata III, mm. 121-122



3. Sonata III, mm. 158-160, p.34



2	Excerpt	Three Voice Excerpts	Comparison
17	4. Sonata V, mm. 154-155, p. 65		
21	5. Sonata, IV, mm. 40-41, p. 46		
25	6. Sonata IV, mm. 63-65, p.47		

Three Voice Excerpts

3

Excerpt

7. Sonata IV, mm. 42-43, p. 51

Comparison

8.Sonata IV, mm. 88-90, p.55

9. Sonata V, m. 16-17, p.57

4


Excerpt

Three Voice Excerpts

Comparison


45

Tr. 1. Sonata V, m. 2, p.56



49

Tr. 2. Sonata III, mm.22-23, p. 37



(This page is intentionally left blank.)



# Bibliography

- Adler, S. (2002). *The study of orchestration* (Third ed.). New York, NY: W. W. Norton & Company.
- Akaike, H. (1977). On entropy maximization. In P. R. Krishnaiah (Ed.), *Applications of statistics* (pp. 27–41). Amsterdam: North-Holland.
- American Standards Association. (1960). *Acoustical terminology, s1.1-1960*. New York: American Standards Association.
- Anstis, S., & Saida, S. (1985). Adaptation to auditory streaming of frequency-modulated tones. *Journal of Experimental Psychology: Human Perception and Performance*, 11, 257–271.
- Apple Computer. (2012). *Logic*. Cupertino, CA.
- Bach, J. S. (1730). *Six trio sonatas for organ, bwv 525–530* (Vols. Urtext der Neue Bach-Ausgabe Serie IV, Band 7; D. Kilian, Ed.). Bärenreiter. (Published in 1984)
- Berger, K. (1964). Some factors in the recognition of timbre. *Journal of the Acoustical Society of America*, 36(10), 1888–1891.
- Berlioz, H. (1855). *Berlioz's orchestration treatise: A translation and commentary (cambridge musical texts and monographs)* (H. Macdonald, Ed.). Cambridge, England: Cambridge University Press. (Republished in 2002)
- Bey, C. (1999). *Reconnaissance de melodies intercalees et formation des flux auditifs: Analyse fonctionnelle et exploration neuropsychologique*. Unpublished doctoral dissertation, IRCAM, Paris, France.
- Bey, C., & McAdams, S. (2003). Postrecognition of interleaved melodies as an indirect measure of auditory stream formation. *Journal of Experimental Psychology: Human Perception and Performance*, 29(2), 267–279.
- Botte, M.-C., Drake, C., Brochard, R., & McAdams, S. (1997). Perceptual attenuation of non focused auditory streams. *Perception & Psychophysics*, 59(3), 419–425.
- Bregman, A. (1990). *Auditory scene analysis: The perceptual organization of sound*. Cambridge, MA: MIT Press.
- Bregman, A., & Campbell, J. (1971). Primary auditory stream segregation and perception of order in rapid sequences of tones. *Journal of Experimental Psychology*, 89(2), 244–249.
- Caclin, A., McAdams, S., Smith, B. K., & Winsberg, S. (2005). Acoustic correlates of

- timbre space dimensions: A confirmatory study using synthetic tones. *Journal of the Acoustical Society of America*, 118(1), 471–482.
- Darwin, C. (1981). Perceptual grouping of speech components differing in fundamental frequency and onset-time. *The Quarterly Journal of Experimental Psychology Section A*, 33, 185–207.
- Deutsch, D. (1972). Octave generalization and tune recognition. *Perception & Psychophysics*, 11, 411–412.
- Donk, M., & Zoest, W. van. (2008). Effects of salience are short-lived. *Psychological Science*, 19(7), 733–739.
- Dowling, W. (1973). The perception of interleaved melodies. *Cognitive Psychology*, 5, 322–337.
- Dowling, W. (1978). Scale and contour: Two components of a theory of memory for melodies. *Psychological Review*, 85, 341–354.
- Dowling, W., & Fujitani, D. (1971). Contour, interval and pitch recognition in memory for melodies. *Journal of the Acoustical Society of America*, 49, 524–531.
- Dowling, W., Lung, K., & Herrbold, S. (1987). Aiming attention in pitch and time in the perception of interleaved melodies. *Perception & Psychophysics*, 41(6), 642–656.
- Elazary, L., & Itti, L. (2008). Interesting objects are visually salient. *Journal of Vision*, 8(3), 1–15.
- Eramudugolla, R., Irvine, D., McAnally, K., Martin, R., & Mattingley, J. (2005). Directed attention eliminates ‘change deafness’ in complex auditory scenes. *Current Biology*, 15, 1108–1113.
- Glasberg, B. R., & Moore, B. C. J. (2002). A model of loudness applicable to time-varying sounds. *Journal of Audio Engineering Society*, 50, 331–342.
- Goodwin, A. (1989). An acoustic study of individual voices in choral blend. *Journal of Research in Singing*, 13(1), 119–128.
- Gregory, A. H. (1990). Listening to polyphonic music. *Psychology of Music*, 18, 163–170.
- Grey, J. M. (1977). Multidimensional perceptual scaling of musical timbres. *Journal of the Acoustical Society of America*, 61, 1270–1277.
- Grey, J. M., & Gordon, J. W. (1978). Perceptual effects of spectral modifications on musical timbres. *Journal of the Acoustical Society of America*, 63, 1493–1500.
- Handel, S. (1995). Timbre perception and auditory object identification. In B. Moore (Ed.), *Hearing* (pp. 425–461). San Diego, California: Academic Press.
- Helmholtz, H. (1877). *On the sensations of tone as a physiological basis for the theory of music*. New York, NY: Dover Publications.
- Huron, D. (1989a). The avoidance of part-crossing in polyphonic music: Perceptual evidence and musical practice. *Music Perception*, 9, 93–104.
- Huron, D. (1989b). Voice denumerability in polyphonic music of homogeneous timbres. *Music Perception*, 6(4), 361–382.
- Huron, D., & Fantini, D. (1989). The avoidance of inner-voice entries: Perceptual evidence

- and musical practice. *Music Perception*, 7, 43–47.
- Industrial Acoustics Company. (n.d.). *Industrial acoustics, model 1203*. Bronx, NY.
- International Organization for Standardization, Geneva. (2004). *Acoustics reference zero for the calibration of audiometric equipment, part 8: Reference equivalent threshold sound pressure levels for pure tones and circumaural earphones*.
- IRCAM. (n.d.). *Audio sculpt*. Paris, France. Available from <http://forumnet.ircam.fr/691.html?L=1>
- Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20, 1254–1259.
- Iverson, P. (1995). Auditory stream segregation by musical timbre: Effects of static and dynamic acoustic attributes. *Journal of Experimental Psychology*, 21(4), 751–763.
- Iverson, P., & Krumhansl, C. (1993). Isolating the dynamic attributes of musical timbre. *Journal of Acoustical Society of America*, 94, 2593–2603.
- Kalinli, O., & Narayanan, S. (2009). Prominence detection using auditory attention cues and task-dependent high level information. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(5), 1009–1024.
- Kayser, C., Petkov, C., Lippert, M., & Logothetis, N. (2005). Mechanisms for allocating auditory attention: An auditory saliency map. *Current Biology*, 15, 1943–1947.
- Kendall, R., & Carterette, E. (1991). Perceptual scaling of simultaneous wind instrument timbres. *Music Perception*, 8(4), 369–404.
- Kendall, R., & Carterette, E. (1993). Identification and blend of timbres as a basis for orchestration. *Contemporary Music Review*, 9, 51–67.
- Krimphoff, J., McAdams, S., & Winsberg, S. (1994). Caractérisation du timbre des sons complexes. ii. analyses acoustiques et quantification psychophysique. *Journal de Physique*, 4, 625–628.
- Krumhansl, C. L. (1989). Why is musical timbre so hard to understand? In S. Nielzen & O. Olsson (Eds.), *Structure and perception of electroacoustic sound and music* (pp. 44–53). Amsterdam: Excerpta Medica.
- Kruskal, J. (1964a). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29, 1–27.
- Kruskal, J. (1964b). Non-metric multidimensional scaling: a numerical method. *Psychometrika*, 29, 115–129.
- Lakatos, S. (2000). A common perceptual space for harmonic and percussive timbres. *Perception & Psychophysics*, 62(7), 1426–1439.
- Lemaitre, G., Susini, P., Winsberg, S., McAdams, S., & Letinturier, B. (2007). The sound quality of car horns: A psychoacoustical study of timbre. *Acta Acustica United With Acustica*, 93, 457–468.
- Levin, D., Momen, N., Drivdahl, S., & Simons, D. (2000). Change blindness blindness: The metacognitive error of overestimating changedetection ability. *Visual Cognition*,

- 7, 397–412.
- Levin, D., & Simons, D. (1997). Failure to detect changes to attended objects in motion pictures. *Psychonomic Bulletin & Review*, 4, 501–506.
- Luck, S., & Vecera, S. (2002). Attention. In *Stevens' handbook of experimental psychology*. Wiley.
- Martin, F. N., & Champlin, C. (2000). Reconsidering the limits of normal hearing. *Journal of the American Academy of Audiology*, 11(2), 64–66.
- Massaro, D., Kallman, H., & Kelly, J. (1980). The role of tone height, melodic contour, and tone chroma in melody recognition. *Journal of Experimental Psychology: Human Perception and Performance*, 6, 77–90.
- 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: Oxford University Press.
- 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. *Psychol Res*, 58, 177–192.
- Moore, B., & Glasberg, B. (1983). Suggested formulae for calculating auditory-filter bandwidths and excitation patterns. *Journal of the Acoustical Society of America*, 74(3), 750–753.
- Paavilainen, P., Arajarvi, P., & Takegata, R. (2007). Preattentive detection of nonsalient contingencies between auditory features. *Cognitive Neuroscience and Neuropsychology*, 18(2), 159–163.
- Patterson, R., Robinson, K., Holdsworth, J., McKeown, D., Zhang, C., & Allerhand, M. (1992). Complex sounds and auditory images. *Auditory Physiology and Perception*, 83, 429–446.
- Peeters, G. (2004). *A large set of audio features for sound description (similarity and classification) in the cuidado project* (CUIDADO IST Project Report).
- Peeters, G., McAdams, S., & Herrera, P. (2000). Instrument sound description in the context of mpeg-7. *Proceedings of International Computer Music Conference*, 166–169.
- Peeters, G., Susini, P., Misdariis, N., Giordano, B. L., & McAdams, S. (2011). The timbre toolbox: Extracting audio descriptors from musical signals. *Journal of the Acoustical Society of America*, 130(5), 2902–2916.
- Plomp, R. (1970). Timbre as a multidimensional attribute of complex tones. In R. Plomp & G. Smoorenburg (Eds.), *Frequency analysis and periodicity detection in hearing* (pp. 397–414). Leiden: Sijthoff.
- Poiese, P., Spalek, T., & Di Lollo, V. (2008). Attention capture by a salient distractor in visual search: The effect of target-distractor similarity. *Canadian Journal of Experimental Psychology*, 62(4), 233–236.
- Pollard, H., & Jansson, E. (1982). A tristimulus method for the specification of musical

- timbre. *Acustica*, 51, 162–171.
- Saldanha, E., & Corso, J. (1964). Timbre cues and the identification of musical instruments. *Journal of the Acoustical Society of America*, 36(11), 2021–2026.
- Saliency. (n.d.). *Merriam-webster dictionary*. Merriam-Webster, Inc. Available from <http://www.merriam-webster.com/dictionary/salience?show=0&t=1354136596> (Retrieved November 28, 2012)
- Sandell, G. J. (1995). Roles for spectral centroid and other factors in determining “blended” instrument pairings in orchestration. *Music Perception*, 13, 209–246.
- Schubö, A. (2009). Salience detection and attention capture. *Psychological Research*, 73, 233–243.
- Shepard, R. (1962a). The analysis of proximities: Multidimensional scaling with an unknown distance function. part i. *Psychometrika*, 27, 125–140.
- Shepard, R. (1962b). The analysis of proximities: Multidimensional scaling with an unknown distance function. part ii. *Psychometrika*, 27, 219–246.
- Simons, D., & Levin, D. (1998). Failure to detect changes to people during a real-world interaction. *Psychonomic Bulletin & Review*, 5, 644–649.
- Slaney, M., Agus, T., Liu, S.-C., Kaya, M., & Elhilali, M. (2012, march). A model of attention-driven scene analysis. In *Acoustics, speech and signal processing (icassp), 2012 ieee international conference on* (p. 145 -148).
- Slaney, M., Lalor, E., Choi, I., Wright, J., Brumberg, J., Ding, N., et al. (2012). *Attention in machine – final report* (Unpublished report from Telluride Neuromorphic Engineering Workshop). Available from <http://neuromorphs.net/nm/wiki/2012/att12/FinalReport>
- Smith, B. K. (1995). *Psiexp: an environment for psychoacoustic experimentation using the ircam musical workstation*. Paper presented at the Society for Music Perception and Cognition, University of California, Berkeley, U.S.A.
- Susini, P., McAdams, S., Winsberg, S., Perry, I., Vieillard, S., & Rodet, X. (2004). Characterizing the sound quality of air-conditioning noise. *Applied Acoustics*, 65, 763–790.
- Tardieu, D., & McAdams, S. (2012). Perception of dyads of impulsive and sustained sounds. *Music Perception*, 30(2), 117–128.
- Theeuwes, J. (1992). Perceptual selectivity for color and form. *Perception & Psychophysics*, 51, 599–606.
- Van Noorden, L. (1977). Minimum difference of level and frequency for perceptual fission of tone sequences abab. *Journal of the Acoustical Society of America*, 61, 1041–1045.
- Vienna Symphonic Library GmbH. (2011). *Vienna symphonic library*. Available from <http://vsl.co.at>
- Vitevitch, M. (2003). Change deafness: The inability to detect changes between two voices. *Journal of Experimental Psychology: Human Perception and Performance*, 29(2), 333–342.
- Winsberg, S., & De Soete, G. (1993). A latent-class approach to fitting the weighted

euclidean model, classical. *Psychometrika*, 58, 315–330.