Timbre 2020

PROCEEDINGS OF THE 2nd INTERNATIONAL CONFERENCE ON TIMBRE

Presented online on 3-4 September 2020

Jointly organised by the

School of Music Studies
Aristotle University of Thessaloniki

School of Electronic Engineering and Computer Science
Queen Mary University of London

Department of Medical Physics and Acoustics
University of Oldenburg

Published by:
The School of Music Studies
Aristotle University of Thessaloniki
Thessaloniki
Greece

Citation:
In A. Zacharakis, C. Saitis, K. Siedenburg (Eds.), Proceedings of the 2nd International Conference on Timbre (Timbre 2020), online conference.

© September 2020 the authors and Aristotle University of Thessaloniki
ISBN: 978-960-99845-7-7

CONFERENCE CHAIRS

Asterios Zacharakis, Aristotle University of Thessaloniki
Charalampos Saitis, Queen Mary University of London
Kai Siedenburg, University of Oldenburg

LOCAL ORGANISING COMMITTEE

Asterios Zacharakis
Konstantinos Pastiadis | Aristotle University of Thessaloniki
Emilios Cambouropoulos

TECHNICAL SUPPORT

Logo design – **Yannis Bartzis**
Proceedings – **Matina Kalaitzidou**
IT support – **Konstantinos Velenis**
Live Music – **Nikos Diminakis**

INTERNATIONAL EVALUATION COMMITTEE

Computer science – **Philippe Esling**, IRCAM/Sorbonne Université
Composition – **Denis Smalley**, City, University of London (Emeritus)
Composition and theory – **Jason Noble**, McGill University
Ethnomusicology – **Cornelia Fales**, Indiana University
Music theory and analysis – **Robert Hasegawa**, McGill University
Musicology – **Emily Dolan**, Brown University
Neuroscience – **Vinoo Alluri**, International Institute of Information Technology
Popular music studies – **Zachary Wallmark**, University of Oregon
Psychoacoustics – **Sven-Amin Lembke**, De Montfort University
Psychology – **Stephen McAdams**, McGill University
Signal processing – **Marcelo Caetano**, McGill University
Sound recording – **Joshua Reiss**, Queen Mary University of London
Voice/Synthesis – **David Howard**, Royal Holloway University of London

WELCOME FROM THE CONFERENCE CHAIRS

It is our great pleasure to welcome you to Timbre 2020: The 2nd International Conference on Timbre. The seeds for this event were planted in Montreal during the summer of 2018. The *Timbre 2018: Timbre Is a Many-Splendored Thing* conference was the second timbre event in two consecutive years, being preceded by the *Berlin Interdisciplinary Workshop on Timbre* in 2017. The scientific discourse on timbre was gaining momentum and it looked as an appropriate timing to establish a periodic gathering that would forge more solid bonds among the members of the emerging community. Thus, we chose 2020 as a year with less meetings on music perception and cognition than 2021, Thessaloniki in early autumn as the perfect location and time, and titled our meeting the 2nd International Conference on Timbre considering it a natural continuation of the Montreal conference and signifying our ambition for this to become a recurrent meeting.

When the COVID-19 pandemic broke out in the beginning of 2020 we examined going virtual, postponing or even cancelling the event altogether. Feeling that life should go on and aiming not to lose momentum we opted for the former. Of course, going online was not quite what we had envisioned in the first place. It might be less easy to form personal relations in a virtual conference compared to a physical one under the early autumn sun in Greece. On the other hand, an online format facilitates participation considerably and as a result we are delighted to report over 320 registrations coming from 41 different countries.

The scientific programme will feature 4 keynote talks and 41 original contributions (18 oral and 23 poster presentations) from a variety of disciplines including music theory and analysis, composition, psychology, computer science, acoustics and cognitive neuroscience. The oral presentations are structured around six themes: Affect, Semantics, Instruments, Perception, Orchestration, and Analysis. A panel discussion at the end of the second day will consider challenging issues of great interest across the different communities with the aim of coming up with novel ideas to help guide future research. We hope to increase engagement and interaction through live music performance in various instances, a speed dating session and a social hour at the end of each day. We would like to thank our four distinguished keynote speakers, all the presenters and the members of the international evaluation committee, without whom the realisation of this conference would not have been possible.

Welcome to Timbre 2020! Καλωσορίσατε! Willkommen!

The conference chairs,
Asteris Zacharakis
Charalampos Saitis
Kai Siedenburg

SCHEDULE

THURSDAY 3 SEPTEMBER 2020

13.40 CET: WELCOME | MUSICAL INTRODUCTION

KEYNOTE TALK 1

14.00 CET: MORWAREAD M. FARBOOD, New York University – Timbre and Musical Tension

ORAL SESSION: AFFECT

CHAIR: SOFIA DAHL

15.00 CET: Caitlyn Trevor, Luc Arnal and Sascha Frühholz | *Scary music mimics alarming acoustic feature of screams*

15.20 CET: Lena Heng and Stephen McAdams | *Timbre's function within a musical phrase in the perception of affective intents*

15.40 CET: Maria Perevedentseva | *Timbre and Affect in Electronic Dance Music Discourse*

16.00 CET: POSTER SESSION

- Kai Siedenburg | *Mapping the interrelation between spectral centroid and fundamental frequency for orchestral instrument sounds*

- Sven-Amin Lembke | *Sound-gesture identification in real-world sounds of varying timbral complexity*

- Cyrus Vahidi, George Fazekas, Charalambos Saitis and Alessandro Palladini | *Timbre Space Representation of a Subtractive Syntheziser*

- Matt Collins | *Timbral Threads: Compositional Strategies for Achieving Timbral Blend in Mixed Electroacoustic Music*

- Lindsey Reymore | *Timbre Trait Analysis: The Semantics of Instrumentation*

- Christos Drouzas and Charalampos Saitis | *Verbal Description of Musical Brightness*

- Ivan Simurra, Patricia Vanzella and João Sato | *Timbre and Visual Forms a crossmodal study relating acoustic features and the Bouba-Kiki Effect*

- Gabrielle Choma | *How Periodicity in Timbre Alters Our Perception of Time: An Analysis of "Prologue" by Gerard Grisey*

- Ryan Anderson, Alyxandria Sundheimer and William Shofner | *Cross-categorical discrimination of simple speech and music sounds based on timbral fidelity in musically experienced and naïve listeners*

- Graeme Noble, Joanna Spyra and Matthew Woolhouse | *Memory for Musical Key Distinguished by Timbre*

- Harin Lee and Daniel Müllensiefen | *A New Test for Measuring Individual's Timbre Perception Ability*

- Kaustuv Kanti Ganguli, Christos Plachouras, Sertan Şentürk, Andrew Eisenberg and Carlos Guedes| *Mapping Timbre Space in Regional Music Collections using Harmonic-Percussive Source Separation (HPSS) Decomposition*

ORAL SESSION: SEMANTICS

CHAIR: KONSTANTINOS PASTIADIS

17.00 CET: Ben Hayes and Charalampos Saitis | *There's more to timbre than musical instruments: semantic dimensions of FM sounds*

17.20 CET: Bodo Winter and Marcus Perlman | *Crossmodal language and onomatopoeia in descriptions of bird vocalization*

17.40 CET: Permagnus Lindborg | *Which timbral features granger-cause colour associations to music?*

18.00 CET: BREAK | MUSICAL INTERLUDE

19.00 CET: SPEED DATING

ORAL SESSION: INSTRUMENTS

CHAIR: ZACHARY WALMARK

20.00 CET: Francesco Bigoni, Sofia Dahl and Michael Grossbach | *Characterizing Subtle Timbre Effects of Drum Strokes Played with Different Technique*

20.20 CET: Claudia Fritz | *On the difficulty to relate the timbral qualities of a bowed-string instrument with its acoustic properties and construction parameters*

20.40 CET: Joshua Albrecht | *One singer, many voices: Distinctive within-singer groupings in Tom Waits*

KEYNOTE TALK 2

21.00 CET: STEFAN BILBAO, University of Edinburgh – Physical Modeling Synthesis: Natural Sound and Timbre

22.00 CET: SOCIAL EVENT | LIVE MUSIC

FRIDAY 4 SEPTEMBER 2020

13.45 CET: MUSICAL INTRODUCTION

KEYNOTE TALK 3

14.00 CET: JENNIFER BIZLEY, University College London – Neural mechanisms of timbre perception

ORAL SESSION: PERCEPTION

CHAIR: SVEN-AMIN LEMBKE

15.00 CET: Braden Maxwell, Johanna Fritzinger and Laurel Carney | *Neural Mechanisms for Timbre: Spectral-Centroid Discrimination based on a Model of Midbrain Neurons*

15.20 CET: Sarah Sauvé, Benjamin Rich Zendel and Jeremy Marozeau | *Age and experience-related use of timbral auditory streaming cues*

15.40 CET: Eddy Savvas Kazazis, Philippe Depalle and Stephen McAdams | *Perceptual ratio scales of timbre-related audio descriptors*

16.00 CET: POSTER SESSION

- Alejandro Delgado, Charalampos Saitis and Mark Sandler | *Spectral and Temporal Timbral Cues of Vocal Imitations of Drum Sounds*

- Islah Ali-MacLachlan, Edmund Hunt and Alastair Jamieson | *Player recognition for traditional Irish flute recordings using K-nearest neighbour classification*

- Thomas Chalkias and Konstantinos Pastiadis | *Perceptual characteristics of spaces of music performance and listening*

- Erica Huynh, Joël Bensoam and Stephen McAdams | *Perception of action and object categories in typical and atypical excitation-resonator interactions of musical instruments*

- Carolina Espinoza, Alonso Arancibia, Gabriel Cartes and Claudio Falcón | *New materials, new sounds: how metamaterials can change the timbre of musical instruments*

- Antoine Caillon, Adrien Bitton and Brice Gatinet, Philippe Esling | *Timbre Latent Space: Exploration and Creative Aspects*

- Victor Rosi, Olivier Houix, Nicolas Misdariis and Patrick Susini | *Uncovering the meaning of four semantic attributes of sound : Bright, Rough, Round and Warm*

- Jake Patten and Michael McBeath | *The difference between shrieks and shrugs: Spectral envelope correlates with changes in pitch and loudness*

- Ivonne Michele Abondano Florez | *Distorted Pieces of Something: A Compositional Approach to Luminance as a Timbral Dimension*

- Asterios Zacharakis, Ben Hayes, Charalampos Saitis and Konstantinos Pastiadis | *Evidence for timbre space robustness to an uncontrolled online stimulus presentation*

- Kaustuv Kanti Ganguli, Akshay Anantapadmanabhan and Carlos Guedes | *Questioning the Fundamental Problem-Definition of Mridangam Transcription*

ORAL SESSION: ORCHESTRATION

CHAIR: LINDSEY REYMORE

17.00 CET: Moe Touizrar and Kai Siedenburg | *The medium is the message: Questioning the necessity of a syntax for timbre*

17.20 CET: Didier Guigue and Charles de Paiva Santana | *Orchestration and Drama in J.-P. Rameau Les Boréades*

17.40 CET: Jason Noble, Kit Soden and Zachary Wallmark | *The Semantics of Orchestration: A Corpus Analysis*

18.00 CET: BREAK | MUSICAL INTERLUDE

ORAL SESSION: ANALYSIS

CHAIR: ROBERT HASEGAWA

19.00 CET: Nathalie Herold | *Towards a Theory and Analysis of Timbre based on Auditory Scene Analysis Principles: A Case Study of Beethoven's Piano Sonata Op. 106, Third Movement*

19.20 CET: Matthew Zeller | *Klangfarbenmelodie in 1911: Anton Webern's Opp. 9 and 10*

19.40 CET: Felipe Pinto-d'Aguiar | *Musical OOPArts: early emergences of timbral objects*

KEYNOTE TALK 4

20.00 CET: DAVID HOWARD, Royal Holloway University of London – The influence of timbre and other matters on unaccompanied choral tuning

21.00 CET: PANEL DISCUSSION with NATHALIE HEROLD (University of Strasbourg), PHILIPPE ESLING (IRCAM/Sorbonne Université) and GARY BROMHAM (Queen Mary University of London)

22.00 CET: CLOSING | LIVE MUSIC

KEYNOTE ADDRESSES

Timbre and Musical Tension

Morwaread M. Farbood
Department of Music and Performing Arts Professions, New York University, NY, US
mfarbood@nyu.edu

Abstract

This talk explores how timbre contributes to the perception of musical tension. Tension is an aggregate of a wide range of musical and auditory features and is a fundamental aspect of how listeners interpret and enjoy music. Timbre as a contributor to musical tension has received relatively little attention from an empirical perspective compared to other musical features such as melodic contour and harmony. The studies described here explore how common timbre descriptors contribute to tension perception. Multiple features including spectral centroid, inharmonicity, and roughness were examined through listener evaluations of tension in both artificially generated stimuli and electroacoustic works by well-known composers such as Stockhausen and Nono. Timbral tension was further examined in an audiovisual context by pairing electroacoustic compositions with abstract animations.

Physical Modeling Synthesis: Natural Sound and Timbre

Stefan Bilbao
Acoustics and Audio Group, University of Edinburgh, Edinburgh, UK
s.bilbao@ed.ac.uk

Abstract

Sound synthesis and explorations of timbre have been intertwined at least as far back as Risset's early experiments with additive synthesis in the 1960s. Particularly in the early days, there was a preoccupation with the notion of "natural" synthetic sound. As the thinking went, a good sound synthesis method should produce sound output with all the attributes of acoustically-produced sound. As Chowning wrote in 1973: "The synthesis of natural sounds has been elusive…" Physical modelling principles offer a partial remedy: natural synthetic sound is no longer elusive. And yet, physical models are posed in a way which makes any scientific exploration of the notion of timbre quite difficult. Physical modeling is a roundabout approach: physical models obey laws of physics and not human perception, and it is expected that any acoustic system that obeys the laws of physics will produce natural sound. As to why they produce natural sound---this is wrapped up in the physical parameters that define a particular model, which often do not relate directly to perceptual definitions of timbre (or even pitch or loudness). The aim of this talk is to give a short qualitative introduction to physical modelling synthesis and to explore, through both mathematical models and sound examples, the way engineers and musicians (and not scientists!) think about the notion of timbre.

Neural mechanisms of timbre perception

Jennifer Bizley
Ear Institute, University College London, London, UK
j.bizley@ucl.ac.uk

Abstract

Timbre is a key perceptual feature of sound, that allows the listener to identify a sound source. Timbral differences enable the recognition of musical instruments and are critical for vowel perception in human speech. In this talk I will present recent work that has explored how the auditory cortex extracts and represents spectral timbre and how neural representations facilitate perceptual constancy. Perceptual constancy requires neural representations that are selective for object identity, but also tolerant across identity-preserving transformations. By combining behavioural testing in ferrets and humans, with neural recordings from the auditory cortex of ferrets actively discriminating sound timbre, we will demonstrate how cortical representations represent timbre across differences in pitch and location and robustly in the presence of background noise.

The influence of timbre and other matters on unaccompanied choral tuning

David Howard
Department of Electronic Engineering Audio, Biosignals and Machine Learning, Royal Holloway University of London, London, UK
David.Howard@rhul.ac.uk

Abstract

Unaccompanied or ‘a cappella’ choral singing is a fine art when done to perfection that involves subtleties in tuning to achieve a high degree of consonance in tuning throughout. Timbre can influence pitch perception – an effect that is not directly obvious and this talk will explore some of the ways this can occur. In addition, it will consider the implications for the tuning of individual notes by singers and what the potential can be for pitch drift as well as audience appreciation of the overall tuning.

ORAL AND POSTER PRESENTATIONS

Scary music mimics alarming acoustic feature of human screams

Caitlyn Trevor[1†], Luc Arnal[2] and Sascha Frühholz[1,3]
[1] Department of Psychology, University of Zurich, Zurich, Switzerland
[2] Department of Fundamental Neuroscience, University of Geneva, Geneva, Switzerland
[3] Department of Psychology, University of Oslo, Oslo, Norway
[†] Corresponding author: caitlyn.trevor@psychologie.uzh.ch

Introduction

Music used to underscore frightening scenes in movies is often described as sounding "scream-like". A well-known example is the music accompanying the infamous shower murder scene in Alfred Hitchcock's film Psycho (1960) with, "screeching, upward glissandi," from the violins (Brown, 1982; p. 46). Although 'scream-like' is a common descriptor, the question remains: do these scary film soundtrack excerpts actually sound like and are perceived similarly to human screams? Research has demonstrated that screams have a unique auditory feature called "roughness" (Arnal et al., 2015; Schwartz et al., 2019). To measure roughness, we employ the modulation power spectrum (MPS) method and parameters used by Arnal et al. (2015). The MPS is a two-dimensional Fourier transformation of a spectrogram that quantifies both temporal and spectral power modulations (Elliott & Theunissen, 2009). Previous research indicates that human screams feature higher MPS values than non-alarming vocalizations in the 30 to 150 Hz range of the temporal modulation rate dimension of the MPS (Arnal et al., 2015). To investigate whether scream-like music has the same roughness feature as, and is perceived similarly to, human screams, we conducted two studies. In the first study, we ran an acoustic analysis to test whether recorded screams and scream-like music exhibit enhanced roughness compared with control recordings. In the second study, we collected valence and arousal ratings for the audio files in order to test whether screams and scream-like music are perceived as sharing similar emotional qualities. We made the following hypotheses. First, we hypothesized that the mean power of the MPS within the roughness region (henceforth "roughness") would be similar for screams and scream-like music, and would be significantly greater for screams compared to non-screaming vocalizations and for scream-like music compared to non-scream-like music. Second, given that roughness may be a universal cue for danger (Arnal et al., 2015), we hypothesized that roughness would correlate negatively with valence ratings and positively with arousal ratings for both music and vocal stimuli. Taken together, these results would demonstrate that scream-like music both sounds like and is perceived similarly to actual human screams.

Method

The audio recordings used in the studies were deployed in a 2x2 factorial design, with one factor corresponding to the sound source (music, voice) and the other one to the scream-likeness of the sounds (scream-like, non-scream-like). Specifically, the four collections included (a) fearful scream vocalizations, (b) scream-like film music excerpts, (c) non-fearful human vocalizations (sounding similar to a held "ah" sound), and (d) non-scream-like film music excerpts as controls. All audio recordings are 800ms in duration, RMS normalized, sampled at 16kHz, and are in wav-file format. The film music excerpts were curated from recently released horror film soundtracks (to download the excerpts, see our Open Science Framework project page at https://osf.io/7d2cy/). The vocalizations were recorded at the University of Zurich. Using MATLAB, the MPS of each excerpt was measured using the same procedure and equations as used by Arnal and colleagues (Arnal et al., 2015). Specifically, the initial spectrograms were obtained using a filter-bank approach with 128 Gaussian windows whose outputs were Hilbert transformed and then log-transformed. Then the modulation power spectra were obtained by applying a two-dimensional Fourier transform to the spectrogram (24 channels/octave) and log-transforming the resulting spectral power density estimates (Arnal et al., 2015). From there, the mean amplitude in the roughness range of 30 to 150 Hz along the temporal modulation range was taken [see Fig. 1(A); see Elliott and Theunissen (2009) for more detailed information on the MPS]. In the second study, 20 healthy

participants (twelve female) listened to each of the 200 audio files and rated the valence and arousal of the conveyed emotion using two continuous sliding scales.

Results

Recall that our first hypothesis predicted that scream-like music and screams would share a similar roughness level that would be higher than their matched controls. To test our first hypothesis, we used a standard general linear regression analysis. We also tested for an interaction effect between sound types and scream-likeness. Finally, another standard general linear regression analysis was used to test for a main effect of scream-likeness on roughness for just the voice (coded as 1) for replicative comparison to the findings of Arnal et al. (2015). The results are reported in Table 1.

Table 1: Roughness as predicted by Scream-likeness and Sound Type

	Roughness (Voice Only)	**Roughness** (Main Effect)	**Roughness** (Interaction)
Intercept	**-0.88*****	**-0.70*****	**-0.39*****
	[-1.00, -0.75]	[-0.89, -0.52]	[-0.59, -0.19]
Scream-likeness (Screaming = 1)	**1.89*****	**1.27*****	**0.65*****
	[1.72, 2.06]	[1.05, 1.48]	[0.37, 0.93]
Sound Type (Voice = 1)		0.14	**-0.48****
		[-0.08, 0.35]	[-0.76, -0.21]
Scream-likeness X Sound Type			**1.24*****
			[0.85, 1.64]
N	100	200	200
R^2 / R^2 *Adjusted*	0.826 / 0.824	0.407 / 0.401	0.503 / 0.495
*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$			CI 95%

Our second hypothesis was that roughness would correlate negatively with valence ratings and positively with arousal ratings for both music and vocal stimuli, supporting its reputation as an aural cue for danger (Arnal et al., 2015). To test this hypothesis, we used emotion ratings (valence and arousal) as the predicted values for two mixed effects linear regression models. Results are reported in part in Figure 1 [for more details on the results, see (Trevor, Arnal, & Frühholz, 2020)].

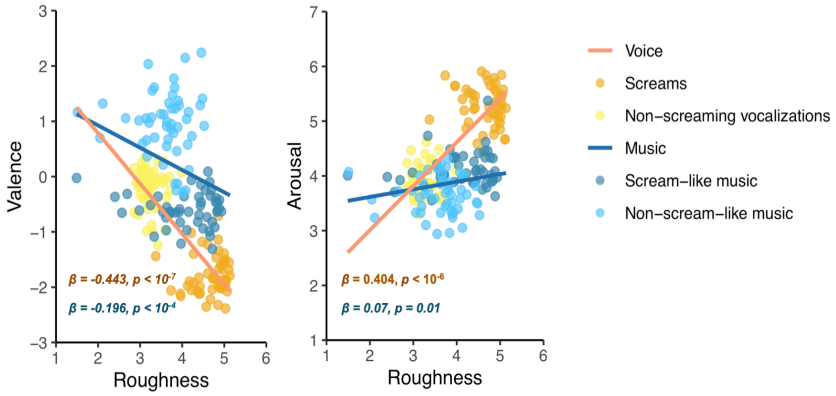


Figure 1: The left plot shows the relationship between the average valence ratings and average roughness of the four stimulus categories. There is a significant negative correlation between valence and roughness for both vocal ($\beta = -0.443$, $p < 10-7$) and musical ($\beta = -0.196$, $p < 10-4$) stimuli.

However, the correlation is significantly more pronounced for vocal stimuli as opposed to musical stimuli ($p < 10-5$). The right plot shows the relationship between the average arousal ratings and average roughness of the four stimulus categories. There is a significant positive correlation between arousal and roughness for both vocal ($\beta = 0.404$, $p < 10-6$) and musical ($\beta = 0.07$, $p = 0.01$) stimuli. However, the correlation is significantly more pronounced for vocal stimuli as opposed to musical stimuli ($p < 10-5$).

Consistent with our hypotheses, we found that both screams and scream-like music exhibited a higher level of roughness and were rated as having a more negative valence and a higher arousal level than their non-screaming counterparts. However, contrary to our hypotheses, screams had a higher roughness level than scream-like music. Overall, the results demonstrated a greater difference in roughness levels and emotion ratings between the vocal stimuli than between the musical stimuli.

Discussion

These results suggest that while scream-like music does seem to sound like and be perceived similarly to human screams, the musical rendition is still a muted version of the real thing and therefore may not provoke as potent of a reaction. Overall, the results suggest that roughness can effectively translate from a vocal cue for danger into a musical cue for danger. It is therefore reasonable to suggest that scream-like music might scare viewers in part because it is evocative of a human scream, a naturally alarming sound. For further details and analyses, please see our published paper in The Journal of the Acoustical Society of America (Trevor, Arnal, & Frühholz, 2020).

Acknowledgments

C.T. received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie Grant Agreement (No. 835682). S.F. received funding from Swiss National Science Foundation (Grants Nos. SNSF PP00P1_157409/1 and PP00P1_183711/1). The authors thank Lawrence Feth for guidance regarding the acoustic analyses and to David Huron for valuable feedback on the project. Finally, the authors thank Arkady Konovalov for helpful input regarding the statistical analyses.

References

Arnal, L. H., Flinker, A., Kleinschmidt, A., Giraud, A.-L., & Poeppel, D. (2015). Human Screams Occupy a Privileged Niche in the Communication Soundscape. *Current Biology, 25*(15), 2051–2056. https://doi.org/10.1016/j.cub.2015.06.043

Brown, R. S. (1982). Herrmann, Hitchcock, and the music of the irrational. *Cinema Journal, 21*(2), 14-49. https://doi.org/10.2307/1225034

Elliott, T. M., & Theunissen, F. E. (2009). The modulation transfer function for speech intelligibility. *PLoS Computational Biology, 5*(3), 1–14.

Schwartz, J. W., Engelberg, J. W., & Gouzoules, H. (2019). What is a scream? Acoustic characteristics of a human call type. *The Journal of the Acoustical Society of America, 145*(3), 1776–1776. https://doi.org/10.1121/1.5101500

Trevor, C., Arnal, L. H., & Frühholz, S. (2020). Terrifying film music mimics alarming acoustic feature of human screams. *The Journal of the Acoustical Society of America, 147*(6), EL540-EL545. https://doi.org/10.1121/10.0001459

Timbre's function within a musical phrase in the perception of affective intents

Lena Heng[1†] and Stephen McAdams[1]
[1] Schulich School of Music, McGill University, Montreal, Quebec, Canada
[†] Corresponding author: lena.heng@mail.mcgill.ca

Introduction

Timbre has been identified by music perception scholars as a component in the communication of affect in music. While its function as a carrier of perceptually useful information about sound source mechanics has been established, studies of whether and how it functions as a carrier of information for communicating affect in music are still in their infancy. It is a "complex auditory attribute... [and] it is also a perceptual property, not a physical one" (McAdams, 2019). However, even as timbre is a psychophysical attribute, it is the perception of the physical properties of a sound that defines timbre, and therefore the physical acoustic properties are important in timbre perception.

If timbre functions as a carrier of affective content, different aspects of the acoustic property of a sound may be implicated for different affective intents. The amount of information timbre carries across different parts of a phrase may vary according to musical context. In addition, how timbre is used for musical communication may also be different across musical traditions. Studies have revealed some differences between listeners from different cultures (Chinese and Western) in the multidimensional space obtained from rating dissimilarities of instrument sounds (e.g., Zhang and Xie, 2017). Although the differences might have been due to different sets of instruments used (Chinese vs. Western instruments), the different dimensions obtained from the multidimensional scaling could also imply a focus on different aspects of a sound by different groups of listeners (McAdams et al., 1995). How these acoustic features are used may also have been learned differently across different musical traditions.

This study therefore aims to find out if specific acoustic features or combinations of them are related to affective intents communicated in a performance, and how increasing musical context might influence this process of understanding. In addition, it also attempts to look at whether differences in musical experience play a role in this decoding process for performances by instruments from a musical tradition listeners have different familiarity with, and whether different acoustic features within a sound are being used differently by each of the groups of listeners in the understanding of affective intents in music.

Method

To investigate these issues, three groups of listeners with different musical backgrounds (Chinese musicians (CHM) and Western musicians (WM) and nonmusicians (NM), n = 30 per group) from Singapore were recruited for listening experiments. The criteria for musicians during participant recruitment was to have more than five years of formal musical training in either the Chinese (mean = 12.00, SD = 2.98) or Western (mean = 12.13, SD = 7.40) music tradition, and the criteria for nonmusicians (mean = 0.2, SD = 0.41) was less than a year of formal training in any type of music. There was no significant difference between the number of years of musical training between the CHM and WM listeners, $F(1, 58) = 1.41$, $p = .24$. None of the WM listeners had any prior training in Chinese music while some CHM listeners had received formal instruction in Western music. All the CHM listeners however self-identified as being more proficient in Chinese music than Western music. All the participants had casual exposure to both Chinese and Western art music, both being ubiquitous musical forms found in Singapore.

One professional musician for each instrument (*dizi*, flute, *erhu*, violin, *pipa*, and guitar) was recruited for the recording. The two-dimensional model of valence and arousal (Russell, 1980) was explained to the performers, and they were asked to interpret the excerpt of music in performance with five different affective intents: low valence and arousal, low valence and high arousal, high valence and arousal, high valence and low arousal, and neutral.

All of the participants took part in two experimental sessions conducted at least a week apart. As the stimuli used for both experiments were obtained from the same recordings, this delay between the first and second experiments was to reduce any memory effects. Experiment A involved participants listening to individual notes extracted from the recorded excerpts, which were interpreted with a variety of affective intents by performers on Western and Chinese instruments, and then making judgements about each stimulus' perceived affective intent within a two-dimensional affective space of valence and arousal. Experiment B involved participants listening to measures and phrases of these same recorded excerpts and making judgements of the affective intents. Half of the participants were randomly assigned to experiment A first while the other half were assigned to experiment B first.

Using the Timbre Toolbox implemented in the MATLAB environment, individual notes were analyzed for their temporal, spectral, and spectrotemporal descriptors. Based on hierarchical clustering analyses done by Peeters and colleagues (2011), 13 acoustic descriptors that represent each cluster were selected. These acoustic descriptors included median and interquartile range of spectral centroid, spectral flatness, and RMS envelope, as well as the median for noisiness, harmonic spectral deviation, spectrotemporal variation, temporal centroid, frequency and amplitude modulations, and log attack time.

Results

The first set of analyses attempts to address the question of how acoustic features may be related to listeners' decoding of perceived affective intents, and whether formal training in different musical traditions influences the ways in which these listeners use the acoustic features. The interest in this study is focused on whether listeners fluent in a particular musical tradition converge on a similar set of acoustic features they use in their decoding process, rather than on the accuracy of this communication process. Instead of looking at the number of "correct" responses from the listeners, all the responses of the listeners in each group were coded into one of the four quadrants on the affective space, regardless of whether they were correct in their judgement of the performer's affective intent. These four quadrants are: low valence and arousal, low valence high arousal, high valence and arousal, and high valence low arousal. The values of each acoustic descriptor for the notes in a particular quadrant are averaged. From this, four different sets of values for each acoustic descriptor are obtained over the 30-note excerpt. Similar procedures are used for listeners' responses from individual notes, measures, and phrases. The Kruskal-Wallis test on ranks was used to test if the acoustic descriptors that are perceived as expressing different affective intents were significantly different between the groups of listeners, given that the sample size for each group of perceived affective intent can be very different and that consequently the assumptions for parametric tests might be violated. Further post-hoc pairwise comparisons were performed using the Mann-Whitney test, and the Bonferroni-Holm method was used to adjust the critical alpha for the multiple pair-wise comparisons. Due to space limitations, graphic representations of the results can be found on the internet at the following address: http://132.206.14.109/supplementaryMaterials/HengTIMBRE2020/1.pdf. With increasing musical context (note to measure to phrase), there was increasing differentiation between the different affective intents, indicating an important function of contextual information in understanding perceived affective intents. It also appears that even when the notes are presented individually in a random order to listeners, listeners are able to use certain acoustic features quite consistently in their attempts at understanding affective intents. This effect is even more pronounced for the CHM listeners where the values for several acoustic descriptors such as the spectral centroid median for *dizi* stimuli were all significantly different between the different affective intents even at the note level. CHM listeners were generally more consistent in the acoustic features they used to determine the perceived affective intents. There was also greater differentiation between the different affective intents in the CHM listeners, followed by the WM listeners, whereas NM listeners were the least consistent and had the least differentiation between the different affective intents. This trend was seen regardless of the musical tradition of the performer: CHM listeners performed with the greatest consistency in excerpts played by both Chinese and Western instruments.

Ordinary least squares linear multiple regression is next used to look at the relationships between the acoustic features and the dimensions of valence and arousal rated by listeners. The 13 acoustic descriptors for each note were regressed onto the valence and arousal values, which ranged from –1 to +1. A high degree of collinearity was present between these acoustic descriptors, and some were excluded from the regression equations in different analyses. The collinearity values were different for the different notes which resulted in different sets of acoustic descriptors being represented in each regression equation. As the sample sizes for each note were the same, a meaningful comparison could be made with the *t*-values of the regression coefficients. Comparing across listener groups, there appears to be more significant (and more highly significant $p < .001$) *t*-values for the CHM listeners as can be seen for the iqr of RMS energy envelope: http://132.206.14.109/supplementaryMaterials/HengTIMBRE2020/2.pdf, followed by WM, and then the NM listeners. This trend can be also found in the other acoustic features. These different groups of listeners also appear to utilize the acoustic features differently for the affective intents. The weight of the acoustic features used in each note over the excerpt also varies across the different listener groups, as well as across different instruments. Spectral centroid median, for instance, is consistently negatively correlated with valence in the CHM listeners for stimuli played by the *dizi*, whereas the relationship is not as consistent for the other two groups of listeners: http://132.206.14.109/supplementaryMaterials/HengTIMBRE2020/3.pdf. However, the correlation of spectral centroid with valence is less strong in all the groups of listeners for stimuli played by the flute. Similarly, when instruments with similar sound-producing mechanisms are compared (*dizi* and flute; *erhu* and violin; *pipa* and guitar), the temporal centroid appears to be implicated more in communicating affective intents in instruments of the Chinese music tradition, although the few notes with temporal centroid that correlated with affective intents for the violin appeared to have higher significance. CHM listeners also appeared to make greater use of spectral flatness iqr and RMS envelope iqr in their judgements of perceived arousal.

Discussion

There appears to be increasing differentiation in judgement of the different affective intents from participants' responses of notes to measures to phrases, indicating an important function of contextual information in understanding perceived affective intents. While this is expected, it also appears that even when the notes are presented individually and in a random order to listeners, listeners who have had musical training (CHM and WM) are able to use certain acoustic features quite consistently in their attempts at understanding affective intents. Although this is not the same as compared with an approach in which participants rate continuous changes in affective intents for an excerpt of music, the relationships emerging from this current experiment suggests that contextual changes to timbre manipulations by performers contribute a certain extent to the timbre quality of the produced sound, and these subtle changes can provide enough information for quite consistent decoding of affective intents, albeit with less and different information as compared to when a listener can hear an entire excerpt in the right sequence. Musical listening therefore appears to be a complex combination of decoding acoustic information from each individual sound and a complimentary ensemble of contextual cues for an increasingly nuanced understanding.

Results also show that listeners trained in the Chinese music tradition are the most consistent in decoding affective intent of a musical performance, for both Chinese and Western instruments, and nonmusicians fared the worst. Differences in musical training could perhaps have led to an increased focus on timbre characteristics in comprehending affective intent in music. There may be differences in the emphasis on timbre use in the musical training of different musical traditions. The divergent ways in which musical parameters are used and interpreted within the Chinese music tradition compared to the Western music tradition may mean that CHM listeners are much more sensitive to minute changes in the way timbre is being manipulated in expressing an affective intent. While this analysis does not reflect whether CHM listeners are more accurate than WM or NM listeners, it does indicate the increased consistency with

which listeners make use of each acoustic feature, as shown in the greater divergences in the acoustic features between the affective intents.

The function of timbre in communicating musical information is a highly complex process with many interactions involving not only different musical parameters, but also interactions between the rich multitude of acoustic features that make up the quality of a sound. It appears that how timbre is used in communicating affective intents in music is also different across different musical traditions, with listeners having experiences in different musical traditions making use of different sets of acoustic features to different extents. Once again, this suggests the importance of learning with respect to conventions regarding timbre manipulations in musical communication. Musical understanding is therefore dependent on the performer, the stylistic characteristics of the composer, the musical tradition, and also on the experience of the listener and the listening process, to name just a few of these complex factors.

Only listeners from Singapore were recruited for this study. Such a sampling means that besides the difference in musical backgrounds, confounding variables from other socio-cultural factors of the listeners were reduced. However, another interesting question will be whether CHM, WM, and NM listeners from other localities might have different responses from those in Singapore. If so, this might imply very subtle differences in musical communication in different parts of the world, even within the same type of musical tradition. This study focuses only on whether listeners make consistent use of particular acoustic features in understanding perceived affective intents, and if differences in musical experience might relate to differences in this process. Future work will attempt to look into how performers utilize these timbre manipulations in expressing their intents, and also at the amount of convergence between the performers' intents and the listeners' comprehension of them. No continuous response was elicited from listeners with respect to changes in affective intents over the course of the excerpt. While the comparisons across responses for notes, measures, and phrases provide an indication of musical context providing increasing cues for understanding, future studies could also attempt to look at continuous responses to better understand the function of timbre over the course of an excerpt of music.

Acknowledgments

We thank CIRMMT for funding the Inter-Centre Research Exchange. We also thank Dr. Dorien Herremans (Singapore University of Technology and Design) for hosting LH for the Inter-Centre Research Exchange and for providing support in data collection. This research was supported by grants to SMc from the Canadian Social Sciences and Humanities Research Council (895-2018-1023) and the Fonds de recherche Québec—Société et culture (017-SE-205667), as well as a Canada Research Chair (950-223484).

References

McAdams, S. (2019). The perceptual representation of timbre. In K. Siedenburg, C. Saitis, S. McAdams, A. Popper, & R. Fay (eds.), *Timbre: Acoustics, perception, and cognition* (pp. 23–57). Springer Handbook of Auditory Research, vol 69. Springer, Cham.

McAdams, S., Winsberg, S., Donnadieu, S., De Soete, G., & Krimphoff, J. (1995). Perceptual scaling of synthesized musical timbres: Common dimensions, specificities, and latent subject classes. *Psychological Research, 58*(3), 177–192.

Peeters, G., Giordano, B. L., Susini, P., Misdariis, N., & McAdams, S. (2011). The timbre toolbox: Extracting audio descriptors from musical signals. *The Journal of the Acoustical Society of America*, *130*(5), 2902–2916.

Russell, J.A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology, 39*(6), 1161–1178.

Zhang, J., & Xie, L. (2017). Analysis of timbre perceptual discrimination for Chinese traditional musical instruments. *Proceedings of 10th International Congress on Image and Signal Processing.* Shanghai: China., 1–4.

Timbre and Affect in Electronic Dance Music Discourse

Maria Perevedentseva
Department of Music, Goldsmiths, University of London, United Kingdom
mpere001@gold.ac.uk

Introduction

Different notions of affect are often portrayed as the singular motivating force behind and distinguishing feature of electronic dance music and culture (e.g. Gilbert & Pearson, 1999; Jasen, 2017). Especially prominent is the Deleuze-Guattari-Massumi strand of affect theory, which locates affect in the midst of things and uses it to probe 'how the "outside" realms of the pre-/para-linguistic intersect with "lower" proximal senses' (Seigworth & Gregg, 2010, p. 8). Ruth Leys (2011, p. 449) summarises the prized attributes of post-Deleuzean affect theorists as including 'the nonsemantic, the nonlinear, […] the vital, the singular, […] the indeterminate, […] and the disruption of fixed or "conventional" meanings', all of which privilege nonconscious bodily becomings over conscious deliberation. Such emphases on ineffability, in-betweenness and vitality also mark some recent musicological work on timbre, which conceives of timbre as 'vital relationality' which supersedes binarisms between the acoustic and the perceptual, the material and the ideal, and between timbre and tone (Van Elferen, 2018, p. 18). At the same time, studies of user-generated EDM discourse (e.g. Jóri, 2020) have noted that EDM fans show a keen awareness and understanding of timbre in their music, despite the language used to express it being marked—like the timbres of EDM—by 'nonspecificity' (Fales, 2018, p. 25) and affectivity. In this paper, I conduct a discourse analysis of what appear to be nonspecific affective verbal descriptors used in EDM record reviews on the online record retailer Boomkat.com, in order to investigate whether the vernacular affective terms used by EDM scene participants really are nonspecific, or whether and to what extent they correlate with timbral features in the tracks they are reviewing. Boomkat's record reviews are well-known for their inventive use of language which plays upon the tacit scene knowledge of their customers, and as such represent an interesting case study into the natural language of the EDM discourse community.

Method

My analysis relies on a mix of quantitative, qualitative and computational methods, bringing approaches from corpus linguistics into the fold of music analysis. This paper forms part of a larger study which analysed language data from multiple EDM subgenre categories on the Boomkat website using the Voyant text analysis software, manually classified frequently occurring terms into trope categories, and compared the relative weightings of the tropes between subgenres in order to explore the 'concealed inflections of taste' (Graham, 2019, p. 533) latent in EDM culture. In this paper, the focus is on the Techno–House genre category only: my corpus is made up of data pertaining to 1027 records released in the year leading up to January 2019 which were scraped from the Boomkat website using the Beautiful Soup package in Python. Figure 1 below shows the distribution of types (individual word-forms) and tokens (instances of word-forms) among the trope categories for all words occurring >4 times in the genre corpus. As the chart shows, unambiguous references to timbre and instrumentation (Trope 3), rhythm, metre and velocity (Trope 4), and other musical features (Trope 8) are considerably less prevalent than references to affect (Trope 7), the trope category which contains emotive and whimsical terms describing loosely specified sonic properties or the general "feel" of the releases. In line with the stated aims of this paper to identify timbral correlates to nonspecific affective descriptors, the rest of the analysis concerns the 797 types derived from the Affect trope.

From this large and diverse lexical pool, it is possible to identify latent sub-tropes, such as words referring to force, action or motion ("pressure", "swerve", "bang"); emotional and evaluative terms ("cranky", "natty", "infectious"); references to mass, texture and luminance ("weightless", "rough", "murky"); cross-modal correspondences ("tangy", "woozy"); stylistic references ("cinematic", "jazzy"); and references to altered or transformed consciousness ("trippy", "sublime", hypnotic"). Many of these categories are

consistent with the types of language commonly used to describe timbre. The prevalence of terms which emphasise materiality, sensation and action in my corpus echoes Wallmark's (2019) identification of these domains as the key dimensions underpinning the cognitive linguistics of timbre, and supports Wallmark and Kendall's (2018) observation that timbral language is often affectively loaded.

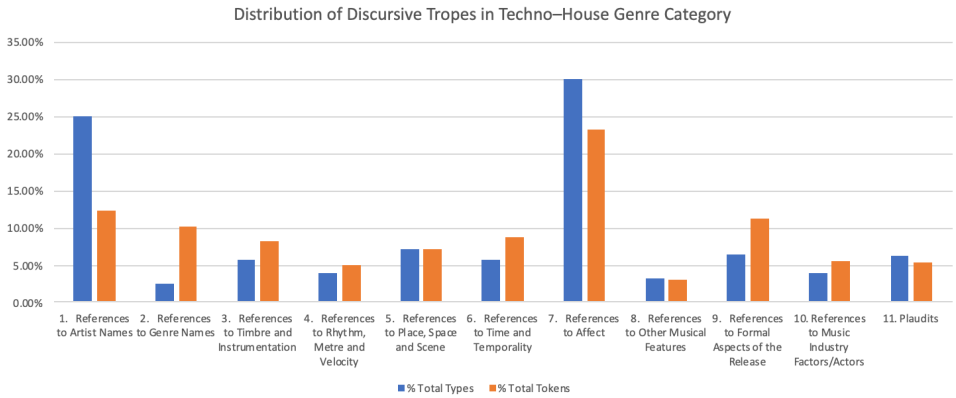


Figure 1. Distribution of types and tokens among 11 trope categories. % of 2650 types and 37262 tokens.

In order to explore how particular words relate to the timbral features of the records being reviewed, two terms—"torque" and "keening"—are selected for further analysis on the basis of their idiosyncrasy, frequency of occurrence, and their affective, mimetic and embodied connotations, which bring them in line with the types of language often used to describe timbre outlined above. The standard definition of "torque" is that it is a force that tends to cause rotation. In terms of embodied image schemata, it implies a downwards force and a sideways, rotary motion, as well as having a haptic dimension through its association with grinding and an affective charge of pressure and tension. "Keening" is a type of solo female singing originating from Ireland and traditionally practiced as part of the mourning ritual, which is characterised by a haunting and often wavering wailing sound either devoid of lyrics altogether, or with barely articulated consonants so that the visceral properties of the wail are foregrounded over semantic meaning.

Following the seven categories of timbre descriptors arrived at in Wallmark's (2019) study of orchestration treatises, "keening" can be understood as straddling mimetic, acoustic and affective categories, while "torque" exhibits traits of affect, action, and cross-modal correspondence. These associations and their relevance to the Techno–House corpus are further evidenced by visualisations created with a word embedding projector tool based on the Fasttext.cc language model which enables the dimensionality reduction and visualisation of semantically related terms. In this corpus, "torque" is represented as closely related to terms like "traction", "downstroke" and "angles", while "keening" is related to descriptors like "ghostly", "lilting" and "wistful". Using the context tool in Voyant, the terms "torque" and "keening" are matched with the record reviews in which they appear. Of its 26 total instantiations in the Techno–House corpus, "torque" is used in reference to 22 named tracks, while "keening" is used a total of 18 times in the corpus, and 13 times in reference to specific tracks. All named tracks containing either of the two terms are then analysed by close listening and the visual inspection of spectrograms in order to investigate whether the tracks share particular timbral features which could, via conceptual metaphor, be matched to the terms used to describe them.

Results

My findings show that, of the 13 tracks whose descriptions included the term "keening", 9 contain synthesiser lines placed well above the other elements of the mix in the middle or high registers, whose spectromorphologies are characterised by extremely smooth amplitude envelopes with almost inaudible attack onsets, slight filter oscillation throughout the duration of the gesture, and subtle detunings, reverb and noise elements. Of the 22 tracks whose descriptions include "torque", 11 contained a specific gesture characterised by a mid-range synthesiser line doubled by strong sub-bass several octaves below, an

internally dynamic temporal envelope achieved by a combination of amplitude, pitch and filter modulation, and a complex, resonant harmonic spectrum with plenty of distortion, especially near the attack onset. Figures 2 and 3 below show spectrograms of examples of "keening" and "torque" gestures created using Sonic Visualiser. Audio examples will be provided.

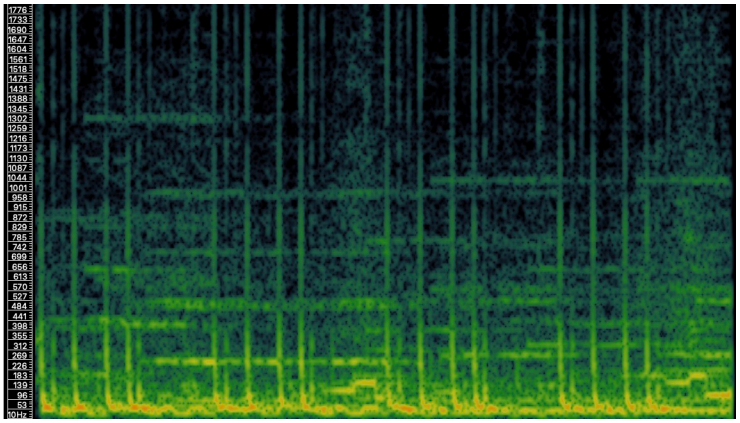


Figure 2. Spectrogram of "keening" gesture from Pär Grindvik 'Trails' (3'11–3'17)

In the Pär Grindvik example, the smooth and fairly clean-sounding top lines "float" faintly above an otherwise bottom-heavy mix, and are doubled by a mid-range line surrounded by a slight halo of distortion. When these two timbres are perceptually grouped together, the effect seems to mimic the sound of a voice breaking as it attempts to reach a higher pitch. The association with keening is further reinforced by the *portato* melodic contour and lack of clear attack onsets which mirrors the lack of consonants in sung keening. The overall prevalence of surface noise and crackle plays on an established "hauntological" trope in EDM discourse which further bolsters the links between keening and the supernatural.

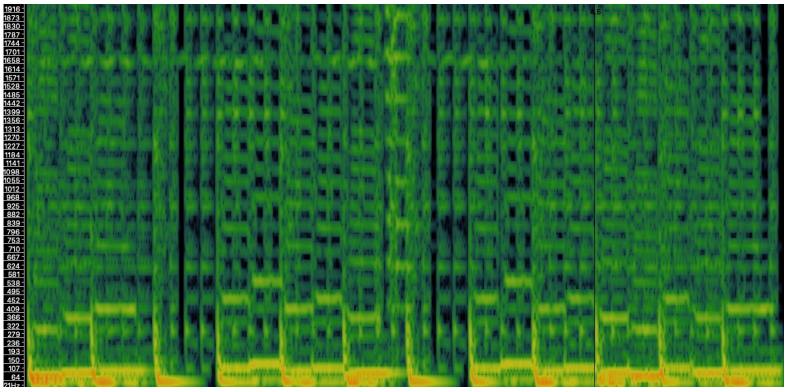


Figure 3. Spectrogram of "torque" gesture from Bambooman 'Ricochet (Matthew Herbert's Milky Dub)' (1'12–1'18)

In the Bambooman example, the resonant and harmonically complex mid-range line is doubled several octaves below by an LFO-modulated bass line and a high-amplitude sub-bass. As Smalley (2007) notes, the presence of sub-bass extends the depth of the sound, and the "weight" of the bass part in comparison with the mid-range line, with which it is perceptually fused when listening, creates a distinct sense of downward pull. Furthermore, each note of the main synth line is marked by a filter envelope modulation which creates a sensation of timbral dilation, schematically mapping onto the rotary dimension of torque. Lastly, the noise distortion present in the sound could, as per the findings of Wallmark et al. (2018), contribute to a cross-modal perception of physical exertion required to produce the sound, which again corresponds to the feeling of force associated with torque as an embodied action.

Discussion

In line with several recent studies surveyed by Wallmark and Kendall (2018), my results show that conceptual metaphors and notions of embodiment play a prominent role in the affective language used to

describe sound in EDM on Boomkat.com, and that particular terms are consistently applied to specific timbral features with which they share certain invariant characteristics. As noted by Fales (2018), pitch and other musical parameters also clearly influence timbre cognition here, and embodied image schemata are complemented by the cultural conventions of EDM, all of which serve to flesh out and make meaningful the effects of timbre in this music. Given the extent of the correlation between affective verbal descriptors and particular timbral features, it seems clear that EDM scene participants have developed a stable and, contrary to previous accounts, relatively specific vernacular lexicon for communicating about timbre. Words like "keening" and "torque", which may appear vague to listeners outside of the EDM discourse community, appear to form a key part of the 'working vocabulary' (ibid., p.29) for committed EDM fans, specifying concrete spectromorphological attributes that are recognisable and communicable. It is possible, then, that the nonspecificity and emphasis on affective becomings over conscious meaning-making ascribed to EDM cultures is founded on ideological rather than musical grounds. My findings tentatively confirm Wallmark's (2019, p. 600) hypothesis that embodied and ecological concerns undergird a broad 'swath of the discursive landscape for musical timbre in many linguistic and cultural contexts'. However, more work is needed to scale up and explore other terms, and establishing inter-rater agreement between multiple researchers would help to counteract any subjective biases that are all but inevitable when working alone.

Acknowledgments

With thanks to Stephen Graham, George Lewis Walker, and the CHASE Doctoral Training Partnership.

References

Fales, C. (2018). Hearing Timbre: Perceptual Learning among Early Bay Area Ravers. In R. Fink, M. Latour, & Z. Wallmark (eds.), *The Relentless Pursuit of Tone. Timbre in Popular Music* (pp. 21–42). New York: Oxford University Press.

Gilbert, J., & Pearson, E. (1999). *Discographies: Dance music, culture, and the politics of sound*. London: Routledge.

Graham, S. (2019). From Microphone to the Wire: Cultural change in 1970s and 1980s music writing. *Twentieth-Century Music*, *16*(3), 531–555.

Jasen, P. C. (2017). *Low end theory: Bass, bodies and the materiality of sonic experience*. New York: Bloomsbury Academic.

Jóri, A. (2020). The Discourse Community of Electronic Dance Music Through the Example of the TB-303 Owners Club. In A. Jóri & M. Lücke (eds), *The New Age of Electronic Dance Music and Club Culture* (pp. 117–131). Cham: Springer International Publishing.

Leys, R. (2011). The Turn to Affect: A Critique. *Critical Inquiry*, *37*(3), 434–472.

Seigworth, G. J., & Gregg, M. (2010). An Inventory of Shimmers. In M. Gregg & G. J. Seigworth (eds.), *The Affect Theory Reader* (pp. 1–25). Durham and London: Duke University Press.

Smalley, D. (2007). Space-form and the acousmatic image. *Organised Sound*, *12*(1), 35–58.

Van Elferen, I. (2018). Timbrality: The Vibrant Aesthetics of Tone Color. In E. Dolan & A. Rehding (eds.), *The Oxford Handbook of Timbre*. New York: Oxford University Press.

Wallmark, Z. (2019). A corpus analysis of timbre semantics in orchestration treatises. *Psychology of Music*, *47*(4), 585–605.

Wallmark, Z., Iacoboni, M., Deblieck, C., & Kendall, R. (2018). Embodied Listening and Timbre: Perceptual, Acoustical, and Neural Correlates. *Music Perception*, *35*(3), 332–363.

Wallmark, Z., & Kendall, R. (2018). Describing Sound: The Cognitive Linguistics of Timbre. In E. Dolan & A. Rehding (eds), *The Oxford Handbook of Timbre*. New York: Oxford University Press.

Mapping the interrelation between spectral centroid and fundamental frequency for orchestral instrument sounds

Kai Siedenburg
Department of Medical Physics and Acoustics and Cluster of Excellence Hearing4all, University of Oldenburg, Oldenburg, Germany
kai.siedenburg@uni-oldenburg.de

Introduction

Spectral envelope and fundamental frequency (F0) are important parameters for the acoustical description of musical sounds and often considered as key acoustical determinants of timbre and pitch perception, respectively. Although both factors are ubiquitous in research on psychoacoustics, little empirical work has addressed their interrelation in acoustical instrument sounds. Considering instruments that adhere to a source-filter model of sound generation, Patterson, Gaudrain, and Walters (2010) outlined that instrument sounds exhibit F0-invariant spectral envelope shapes that are characteristic of the respective instrument family. Furthermore, it was hypothesized that instrumental register (or size) is determined by the position of the spectral envelope along the log-frequency axis. The latter hypothesis would imply that smaller instruments from the same family (e.g., tenor compared to alto) generate sounds with higher spectral centroids (SC), that is, yielding a brighter sound. The present study sought to test this hypothesis by characterizing the relation between F0 and SC for acoustic instrument sounds from six different instrument families of the Western orchestra.

Method

Sounds were obtained from the Vienna Symphonic Library (www.vsl.co.at) and were sustained for durations of 250 ms plus instrument-specific decay times. For this analysis, we considered sounds from 6 different instrument families (brass, double-reeds, saxophones, strings, woodwinds, and voices). Sounds were played with their regular articulation (strings were bowed). Each instrument contributed with its full playing range at 3 different dynamic levels (pp, mf, ff). The SC was measured based on an ERB-spaced spectral binning and plotted as a function of F0.

Results

Figure 1 shows SC trajectories for the 29 instruments (or voices) involved, averaged across dynamic levels. For brass and double-reed instruments, SC is relatively constant across F0 and there is a clear increase of SC for higher registers (e.g., trumpet compared to tenor trombone). SCs from saxophones, woodwinds, and string sounds vary more linearly with F0 and do not seem to show any systematic variation of SCs for instruments from different registers. For instance, the clarinet, surprisingly, does not show consistently higher SCs compared to the bass clarinet. Sounds from the voice (the archetypal source-filter system) appear to have relatively fixed SC values at least up until the tenor register, but do not show any consistent separation of SC for the different vocal registers.

Discussion

This study seeks to empirically characterize the relation between F0 and SC in a large set of instrument sounds. The SC is a primitive yet useful measure of the center of gravity of the spectral envelope, and correlating F0 and SC revealed several differences between instrument families. Specifically, the hypothesis that instrumental size (or register) is associated with a shift of spectral envelope along log-frequency (Patterson et al., 2010) could only be confirmed for the brass and double-reed families. For sounds from the saxophones, woodwinds, strings, and voices, no consistent shift of SC was observed. Further exploring the relation between these elementary acoustical parameters may help to clarify the

acoustical underpinnings of several perceptual phenomena related to the interaction of pitch and timbre, such as instrument identification (Steele & Williams, 2008), dissimilarity perception (Marozeau, de Cheveigne, McAdams, & Winsberg, 2003), pitch interval perception (Russo & Thomson, 2005), or auditory short-term memory (Siedenburg & McAdams, 2018).

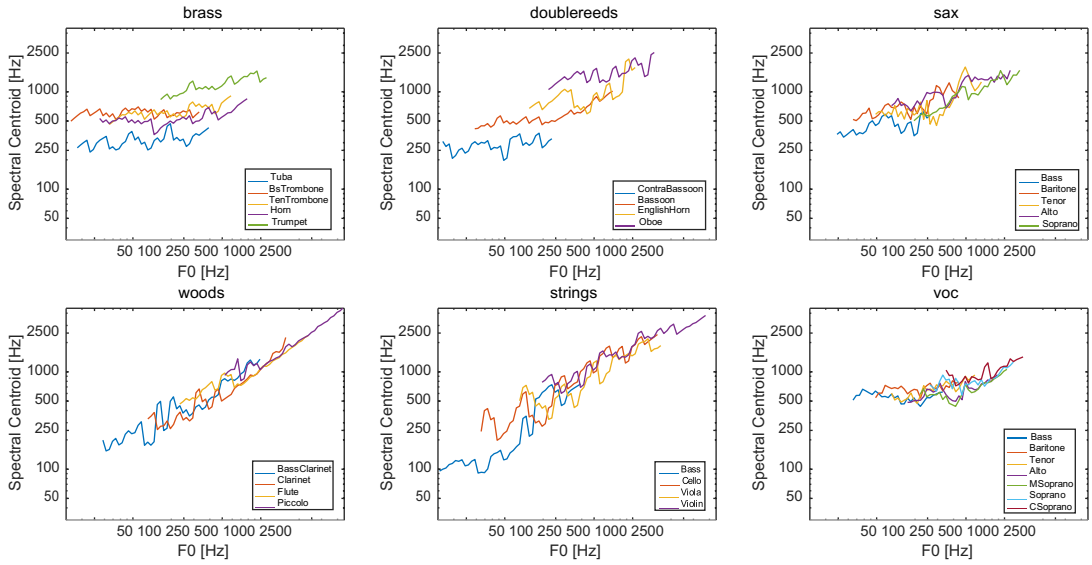


Figure 1: Relation between spectral centroid and fundamental frequency for sounds from six instrument families.

Acknowledgments

KS is supported by a Freigeist Fellowship of the Volkswagen Stiftung.

References

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

Patterson, R. D., Gaudrain, E., and Walters, T. C. (2010). The perception of family and register in musical tones. In Riess Jones, M., Fay, R. R., and Popper, A., *Music Perception*. Springer Handbook of Auditory Research (pp. 13–50). Springer, New York, NY.

Russo, F. and Thompson, W. (2005). An interval size illusion: The influence of timbre on the perceived size of melodic intervals. Attention, Perception, & Psychophysics, 67(4), 559–568.

Siedenburg, K. and McAdams, S. (2018). Short-term recognition of timbre sequences: Music training, pitch variability, and timbral similarity. Music Perception, 36(1), 24–39.

Steele, K. M. and Williams, A. K. (2006). Is the bandwidth for timbre invariance only one octave? Music Perception, 23(3), 215–220.

Sound-gesture identification in real-world sounds of varying timbral complexity

Sven-Amin Lembke
Music, Technology and Innovation - Institute for Sonic Creativity (MTI),
De Montfort University, Leicester, United Kingdom
sven-amin.lembke@dmu.ac.uk

Introduction

Real-world sounds commonly allow the identification of their physical source or cause, whereas they also convey qualitative timbral features. Source-related *categorical* properties of timbre can detract listeners' evaluation of timbral *qualities* (e.g., Lemaitre et al., 2010), where it can be assumed that both aspects of timbre may compete for listeners' attention.

In electroacoustic music, the acousmatic tradition and also the theory of spectromorphology (Smalley, 1997) place greater importance on the intrinsic sound qualities, based on which morphologies like *sound gestures* or *textures* can arise. A sound gesture concerns a process in which one or more spectral properties (e.g., spectral frequency or amplitude) vary over time, bearing either direct or metaphorical associations to underlying spatio-kinetic action(s) of gestural kind.

Spectromorphology also considers the notion of *source bonding* (Smalley, 1997), which acknowledges listeners' natural tendency to focus on extrinsic links, like the source or cause when they are identifiable, which in turn might impair the perception of the more important intrinsic features. The current study investigated the role of source bonding on the perception of sound gestures based on whether the source or cause could be identified and also across range of real-world sounds of varying timbral complexity.

Method

Twenty listeners (age range: 18-65, 12 female, 8 male) were asked to identify sound gestures inherent in real-world sounds. Gestures concerned variations along the auditory parameters timbral brightness or loudness. The latter varied along the temporal amplitude envelope (e.g., tennis ball bouncing, metal coin spinning). In many real-world sounds, loudness variation may also affect timbre. Gestures along timbral brightness varied along frequency trajectories, which could concern tonal or filtered noise components in the original real-world sounds (e.g., vacuum cleaner turned on and off, filling a water carafe). Although the underlying spectral features may have affected both timbral brightness and a more abstract notion of pitch, in the context of gestural contours, both can be assumed equivalent (e.g., McDermott et al., 2008).

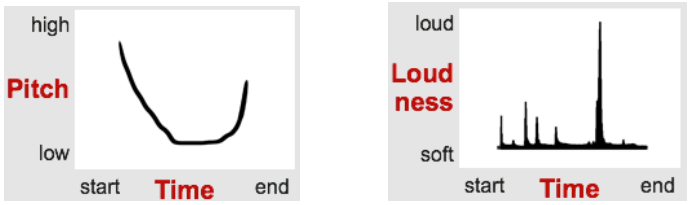


Figure 1: Examples of visual sketches based on which brightness gestures (left, here labeled pitch) or loudness gestures (right) gestures were identified.

For each parameter, participants evaluated 14 different sounds to 1) identify the embedded sound gesture and 2) also attempt to identify the underlying real-world source or cause by both providing verbal descriptions for source and cause and rating their confidence on this source/cause identity. The identification of sound gestures was achieved by selecting one correct visual sketch out of four options. As illustrated in Figure 1, the sketches were visual analogues of the underlying gestural brightness (labeled "pitch" for greater clarity to untrained listeners) or amplitude features extracted from the real-world sounds. Incorrect visual-sketch options concerned reversals or inversions of the correct gesture and/or alternatives from similar sounds.

Furthermore, gesture-identification accuracy was studied by taking the role of source bonding and repeated listening into account. To remove source bonds while retaining gestural cues, gestures were resynthesised as bandpass-filtered noise. For brightness, the filter's centre frequency followed the gestural shape, whereas for loudness, the gesture concerned variations of the amplitude envelope. Participants listened to the same gestures three times and in the following order: I. as the *original* real-world sounds, II. as the *noise*-based variants, and III. as a repeated presentation of the original. This III. presentation, however, either involved the *original* sound or a *hybrid* between the original and noise-based sounds.

Results

As shown in Figure 2, brightness (left) and loudness (left) gestures yielded 50% and more correct identifications for half the gestures investigated, suggesting that gestures can indeed be perceived in real-world sounds. Compared to gestures in original sounds (I.), however, noise-based gestures (II.) appeared to be identified more accurately, as the entire distribution of 14 gestures yielded 5-10% higher identification. Overall, these trends suggest that gestures presented in isolation were identified more reliably, i.e., when the source or cause was less identifiable (see relatively lower confidence for II. in Figure 3). As the noise-based gestures (II.) also represented a repeated listening, however, familiarity with the gesture could have also factored in. If repeated listening alone increased identification accuracy, then the identification rate in condition III. would have surpassed both previous presentations in I. and II. Indeed, the III. presentation of *hybrid* gestures appeared to increase identification rate, whereas the repetition of *original* sounds exhibited a lower median identification rate compared to the noise-based sounds in II. In sum, although gesture identification appeared to be aided by listening repetition, identification improved further when gestural cues were readily apparent, supported by the greater salience of gestural cues in hybrid sounds compared to original sounds.

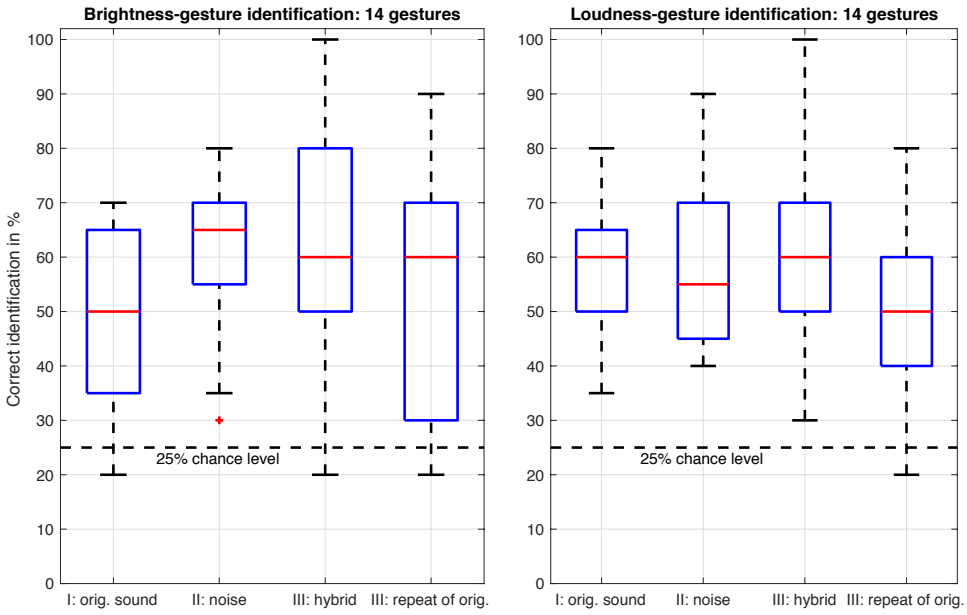


Figure 2: Identification accuracy of 14 brightness (left) and 14 loudness gestures (right).

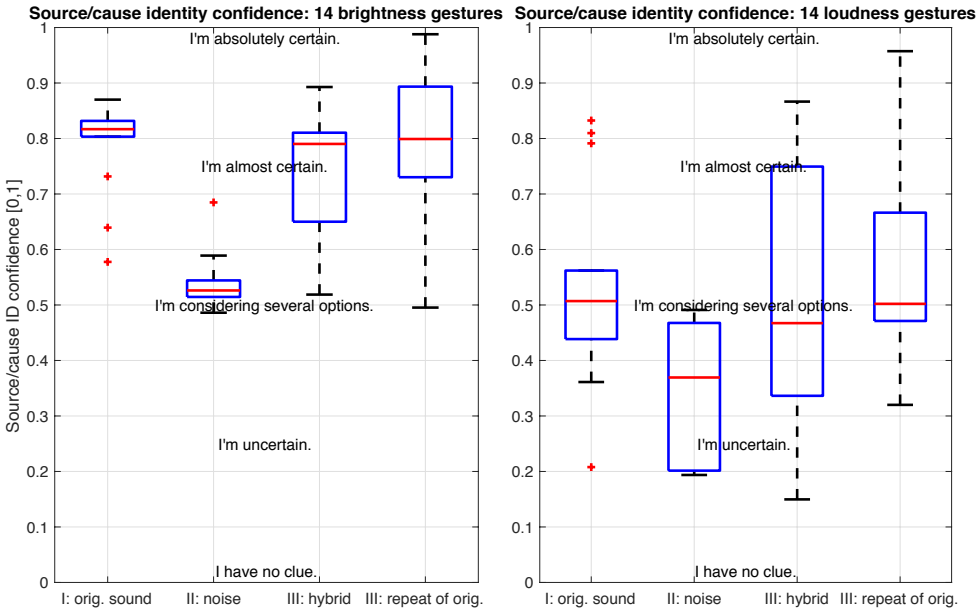


Figure 3: Ratings of confidence on source/cause identity of 14 brightness (left) and 14 loudness gestures (right).

Across all gestures, a wider variation of identification accuracy still became apparent and may relate to a range of factors, which will be studied using partial least-squares regression in time for the conference. A wider set of predictors will be considered, spanning source/cause confidence, timbre descriptors, and variables related to gestural features (e.g., shape, orientation, duration).

Discussion

The perception of time-variant timbral features like sound gestures relates to theories of embodied perception (Godøy, 2006), which presumably could extend to the perception of conventional instruments (e.g., note articulations involving timbral brightness contours). With regard to electroacoustic music, the central role of timbre has motivated the review of established distinctions between parameters (e.g., "the pitch *within* timbre", Smalley, 1994, p. 40), while this broader notion of timbre may also benefit the study of other, more timbre-reliant musical practices (e.g., popular, non-Western).

References

Godøy, R. I. (2006). Gestural-Sonorous Objects: embodied extensions of Schaeffer's conceptual apparatus. *Organised Sound*, *11*(2), 149–157.

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

McDermott, J. H., Lehr, A. J., & Oxenham, A. J. (2008). Is relative pitch specific to pitch? *Psychological Science*, *19*(12), 1263–1271.

Smalley, D. (1994). Defining timbre - Refining timbre. *Contemporary Music Review*, *10*(2), 35–48.

Smalley, D. (1997). Spectromorphology: explaining sound-shapes. *Organised Sound*, *2*(2), 107–126.

Timbre Space Representation of a Subtractive Synthesizer

Cyrus Vahidi[1†], George Fazekas[1], Charalampos Saitis[1] and Alessandro Palladini[2]
[1] Centre for Digital Music, Queen Mary University of London, London, United Kingdom
[2] Music Tribe Research, Manchester, United Kingdom
[†] Corresponding author: c.vahidi@qmul.ac.uk

Introduction

Sound synthesis modules include oscillators, filters and amplifiers, that can be routed to form a subtractive synthesizer (Moog, 1964). Sound synthesizers provide a rich, abstract timbre palette, that is not achievable under the physical constraints of acoustic musical instruments. Synthesis modules are controlled by low-level signal-processing parameters; this parameterization does not allow control of perceptual timbre dimensions and requires expert knowledge to smoothly navigate the space of timbres. Acoustic descriptions of perceptual timbre spaces have been discovered, with multidimensional scaling (MDS) of pairwise dissimilarity ratings, obtained for real and synthetic orchestral musical instrument stimuli (McAdams et al., 2019). Timbre space studies have been conducted jointly between acoustic and synthesized sounds, where the synthesized stimuli has clear definition (Zacharakis et al., 2015). In contrast to instrument sounds, synthesized sounds are not generated by a physical body; familiarity with the object may have less influence on the dissimilarity judgements of listeners (Siedenburg et al., 2016).

Abstractions for interaction with sound synthesis have been advanced with deep neural networks. These generative systems are capable of high-level characterization of timbre as a control structure (Esling et al., 2020). It has been shown that the alignment of generative synthesis spaces with timbre dissimilarity distances, produces sparser latent representation spaces (Esling et al., 2018). We aim to generate new domain knowledge for generative modelling of sound synthesizers.

In this study, we produce a geometrically scaled perceptual timbre space from dissimilarity ratings of subtractive synthesized sounds and correlate the resulting dimensions with a set of acoustic descriptors. We curate a set of 15 sounds, produced by a synthesis model that uses varying source waveforms, frequency modulation (FM) and a lowpass filter with an enveloped cutoff frequency. Pairwise dissimilarity ratings were collected within an online browser-based experiment. We hypothesized that a varied waveform input source and enveloped filter would act as the main vehicles for timbral variation, providing novel acoustic correlates for the perception of synthesized timbres.

Method

Participants

35 participants (age 28.2 mean, 5.8 std, 21-46) for this experiment consisted of domain experts and musically critical listeners. A total of 41 responses were received. Participants who were not included in the final analysis either reported hearing issues or excessively violated the experiment control. Participants were recruited through the Centre for Digital Music and music informatics mailing lists.

Stimuli and Presentation

Figure 1 shows a block diagram of the digital subtractive synthesizer which was implemented in SuperCollider to generate the set of 15 stimuli. All oscillators were fixed at 440 Hz (A4) with an initial phase of 0. FM was applied with a modulator frequency that is an integer multiple of the carrier. Tones were normalized for A-weighted RMS sound level and constrained to 1000ms in duration. The stimuli were composed from either of two oscillators, a pulse wave or a sawtooth. The source was optionally frequency modulated and then passed into a resonant low pass filter. The lowpass filter's cutoff frequency, ω_c, was time-varied by an attack-decay-sustain-release (ADSR) envelope. The master gain, g, was also time-varied with a separate ADSR.

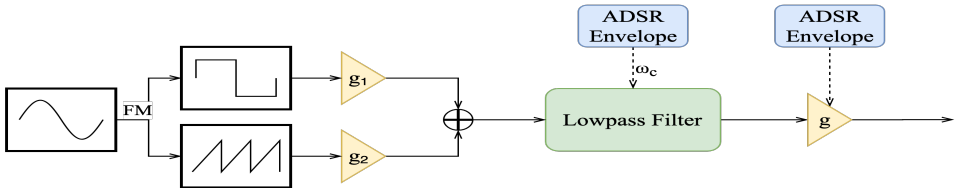


Figure 1: Block diagram of a subtractive synthesizer.

Procedure

The experimental procedure was approved by Queen Mary University of London Ethics of Research Committee (ref QMREC2445).

Conducting a listening experiment in the browser presents challenges, such as varying listening environments. In order to mitigate the effects of insufficient audition quality, we used a headphone screening task (Woods et al., 2017) to ensure greater control over stimuli presentation.

Each participant provided a single rating for each sound pair A-B. The direction of presentation, A-B vs B-A, was selected randomly for each pair and each participant. This resulted in a total of 120 pairwise ratings per participant. Participants were instructed to provide a dissimilarity rating between 0 (identical) and 9 (very dissimilar), via a slider that was discretized in steps of 0.5. As a control, participants were instructed to provide a 0 rating for identical sounds. Participants were also given the opportunity to familiarize with the sound set and task. A pair of sounds could be replayed infinitely before submitting a rating.

Non-metric Multidimensional Scaling

Multidimensional scaling (MDS) is a statistical dimensionality reduction technique that aims to preserve the geometric relationships between data objects (Kruskal, 1964a). The dissimilarity data were analysed by means of MDS, and the dimensions of the timbre space were examined in terms of audio content descriptors that were computed from the acoustic signals. Non-metric multidimensional scaling (Kruskal, 1964b), provided by MATLAB, was applied to the mean dissimilarity matrix across participants.

Feature Selection

Suitable acoustic descriptors were identified from literature on acoustic feature extraction (Peeters et al., 2004, 2011) and extracted using Essentia (Bogdanov et al., 2013). A Hann window and hop length of 2048 and 512 samples, at a sampling rate of 44.1 kHz, were used for the magnitude Short-time Fourier transform (STFT) spectral feature extraction. The mean statistic was computed across time frames. A Spearman correlation matrix was computed for the initial set of acoustic descriptors, discarding features that showed a strong collinearity with others ($r > 0.8$), for example spectral centroid and spectral roll-off[1].

Results

Figure 2 shows the metrics for MDS solutions of varying dimensions. *Stress-1* and R^2 goodness-of-fit metrics were computed between the original dissimilarities and new MDS disparities. There was a significant improvement in fit when moving from 3 to 4 dimensions and greater interpretability of the resulting acoustic correlations. Hence, a four-dimensional MDS solution was used for analysis (*Stress-1 = .047,* $R^2 = .91$*)*.

[1] Collinear features were eliminated based on their relevance in timbre literature and interpretability.

Table 1 shows the Pearson correlation matrix, computed between the final set of acoustic content descriptors, and the 4 perceptual dimensions derived from MDS analysis. *Dimension 1* indicated significant correlations with *spectral complexity* ($p < 0.01$), *spectral flux* ($p < 0.01$), *log-attack time* ($p < 0.05$) and *tristimulus 3* ($p < 0.01$). *Dimension 2* indicated a significant positive correlation with *spectral decrease* ($p < 0.05$). *Dimension 3* demonstrated significant positive correlations with *spectral kurtosis* ($p < 0.05$) and

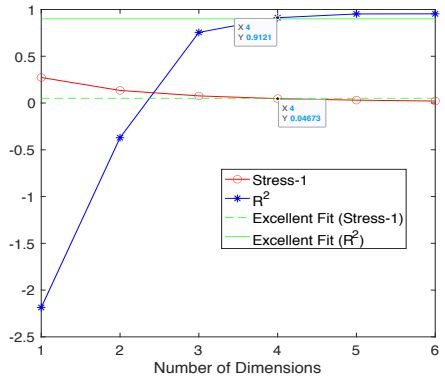


Figure 2: Stress-1 and R^2 metrics

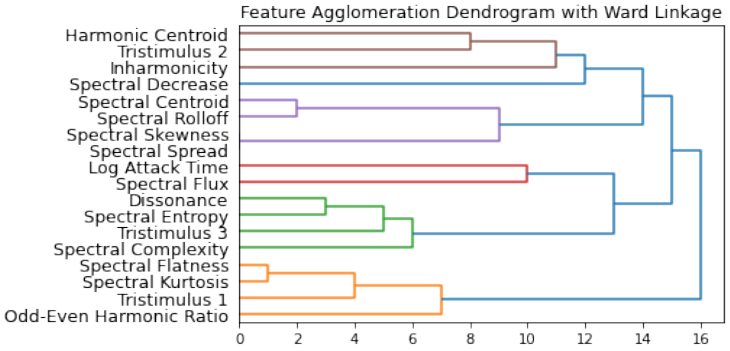


Figure 3: Feature Agglomeration Dendrogram

odd-even harmonic ratio ($p < 0.05$), and negative correlations with *tristimulus 2* ($p < 0.05$) and *spectral centroid* ($p < 0.05$). *Dimension 4* demonstrated a significant negative correlation with *spectral centroid* ($p < 0.01$).

	MDS Dimension			
Acoustic Descriptor	**1**	**2**	**3**	**4**
Spectral Complexity	***.75****	*.39*	*-.23*	*.10*
Spectral Flux	***.68****	*-.24*	*-.08*	*.41*
Log Attack Time	***.60****	*-.39*	*-.26*	*-.37*
Tristimulus 3	***.75****	*.31*	*-.23*	*-.37*
Spectral Decrease	*-.23*	***.58****	*.24*	*-.43*
Tristimulus 2	*-.44*	*.14*	***-.51****	*.30*
Spectral Kurtosis	*-.01*	*-.43*	***.61****	*.35*
Odd-Even Ratio	*-.34*	*-.30*	***.56****	*.14*
Spectral Centroid	*.05*	*.28*	***-.52****	***-.65****

*Table 1: Acoustic Feature vs Perceptual Dimension Pearson Correlations. * - $p < 0.05$, ** - $p < 0.01$*

Discussion

The results indicate some anticipated acoustical correlates in the context of the literature, as well as additional dimensions that can be explained by the design of the synthesis model.

The first perceptual dimension appears to be explained by the content of spectral peaks and their change over time. Note that spectral flux and log-attack are entangled. A plausible explanation for this is the design of the sound set. The attack of several sounds is implicit in the rising cutoff of the envelope filter, that moves from 0 Hz to N Hz over a time determined by the filter envelope's attack time. Hence, we define *dimension 1* to relate to spectrotemporal variation and harmonic peaks.

Dimension 2 can be interpreted to relate to the varying cutoff of the lowpass filter, as the spectral decrease indicates the ratio between low and high frequency components. *Dimension 3's* acoustical correlates can be interpreted to be related to the (a) strength of the FM modulation index, which when increased reduces flatness around the centroid, and (b) the input waveforms. This accounts for spectral detail and distribution of harmonic partials. *Dimension 4's* significant correlation with the spectral centroid is an anticipated result,

given the ubiquity of the spectral centroid in previous perceptual timbre space studies, and the effects of harmonic FM and resonant filters on shaping timbral brightness.

Conclusions and Future Work

In the present study, additional dimensions explaining perception of synthesized timbres were observed, relative to existing timbre studies on acoustic and synthesized sounds. A study with a larger, more diverse sound set is necessary to disentangle log-attack time and spectral flux. The results of this study shall contribute to future large-scale data generation and perceptual domain knowledge for generative modelling of sound synthesizer timbre spaces. With these results, we can move further towards perceptual representation of synthesized timbres.

Acknowledgements

We acknowledge Benjamin Hayes for his development of the framework used for online dissimilarity data collection. We thank all who participated in the listening experiment. The author is a research student at the UKRI CDT in AI and Music, supported jointly by the UKRI [grant number EP/S022694/1] and Music Tribe.

References

Bogdanov, D., Wack, N., Gómez Gutiérrez, E., Gulati, S., Boyer, H., Mayor, O., & Serra, X. (2013). Essentia: An audio analysis library for music information retrieval. In Britto A, Gouyon F, Dixon S, (eds), *14th Conference of the International Society for Music Information Retrieval (ISMIR*) (pp. 493-8). Curitiba, Brazil.

Esling, P., Chemla-Romeu-Santos, A., & Bitton, A. (2018, September). Generative timbre spaces with variational audio synthesis. In Matthew Davies, Aníbal Ferreira, Guilherme Campos, Nuno Fonseca (eds), *Proceedings of the International Conference on Digital Audio Effects (DAFx)* (pp. 175–181). Aveiro, Portugal.

Esling, P., Masuda, N., Bardet, A., Despres, R., & Chemla-Romeu-Santos, A. (2020). Flow Synthesizer: Universal Audio Synthesizer Control with Normalizing Flows. *Applied Sciences*, *10*(1), 302.

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

Kruskal, J. B. (1964b). Nonmetric multidimensional scaling: a numerical method. *Psychometrika*, *29*(2), 115-129.

McAdams S. (2019). The Perceptual Representation of Timbre. In: Siedenburg K., Saitis C., McAdams S., Popper A., Fay R. (eds), *Timbre: Acoustics, Perception, and Cognition* (pp. 23-57). Springer Handbook of Auditory Research, vol 69. Springer, Cham.

Moog, R. (1965) 'Voltage-controlled electronic music modules', *Journal of the Audio Engineering Society*, *13*, 200–206.

Peeters, G. (2004). A large set of audio features for sound description (similarity and classification) in the CUIDADO project. *CUIDADO IST Project Report*, *54*(0), 1-25.

Peeters, G., Giordano, B. L., Susini, P., Misdariis, N., & McAdams, S. (2011). The timbre toolbox: Extracting audio descriptors from musical signals. *The Journal of the Acoustical Society of America*, *130*(5), 2902-2916.

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

Woods, K. J., Siegel, M. H., Traer, J., & McDermott, J. H. (2017). Headphone screening to facilitate web-based auditory experiments. *Attention, Perception, & Psychophysics*, *79*(7), 2064-2072.

Zacharakis, A., Pastiadis, K., & Reiss, J. D. (2015). An interlanguage unification of musical timbre: Bridging semantic, perceptual, and acoustic dimensions. *Music Perception*, *32*(4), 394-412.

Timbral Threads: Compositional Strategies for Achieving Timbral Blend in Mixed Electroacoustic Music

Matt Collins
Faculty of Music, University of Oxford, Oxford, United Kingdom
matt.collins@music.ox.ac.uk

Introduction

Within the discursive paradigms of electroacoustic music, an issue that many composers have confronted is the difficulty of producing works that feature both acoustic and electronic forces without some sense that there is an aesthetic disconnect between them. Recent literature in this area (e.g. Croft, 2007; McLaughlin, 2012) has often attempted to tackle this sense of disconnect by appealing to concerns relating to retaining or avoiding a sense of liveness and/or instrumentality, while the role that timbre may play in this issue has not been the subject of much discussion to date.

To begin to investigate the role that timbre may play in this sense of disconnect, this paper examines, specifically from the perspective of an unaided listener and reader, notions of timbral blend between the acoustic and electronic forces in two mixed electroacoustic compositions: 'Salt Canyons' from *The Inner Universe* (1983) by Priscilla McLean; and *RAS* (2000, rev. 2011) by Natasha Barrett. These works have been chosen because of what I believe to be the successful blending of their acoustic and electronic forces. This discussion will primarily take the form of a set of timbral analyses that establish common compositional strategies for achieving timbral blend. These analyses are then contextualised and situated in relation to the broader discourses on timbre and timbre analysis.

Method

For the purposes of this paper, 'timbral blend' may be defined as the capacity of sounds from two or more physical sources to coalesce together in such a way so that the listener is not likely to determine that there is an aesthetic disconnect between them. This paper also introduces the idea of 'timbral threads'. A thread of timbre is defined as a sound or set of sounds, which can either occur once, repeat a number of times, or sustain continuously for a given period, that could reasonably be perceptually categorised as being produced by one source. Importantly, the sound(s) in a single thread do not have to be only produced by one physical source – they need only be perceived by the listener as if they are originating from the same physical or imagined source, even if the nature of that source cannot be identified.

The two compositions are analysed by defining and describing the timbral threads that can be delineated within them. The language used to describe the sounds within these threads is as elementary as possible and invokes metaphor where it is necessary to do so. By examining these threads and exploring how they develop and interlink with each other, the compositional strategies used to achieve timbral blend between the acoustic and electronic forces within each work individually as well as across both works are ascertained. Each analysis is also preceded by an exposition of any relevant ideas, philosophies, aesthetic ideals, or compositional preferences of the composers concerned, so that the analyses themselves may also explore whether any of these have an effect on the compositional strategies used for the purposes of timbral blend.

Results

The poster connected to this paper depicts graphical representations of the timbral threads within each work and how they link together. These threads are briefly summarised in written form below.

'**Salt Canyons**', by American composer Priscilla McLean (b. 1942), lasts ca. 7'52'' and forms part of a group of compositions titled *The Inner Universe* (1983) for solo piano or piano and tape (McLean, 1983, pp. i, 48–49). McLean (1977, pp. 205–207) contrasts two different approaches to timbres used in

electroacoustic music: using abstract, non-referential sounds; and environmental/instrumental sounds. She problematises the exclusive use of these kinds of sound, before proposing a third approach, which she terms the use of 'imago-abstract' sound. This type of sound has a timbre that is not entirely abstract/non-referential, but is also not a direct reference to an environmental sound. This concept could help to suggest a contributor to this sense of a lack of acoustic-electronic timbral blend: that it is actually caused by a lack of integration between referential and non-referential sounds, and that a solution to this may involve using 'imago-abstract' sounds to blur the distinction between these two forces.

Thread 'A': Contains a 'choir' of piano overtones on the tape. This drone chord is then timbrally blended with a live piano sound of a mug being glided up piano strings. The blend is further created by sustaining the acoustic-sounding tape chord in a way that could not be done live, and by timbrally extending upon it using subtle electronic manipulations and the introduction of complimentary environmental sounds.

Thread 'B': Features a tape sound that sounds like a car braking heavily, in combination with harsh overtone pitches produced on the live piano. Here, both sounds used in isolation would likely be assumed to be of environmental and instrumental origin respectively – however, the timbral similarities between the sounds lead to doubt about their respective origins when they are used together.

Threads 'C' and 'D': Both of these comprise tape sounds that are timbrally identical to their live piano counterparts – the former featuring samples of low-pitched glissandi on the piano strings; the latter featuring samples of rapidly repeating prepared and unprepared piano notes. Because the tape sounds are heard at the same time as their live versions, the acoustic/electronic distinction is entirely removed.

Thread 'E': Contains short rapid runs of piano note samples on tape, electronically altered to sound like muted versions of some prepared pitches played on the live piano earlier in the work. As both the piano and tape samples sound more like a mandolin being plucked, they create ambiguity about which, if any, of these sounds are live and which could be, for example, pre-recorded samples of a mandolin on tape.

RAS (2000, rev. 2011), by British composer Natasha Barrett, lasts ca. 9'32'' and is written for 'percussion quartet, electroacoustic sound and live electronics' (Barrett, n.d.). Barrett (1997, pp. 30–31, 34–36) writes about what she terms as a 'sound-world realism' frame of reference within acousmatic music. She argues that this realism frame of reference exists on a continuum. A sense of 'maximum' realism exists when a 'real' sound is in a 'familiar' virtual space. Conversely, a state of 'minimum' realism is achieved when real and 'synthetic' sounds are used together in a familiar space. Finally, in between these extremes are real and synthetic sounds in 'unfamiliar' spaces, and synthetic sounds in familiar spaces. These ideas could be helpful in suggesting a contributor to a lack of timbral blend in mixed electroacoustic works – perhaps sounds used together in spaces which perceptually increase a sense of realism are likely to also give more of a sense of timbral blend.

Thread 'A': Mostly features isolated and short acoustic sounds (e.g. timpani or rototom hits), which serve as timbral springboards for what are usually much longer and denser expanses of (mostly) electronic sound. These springboards, due to their clear gestural development and frequent repetition, provide a sense of timbral blend despite the considerable timbral shifts that occur within them.

Thread 'B': Begins with sparse acoustic sounds from the matrix that sound metal-like. These sounds timbrally evolve over time, with an increasing use of both live electronic manipulation and similar-sounding pre-recorded sounds, but this evolution is gradual and subtle enough to ensure that how much of what we are hearing is acoustic, electronic, or a combination thereof is ambiguous.

Thread 'C': Contains an electronic pad-like sound that persists throughout the work, beginning initially as a timbrally extended version of the timpani and rototom glissandi. It is not clear, however, at what point the acoustic glissandi sonically end and the electronic sound begins due to their timbral similarities.

Threads 'D', 'E', 'F', 'G', 'H', and 'I': All of these threads feature pre-recorded environmental sounds that at some point are electronically manipulated. The origins of these sounds are usually ambiguous, but

more importantly, the electronic qualities of these sounds serve to blend very well with the live acoustic sounds they often sit alongside with (mostly sounds from thread 'A').

A large majority of the live acoustic sounds in both of these compositions either have origins that are unfamiliar to the majority of the listeners but are still likely to be identifiable, are ambiguous enough to be unrecognisable even when heard in isolation, or would be recognisable if not for the fact that they are disguised in some way within the prevailing texture. This potentially suggests that a lack of source recognition or familiarity enables a variety of timbral possibilities to become available. If the acoustic sound is completely unrecognisable, then the sound has a much larger potential to be used alongside electronic sounds of any kind and retain a sense of timbral blend.

Common strategies used to influence the nature of most of the electronic sounds in the works include either the use of pre-recorded or live delayed acoustic sounds that are subject to either no or minimal electronic manipulation, or pre-recorded environmental sounds that are subject to different levels of perceptible electronic manipulation depending on the musical context. Passages that feature environmental sounds with little or no manipulation, so that their sources are recognisable, do not feature any familiar or recognisable acoustic sounds at all, and are often reached through very gradual timbral evolutions from non-environmental soundworlds. Passages which feature environmental sounds with a very noticeable amount of manipulation tend to transform the sounds in such a way so that their origins become ambiguous, and often start to sound more like the acoustic sounds they sit alongside. Importantly, there are very few sounds across both works that are obviously synthesised. These strategies suggest that, similar to the acoustic sounds, the less recognisable or familiar an electronic sound is, the more potential that sound has to blend with its acoustic counterpart.

Turning to the composers' aesthetic philosophies, it is significant that Barrett's 'sound-world realism' concept plays an important role in both compositions. The vast majority of the sounds in both works have, I would contend, a 'real' origin. The spaces which these sounds are situated in are initially unfamiliar, as the soundworlds throughout are not something we would find in the real world. However, these spaces remain the kinds of spaces we experience throughout the works; thus, by Barrett's definition, the spaces become non-real-world spaces over time. Consequently, the combination of real sounds and spaces that become familiar over time mean that the perception of realism increases over the course of both works. This lens through which to look at these works, therefore, may give a real insight into how mixed electroacoustic compositions construct a sense of successful timbral blend.

In relation to McLean's idea of 'imago-abstract' sound however, I would argue that this kind of sound does not seem to be synonymous with the idea of timbrally blending acoustic and electronic sounds. Both works are successful at achieving the latter despite the prolific use of referential sounds. Particularly in 'Salt Canyons', while there are several sounds in the work that arguably do meet the criteria of being 'imago-abstract', all of these sounds are still first heard in a referential context, thus endangering their ambiguity of origin when later heard in their 'imago-abstract' form.

In conclusion, the compositional strategies used to achieve timbral blend in these works are multifarious and varied. Nevertheless, there are a small number of strategies which can be assimilated into one unifying strategy that can be found across both works. Sounds, regardless of origin, whose sources are less easy to identify, or less familiar to the listener, are more likely to blend and integrate with each other. Equally however, it must also be noted that the means used to make sounds less familiar are manifold in these works, and that any strategies unique to individual works are not any less important or useful.

Discussion

These results and the methods used to obtain them will require further investigation, as the issues discussed in this paper engage not just with compositional questions, but questions across the various discourses of timbre. Furthermore, a longer study on the role of timbre in mixed electroacoustic music would need to examine a much wider and more representative range of works, as a larger sample size

would allow for a greater certainty over which, if any, compositional strategies used for ensuring timbral blend are common and shared. A more comprehensive study should also attempt to account for any musical factors, such as harmony, gesture, and structure, that might act in dialogue with timbre during any given composition. It might also seek to investigate whether timbre may play a role in the practical challenges performers often face when playing mixed electroacoustic works.

An examination of conceptual metaphor theory as it can be applied to the semantics of timbre, by Wallmark and Kendall (2018), was the primary influence on the methodology used in this paper. The theory proposes that language acts as a system for human embodiment. Highlighting two metaphors that are very commonly used – 'sound is light' and 'sound is texture' – the authors link these into their claim that there is concrete evidence for '"weak synesthesia" between auditory, visual and tactile modalities in non-synesthetic individuals'. They argue that one possible explanation for this, based on a theory by Lawrence Marks (1975), is that timbre-based synesthetic metaphors provide a shortcut by which we can meaningfully distinguish cognitive stimuli for ourselves and to others. This theory, therefore, suggests that the use of certain metaphors to describe timbre does not depend entirely on individual subjectivity, but rather reflects a deeply embodied, culturally situated intersubjective form of communication.

Selecting the methodology most appropriate for identifying different compositional approaches to timbre is a perilous task. I would argue that a phenomenologically grounded approach to the analysis of timbre should be curated on the basis of how listeners, across different cultures, directly experience timbre both in the moment(s) of listening, and how they reflect upon the phenomenon and its effects on the music they have previously listened to. The conceptual metaphor theory described above, as applied to our experience of timbre, appears to closely build upon this phenomenological notion of timbre, because it theoretically allows for the possibility of a kind of, albeit culturally situated, timbral analysis that uses descriptive and metaphorical language to convey the timbral devices employed in a musical work to the reader in a way that potentially allows said devices to be meaningfully understood.

Acknowledgments

This research is supported by the UK Arts and Humanities Research Council.

References

Barrett, N. (n.d.). *Instruments and Live Electronics*. Natasha Barrett http://www.natashabarrett.org/live_electronics.html

Barrett, N. (1997). *Structuring Processes in Electroacoustic Composition* (Doctoral thesis), City University: London. http://openaccess.city.ac.uk/7468/

Croft, J. (2007). Theses on liveness. *Organised Sound*, *12* (1), 59–66.

Marks, L. E. (1975). On Colored-Hearing Synesthesia: Cross-Model Translations of Sensory Dimensions. *Psychological Bulletin*, *82* (3), 303–331.

McLaughlin, S. (2012). If a tree falls in an empty forest: Problematization of liveness in mixed-music performance. *Journal of Music, Technology and Education*, *5* (1), 17–27.

McLean, P. (1977). Fire and Ice: A Query. *Perspectives of New Music*, *16* (1), 205–211. https://doi.org/10.2307/832858

McLean, P. (1983). *The Inner Universe*. MLC Publications.

Wallmark, Z., & Kendall, R. A. (2018). Describing Sound: The cognitive linguistics of timbre. In A. Rehding & E. I. Dolan (eds), *The Oxford handbook of timbre*. New York: Oxford University Press. https://doi.org/10.1093/oxfordhb/9780190637224.013.14

Timbre Trait Analysis: The Semantics of Instrumentation

Lindsey Reymore
Schulich School of Music, McGill University, Montréal, Quebec, Canada
reymore.1@osu.edu

Introduction

The cognitive linguistics of timbre—that is, the study of the interaction among language, thought, and perception of timbre—has recently emerged as a promising sub-field in timbre studies, and research supports a strong link between timbre perception and semantics (Saitis & Weinzierl, 2019). The current proceedings article provides an overview of some of the work reported in my dissertation (Reymore, 2020), which addresses how study of the cognitive linguistics of timbre can inform music theoretical discourse and analysis. I review the studies of timbre semantics that were used to generate a cognitive linguistic model of musical instrument timbre *qualia* and describe how this model can be used in timbre and orchestration analysis through a method I call Timbre Trait Analysis.

Method

Building a cognitive linguistic model of musical instrument timbre qualia

The model-building process began through interviews with 23 musicians who were asked to imagine and describe the sounds of 20 musical instruments. The use of imagined stimuli was motivated by the goal of characterizing "prototypical" sounds of each instrument rather than any specific recorded instantiation. A pile sort analysis (deMunck, 2009) of the interviews yielded 77 categories of timbre descriptors.

These 77 categories were used in a subsequent online rating task (Rating Task #1) in which 460 musician participants were asked to rate imagined instrument sounds using 7-point Likert scales. The resulting data were subjected to Principle Components Analysis (PCA); creating the final model from the PCA involved further input from musicians (for more detail, see Reymore & Huron, 2020). The final 20-dimensional model is reported in Table 1 in the Results section.

Generating Timbre Trait Profiles

Next, the 20-dimensional model was used in a second online rating task (Rating Task #2) of imagined timbres with the goal of providing characterizations, or "Timbre Trait Profiles" of 34 common Western large ensemble musical instruments. In this study, 243 musicians rated 11 of the 34 instruments on the 20 dimensions. The Timbre Trait Profiles comprise the average means and standard deviations of the ratings for each instrument. These data can be visualized through radar plots, as illustrated in Figures 1–3.

Illustrating the semantics of instrumentation

The Timbre Trait Profiles were then used as the framework for a computational program which generates semantic orchestration plots given a musical piece as input. This "Timbre Trait Analysis" provides information on how the semantic dimensions of timbre evolve throughout a piece of notated music. At the moment, only the information from the prototypical profiles is used; however, the eventual goal is to incorporate refinements to each of the profiles with respect to dynamics, register, and other factors (initial work is described below). For each beat of the piece, the program calculates an average value for the model's 20 semantic dimensions. The total value of a given dimension is divided by how many instruments are playing at that moment so that semantic values are not skewed by the number of sound source types.

Mapping the effects of register and dynamics

The Timbre Trait Profiles provide characterizations of the prototypical sounds of musical instruments. As such, they have already proved useful in musical analysis (see "Illustrating the semantics of instrumentation" and Results sections); however, more nuance may be achieved in future analyses by

incorporating information on timbral variation within instruments relative to factors such as dynamics, articulation, range, register, or duration. These variations are huge in scope and will require further research to map comprehensively. As an initial step, Rating Task #3 mapped timbre *qualia* across pitch range and dynamics on the oboe and the French horn. Participants (47 Ohio State music majors) used the 20-dimensional model to rate 36 two-second recordings of single tones played by either the oboe or the horn in four registers, where each register was represented by three pitches, and each pitch was recorded at three dynamic levels (*pp*, *mm*, and *ff*).

Results

Table 1: 20-dimensional cognitive linguistic model of instrument timbre qualia*, Rating Task #1. The left column lists all terms belonging to a given dimension, while the right column contains shorthand labels.*

Dimension descriptors	Shorthand label
rumbling, booming, low, deep, thick, fat, heavy	*rumbling/low*
soft, smooth, singing, voice-like, sweet, gentle, calm	*soft/singing*
direct, projecting, loud, aggressive, commanding, assertive, powerful	*direct/loud*
nasal, reedy, buzzy, pinched, constrained	*nasal/reedy*
shrill, harsh, noisy	*shrill/noisy*
percussive	*percussive*
pure, clear, precise, clean	*pure/clear*
brassy, metallic	*brassy/metallic*
raspy, guttural, grainy, gravelly	*raspy/grainy*
ringing, long decay	*ringing/long decay*
sparkling, shimmering, brilliant, bright	*sparkling/brilliant*
airy, breathy	*airy/breathy*
resonant, vibrant	*resonant/vibrant*
hollow	*hollow*
woody	*woody*
muted, veiled	*muted/veiled*
sustained, even	*sustained/even*
open	*open*
focused, compact	*focused/compact*
watery, fluid	*watery/fluid*

Timbre Trait Profiles were generated for 34 large ensemble instruments. Figs 1–2 below provide example illustrations of the Profiles. Average ratings are indicated on each of the dimensions around the circle via the circles' radii, which are connected to create timbral "thumbprints" for each instrument.

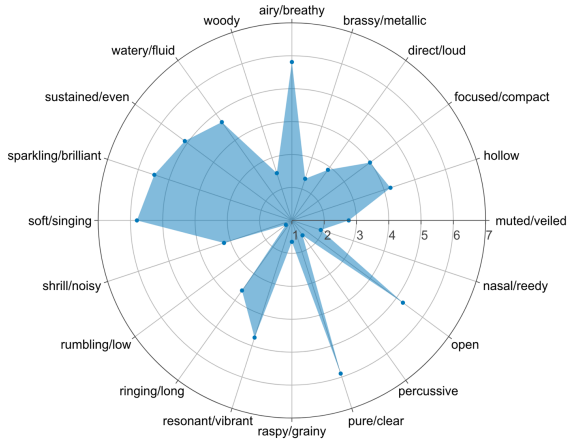


*Fig 1: The **FLUTE** was rated highly on dimensions including pure/clear, airy/breathy, and soft/singing.*

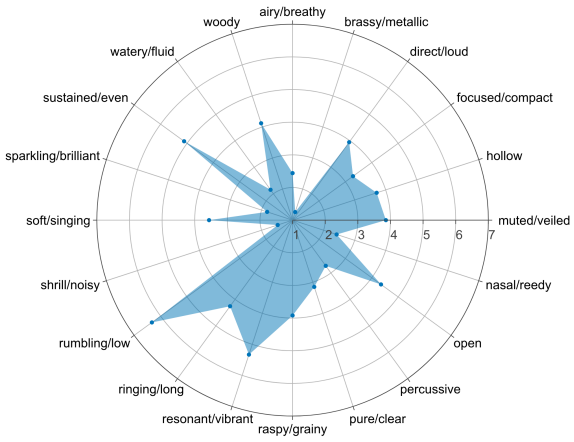


*Fig 2: The **DOUBLE BASS** was rated highly on dimensions including rumbling/low, resonant/vibrant, and sustained/even.*

In Rating Task #3, in which participants rated recorded tones varied in pitch and dynamic, dimensions implicating the manipulated factors (e.g. *rumbling/low*, *shrill/noisy*, *direct/loud*) tended to demonstrate similar, roughly linear trends in both oboe and French horn. Some dimensions varied greatly across dynamics and/or register for one instrument but not for the other. For example, the total range (max-min, on a 7-point scale) of average values on oboe stimuli for *airy/breathy* was 3.35, while the range for the horn was only 1.46. On the other hand, horn varied much more on *raspy/grainy* (range = 4.07) than the oboe (1.37). Some dimensions, notably *soft/singing*, yielded arch-shaped graphs, where the middle ranges of the instruments were on average rated more highly than the extreme registers. Complete results from this study can be found in Reymore (2020).

Model consistency & reliability

Rating Tasks #1–3 provided the opportunity to test the consistency and reliability of the 20-dimensional model for representing timbre semantics. To judge whether the 20-dimensional model is a reliable reduction of the initial 77 categories, data from Rating Task #1 was used to predict data from Rating Task #2. The aggregate correlation between predicted and actual data was $r = .96$, indicating a high degree of reliability. Next, Timbre Trait Profiles generated from Rating Task #2 for the oboe and French horn were compared to the data collected for recorded sounds on these instruments in Rating Task #3. Correlations for both instruments were positive, with a very strong correlation between datasets for the oboe ($r = .92$) and a moderate correlation between those of the horn ($r = .42$). I conjecture that the different correlation strengths may relate to differences in timbral flexibility between the oboe and horn with respect to the variables manipulated in task #3 (dynamics and register). That is, because the horn can produce a wider range of timbres with respect to dynamic and pitch variability, the collection of recorded tones in Rating Task #3 for the oboe was much closer to the prototypical sound of the oboe imagined by participants in Rating Task #2 than were the recorded horn tones to the prototypical horn sound.

Timbre Trait Profiles in action

The music analytical method described above in the section "Illustrating the semantics of instrumentation" was applied to the first movement of Mahler's Symphony No. 1. This process resulted in a series of graphs mapping each of the 20 semantic dimensions over time in the piece. A sample graph for the *sparkling/brilliant* dimension is produced in Fig 3 below.

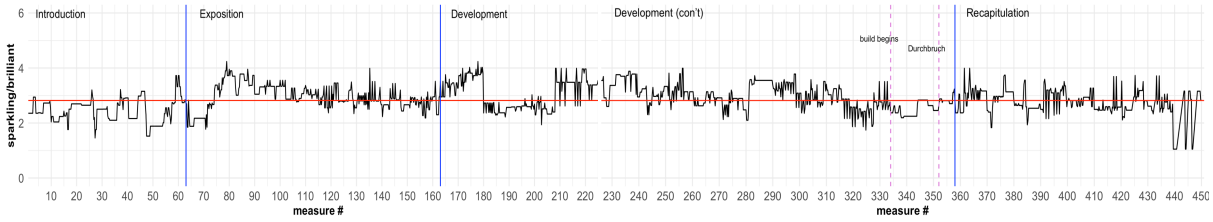


Fig 3: Semantic graph of the sparkling/brilliant *dimension of the instrumentation in the first movement of Mahler's first symphony. Formal boundaries are given in solid blue, while the horizontal red line represents the average value for sparkling/brilliant over the course of the piece.*

For example, we can observe that the instrumentation in the introduction is relatively low on the *sparkling/brilliant* dimension as compared to other formal sections and to the average value for this dimension across the pieces. Computationally, the data suggest sets of *qualia* that are especially relevant for the instrumentation of particular formal sections of the piece, shown in Table 2 below. In this movement, the semantics of instrumentation considered by formal section seems to tell a story that is consistent with the narrative of the musical work.

Table 2: Prominent qualia *of instrumentation by formal section, Mahler, Symphony No.1, first movement.*

Introduction	**Exposition**	**Development**	**Recapitulation**
rumbling/low	*soft/singing*	*sparkling/brilliant*	*shrill/noisy*
raspy/grainy	*watery/fluid*	*brassy/metallic*	*direct/loud*
muted/veiled	*pure/clear*		*nasal/reedy*
hollow	*sparkling/brilliant*		*airy/breathy*
direct/loud	*focused/compact*		*brassy/metallic*

Discussion

The semantic graphs provided in this analytical approach are not intended as ends in themselves, but rather as navigable summaries of a piece that can be used for both close and distant readings. They are intended as tools to help with musical analysis. The metaphor INSTRUMENTS ARE VOICES has been identified as critical for timbre semantics (Wallmark, 2014), and it is an important metaphor for the approach described here to applying Timbre Trait Profiles in musical analysis. The profiles tell us about the characters that are speaking (musically) at any given moment. In much music, considering a musical idea divorced from the voice that speaks it may be like analyzing the line of a play without taking into account information about the character who speaks it. Timbre Trait Profiles may help us understand how the "conversation" goes in any given piece; we can consider the ways in which composers tend to combine different types of instrumental characters.

It should be noted that statements about the timbre *qualia* in the movement are not made about musical perception. The point is not, for example, to claim that this or that moment as a whole is perceived as especially *sparkling/brilliant*. Rather, it is to note that the composer used instruments with conventionally *sparkling/brilliant* traits in a given passage—that is, the collection of voices comes from a combination of individual instruments that tend on average to be rated relatively highly on *sparkling/brilliant*. Furthermore, because this analytical approach is rooted in cognitive representations of timbre, it is not intended to comment on timbral blends or other perceptual effects, nor does it account for factors such as pitch or intensity. In the future, I plan to incorporate data regarding timbral variability in range and dynamic, such as that gathered in Rating Task #3, providing a more refined analysis. Following additional research, the approach to musical analysis with the Timbre Trait Profiles can be modified to reflect updated understanding of the semantics of instrumental blend, registral/dynamic variability, and more.

The vocabulary and profiles of the timbre *qualia* model confer the advantage of enabling the comparative discussion of diverse instruments using the same set of 20 measures. The Timbre Trait Profiles allow us to use consistent language to point to specific ways in which instrumental characters are considered to be the same or different: working from the consistent vocabulary of the timbre *qualia* model has advantages for clarity of communication that may prove beneficial in music theoretical analysis.

References

De Munck, V. (2009). *Research Design and Methods for Studying Cultures*. Plymouth, UK: AltaMira Press.

Reymore, L. & Huron, D. (2020). Using Auditory Imagery Tasks to Map the Cognitive Linguistics Dimensions of Musical Instrument Timbre Qualia. *Psychomusicology*.

Reymore, L. (2020). Empirical approaches to timbre semantics as a foundation for musical analysis. (Doctoral dissertation), The Ohio State University, Ohio.

Saitis, C. & Weinzierl, S. (2019). The Semantics of Timbre. In Siedenburg, K., Saitis, C., McAdams, S., Popper, A.N., Fay, R.R. (eds.), *Timbre: Acoustics, Perception and Cognition* (pp.119–149). Springer Handbook of Auditory Research, vol 69. Springer, Cham.

Wallmark, Z. (2014). *Appraising timbre: embodiment and affect at the threshold of music and noise*. (Doctoral dissertation), University of California, Los Angeles.

Verbal Description of Musical Brightness

Christos Drouzas[1†] and Charalampos Saitis[2]
[1] Audio Communications Group, Technical University of Berlin, Berlin, Germany
[2] Centre for Digital Music, Queen Mary University of London, London, UK
[†] Corresponding author: drouzaschristos@gmail.com

Introduction

Amongst the most common descriptive expressions of timbre used by musicians, music engineers, audio researchers as well as everyday listeners are words related to the notion of brightness (e.g., bright, dark, dull, brilliant, shining). From a psychoacoustic perspective, brightness ratings of instrumental timbres as well as music excerpts systematically correlate with the centre of gravity of the spectral envelope and thus brightness as a semantic descriptor of musical sound has come to denote a prevalence of high-frequency over low-frequency energy. However, relatively little is known about the higher-level cognitive processes underpinning musical brightness ratings. Psycholinguistic investigations of verbal descriptions of timbre suggest a more complex, polysemic picture (Saitis & Weinzierl 2019) that warrants further research. To better understand how musical brightness is conceptualised by listeners, here we analysed free verbal descriptions collected along brightness ratings of short music snippets (involving 69 listeners) and brightness ratings of orchestral instrument notes (involving 68 listeners). Such knowledge can help delineate the intrinsic structure of brightness as a perceptual attribute of musical sounds, and has broad implications and applications in orchestration, audio engineering, and music psychology.

Method

Corpus 1: Sixty-nine musically naive listeners (average age = 29.8 yrs, SD = 7.2 yrs, range = 17–51 yrs) provided written responses to the question "What is a bright sound for you?" following ratings of 30 sec long music snippets on perceived brightness. Stimuli were taken from diverse music genres and were also rated on complexity and likeness. Listeners were mainly of Greek, German, and Turkish background. Sixty-four of them responded in English and five in German.

Corpus 2: Sixty-eight musically experienced listeners were recruited from audio technology and musicology programmes in Berlin and Vienna (average age = 30.6 years; SD = 9.3 years; range = 18–66 years). They were German native speakers or spoke German fluently. After the completion of two experiments involving pairwise dissimilarity and direct ratings of brightness of orchestral instrument sounds (all had a fundamental frequency of 311 Hz and a duration of 500 ms) listeners provided written responses (in German) to the question "How and according to which criteria did you compare the sounds in terms of their brightness?" (Original in German: "Wie und anhand welcher Kriterien haben Sie die Klänge bzgl. ihrer Helligkeit verglichen?")

German responses in both corpora were translated into English by the two authors who each speak both languages fluently. Verbalizations were analysed on the basis of semantic proximities in order to identify emerging concepts (thereafter denoted verbal units) that could be coded under broader cognitive categories (see Saitis et al., 2017; 2019a, for a similar psycholinguistic analysis of violin quality descriptions). For example, the phrase "giving me uplifting, happy vibes" contained two verbal units, namely "uplifting" and "happy," whereas the phrase "greater proportion of higher frequencies" constituted a single unit.

Results

In total, 162 verbal units were extracted from the responses in Corpus 1 (2.3 units per respondent on average) and 160 units in Corpus 2 (2.4 units per respondent on average) and were classified in 12 distinct semantic categories (Tables 1 and 2). These appeared to span four central themes: *acoustics* (descriptions of spectral, temporal, and loudness characteristics using acoustical terminology); *affect* (judgments of

aesthetic and emotional value), *musical structure* (references to chords, intervals, melodic lines, and instrumentation; e.g., "major chords are brighter"), and *crossmodal correspondence* (descriptions referencing other sensory modalities).

The musical structure theme emerged primarily in corpus 1, where stimuli comprised multi-instrumental musical sounds, with a small number of listeners in corpus 2 (isolated instrument notes only) also citing instrument type/family as a criterion for determining brightness ratings. Corpus 1 further revealed strong relationships between perceived brightness and valence (e.g., happy, cheerful) or arousal (e.g., uplifiting, energetic) emotions in music perception, which warrants further investigation. Verbal units were assigned to the emotional categories of Valence and Arousal consulting (Kolias et al. 2019).

Verbal descriptions of brightness for isolated instrument notes (Corpus 2) predominantly referred to sensory cues, largely through acoustical terminology but also by employing crossmodal metaphors. We classified the latter into three types: words related to the concept of clarity, with dull used sometimes as semantically opposite; direct synonyms of brightness in the visual domain, such as light and brilliance, with dark but also dull used as antonyms; and other visual and nonvisual descriptions that appeared more idiosyncratically, such as rough, soft, sharp, deep, metallic, and cutting, among others. For a more detailed overview see Table 2. Interestingly, a little more than 16% of Corpus 2 related brightness to the attack portion of instrumental sounds (cf. Saitis et al., 2019b).

Table 1: Distribution of categories within and across corpora
(N=total verbal units; parentheses report proportion over N)

Categories	***Corpus 1 (N=162)***	***Corpus 2 (N=160)***	***Categories***	***Corpus 1 (N=162)***	***Corpus 2 (N=160)***
Spectral char.	*14 (8.6)*	*58 (36.3)*	*Harmony*	*15 (9.3)*	*-*
Temporal char.	*-*	*26 (16.3)*	*Instrument/ation*	*9 (5.6)*	*7 (4.4)*
Loudness	*4 (2.5)*	*3 (1.9)*	*Rhythm/melody*	*10 (6.2)*	*1 (0.6)*
Valence	*65 (40.1)*	*4 (2.5)*	*Clarity*	*15 (9.3)*	*13 (8.1)*
Arousal	*15 (9.3)*	*-*	*Light*	*5 (3.1)*	*19 (11.9)*
Aesthetics	*7 (4.3)*	*2 (1.3)*	*Other crossmodal*	*3 (1.9)*	*29 (16.9)*

Discussion

These findings would appear to suggest that non-expert listeners relied more on affective connotations of the word brightness than their expert counterparts, whereas the latter focused almost exclusively on timbral characteristics (via acoustical or metaphorical language). However, it is not clear whether these can be ascribed to effects of acoustical material (multi-instrumental excerpts versus solo instrument sounds), of musical experience (naïve versus expert listeners), or of language and culture (see Method for differences between the two corpora). More research is needed to understand how brightness and emotion interact in musical contexts (cf. Wallmark et al., 2019). Such knowledge can help improve orchestration strategies for conveying emotional intention in music.

Table 2: Distribution of verbal units across semantic categories

Spectral char.	*high frequencies (15), high-pitch (5), pitch (6), overtones (5), overtone spectrum (3), tone frequency (2), spectral distribution (2), proportion of high frequencies (2), frequency spectrum (2), higher partials, everything up 3Khz-4Khz, higher note, higher register, high harmonics, upper frequency spectrum, dominant frequency range in sustain, spectral energy, dominance of harmonics to fundamental, tonal spectrum, proportion of high noise, no bass, time course of the harmonics, comparison to low frequencies, high cut, low frequencies, harmonics of reverb, frequency, ratio of high to low frequencies, spectral components, register, clear frequency components, proportion of highs, low sound, share of high frequencies in the frequency spectrum, percentage of low frequencies, dominance of certain particularly noticeable frequency ranges, noise level/filter, sound higher, overtone richness*
Temporal char.	*attack (3), attack time (2), duration of impulse, decay of sound, dominant frequency range in sustain, short touch for more brightness, attack of tone, tone time, time course of the harmonics, impulse character, damping, envelope, harmonics of reverb, duration, envelope curve, reverberation, tail of the sound, with reverb, dry, the speed of the attack, the speed of the transition, the speed of the response, transient phase*
Loudness	*louder playing, playing in ff (fortissimo), louder, full of power, impact strength, resonance, nuances*
Valence	*happy (16), positive (9), cheerful (6), uplifting (4), joyful (3), not melancholic (2), makes you smile (2), not sad (2), cheer you up (2), pleasant (2), "pure" (2), joy, "peace in mind", "open", "not hiding", "not pretending", motivating, not depressing, not negative, bright emotions, hope, innocence, peaceful, serene, not annoying, comfortable, chirpiness, without fear, feel good, "beautiful"*
Arousal	*uplifting (4), dance (2), energetic, vivid, motivating, not depressing, fast movement, peaceful, serene, resonates inside me, chirpiness*
Aesthetics	*good sound, natural, mysterious, digital, no scratches, projected, quality, presence, better/worse audible*
Harmony	*major chords (4), major scales (2), major key (2), major mode, not minor mode, harmony (minor, major, atonal or modal), harmony in background, simple harmonic structure, more major than minor, harmony*
Instrument/ ation	*marimba, xylophone, vibraphone, solo, orchestration, light orchestration, high-pitched instruments, high-frequency instruments, drums, a full symphonic orchestra, timbre of the instrument, timbre, light vs dark instruments, from instrumental knowledge, sound of instruments, instrument*
Rhythm/ melody	*clear melody, rhythmic beats, bassline, simple melodies, uptempo, interval succession, melodic line, fast beats, good tempo, speed/rhythm*
Clarity	*clear sound (8), clarity (7), dullness (as opposed to clarity) (2), clean sound (2), purity of sound, clear melody, crystal clear, clear frequency components, without noise, listen clearly, dull (as opposed to clear), clarity vs dullness, blurriness*
Light	*dull (as opposed to bright) (6), dullness (as opposed to brightness) (4), dark (4), sun (2), shining, sunny, bright colours, light vs dark instruments, brilliance, radiant, nitid, tonal brightness*
Other crossmodal	*tone colour (3), soft (2), sound colour (2), sharp (2), deep (2), not deep tones, bright colours, spectral colour, texture, hardness, cutting, metallic, shrill, dry, tonal sharpness, presence/absence of depth, narrowness, width, material, roughness, height/depth, height of note, high cut*

Acknowledgements

CD would like to thank Miguel Reyes and Canberg Turan for assisting with Corpus 1 data collection. Corpus 2 was collected by CS in collaboration with Kai Siedenburg and Christoph Reuter.

References

Kollias, D., Tzirakis, P., Nicolaou, M.A. et al. (2019). Deep affect prediction in-the-wild: Aff-Wild database and challenge, deep architectures, and beyond. International Journal of Computer Visual, *127*, 907–929.

Saitis, C., Fritz, C., and Scavone, G. P. (2019a). Sounds like melted chocolate: how musicians conceptualize violin sound richness. In: Kob M., *Proceedings 2019* International Symposium on Musical Acoustics, (pp. 50–57). Detmold, Germany.

Saitis, C., Fritz, C., Scavone, G. P., Guastavino, C., & Dubois, D. (2017). Perceptual evaluation of violins: A psycholinguistic analysis of preference verbal descriptions by experienced musicians. *Journal of the Acoustical Society of America*, *141*(4), 2746–2757.

Saitis, C., Siedenburg, K., Schuladen, P., and Reuter, C. (2019b). The role of attack transients in timbral brightness perception. In: Ochmann M., Vorländer M., Fels J. (eds), *Proceedings of the 23rd International Congress on Acoustics* (pp. 5505-5543). Aachen, Germany.

Saitis, C., & Weinzierl, S. (2019). The semantics of timbre. In: K. Siedenburg, C. Saitis, S. McAdams, et al. (eds.), *Timbre: Acoustics, Perception, and Cognition* (pp. 119–149). Springer Handbook of Auditory Research, vol 69. Springer, Cham.

Wallmark, Z., Frank, R. J., & Nghiem, L. (2019). Creating novel tones from adjectives: An exploratory study using FM synthesis. *Psychomusicology*, *29*(4), 188–199.

Timbre and Visual Forms: a crossmodal study relating acoustic features and the Bouba-Kiki Effect

Ivan Eiji Simurra[1,2†], Patrícia Vanzella[2] and João Ricardo Sato[2]
[1] Music Department, Federal University of Acre, Rio Branco, Acre, Brazil
[2] Center of Mathematics, Computing and Cognition, Federal University of ABC, São Bernardo, São Paulo, Brazil
[†] Corresponding author: ieysimurra@gmail.com

Introduction

Music has a multidimensional nature with a myriad of features set over time that vary in a multitude of acoustic profiles. It has been shown, for instance, that music listening may be a multimodal experience where musical sounds can evoke abstract visual forms. Particularly, we are interested in a well-described effect known as the Bouba-Kiki effect (Köhler, 1929). This phenomenon relates to a non-arbitrary tendency to associate abstract words whose utterance demand rounding sounds (as in Bouba) with rounded shapes, while sharp words (as in Kiki*)* are usually associated with angular shapes. The studies based on the bouba-kik effect provide the first vital clues to understand the origins of proto-language, as it suggests that there may be natural restrictions on the way sounds are mapped on objects. Previous research suggests that this cross-modal phenomenon may also be found between musical timbre and shapes (Adeli et al, 2014). The authors studied the cross-modal correspondences between musical timbre and visual forms. Basically, using visual stimuli based on the literature about the Bouba-Kiki effect, each sound stimulus with timbre variation was related with some peculiarities of shapes, that is, rounded or angular. For the music orchestration and timbre studies, contemporary music, particularly from the Second Half of the 20th Century and the 21st Century music compositions, have made extensive use of technical procedures to draw attention to novel sound characterization features, such as texture and the presence of noisy sounds as relevant sound events. Such prospect is delved in the context of non-standard instrumental techniques concurrent to the usage of alternative musical orchestration settings. The present study focuses on the cross-modal correspondences between the acoustic correlates of contemporary orchestral music and the visual forms from the Bouba-Kiki Effect. We carried out an online experiment to collect ratings from subjects listening to contemporary music excerpts. Then, we cluster the classification to summarise results and then we analyze them by the acoustic features from auditory stimuli.

Method

The sound stimuli database was designed to address contemporary music techniques and practices aimed at the creation of new sounds and textures in orchestral writing (Griffths, 1978). A total of eleven sound stimuli were selected each with a duration of 5.0 second. Audio mixings were created using Audacity and the instrumental audio samples used to generate the orchestral sound textures belong to three sound databases (Ballet et al, 1999). The audio fragments were chosen from excerpts of chamber music and orchestral works by composers such as Ravel, Debussy, Stravinsky, Messiaen, Schoenberg, Ligeti, Grisey, Scelsi, Lachenmann, Xenakis and Sciarrino. Accordingly, such compositions and instrumental techniques may be explored to create unexpected sound effect modifying the global timbre perception. For the visual shapes we selected forms with different structures between rounded and jagged features. For that, we settled our assortment based on the study performed by Nielsen and Rendall (2011). On the first step of the experiment, we gathered data from an online survey in order to select the most appropriate visual shapes by listening each of the auditory stimulus. Fifty-one volunteers (30 women, age average = 34.79, sd ± 9.80) rated each auditory stimulus with one of the ten abstract shapes alternatives. To avoid fatigue effects and

other biases responses both the sound events and the visual shapes were randomized presented for every subject. Moreover, participants were allowed to listen to each audio excerpt many times while they filled their ratings. The experiment lasted about nine minutes on average, according to log data retrieve from the online experiment. Figure 1 depicts the online interface.

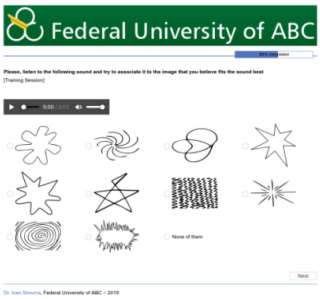


Figure 1: online interface for each auditory stimulus.

We then applied K-means to cluster the visual forms according to the subject ratings. A total of 4 groups of clusters was achieved. Chi-squared tests were conducted to determine whether shapes had been selected randomly, with results suggesting that shape selection was very unlikely to have arisen by chance: cluster 1, $p < 2.2e-16$; 2, $p = 9.365e-09$; 3, $p = 1.202e-04$ and 4, $p = 4.086e-06$. Figure 2 depicts the 4 most consistent visual forms selected based on the K-means results.

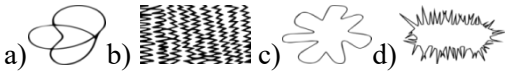


Figure 2: The four visual forms selected from K-means cluster techniques

Additionally, we also grouped each of the audio stimuli presented by the K-means indexes. Figure 3 displays all the audio stimuli according to the K-means plot visualization.

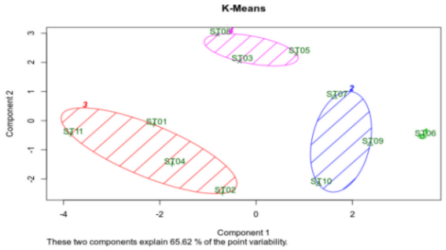


Figure 3: K-means plot featuring all the 11 audio stimuli

Next, we examined the acoustic features of the audio stimuli presented to participants in order to determine any relationship with the visual forms selected. For that, we used Sonic Visualiser (Cannam et. al, 2010) application with specific function libraries to the following acoustic features processing: Spectral Centroid, Spectral Standard Deviation, Spectral Kurtosis, Spectral Flatness, Spectral Flux, Spectral Irregularity, Odd to Even Ratio, Zero Crossing Rate, Energy RMS and Spectral Rolloff (Bullock, 2007). Therefore, for each of the 11 audio stimuli, we retrieved a vector of 10 acoustic features. We then analyzed the acoustic data to examine the presence of multicollinearity among the different features. We assume that on the used acoustic features analysis vectors there might be a some similarity among them. This can result in artifacts data and consequently in more noisy results. To verify if it is occurs we analyzed the multicollinearity. For that, we performed the Kaiser-Meyer-Olkin (KMO) model to predict whether data are likely to factor well, based on correlation and partial correlation (Zacharakis et al, 2014). First, all variables were normalized from the whole 11 audio stimuli between range [0,1]. Then we performed a KMO test to verify collinearity. On the acoustic features vector, we applied Principal Components Analysis (PCA) to reduce dimensionality from variables. Finally, we calculated Factor Analysis (FA) to determine the most expressive acoustic features by each component. Next Section evinces results from the method described.

Results

Each of the acoustic features retrieved from each sound stimulus resulted in a overal KMO Measure of Sampling Adequacy (MSA) index greater than 0.69 (0.76, 0.85, 0.78, 0.69, 0.75, 0.79, 0.78, 0.76, 0.84, 0.73, 0.85). According to Zacharakis, KMO overall should be .60 or higher to proceed with factor analysis. The average of the overall KMO for the 11 sound events was 0.76. Following Henry Kaiser evaluation (1974) this result in middling and its necessary to discuss it with caution. After, we calculated PCA to determine the number of components for the FA (rotation: 'varimax' and score: 'Bartlett'). We defined the number of 04 principal components which explain a total average of 88% of the cumulative proportion of the data variance. The prominent descriptors over the four factors for each groupings by visual shapes are shown in Tables 1.

Table 1: Factor loadings by Figures 1a, 1b, 1c and 1d.

	Figure 1a				**Figure 1b**			
Acoutic Feature	**Factor 1**	**Factor 2**	**Factor 3**	**Factor 4**	**Factor 1**	**Factor 2**	**Factor 3**	**Factor 4**
SpecCentroid		*0.850*	*0.504*	*0.128*	*0.908*	*-0.114*	*-0.204*	*-0.285*
SpecFlatness	*-0.503*		*0.858*		*-0.310*	*0.183*	*0.913*	*0.179*
SpecFlux		*0.203*		*0.640*	*0.519*	*0.340*	*-0.170*	
SpecIrregularity	*0.970*			*-0.113*	*0.450*	*0.808*	*-0.107*	*0.194*
Odd to Even	*0.377*	*-0.178*	*0.231*	*0.181*		*-0.117*	*0.145*	*-0.127*
SpecStandardDev	*-0.243*	*0.615*	*0.406*	*0.524*	*0.963*		*-0.147*	
Zero Crossing	*0.113*	*0.263*	*0.595*		*0.379*	*-0.102*	*0.118*	*-0.545*
RMS	*0.985*					*0.677*		*0.650*
SpecKurtosis	*0.907*	*-0.130*	*0.338*			*0.921*	*0.153*	*0.124*
SpecRoll Off		*0.879*			*0.916*	*0.286*	*-0.105*	

	Figure 1c				**Figure 1d**			
Acoutic Feature	**Factor 1**	**Factor 2**	**Factor 3**	**Factor 4**	**Factor 1**	**Factor 2**	**Factor 3**	**Factor 4**
SpecCentroid	*0.893*	*-0.266*	*-0.280*	*-0.176*	*-0.286*	*0.929*	*0.123*	*-0.189*
SpecFlatness	*-0.233*		*0.581*	*0.578*	*-0.913*	*0.337*		
SpecFlux		*0.296*	*0.587*	*0.188*	*0.140*		*0.480*	*0.347*
SpecIrregularity	*-0.115*	*0.818*	*0.402*		*0.863*	*-0.231*		
Odd to Even		*0.111*	*0.427*				*0.471*	
SpecStandardDev	*0.979*	*-0.168*			*-0.302*	*0.950*		
Zero Crossing	*0.580*	*-0.343*	*-0.388*	*-0.546*	*-0.363*	*0.264*	*0.699*	
RMS	*-0.286*	*0.921*	*0.168*	*0.191*	*0.827*	*-0.268*	*-0.463*	
SpecKurtosis	*-0.628*	*0.433*	*0.260*	*0.218*	*0.758*	*-0.469*		*0.151*
SpecRoll Off	*0.948*		*0.139*		*-0.341*	*0.891*		

Discussion

The results indicated that, irrespective of the contrasting visual shape attributes, the prevalence of the spectral magnitude, mainly associated with noise content and the specific magnitude on spectrum region, is a substantial acoustic feature for all four groupings. However, some peculiarities inherent to the sound qualities associated with each figure were observed. For jagged and sharp forms, Spectral Kurtosis and Spectral Flatness could be highlighted as the acoustic features that were most associated with angular shapes. Interestingly, previous findings suggest that these acoustic features are generally associated with noise sounds (Peeters, 2003; Krimphoff et al, 1994; Simurra and Manzolli, 2016). On the other hand, the most prominent acoustic features associated with rounded and smooth shapes were Energy RMS and the Spectral Roll-Off, which are often related to the brightness of the sound (Peeters, 2003; Krimphoff et al, 1994; Simurra and Manzolli, 2016). These results thus suggest that certain symbolic visual features seem to share acoustic resources at some stage, such as those centered on noisy sound. Moreover, by using only spectral features content with only a small portion of spectro-temporal features results centered only on the spectral magnitude output. Some other audio descriptors will be necessary to test spectro-temporal and temporal features effect. It is interesting to pinpoint for contemporary music repertoire that presents, among multiple variables, non-harmonic spectral content. Further studies would be necessary to compare or correlate groupings sharing similar features.

Acknowledgments

The authors gratefully acknowledge financial support from (CAPES-88887.341537/2019-00). We thank Thenille Braun Janzen for her suggestions on earlier versions of the manuscript.

References

Adeli, M., Rouat, J., & Molotchnikoff, S. (2014). Audiovisual correspondence between musical timbre and visual shapes. *Frontiers in human neuroscience, 8*, 352.

Ballet, G., Borghesi, R., Hoffmann, P., & Lévy, F. (1999). Studio online 3.0: An internet" killer application" for remote access to IRCAM sounds and processing tools. *Journées d'informatique musicale (JIM).* (pp. 123–131). Issy-les-Moulineaux, France.

Bullock, J. (2007). Libxtract: A Lightweight Library for audio Feature Extraction. Conservatoire. In *Proceedings of the International Computer Music Conference (ICMC)*, *43,* 25-28.

Cannam, C., Landone, C., & Sandler, M. (2010, October). Sonic visualiser: An open source application for viewing, analysing, and annotating music audio files. ACM

Griffiths, P. (1978). *A concise history of avant-garde music: from Debussy to Boulez*. New York: Oxford University Press.

Kaiser, H. F. (1974). An index of factorial simplicity. *Psychometrika*, *39*(1), 31-36.

Köhler, W. (1929). *Gestalt psychology. Psychological research*, Springer, v. 31, n. 1, p.XVIII–XXX

Krimphoff, J., Mc Adams, S., and Winsberg, S. (1994). Caractérisation du timbre dessons complexes. ii. analyses acoustiques et quantification psychophysique. *Le Journal de Physique IV*, vol. 4.

Nielsen, A., & Rendall, D. (2011). The sound of round: evaluating the sound-symbolic role of consonants in the classic Takete-Maluma phenomenon. *Canadian Journal of Experimental Psychology,* 115.

Peeters, G. (2004). A large set of audio features for sound description in the cuidado project*. IRCAM*.

Simurra, I. E., Manzolli, J. (2016). Sound Shizuku Composition: a Computer-Aided Composition Systems for Extended Music Techniques. *Brazilian Journal of Music and Mathematics*, 86-101.

Zacharakis, A. I., Pastiadis, K., Papadelis, G., & Reiss, J. D. (2011). An Investigation of Musical Timbre:Uncovering Salient Semantic Descriptors and Perceptual Dimensions. In *ISMIR,* 807-812.

Zacharakis, A., Pastiadis, K., & Reiss, J. D. (2014). An interlanguage study of musical timbre semantic dimensions and their acoustic correlates. *Music Perception*, *31*(4), 339-358.

How Periodicity in Timbre Alters Our Perception of Time: An Analysis of "Prologue" by Gérard Grisey

Gabrielle Choma
School of Music and Dance, University of Oregon, Eugene, Oregon, United States
gabrielle_choma@yahoo.com

Introduction

Early spectralism remains under-explored as a treasure trove of new and exciting structures of sound and timbre. As a proponent of this genre, Gérard Grisey has written extensively about his philosophies of time, periodicity, and subjective musical experience. The few scholars that have written about this music have done so with the intent of either understanding how pitch is organized (Rose, 1996), how Grisey organizes his musical gestures to manipulate time (Hennessey, 2009), or to understand the process of instrumental synthesis and spectral harmony (Hasegawa, 2009). Grisey himself, in writing about his own music, discusses how extra-musical factors such as periodicity manipulate how one experiences time in music (Grisey, 1987). This literature opens a door into the philosophy of sound and time, and my presentation hopes to connect Grisey's temporal concepts with an exploration of timbre in "Prologue," from "Les Espaces Acoustiques." With my approach, I hope to shed light on how listeners experience timbre and time in "Prologue."

Method

My methods for this presentation include an analysis of a few key moments of timbral fluctuations in "Prologue" using Grisey's own classification of periodicity from his 1987 article "Tempus Ex Machina: A Composer's Reflection on Musical Time." The score does not include bar lines, so relative location in the score will be referenced to by time. For this presentation, I will be referencing the recording made by violist Gérard Causé, for whom the piece was written. (https://youtu.be/Wy0DqvmMzQE). According to Grisey, periodicity expands a listener's sense of time, and unexpected events, or "chaos," contract time by alerting the listener and drawing in their attention. By examining the distance and differences between timbral events in relation to periodicity, we can infer how expanded or contracted time is to the listener.

Results

This piece features timbral fluctuations that accelerate to a climax in the middle of the piece and then gradually decelerate. Within this grand scheme are smaller systems of timbral tension and release. These smaller systems contribute to the whole by introducing new timbral events that ultimately arrive at a complete destruction of pitch and tone as the performer is instructed to make "scrubbing" sounds. From this, tone and pitch are gradually reintroduced in a way that marries the pitch-less timbres with melodic gestures that preceded them. Time can be said to feel the longest right after this aforementioned climax due to the minute changes in periodicity that are less perceptible after such a shocking timbral event. This piece subjects the listener to a variety of timbres, both comfortable and uncomfortable, and my research has found that changes in periodicity create the expanded and contracted senses of time more so than the timbral changes themselves.

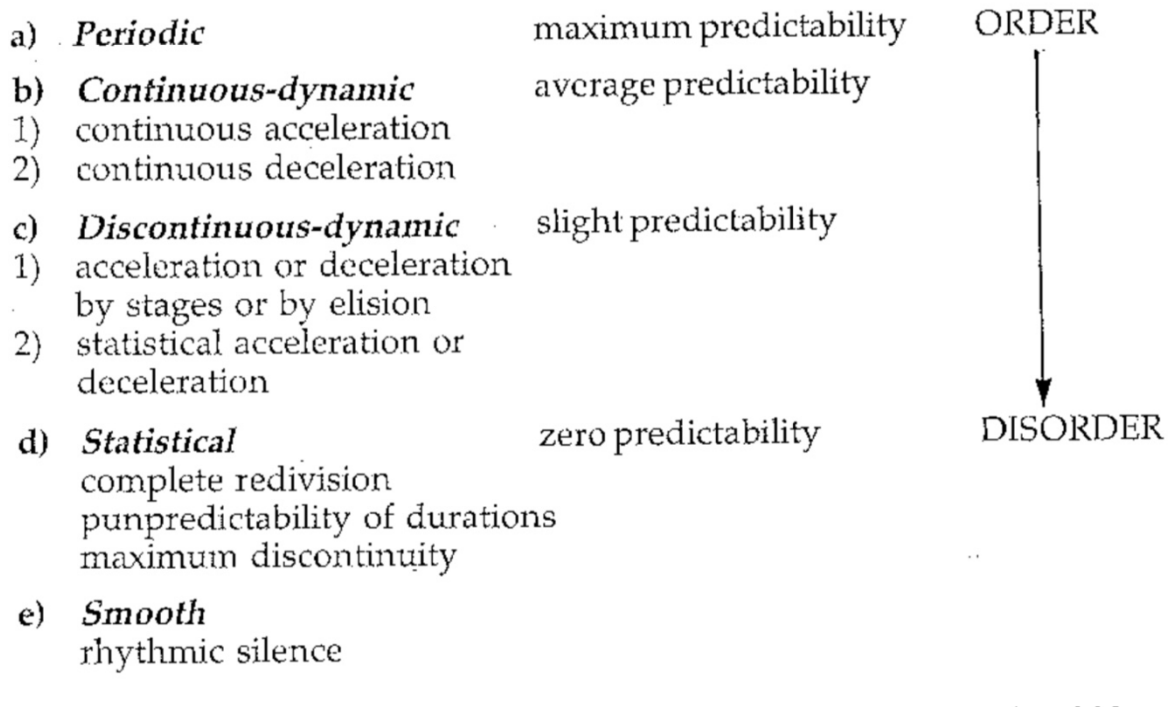


Fig 1: Grisey's table of chaos and disorder from "Tempus Ex Machina: A Composer's Reflection on Musical Time" (1987). Though he references rhythm and time in this table, he admits that this can be translated into other musical parameters, including timbre.

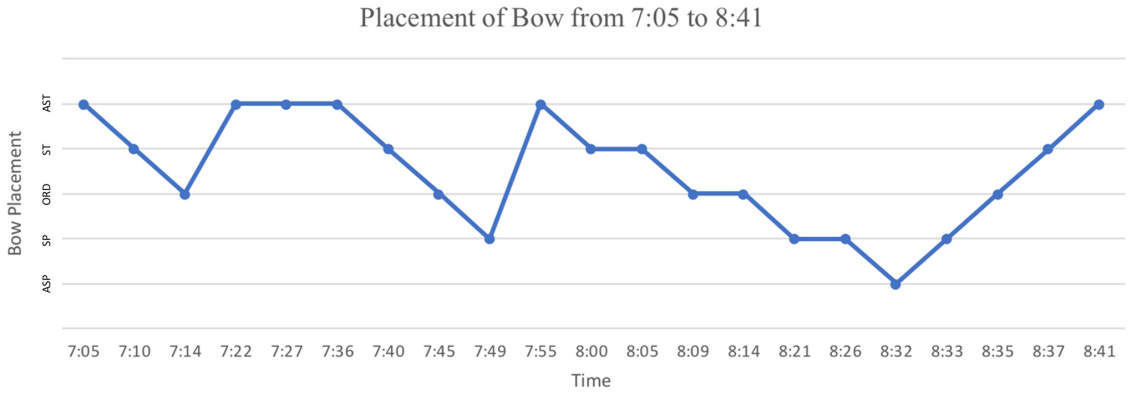


Fig 2: Placement of bow from 7:05 to 8:41. Time can be felt as expanded during 7:55-8:26 due to the doubled bow placements, meanwhile it can feel contracted from 8:32 – 8:41 due to the rapid bow placement changes. Here, AST stands for "alto sul tasto", ST stands for "sul tasto," ORD is short for ordinary placement, SP stands for "sul ponticello" and ASP stands for "alto sul ponticello."

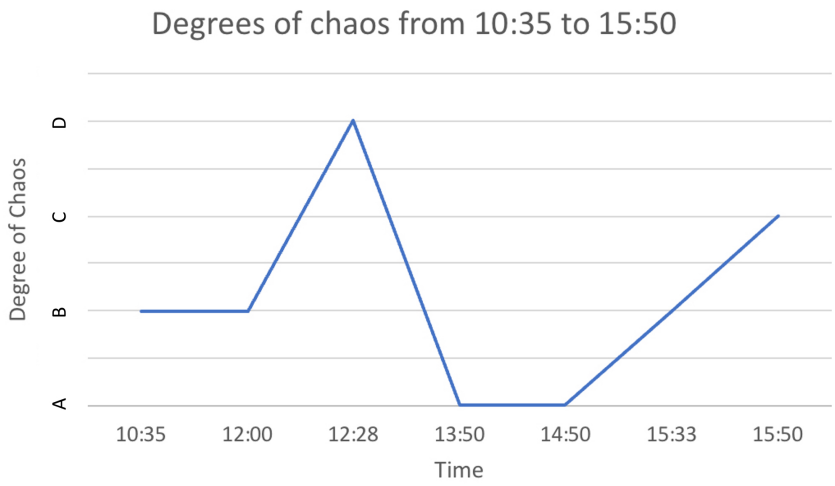


Fig 3: Degrees of Chaos from 10:35 to 15:50. The letters A-D are taken from Grisey's 1987 article, with "A" representing pure periodicity, and "D" representing non-linear changes in timbral periodicity. Time can be felt as slightly contracted between 12:00 and 13:50 due to changes in timbre that shock the listener. The periodicity that follows this from 13:50 to 14:50 can make the listener feel as if time is greatly expanded due to the small, imperceptible changes in pitch and timbre.

Discussion

An understanding of how timbral fluctuations and periodicity play a role in the manipulation of time will allow us to understand Grisey's temporal concepts and how they appear in his early works. Though this concept can be understood through the lens of pitch or rhythm, rich timbral exploration is a key component of this genre, and understanding it together with other musical parameters is essential for a holistic understanding of this genre. This concept can be expanded into discussions about other spectral compositions as well as music outside of this genre.

Acknowledgements

I'd like to thank Dr. Robert Hasegawa for introducing me to spectral studies, as well as Dr. Jack Boss for his academic support.

References

Grisey, G. (1987). Tempus Ex Machina: A Composer's Reflections on Musical Time. Contemporary Music Review, 2, 239-275.

Hasegawa, R. (2009). Gérard Grisey and The 'Nature' of Harmony. Music Analysis, 28(2/3), 349-371.

Hennessy, J. (2009). Beneath the Skin of Time: Alternative Temporalities in Grisey's "Prologue for Solo Viola". Perspectives of New Music, 47(2), 36-58. Retrieved July 15, 2020, from www.jstor.org/stable/25753696

Rose, F. (1996). Introduction to the Pitch Organization of French Spectral Music. Perspectives of New Music, 34(2), 6-39. doi:10.2307/833469

Cross-categorical discrimination of simple speech and music sounds based on timbral fidelity in musically experienced and naïve listeners

Ryan Anderson[1†], Alyxandria Sundheimer[1] and William P. Shofner[1]
[1]Indiana University, Bloomington Indiana, United States
[†]Corresponding author: anderyan@iu.edu

Introduction

Psychoacoustic approaches to complex sound perception suggest that there are differences in how normal hearing humans use spectral information in speech and music signals. Studies applying these approaches conclude that music perception requires greater spectral resolution than speech perception (Shannon, 2005). Intelligibility of noise-vocoded speech is high with as few as 4 vocoder channels (Smith et. al, 2002). Conversely, music perception studies with noise vocoded signals suggest that upwards of 32 channels are necessary for recognition (Mehta and Oxenham, 2014). Analyzing these results together, it seems that music perception is more susceptible to spectral degradation than speech perception. That is, a high level of speech perception performance can be achieved with fewer noise-vocoded channels than required for music perception performance. Such conclusions support arguments for specialized cognitive processes in which the brain uses acoustic information differently depending on the type of sound its processing. This logic propels popular theories regarding auditory processing specialization at various cognitive levels (Liberman, 1984; Zatorre, 2002). However, conclusions from these metadata are problematic given that they aggregate results from several different studies using different methodologies and therefore different cues. In particular, music perception as represented by melody recognition relies on changes in pitch information and the structure of harmonic information across the duration of the stimulus. Conversely, word identification tasks use envelope cues generated by the spectral information in consonants and formant structure of vowels. Given that noise vocoding is used to introduce spectral content manipulations, it is important to note that these stimuli are influenced differently.

Differences in speech and music perception are also prevalent in studies regarding subjects' musical experience. These studies generally demonstrate that musical experience correlates with better speech perception in degraded or challenging conditions (Parbery-Clark et al. 2012) as well as frequency and pitch discrimination (Tervaniemi et. al, 2005) compared to musically inexperienced peers. Based on these differences, the level of musical experience in participants should be considered when exploring differences in categorical sound perception.

Vocoding affects speech and music differently depending on their task context. Furthermore, it is unclear as to whether specific or general mechanisms are driving decisions due to different task demands. To eliminate task differences, the experiment at hand evaluates the perception of speech and music sounds using a single task in which the acoustic cues are equivalent in all conditions. By using natural representations of spoken vowels and music notes, respective spectral structures serve as the primary differentiating acoustic feature across stimuli. Spectral profiles of harmonic structure in musical instruments and formant distribution in vowels provide a common dimension of timbre between speech and music. Therefore, a behavioral task in which the participants make assessments of timbral differences across categories of speech and music in natural and vocoded conditions is used to determine processing differences as dependent on the spectral quality of the signal. The present study expands on preliminary data using a single discrimination paradigm to compare speech and music perception based on similar perceptual dimensions, namely timbre.

Method

28 subjects with normal hearing thresholds (< 20 dB HL across audiological test frequencies) completed experiment 1. Of these participants, 17 were naïve listeners and 11 were musically experienced. Musically experienced participants were considered as those who reported 3 or more years of practiced musical

experience. Regardless of classification, all participants completed all trials of each experiment. 6 subjects completed experiment 2; three of which completed experiment 1.

Stimuli consisted of naturally spoken vowels and notes played on musical instruments as well as 32-, 8-, and 4- channel noise-vocoded versions. Music note samples consisted of bassoon, cello, clarinet, trombone, trumpet, and viola playing either G3 (196 Hz), or B2 (123 Hz). Vowels */a/, /ae/, /i/, /ou/, /u/* were recorded from male and female speakers. All stimuli were equalized in RMS amplitude and presented for 500 ms at 73 dB SPL. Fundamental frequencies for music notes and vowels were closely paired within trials to remove potential pitch cues. Listeners discriminated either instruments from vowels or vowels from instruments via button release in a go/no-go task. Before the participants initiated a trial, a 500 ms standard token was repeatedly presented with a 500 ms interstimulus interval. Participants prompted a trial by pressing and holding down a button. While holding down the button, an 1850 ms response window was placed after a randomly selected hold time (1150 – 8150 ms). During the response window, two 500 ms sounds were presented with a 500 ms interstimulus interval. Participants were instructed to only release the button if the sound source in the response window was from the other category than the category containing the standard token (i.e. if the standard token is a vowel, only release the button if a music note plays). A button release to a within category change indicates a false alarm. The discrimination paradigm offered insight as to how timbral fidelity influenced perception *between* stimulus sound categories. Reaction times and accuracy were measured and organized as a function of signal degradation level to consider potential differences in how normal hearing listeners utilize spectral information. To clarify the cause of changes in task performance, a second experiment used the same procedure as experiment one but tasked the participants with responding to any perceived change from the standard stimulus, regardless of category. If participants are unable to accurately discriminate targets from other stimuli solely due to the general spectral quality of the sound, one would expect for performance in both experiments to be similar. Trials consisted of just 8- and 4-channel vocoded conditions, as these are the most challenging and therefore had the largest chance to influence categorization ability.

Results

Experiment 1

Grand average reaction times in response to 4 channel, 32 channel, and natural stimuli showed no difference beyond error in both conditions (figure 1). Like in preliminary data, there is a significant effect of spectral quality of the signal on accuracy measures with decreasing accuracy as the number of vocoding channels decreases in conditions where discrimination occurs across categories (figure 2). This is

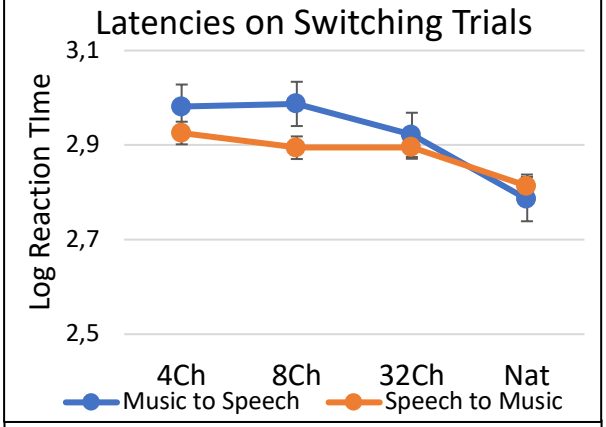
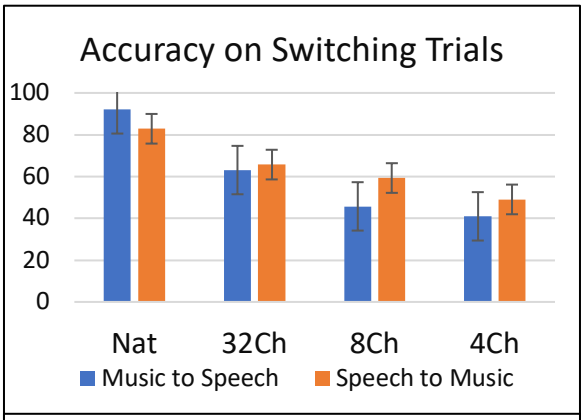


Fig 1. Grand average log transformed reaction times for both conditions as a function of number of vocoding channels.



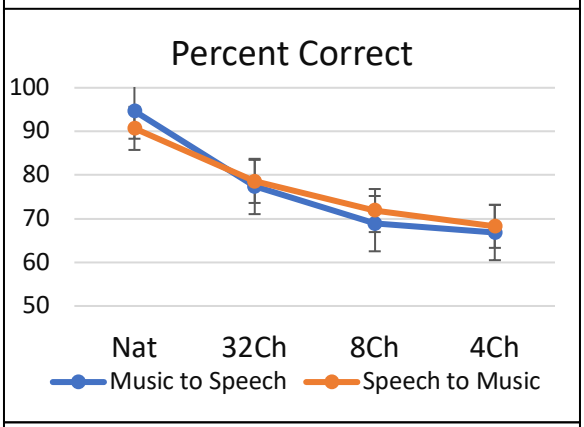


Fig 2. Grand average accuracy on cross-category discrimination trials as a function of vocoding channels. Hit rates in both conditions (top panel) and percent correct (bottom panel) compare performance in both category conditions.

expected, as fewer channels generate a more ambiguous signal. Within vocoding conditions however, there are no significant differences in accuracy measures when discriminating across categories. Using discrimination trials across all listeners as an indication, there are no substantial differences beyond error in discrimination accuracy between perception of basic speech and music sounds, regardless of spectral richness.

Preliminary data suggest that musically trained listeners are better than naïve listeners at discriminating vowels from musical instruments (Anderson et. al, 2019). After including more musically trained subjects in analysis, this asymmetry was still observed, but to a smaller effect compared to the previous study. Two-factor ANOVA on arcsine transformed percent correct with Bonferroni corrections showed a significant effect of group in the music-to-speech condition, $F(1,104) = 7.63$, $p < .01$; $\eta^2 = .07$), while there was no significant effect of group in the speech-to-music condition (figure 3).

Experiment 2

In the general task, participants demonstrated a clear ability to discriminate between the stimuli in both vocoding representations. Percent correct scores were near ceiling for 8 channel noise vocoded stimuli (99.68%), and at ceiling for 4 channel stimuli. These data ensure that participants are indeed able to differentiate the stimuli from experiment 1 even with the largest amount of spectral degradation in the stimuli set. All effects observed in experiment 1 should therefore be attributed to differences in category discrimination ability.

Discussion

Using a single categorical discrimination task in which timbre was the primary decision criteria allows for a more justifiable basis of comparison between the role of spectral processing in speech and music sounds. This type of design ensures that the acoustic cues as well as the task demand are equivalent. Given the balancing of acoustic cues in this design, asymmetrical performance potentially reflects differences in perceptual modes that depend on the type of stimuli being processed. Lower vocoding channel stimuli are of interest, as these localize the previously reported split in performance between music and speech processing. Slight differences in reaction times for spectrally ambiguous stimuli (8-channels) suggest a potential deviation in speech and music processing, but accuracy measures do not currently support this possibility, as there were no substantial differences in average responses in any vocoding conditions. Similarities across listeners in accuracy and reaction time measures when discriminating between sound categories indicate that vowel and musical instrument identification do not involve mode-specific mechanisms. When accounting for musical experience of the listeners, there is a persistent trend in which musically trained listeners demonstrate higher percent correct responses than naïve listeners at

discriminating vowels from musical instruments. This is a curious finding as one would expect the musically trained group to exhibit similar reaction times and percent correct measures in trials where the target switches across conditions regardless of discrimination direction. Reaction times should be significantly lower, and percent correct significantly higher compared to naïve listening peers if this were the case. Given the decrease in effect size from the preliminary study and the current study as the number of subjects increased, it is not unlikely that this asymmetrical trend might be eliminated with larger, matched group sizes. Future studies using more stimuli in each category and larger matches samples may provide greater validity and statistical power to investigate this possibility.

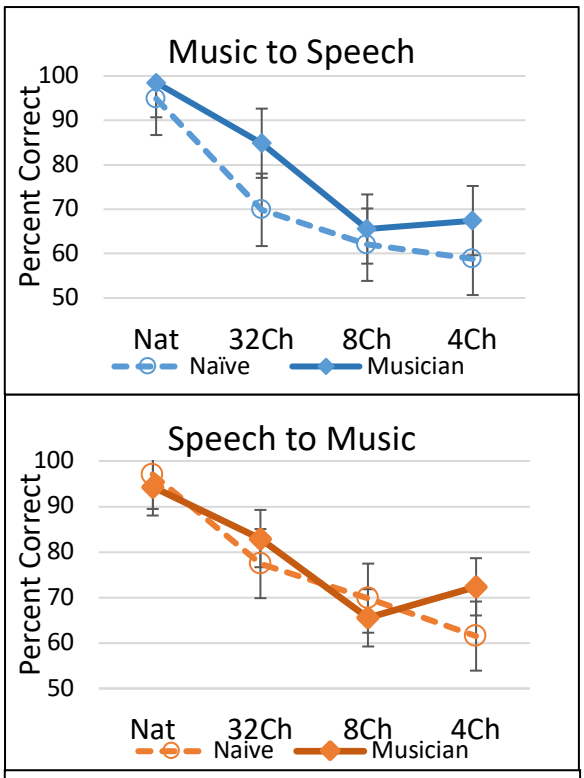


Fig 3. Group level performance as indicated by percent correct in music to speech (top) and speech to music (bottom) conditions.

References

Anderson, R., Sundheimer, A., & Shofner, W. (2019). Ability of normal hearing listeners to recognize vowels and musical instruments under spectrally-degraded conditions. *The Journal of the Acoustical Society of America*, *145*(3), 1720-1720.

Liberman, A. M. (1984). On finding that speech is special. In *Handbook of Cognitive Neuroscience* (pp. 169-197). Springer, Boston, MA.

Mehta, A. H. & Oxenham, A. J. (2017). Vocoder simulations explain complex pitch perception limitations experienced by cochlear implant users. *Journal of the Association for Research in Otolaryngology, 18,* 789 – 802.

Parbery-Clark, A., Tierniey, A., Strait, D.L. & Kraus, N. (2012). Musicians have fine-tuned neural distinction of speech syllables. *Neuroscience, 219,* 111-119.

Shannon, R. V. (2005). Speech and music have different requirements for spectral resolution. *International Review of Neurobiology, 70,* 121-134.

Smith, Z. M., Delgutte, B., & Oxenham, A. J. (2002). Chimaeric sounds reveal dichotomies in auditory perception. *Nature*, *416*(6876), 87-90.

Tervaniemi, M., Just, V., Koelsch, S., Widmann, A., & Schröger, E. (2005). Pitch discrimination accuracy in musicians vs nonmusicians: an event-related potential and behavioral study. *Experimental brain research*, *161*(1), 1-10.

Zatorre, R. J., Belin, P., & Penhune, V. B. (2002). Structure and function of auditory cortex: music and speech. *Trends in cognitive sciences*, *6*(1), 37-46.

Memory for Musical Key Distinguished by Timbre

Graeme Noble[1†], Joanna Spyra[1] and Matthew Woolhouse[2]
[1] Psychology, Neuroscience & Behaviour, McMaster University, Hamilton, Ontario, Canada
[2] School of the Arts, McMaster University, Hamilton, Ontario, Canada
[†] Corresponding author: graemesnoble@gmail.com

Introduction

Previous research has investigated the extent to which an initial key or pitch information is retained in short-term memory (e.g., Krumhansl & Iverson, 1992). Woolhouse, Horton & Cross (2016) empirically estimated the effect of the initial key to last between 10–12 seconds following modulation; Farbood (2016) extended this retention period to roughly 20 seconds. In a subsequent study, Spyra, Stodolak & Woolhouse (2019) evaluated whether the duration or number of chords constituting the newly modulated key was responsible for memory decays of the initial key. Their findings showed that the duration of the second, modulated key was the critical factor, rather than the number of chords. In a similar vein, recent research suggests that surface features enhance retention of nonadjacent keys in short-term memory when we listen to modulating musical sequences. For example, Spyra & Woolhouse (2018) found that the addition of figuration (e.g., passing tones and suspensions) and rhythmical activity (i.e., note density) enhanced memory for an initial key after modulation.

The studies referred to above assessed the effect of the initial key via a probe chord or probe cadence paradigm. Simply put, participants were asked to rate the probe segment's goodness-of-completion with respect to the preceding musical sequence, where the probe had either a tonic- or non-tonic-key relationship to the initial key (Farbood, 2016, used harmonic tension ratings). This paradigm has been repeatedly shown to provide a robust subjective measure of nonadjacent key effects and is assumed to reflect the retention of keys in short-term memory post-modulation. While the above research has covered several distinct aspects of music (e.g., number of chords, figuration), the contribution of timbre in this regard has yet to be studied. The research reported here addresses this lacuna.

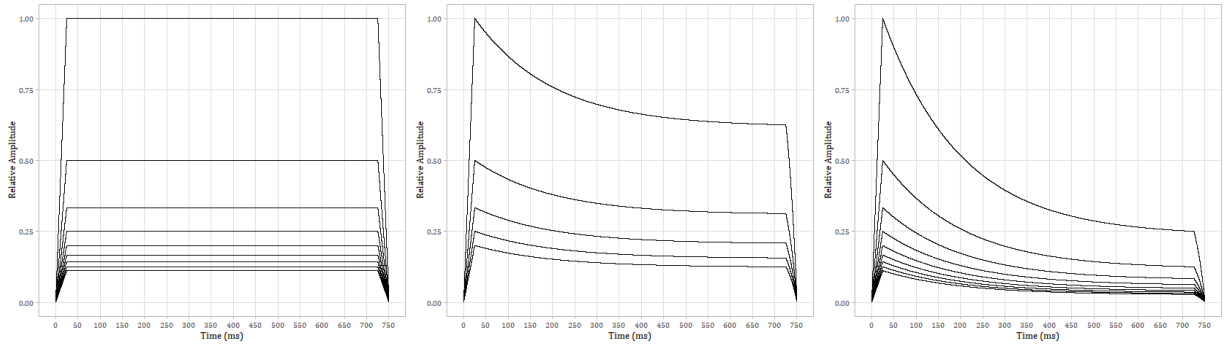
Our study operationalized timbre as consisting of pure-tone components with independent amplitude envelopes (AEs) varying in duration and intensity. The study explored these two aspects of timbre with respect to number of harmonics and AE shape. To validate an underlying assumption of the study—that artificial qualities of sounds are distinguishable—participants in Experiment 1 were asked to distinguish between three timbres constructed with varying degrees of 'naturalness'. In Experiment 2, the sounds from Experiment 1 were used to address whether timbral naturalness impacted key memorization. Using the sequences from Experiment 2, Experiment 3 explored the extent to which individual components of timbre influenced key memorization. Three hypotheses underpinned the study: (1) that timbres are distinguishable in terms of naturalness (Experiment 1); (2) that matching timbres between nonadjacent sections elicit higher goodness-of-completion ratings (Experiment 2); and (3) that distinct elements of timbre, as operationalized above, elicit different ratings (Experiment 3).

Method

Participants in Experiment 1 (n = 64; gender: 14 male, 48 female, 2 non-binary; age: range [*R*] = 17–37, mean [*M*] = 18.38, standard deviation [*SD*] = 2.43) were asked to rate the naturalness of sampled vs. subtractive vs. synthesized sounds. Typically, sounds in nature have a complex mix of harmonic and inharmonic components with independent AEs. Similarly, sampled sounds of acoustic instruments, in our case a grand piano, possess multiple harmonics with independent AEs, many of which have non-integer ratio relationships (Deutsch, 2013). The subtractive timbre was constructed as a fast Fourier transform of a sampled sound, in which inharmonic and upper harmonic elements were omitted. In our case, the subtractive timbre functioned as an intermediate between juxtaposed sampled and additively synthesized timbres. The synthesized sound was comprised of a single sine-tone played with a flat AE.

In Experiment 2 (n = 60; gender: 12 male, 46 female, 2 non-binary; age: *R* = 17–50, *M* = 18.97, *SD* = 4.78), three timbres—as described above—were systematically distributed across three distinct major-key stimulus segments, using the nonadjacency paradigm: (1) an initial 8-chord sequence that complied with voice-leading principles of the Common Practice period (c. 1600–1900); (2) an intervening, 8-chord modulated key sequence; and (3) a three-chord probe cadence (ii-V-I), nonadjacent to the initial key. The tempo of the stimuli was 80 BPM; each chord was 750 ms in duration. The overall duration of the nonadjacent section was therefore 6 s, followed by a 6 s intervening section, 0.75 s rest, and 2.25 s probe cadence. The timbre of the initial segment and probe cadence either matched or differed from each other, with timbres consisting of sampled audio (i.e., most natural; complex), and/or a sine tone with a flat envelope (i.e., least natural; simple). The intervening sequence consistently used the subtractive timbre.

Using the nonadjacency paradigm, in Experiment 3 (n = 33; gender: 6 male, 27 female; age: *R* = 17–19, *M* = 18.24, *SD* = 0.55), AEs and the number of harmonics were manipulated independently such that the initial segment and probe cadence were played with flat or dynamic AEs and either 1 or 9 harmonics. The intervening sequence had a timbre of intermediate complexity (5 harmonics with quasi-dynamic AEs; H_5–AE_{Int}; see Fig. 1b). As in previous studies that employ this paradigm, participants in Experiments 2 and 3 provided goodness-of-completion ratings for the probe cadence using a 7-point Likert-type sliding scale.



Figures 1a, 1b, and 1c: H_9–AE_{Flat}, H_5–AE_{Int}, and H_9–AE_{Dyn} timbres from Experiment 3.

Analysis

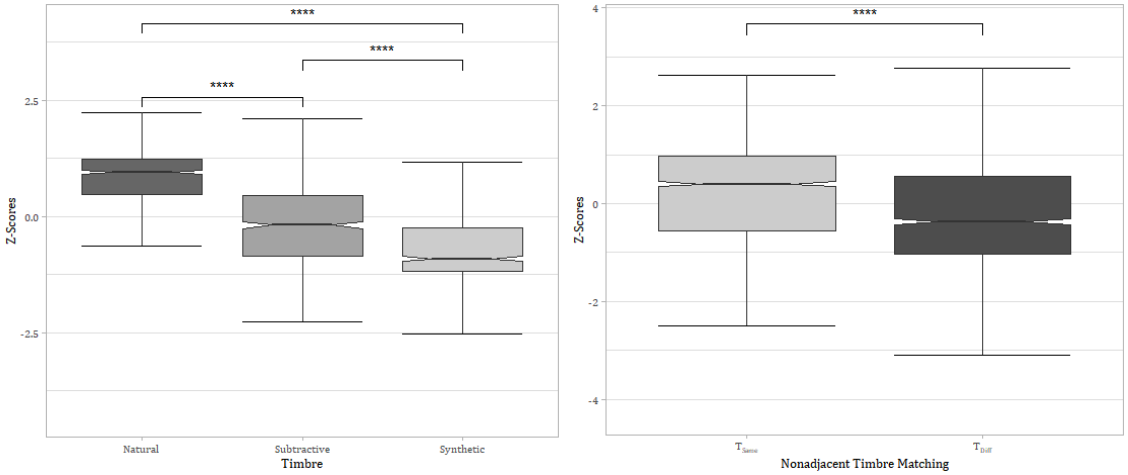
Rating data were normalized using each participant's mean and standard deviation; this process was applied to each experiment. In Experiment 1, we used a two-factor within-subjects ANOVA with *Timbre* and *Pitch Height*. *Timbre* had three levels: sampled, subtractive, and synthesized. *Pitch Height* had 12 levels, the 12-semitone transpositions of the stimuli from F3–F4 (i.e., 175 Hz to 349 Hz). Experiment 2 was assessed with a one-factor within-subjects ANOVA with *Timbre Matching* (T_{Same} vs. T_{Diff}). For Experiment 3, a three-factor within-subjects ANOVA was conducted with *Probe Timbre, Timbral-Component Matching*, and *Key Relationship*. *Probe Timbre* had four levels produced through a combination of 1 or 9 harmonics (H) and flat or dynamic AEs: H_1–AE_{Flat}; H_1–AE_{Dyn}; H_9–AE_{Flat} (see Fig. 1a); H_9–AE_{Dyn} (see Fig. 1c). Using multiple pair-wise comparisons (Tukey HSD), *Timbral-Component Matching* isolated the effects of harmonics vs. AEs. The four levels within this factor were produced through a combination of same or different number of harmonics and same or different AEs between the initial section and probe cadence: H_{Same}–AE_{Same}; H_{Same}–AE_{Diff}; H_{Diff}–AE_{Same}; H_{Diff}–AE_{Diff}. *Key Relationship* had two levels: either tonic or non-tonic between the initial section and probe cadence.

Results

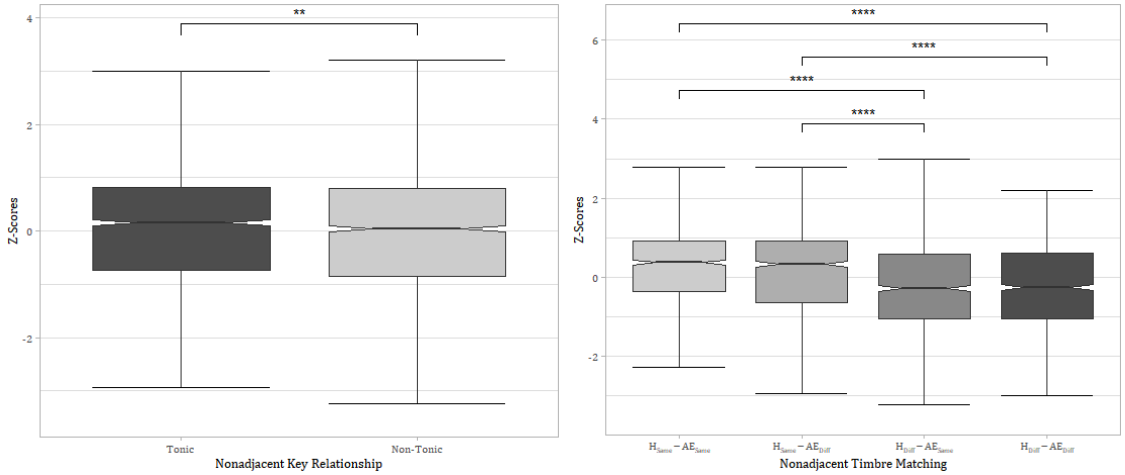
Data from Experiment 1 were consistent with the hypothesis: higher ratings of naturalness were obtained for sampled audio over subtractive over additive synthesized sounds ($F_{2, 64}$ = 776.23, p < .0001, ω^2p = .068); see Fig. 2a. These observations were consistent across pitch height transformations. In Experiment 2, participants' judgement of goodness-of-completion was greatest when the probe cadence used a timbre that matched that of the nonadjacent segment ($F_{1, 60}$ = 121.8, p < .0001, ω^2p = .039); see Fig. 2b.

Results from Experiment 3 revealed a significant effect of tonic-key relationship between the initial segment and probe cadence, independent of timbral manipulation, ($F_{1, 33}$ = 7.85, p = .005, ω^2p = .002); see Fig. 3a. Participants' judgement of goodness-of-completion was greatest when the probe cadence used the most natural timbre ($F_{3, 33}$ = 27.00, p < .0001, ω^2p = .018) or a matching timbre to the initial segment, ($F_{3, 33}$ = 77.50, p < .0001, ω^2p = .051); see Fig. 3b. There were no significant interactions.

Pairwise comparisons for *Timbral-Component Matching* revealed a significant contribution of number of harmonics, but not of AE. The rationale for this deduction is as follows: there was no significant difference between H_{Same} conditions (p > .1), whereas there was a significant difference between AE_{Same} conditions (p < .0001), suggesting that number of harmonics between nonadjacent sections led to consistent ratings whereas AE matching did not. We thus conclude that the harmonic component of timbre within our stimuli was critical with respect to observed nonadjacent effects.



Figures 2a and 2b: Boxplots of Experiment 1 timbre and Experiment 2 timbre matching.



Figures 3a and 3b: Boxplots of relationship and timbre matching, Experiment 3.

Discussion

The results of all three experiments demonstrate the importance of timbre for the perception of musical form involving discrete, nonadjacent sections. Once the groundwork was laid in Experiment 1, by which natural vs. synthetic sounds were shown as easily distinguishable, Experiment 2 conveyed the significance of timbral matching between nonadjacent sections within a piece of music. Irrespective of

whether the nonadjacent timbres were natural or synthetic, the results showed that timbral matching across the 15 s span of each stimulus was rated higher than non-matching. Experiment 3 sought to discover whether the harmonic component or AE shape contributed to nonadjacent memory effects. Via pairwise comparisons, harmonics were found to be largely responsible for the significant variance in the data, and thus we concluded that the number of harmonics played a crucial role within our stimuli.

The results of the discrimination task in Experiment 1 somewhat echo the findings of Siedenburg and McAdams (2017), in which musicians better recognized familiar natural sounds compared to unfamiliar synthetic ones. While our task was not based on explicit recognition, it clearly showed that participants were acutely sensitive to sounds which had an acoustic origin vs. those which did not.

Data from Experiments 2 and 3 expand upon the findings of previous research using the nonadjacency paradigm (Farbood, 2016; Spyra, Stodolak, & Woolhouse, 2019; Spyra & Woolhouse, 2018; Woolhouse, Cross, & Horton, 2016). Although the 6 s intervening section used within our experiments was well below the memory limits found by the above researchers, our study is the first to investigate the influence of timbre on the perception of key nonadjacency in music. The tripartite structure of the nonadjacency paradigm is analogous to the stimuli used in Mercer and McKeown (2010), in which participants compared initial and comparison probe tones over a 10 s interval, separated by a distractor tone. The distractor varied in the number of features it shared with the initial and probe tones; performance was degraded when the distractor either consisted of novel, unshared features or contained the distinguishing feature of the probe. The critical role of the intervening distractor tone suggests that a future study should analyze the potential effects of the intervening timbre on memory for key.

The multiple comparisons conducted in Experiment 3 indicate that the harmonic content of sound is encoded in memory with greater saliency than AEs. That said, it should be noted that the comparison in Experiment 3 between number of harmonics and AEs is not a fair test. Future studies might investigate the relative influences of harmonics and AEs to the perception of timbral difference—for instance, our harmonic manipulations may have simply outweighed those chosen for AEs, resulting in imbalanced contributions to the perceived timbre. Understanding the perceptual balance between harmonic and AE manipulations is crucial for the further refinement of our experimental paradigm. One might also expand our research through alternative dependent measures (e.g., musical tension, timbre recognition or discrimination) to parse the independent effects of timbral components on music perception.

References

Deutsch, D. (ed). (2013). *Psychology of music. (pp. 26–58)*. San Diego: Elsevier Academic Press. Elsevier.

Farbood, M. (2016). Memory of a tonal center after modulation. *Music Perception*, *34*(1), 71–93. doi:10.1525/mp.2016.34.1.71

Mercer, T., & McKeown, D. (2010). Updating and feature overwriting in short-term memory for timbre. *Attention, Perception, & Psychophysics*, *72*(8), 2289–2303. doi:10.3758/BF03196702

Siedenburg, K., & McAdams, S. (2017). The role of long-term familiarity and attentional maintenance in short-term memory for timbre. *Memory*, *25*(4), 550–564. doi:10.1080/09658211.2016.1197945

Spyra, J., Stodolak, M., & Woolhouse, M. (2019). Events versus time in the perception of nonadjacent key relationships. *Musicea Scientiae*, 1–14. doi:102986491986746

Spyra, J., & Woolhouse, M.H. (2018). Effect of melody and rhythm on the perception of nonadjacent harmonic relationships. In Parncutt, R., & Sattmann, S. (eds) *Proceedings of ICMPC15/ESCOM10*, 421–425. Graz, Austria: Centre for Systematic Musicology, University of Graz.

Tillmann, B., & McAdams, S. (2004). Implicit learning of musical timbre sequences: Statistical regularities confronted with acoustical (dis)similarities. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(5), 1131–1142. doi:10.1037/0278-7393.30.5.1131

Woolhouse, M., Cross, I., & Horton, T. (2016). Perception of nonadjacent tonic-key relationships. *Psychology of Music*, *44* (4), 802–815. doi:10.1177/0305735615593409

A New Test for Measuring Individual's Timbre Perception Ability

Harin Lee[1†] and Daniel Müllensiefen[1]
[1] Department of Psychology, Goldsmiths, University of London, London, UK
[†] Corresponding author: mu301hl@gold.ac.uk

Introduction

To date, tests that measure individual differences in the ability to perceive musical timbre are scarce in the published literature. The lack of such tool limits research on how timbre, a primary attribute of sound, is perceived and processed among individuals. We present a novel psychoacoustic assessment tool, the 'Timbre Perception Test' (TPT) (Lee & Müllensiefen, 2020), to fill the gap in the literature and to provide a robust measure that is specific to timbre and its distinct dimensions. This new test aims to examine an individual's perceptual abilities on three important dimensions of timbre (temporal envelope, spectral centroid, and spectral flux) initially proposed by McAdams et al. (1995). We employed a production adjustment task using a new interactive software interface and examined the validity and reliability of our newly developed test.

Method

The TPT was designed to measure participants' ability to reproduce a heard sound as closely as possible by utilising a movable slider that affects one sound dimension at a time (see Figure 1). The test was composed of three Blocks that correspond to the three dimensions of timbre (Temporal Envelope, Spectral Flux, and Spectral Centroid; presented in respective order) with each Block containing a learning trial, 5 match trials, and 10 memory trials.

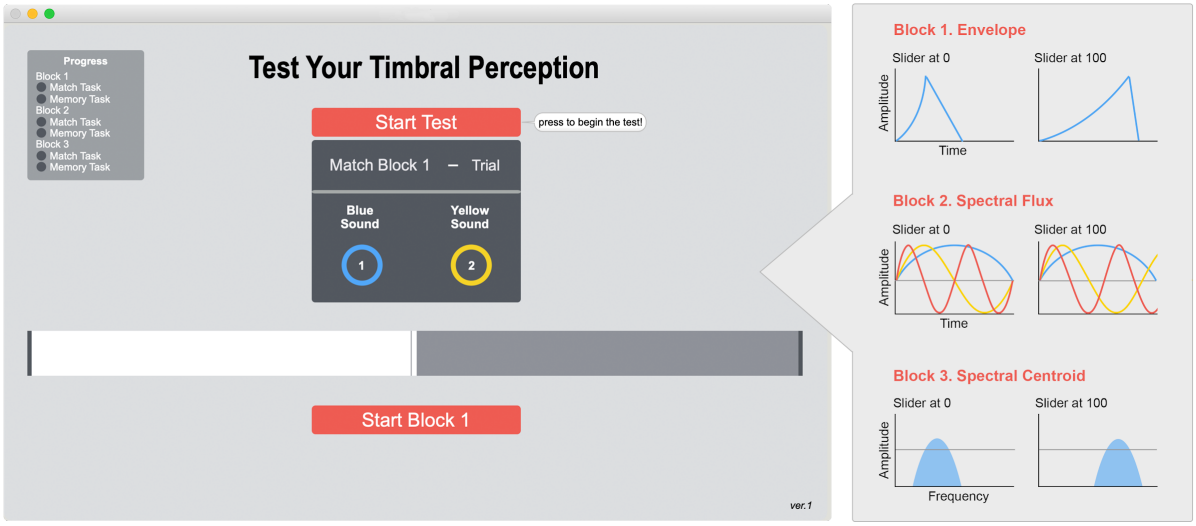


Figure 1. The layout of the TPT (left) and its testing dimensions (right). Graphic figures for the testing dimensions show how the reproduction tone is manipulated when the slider is positioned at '0' (far left) or positioned at '100' (far right). Envelope represent rise and fall time in amplitude, Spectral Flux represent the alignment of harmonics that results as more consonant when aligned in-phase, Spectral Centroid represent the filtered frequency area in the frequency spectrum.

The participant's task per trial was to first listen to the stimulus sound and then move the slider to adjust the reproduction sound to match the stimulus. Unlimited playback opportunities were given for the match trials, whereas only a single playback of stimulus was allowed for the memory trials.

In the Temporal Envelope subtask (Block 1), the slider bar altered the log attack time which also inversely influenced the decay time of the reproduction tone. In the Spectral Flux (Block 2) subtask, the ratios of harmonics to the fundamental frequency were altered to introduce dissonance caused by the beatings of frequency, characterised as *roughness*. Although such method of inducing roughness is strictly speaking - generating *inharmonicity* - we propose that this technique can be applied as one of the approaches to systematically vary the amount of spectral flux when working with synthetic sounds (see Appendix A for a spectral analysis). In the Spectral Centroid subtask (Block 3), a bandpass filter was applied to the source sound to alter its spectral centroid, which has shown to be a good predictor of the perceptual *brightness* of a sound.

A sample of 95 participants performed the TPT and also completed existing tests and questionnaire related to timbre perception, namely the auditory threshold tasks included in the PSYCHOACOUSTICS toolbox (Soranzo & Grassi, 2014), the Timbre subtest from PROMS (Law & Zentner, 2012), and the Gold-MSI self-report inventory (Müllensiefen, Gingras, Musil & Stewart, 2014).

Results

Factor analysis was conducted to determine whether the match and memory variants of the three testing dimensions of timbre can be summarised to measure the same construct. This revealed that match variants of all three subtasks loaded on a single factor, whereas memory variants showed heterogeneous and weaker factor loadings. Interpreting these results, we additionally constructed a short-version of the TPT that excludes all memory subtasks as well as several items from the match variants with low discriminatory power.

The results indicated that the short-version of the TPT has acceptable internal reliability (α = .69, ωt = .70) and good test-retest reliability (r = .79). Moreover, confirming its convergent validity, individuals' TPT scores were correlated with other related auditory tasks involving pitch discrimination (ρ = .56), duration discrimination (ρ = .27), and musical instrument discrimination abilities (i.e., Timbre subtest from PROMS, ρ = .33). Among all, the overall TPT performance showed the strongest relationship with self-reported levels of musical training (ρ = .64) and perceptual abilities (ρ = .56) that are hypothesised to be most closely related to timbre perception ability, and somewhat lower correlations with less related aspects of musical sophistication such as emotions (ρ = .40) and singing abilities (ρ =.45).

Performance accuracy increased when participants chose to listen to more repetitions of the target stimulus and the reproduction tones in the match condition. Considering the strong correlations with the self-reported perceptual ability, the observed correlations between stimulus repetitions and task performance could imply that the participants who were able to hear finer differences between the two tones repeated the tones a greater number of times to make more fine-grained adjustments to the slider position.

Assessed by the absolute slider distance between participant's slider position and target position, participants found the Temporal Envelope subtask to be easiest while Spectral Flux and Spectral Centroid subtasks to be of comparable difficulty. The accuracy to reproduce the timbre of tones was substantially reduced across all subtasks when performing the task from memory. More specifically, Temporal Envelope and Spectral Flux subtasks fell in accuracy by a similar amount while a smaller decrease was observed for the Spectral Centroid subtask (see Table 3 for acoustic values, Lee & Müllensiefen, 2020).

Discussion

Overall, the TPT has shown to be a promising tool for measuring individuals' timbre perception ability (see Lee & Müllensiefen, 2020 for an extensive discussion on correlations between TPT and related measures). Additionally, its use of a reproduction test paradigm and sliders to adjust timbral dimensions has the practical potential to combine short testing times (~8 minutes) with good measurement precision. We propose that the TPT can be broadly applied in the field of perceptual psychology to address outstanding questions on the individual differences on timbre perception. The current versions (full-version: including match and memory trials; short-version: including match trials only) of the TPT is openly available for

research purpose (download at www.osf.io/9c8qz) and its use does not require any coding skills, running as a standalone application on both Windows and Mac operating systems.

Acknowledgements

We thank all student assistance for testing participants and Kai Siedenburg for providing feedback during the early developmental stages of the TPT.

References

Law, L. N. C., & Zentner, M. (2012). Assessing musical abilities objectively: Construction and validation of the profile of music perception skills. *PLoS ONE*, *7*(12). https://doi.org/10.1371/journal.pone.0052508

Lee, H., & Müllensiefen, D. (2020). The Timbre Perception Test (TPT): A new interactive musical assessment tool to measure timbre perception ability. *Attention, Perception, & Psychophysics*. https://doi.org/10.3758/s13414-020-02058-3

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

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

Soranzo, A., & Grassi, M. (2014). PSYCHOACOUSTICS: A comprehensive MATLAB toolbox for auditory testing. *Frontiers in Psychology*, *5*. https://doi.org/10.3389/fpsyg.2014.00712

Appendix A. Graphical representation showing variations in spectral flux induced by the alteration of harmonic phase alignments

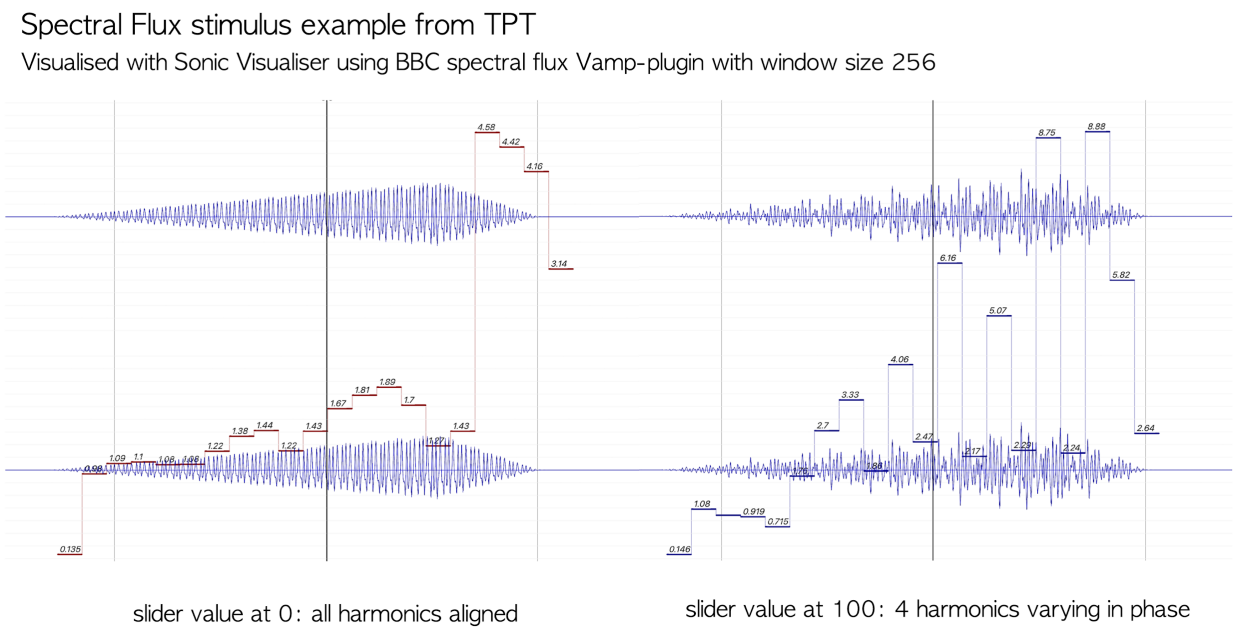


Figure A1. Variations in the level of spectral flux of an example stimulus used for testing Spectral Flux dimension in the TPT. Slider position at '0' (left sound wave) aligns all eight harmonics to the integer multiples of *f*0. Slider position at '100' (right sound wave) shifts the integer multiple ratios of 4 harmonics by a small degree.

Mapping Timbre Space in Regional Music Collections using Harmonic-Percussive Source Separation (HPSS) Decomposition

Kaustuv Kanti Ganguli[1†], Christos Plachouras[1], Sertan Şentürk[2], Andrew Eisenberg[1] and Carlos Guedes[1]
[1] Music and Sound Cultures research group, New York University Abu Dhabi, UAE
[2] Independent Researcher, UK
[†] Corresponding author: kaustuvkanti@nyu.edu

Introduction

Timbre — tonal qualities that define a particular sound/source — can refer to an instrument class (violin, piano) or quality (bright, rough), often defined comparatively as an attribute that allows us to differentiate sounds of the same pitch, loudness, duration, and spatial location (Grey, 1975). Characterizing musical timbre is essential for tasks such as automatic database indexing, measuring similarities, and for automatic sound recognition (Fourer et al., 2014). Peeters et al. (2011) proposed a large set of audio features descriptors for quantifying timbre, which can be categorized into four broad classes, namely temporal, harmonic, spectral, and perceptual. The paradigms of auditory modeling (Cosi et al., 1994) and acoustic scene analysis (Abeßer et al., 2017; Huzaifah, 2017) also have extensively used timbral features for the classification task. Timbre spaces, in the typical connotation (Bello, 2010), empirically measure the perceived (dis)similarity between sounds and project to a low-dimensional space where dimensions are assigned a semantic interpretation (brightness, temporal variation, synchronicity, etc.). We recreate timbre spaces in the acoustic domain by extracting low-level features with similar interpretations (centroid, spectral flux, attack time, etc.) by employing audio analysis and machine learning.

Based on our previous work (Trochidis et al., 2019), in this paper, we decompose the traditional mel-frequency cepstral coefficients (MFCC) features into harmonic and percussive components, as well as introduce temporal context (De Leon & Martinez, 2012) in the analysis of the timbre spaces. We will discuss the advantages of obtaining the stationary and transient components over the original MFCC features in terms of clustering and visualizations. The rest of the paper is structured in terms of the proposed methodology, experimental results, and finally, the obtained insights.

Method

Ganguli et al. (2020) explored cross-cultural similarities, interactions, and patterns of the music excerpts from the New York University Abu Dhabi Library Collection — the NYUAD music compendium — a growing collection with approximately 3000 recordings from the Arab world and neighboring regions, with a view to understanding the (dis)similarities by employing visualization and dimensionality reduction techniques. This study was limited to unsupervised clustering, and no experiments were carried out on derived temporal features (except log-STFT in entirety). Nevertheless, the timbre space model we applied successfully separated the data into meaningful clusters. A data-driven optimization showed that K=5 clusters (via K-Means clustering) captured the diversity of the corpus without over-fitting. The qualitative evaluation revealed interesting structures. For instance, one cluster included traditional instrumental string music and two other traditional Arab vocal, electronic, and pop music. Folk music excerpts with similar instrumentation from both the two archives were clustered together in the mapping. Figure 1 (left) shows the 2-dimensional t-Stochastic Neighborhood Embedding (t-SNE) representation of the timbre space for K=5, the encircled regions are similar in terms of the above descriptors. Figure 1 (right) shows that the intensity feature has a linear gradient. This trend indicates that there is a systematic variation of the intensity value along one axis which could help in interpreting the other axis by reverse engineering. The computation of intensity depends on the power spectrogram frames and differs upon decomposition of the spectrogram into its stationary and transient components.

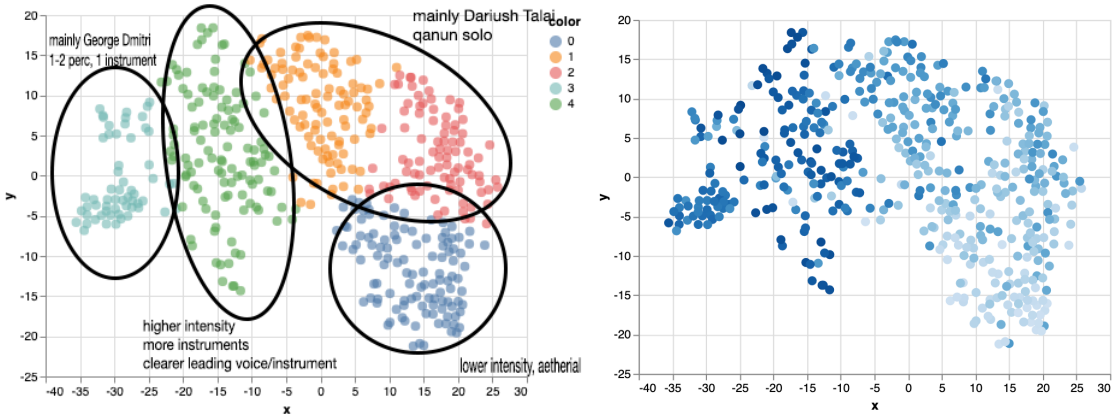


Figure 1: 2D t-SNE representation of the timbre space with K-Means (K=5) clustering on MFCC features, the encircled regions are marked as similar during qualitative evaluation (left). The intensity feature (computed as time-average of frame-wise energy sum) shows a linear gradient (right).

Fitzgerald (2010) proposed a harmonic-percussive source separation (HPSS) method commonly used in music information retrieval (MIR) to suppress transients when analyzing pitch content or suppress stationary signals when detecting onsets or other rhythmic elements. As the music corpus under study comprises a balanced mixture of harmonic and percussive instruments, we employ HPSS[1] to obtain two power spectrograms from each audio excerpt. Figure 2 (left) shows the power spectrogram and the derived harmonic/percussive components for a case-study excerpt. The corresponding MFCC vectors (n_mfcc=13) and the delta feature for the percussive feature vector are also shown in Figure 2 (right).

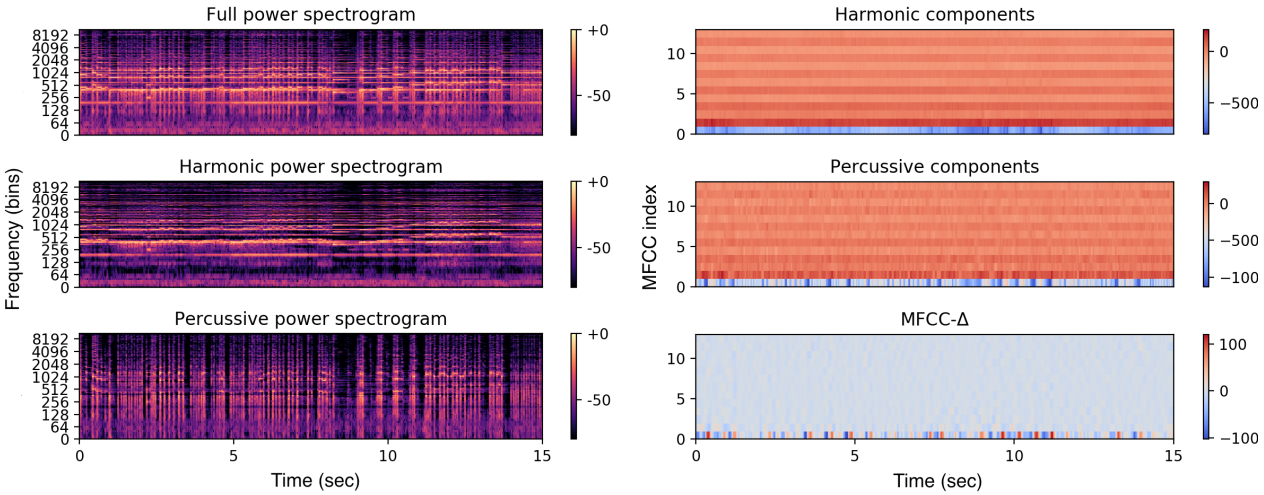


Figure 2: The power spectrogram and the derived harmonic/percussive components for a case-study excerpt (left). The MFCC vectors and the delta feature for the percussive feature vector (right).

The given case-study excerpt consists of both vocal and both melodic and percussive instruments. Some of the melodic instruments are also plucked-string in nature which raises an interesting scenario for the analysis — these instruments produce a wide-band attack at the onset and a stationary sustain before a fading release. The bass drums of the percussive instruments are not pronounced; we can thus safely assume these to be purely transient signals. This phenomenon is visible in Figure 2 (left) in the percussive power spectrogram where the transients are caused by both percussion and plucked-string melodic instruments.

[1] Using LibROSA (McFee et al. (2015)) version 0.8.0, DOI; 10.5281/zenodo.3955228 :: librosa.decompose.hpss

Results

We evaluate the clustering performance in terms of a homogeneity metric for the 2D t-SNE rendering for different spectrogram components. The cluster purity, which is defined as a measure of the extent to which clusters contain a single homogeneous class (Manning et al., 2008), is indicative of the timbre space being able to capture the intended groupings. The delta (differential coefficients, denoted as Δ) and delta-delta (acceleration coefficients, denoted as ΔΔ) features capture the spectrogram dynamics which is a proxy for a pseudo-temporal evolution of the (speech-like) music signal. Table 1 shows that adding delta features improve clustering performance. The harmonically enhanced features have shown overall better performance compared to its percussive counterpart.

Table 1: Cluster Purity for 2D t-SNE rendering for different components of the spectrogram.

No. of clusters (k)	**Full**		**Decomposed**	
	W/o delta	**W delta**	**Harmonic**	**Percussive**
2	*.84*	*.85*	*.85*	*.81*
4	*.86*	*.88*	*.87*	*.84*
6	*.89*	*.92*	*.91*	*.88*

This is, to some extent, intuitive as timbre modeling broadly captures the spectral envelope. However, there is further scope to experiment on hyperparameter tuning to investigate the percussive components. The double-delta parameters did not show significant differences. Table 1 also shows that the stationary components (harmonic MFCC) yield similar performance compared to the full spectrograms with delta features. It is, however, difficult to infer individual phenomena from the cluster purity metric, which involves a series of transformations and a machine learning model. On a corpus level, our proposed methods show better performance.

Discussion

We reported a modified timbre space for the NYUAD music compendium, a collection of recordings from the Arab world and neighboring regions, where a harmonic-percussive decomposition along with delta MFCC features show improvement in the clustering performance. The harmonic components outperformed the percussive components for the given metric; however, the percussive power spectrogram leads to a better tempo estimation and serves as a more reliable feature in rhythm-based studies. As mentioned before, the plucked-string melodic instruments produce both stationary and transient components, which may be utilized as complementary features. This is particularly important because there is hardly any timbre model found for non-Eurogenetic music cultures, whereas it might be easy to obtain a template for Western instruments. Hence, the proposed framework can be beneficial for regional music collections involving folk instruments. Extending from our previous work (Ganguli et al., 2020), we also plan to regenerate the timbre space model in the VR (virtual reality) space with the 3D t-SNE realization of the harmonic and percussive components obtained from the HPSS. This will lead to two independent timbre spaces that can aid in pedagogical as well as community engagement applications.

One drawback of the proposed approach is that it can only serve as a high-level exploratory data analysis tool since there are still not enough metadata regarding the style, genre, instrumentation, and structure of the archives. However, the HPSS decomposition can be particularly useful for an instrument recognition study. This can be better augmented with an audio thumbnailing task by discovering the predominant lead instrument of an excerpt followed by template matching of the representative frames. Finally, analysis of the perceptual timbre space (McAdams et al., 1995) is proposed as future work.

Acknowledgments

This research is part of the project "Computationally engaged approaches to rhythm and musical heritage: Generation, analysis, and performance practice," funded through a grant from the Research Enhancement Fund at the New York University Abu Dhabi.

References

Abeßer, J., Mimilakis, S. I., Gräfe, R., Lukashevich, H., & Fraunhofer, I. D. M. T. (2017). Acoustic scene classification by combining autoencoder-based dimensionality reduction and convolutional neural networks. in *Proceedings of the 2nd DCASE Workshop on Detection and Classification of Acoustic Scenes and Events* (pp.7-11). Munich, Germany.

Bello, J. P. (2010). Low-level features and timbre. Lecture notes MPATE-GE 2623 *Music Information Retrieval*, New York University.

Cosi, P., De Poli, G., & Lauzzana, G. (1994). Auditory modelling and self-organizing neural networks for timbre classification. *Journal of New Music Research, 23*(1), 71-98.

De Leon, F., & Martinez, K. (2012). Enhancing timbre model using MFCC and its time derivatives for music similarity estimation. In *Proceedings of the 20th European Signal Processing Conference (EUSIPCO),* (pp. 2005-2009). Bucharest, Romania.

Fitzgerald, D. (2010, September). Harmonic/percussive separation using median filtering. In *Proceedings of Digital Audio Effects (DAFX), Vol. 10*(4), 10–13.

Fourer, D., Rouas, J. L., Hanna, P., & Robine, M. (2014). Automatic timbre classification of ethnomusicological audio recordings. In *Proceedings of the 15th International Society for Music Information Retrieval (ISMIR) conference, (pp.*133-150).Tapei, Taiwan.

Ganguli, K. K., Gomez, O., Kuzmenko, L., & Guedes, C. (2020). Developing immersive VR experience for visualizing cross-cultural relationships in music. In *Proceedings of 2020 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops*, 401-406.Atlanta, Georgia.

Grey, J. M. (1975). An exploration of musical timbre. (Ph.D dissertation), Stanford University: Standford.

Huzaifah, M. (2017). Comparison of time-frequency representations for environmental sound classification using convolutional neural networks. arXiv preprint arXiv:1706.07156..

Manning, C. D., Raghavan, P., & Schütze, H. (2008). Evaluation in information retrieval. *Introduction to information retrieval, 1,* 188-210.

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

McFee, B., Raffel, C., Liang, D., Ellis, D. P., McVicar, M., Battenberg, E., & Nieto, O. (2015). LibROSA: Audio and music signal analysis in python. In *Proceedings of the 14th Python in science conference Vol. 8*, 18-25. Austin, Texas.

Peeters, G., Giordano, B. L., Susini, P., Misdariis, N., & McAdams, S. (2011). The timbre toolbox: Extracting audio descriptors from musical signals. *The Journal of the Acoustical Society of America, 130(5),* 2902-2916.

Trochidis, K., Russell, B., Eisenberg, A., Ganguli, K. K., Gomez, O., Plachouras, C., Guedes, C., & Danielson, V. (2019). Mapping the Sounds of the Swahili coast and the Arab Mashriq: Music research at the intersection of computational analysis and cultural heritage preservation, In *Proceedings of the Digital Libraries for Music (DLfM) conference*, The Hague, Netherlands.

There's more to timbre than musical instruments: semantic dimensions of FM sounds

Ben Hayes[1†], Charalampos Saitis[1]
[1] Centre for Digital Music, Queen Mary University of London, London, United Kingdom
[†] Corresponding author: b.j.hayes@se19.qmul.ac.uk

Introduction

Much previous research into timbre semantics (such as when an oboe is described as "hollow") has focused on sounds produced by acoustic instruments, particularly those associated with western tonal music (Saitis & Weinzierl, 2019). Many synthesisers are capable of producing sounds outside the timbral range of physical instruments, but which are still discriminable by their timbre. Research into the perception of such sounds, therefore, may help elucidate further the mechanisms underpinning our experience of timbre in the broader sense. In most timbre semantics research, listeners rate a set of sounds along scales defined by descriptive adjectives. By reverse engineering the standard paradigm, a smaller number of studies have provided evidence that musicians can manipulate timbre in abstract synthesis scenarios to match certain adjective descriptions. For example, Wallmark *et al.* (2019) employed a simplified FM (Frequency Modulation) synthesis interface to study the relationship between semantic descriptors and sound creation, showing an association between word valence and specific acoustic features.

In this paper, we present a novel paradigm on the application of semantic descriptors to sounds produced by experienced sound designers using an FM synthesiser with a full set of controls. FM synthesis generates rich and complex timbres via time-varying phase modulation of sinusoidal oscillators (Chowning, 1973), and is amenable to statistical analysis as broad timbral palettes can be expressed as a function of a completely continuous parameter space. Our aim with this work is twofold. First, we intend to ascertain whether the luminance-texture-mass (LTM) model of timbre semantics (Zacharakis *et al.*, 2014) is sufficient to describe the semantic dimensions of sounds produced through FM synthesis. Secondly, we hope the collected data and subsequent analysis will form a basis for future work into perceptually informed deep-learning based semantic synthesis control schemes.

Method

Thirty participants[1] completed the experiment (mean age 28.7 years; range 21-55 years). All spent their formative years in an English speaking country and self-reported having prior experience with synthesis through either music production or sound design backgrounds. Owing to the infeasibility of conducting an in-person study during the COVID-19 pandemic, the study took place online using the WebAudio API to generate sounds and the *lab.js* framework to collect data (Henniger *et al.*, 2020)[2].

The synthesiser employed a three operator architecture, with operators 2 and 3 modulating the phase of operator 1 in linear combination. Each operator had an independent *attack-decay-sustain-release* envelope. Participants were presented with a browser-based synthesiser interface with controls pre-set to generate a reference sound. An instruction was given to adjust the parameters such that the synthesiser produced a new sound matching a given comparative prompt (e.g. 'brighter' or 'less thick'). Each participant undertook nine trials covering each combination of three LTM prompts (*bright*, *rough, thick*) and three pitches (E2, A3, D5). Each trial, the positive or negative comparative form of the relevant prompt was selected randomly. Participants were then asked to rate the magnitude of the difference between the two sounds in terms of the prompt, as well as the difference between the created sound and the reference sound in terms of the remaining two LTM descriptors and an additional set of 24 semantic descriptors.

[1] Forty took part in total, but 10 were not used for analysis due to not meeting age or language restrictions
[2] Source code for the study is available in a GitHub repository: https://github.com/ben-hayes/fm-synth-study

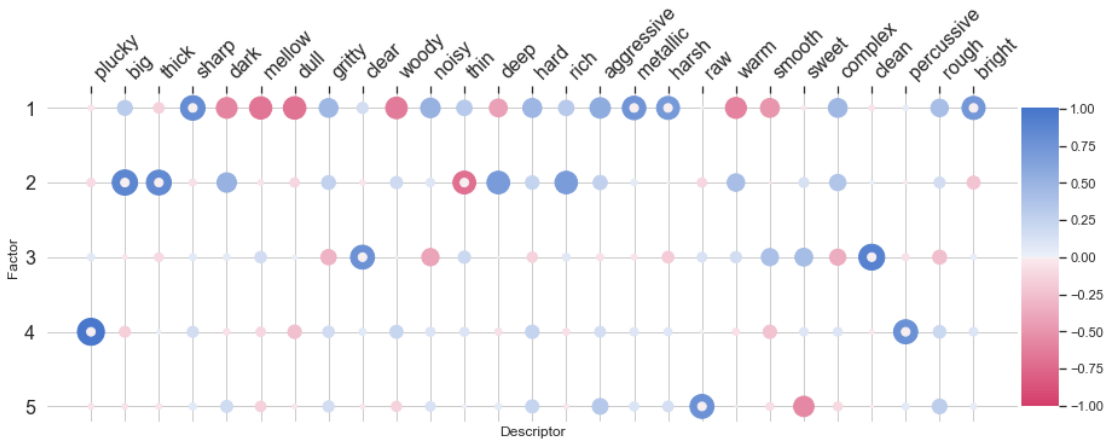


Figure 1: The factor loadings of the five factor solution across all descriptors. A white dot in the centre of a point indicates an absolute loading ≥ 0.7

Descriptors were selected for the experiment by mining and analysing a text corpus from the popular synthesis forum *MuffWiggler*. This approach was selected to maximise appropriateness to the sonic domain of synthesised sounds. The corpus was filtered to a frequency-sorted list of words co-occurring in bigrams with the terms 'sound', 'sounding', 'tone', and 'timbre', and this list was filtered so that only the top 100 adjectives remained. These were independently pruned by two raters according to a set of criteria, resulting in the final set of 27 descriptors. The LTM prompts were selected as the descriptors with the highest corpus frequencies that also showed significant loadings onto the English LTM factors in Zacharakis *et al.* (2014). To avoid biasing the semantic responses towards the characteristics of a fixed set of starting sounds, it was deemed advantageous to explore participants' responses across as much of the synthesiser's parameter space as possible. To this end, the reference sounds presented in each trial were randomly selected from the database of sounds created by previous participants, with the proviso that sounds may not traverse pitch conditions. As well as presenting a more balanced representation of the sonic properties of the synthesis method and participants' responses to LTM prompts, this approach confers the additional benefit that the resulting dataset is more amenable to future use in deep learning synthesis models.

An exploratory factor analysis with non-orthogonal oblimin rotation was performed on the resulting comparative descriptor ratings, using the maximum-likelihood method. The number of factors was selected using parallel analysis (Horn, 1965), a procedure in which the eigenvalues of the data correlation matrix are compared to those of a large number of correlation matrices generated from normally distributed random datasets via a Monte-Carlo simulation. The number of factors then corresponds to the number of eigenvalues from the real data's correlation matrix that exceed a given percentile (typically the 95th) of the synthetic data's eigenvalues. Parallel analysis has been shown to outperform the Kaiser method of retaining factors with eigenvalues greater than 1.0 (Zwick & Velicer, 1986). Finally, the monotonicity of the relationships between synthesiser parameter changes and descriptive prompts were studied by computing the Spearman rank correlation.

Results

Factor Analysis

Performed on all descriptors across all prompts, parallel analysis supported a five factor solution, using the 95th percentile as a threshold. The resulting factors cumulatively accounted for 74.36% of data variance. Fig. 1 illustrates descriptor loadings onto the rotated factors. Notably, factor 1 shows strong loadings onto terms associated with luminance (including *sharp*) as well as terms associated with texture (*metallic, harsh)*. Factor 2 shows strong loadings onto terms related to mass (*big*, *thick*, and negatively *thin*). Factor

Figure 2: Correlations between semantic factors.

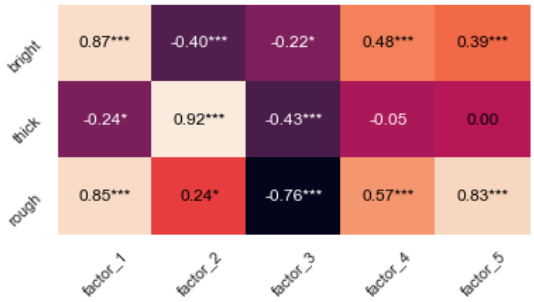


Figure 3: Pearson correlation coefficients between prompts and semantic factors. $p < 0.05$ (*), $p < 0.01$ (**), $p < 0.001$ (***).

3 shows strong loadings for words associated with clarity (*clean, clear*), factor 4 for "pluckiness" (*plucky, percussive*), and factor 5 for *raw*. Fig. 2 shows the inter-factor correlations after non-orthogonal Oblimin rotation. Here we see moderate collinearity between factors, most notably between factor 1 and factors 3-5. Fig. 3 shows the correlations between reported prompt magnitudes and semantic factors within each prompt condition. Each row, therefore, represents a non-overlapping subset of the dataset as each created sound was prompted by only one of the three LTM prompts. Factor 1 shows strong and significant correlations with *bright* and *rough* prompt magnitudes, and factor 2 with *thick.* Factors 3-5, however, all exhibit markedly different relationships with the prompts.

Parameter Correlations

Fig. 4 illustrates the Spearman rank correlation coefficients between reported prompt magnitudes and changes to synthesiser parameters within each prompt condition. The *bright* and *rough* prompts express very similar patterns of correlations, which both imply a tendency to, in response a positive prompt, increase the gains and tuning ratios of modulating operators (thereby increasing the energy and frequency of sidebands, respectively), and to decrease the attack time of the carrier operator (which decreases the overall attack time of the sound; cf. Saitis *et al.*, 2019). The most significant correlations observed for the *thick* prompt occur with the three sustain parameters, with the strongest of which was with the carrier operator sustain (which decreases the attenuation of the sustain portion of the sound).

Discussion

Comparing the loadings (Fig. 1) of factor 1 to those found by Zacharakis *et al.*, (2014) for terms present in both studies suggests it is an amalgamation of luminance and texture dimensions. The patterns of parameter delta correlations shared by the *bright* and *rough* prompts (Fig. 2) suggest that this is a direct result of the properties of FM synthesis: it is challenging to increase the energy in high frequency components (by increasing the modulator tuning or gain) without also increasing inharmonicity. The

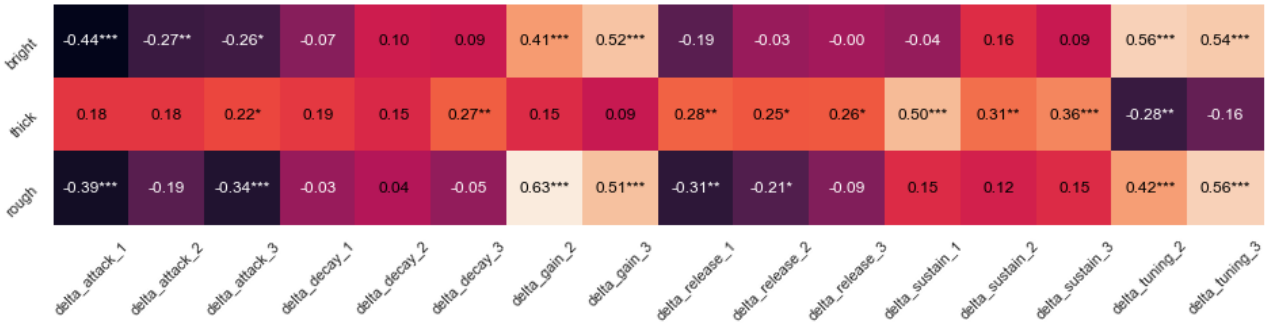


Figure 4: The Spearman rank correlation coefficients between prompt magnitudes and changes to synthesiser parameters. $p < 0.05$ (*), $p < 0.01$ (**), $p < 0.001$ (***).

loadings (Fig. 1) of factor 2 resemble the mass dimension of Zacharakis *et al.*, (2014). Factors 1-2 therefore appear to reflect the LTM prompts used for sound design and, indeed, the strong and significant correlations between these factors and the corresponding prompts (Fig. 3) support this hypothesis.

The loadings (Fig. 1) of factors 3-5 (*clean*, *plucky*, *raw*) suggest the created sounds exhibit attributes not entirely accounted for by LTM factors, and the correlations (Fig 3.) between these factors and prompt magnitudes support this. However, the moderate correlations between these factors and factors 1-2 (Fig. 2) suggest that these attributes are not entirely independent. Ascertaining whether this is due to an inherent property of the synthesiser or the interpretation of the descriptors themselves is left to future analysis.

Conclusions & Future Work

In this study we presented a novel paradigm for studying both the response of experienced sound designers to semantic prompts, and the semantic dimensions of the sounds they created. Exploratory factor analysis yielded a five factor model of which the first two factors correspond to the factors of the LTM model (factor 1: joint luminance-texture, factor 2: mass). The extra three factors appear to correspond, respectively, to clarity, pluckiness, and rawness. In subsequent analysis, acoustic features will be extracted from all synthesiser patches created in the study, enabling the psychoacoustic underpinnings of the semantic space to be analysed and, in particular, the relationship between factors 3-5 and factors 1-2. Owing to the design of this experiment — in particular, the fact that each stimulus is rated by only a single participant, and the use of comparative semantic ratings — it will be necessary to confirm its efficacy and the structure of the resulting semantic space with a classical semantic rating design (Zacharakis *et al.,* 2014) in order to compare the resulting factors to those found in previous studies.

Research into the semantics and perception of synthesised sounds provides a basis for future work into enhanced approaches for the control of synthesisers. Integration with semantic audio technologies and application of neural audio synthesis techniques will enable the intuitive generation and manipulation of novel timbres. Further, continued study of the broad and abstract sonic palettes afforded by synthesis methods such as FM will enable deeper insight into the intrinsic properties of timbre, as opposed to only those associated with physical sources, allowing for a more complete conception of the mechanisms underpinning its perception.

References

Chowning, J. M. (1973). The Synthesis of Complex Audio Spectra by Means of Frequency Modulation. *Journal of the Audio Engineering Society*, *21*(7), 526–534.

Henninger, F., Shevchenko, Y., Mertens, U., Kieslich, P. J., & Hilbig, B. E. (2020). *lab.js: A free, open, online experiment builder*. Zenodo.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, *30*(2), 179–185.

Saitis, C., Siedenburg, K., Schuladen, P., and Reuter, C. (2019). The role of attack transients in timbral brightness perception. In: Vorländer M., Fels J. (eds), *Proceedings of the 23rd International Congress on Acoustics*, (pp. 5506-5543). Aachen, Germany.

Saitis, C., & Weinzierl, S. (2019). The semantics of timbre. In K. Siedenburg, C. Saitis, S. McAdams, *et al.* (eds.), *Timbre: Acoustics, Perception, and Cognition* (pp. 119–149). Springer Handbook of Auditory Research, vol 69. Springer, Cham.

Wallmark, Z., Frank, R. J., & Nghiem, L. (2019). Creating novel tones from adjectives: An exploratory study using FM synthesis. *Psychomusicology*, 29 (4), 188–199.

Zacharakis, A., Pastiadis, K., & Reiss, J. D. (2014). An Interlanguage Study of Musical Timbre Semantic Dimensions and Their Acoustic Correlates. *Music Percept., 31 (4)*, 339–358.

Zwick, W. R., & Velicer, W. F. (1986). Comparison of five rules for determining the number of components to retain. *Psychological Bulletin*, *99* (3), 432–442.

Crossmodal language and onomatopoeia in descriptions of bird vocalization

Bodo Winter[1†] and Marcus Perlman[1]
[1] University of Birmingham, Dept. of English Language & Linguistics
[†] Corresponding author: bodo@bodowinter.com

Introduction

Timbre is often described using a rich inventory of words to express qualities of sound (e.g., onomatopoeia), as well as crossmodal language borrowed from senses of touch, vision and taste (e.g., Saitis et al., 2020; Wallmark & Kendall, 2018). In this study, we look at the use of onomatopoeia and other sensory language in texts describing bird vocalizations. Birders and ornithologists frequently engage in field trips to "spot" particular birds, where they often cannot see the birds, but have to rely on sound to identify the species. This practice has resulted in a long-lasting tradition of field guides which provide detailed descriptions of bird vocalizations. For example, the Harris's Hawk (*Parabuteo unicinctus*) is described in one guide as having a call that is a *harsh, grating "eeaarr"*, combining onomatopoeia (*eeaarr*) with words denoting its timbre (*harsh, grating*). The current study looks at the description of timbre in bird vocalizations in a corpus of bird field guides, drawing parallels to the description of timbre in music.

Method

We transcribed the 'voice' sections for 14 field guides, resulting in a corpus of 4,184 voice descriptions for 2159 bird species from all around the world, totaling 44,966 word tokens. To analyze the crossmodal nature of timbre descriptors, we used data from Lynott and Connell (2009), who asked native British English speakers to rate 423 property descriptors for visual, auditory, tactile/haptic, gustatory and olfactory content.

Results

Figure 1 shows a word cloud of the most frequent adjectives in our dataset. Many of these adjectives also occur frequently in studies of musical timbre. 73% of the instrument descriptors from Zacharakis et al. (2014) were attested in our corpus, such as *rough* and *dull*; 70% of the violin timbre descriptors from Fritz et al. (2012), such as *hard* and *raspy*; and 66% of the descriptors from Wallmark's (2019) analysis of orchestration pieces, such as *mellow* and *rich*. However, there are also differences to musical language. A comparison with the corpus frequencies of Wallmark (2019) shows that the words *sharp, soft, shrill, deep, thin, nasal, sweet, metallic*, and *hard* are significantly over-represented in descriptions of bird vocalizations compared to musical instrument descriptions (standardized Pearson's residuals > 2). On the other hand, *brilliant, dark, full, bright, sonorous, penetrating, clear, warm, dry,* and *piercing* are significantly under-represented.

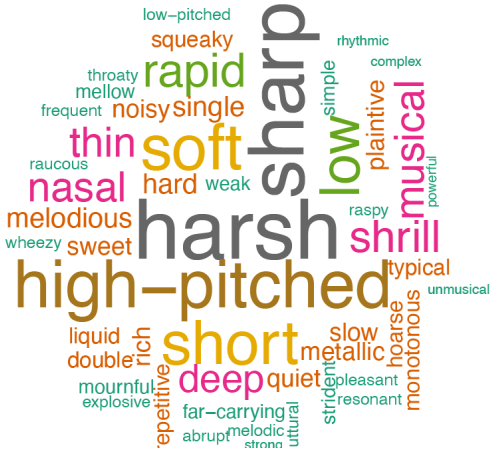


Figure 1: Word cloud of the most frequent adjectives in the bird vocalization corpus.

Besides the overlap in timbre descriptors, the parallels between bird vocalization descriptions and musical language run deep. Bird vocalizations are described as *song*, *carols*, *duets*, and *chorus*; there are references to scales and major/minor chords, as well as direct comparisons to musical instruments (*flute-like, trumpeting*). Also, the same way that the timbre features of instruments are commonly compared to human voices (what Wallmark, 2014, calls the INSTRUMENTS ARE VOICES conceptual metaphor), we find rampant comparison of bird vocalizations to human vocal production (*throaty, ventriloquial, nasal*).

Matching the words in our corpus to the sensory modality ratings shows that, predictably, auditory language is significantly over-represented (Pearson's residual = +19.4), but so is tactile language (+10.5) and visual language (+14.7). Taste (-20.56) and smell words (-24.05) were significantly under-represented, with the only frequent taste word being *sweet*. Lynott and Connell (2009) also provide a measure of 'modality exclusivity' that quantifies how crossmodal a word is. The language used to describe bird vocalizations is overall more crossmodal than the general list of sensory descriptors from their study ($t(2733) = 5.7$, $p < 0.001$), although this difference was not stark (40% 'exclusive' for our corpus, compared to 46% baseline, Cohen's $d = 0.3$).

Finally, we observed that 2,608 out of the 4,184 entries (62%) contained onomatopoeias, which are often combined with timbre words to describe a particular call or song, e.g., *a sharp "twissi-vit", a crisp "pik"*, and *a ringing "krrit"*. We suggest that in this context, timbre words make up for the absence of voice quality, which is a crucial feature of direct imitations, but which cannot be rendered easily in the written form without these words. Thus, onomatopoeias and timbre descriptors work together to signal different aspects of the complex multidimensional nature of bird vocalizations.

Discussion

Whereas most investigations of timbre have focused on music or environmental sounds, here we show that the language of bird vocalizations is a fruitful domain for studying the crossmodal and metaphorical nature of timbre descriptions. We show that the description of bird vocalizations has much overlap with descriptions of music. We furthermore advance the study of timbre semantics by borrowing methods from psycholinguistics, specifically, the use of modality rating studies to quantify the crossmodal nature of timbre language. Our analysis of how timbre descriptors are combined with onomatopoeia suggest that different linguistic strategies in this domain may communicate different aspects of the complex sound of birds, thus demonstrating the potential for trade-offs between different communicative practices.

References

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

Lynott, D., & Connell, L. (2009). Modality exclusivity norms for 423 object properties. *Behavior Research Methods*, *41*(2), 558–564.

Saitis, C., Weinzierl, S., von Kriegstein, K., Ystad, S., & Cuskley, C. (2020). Timbre semantics through the lens of crossmodal correspondences: A new way of asking old questions. *Acoustical Science and Technology*, *41*(1), 365–368.

Wallmark, Z. (2014). *Appraising timbre: Embodiment and affect at the threshold of music and noise* [PhD Thesis]. UCLA.

Wallmark, Z. (2019). A corpus analysis of timbre semantics in orchestration treatises. *Psychology of Music*, *47*(4), 585–605.

Wallmark, Z., & Kendall, R. A. (2018). Describing sound: The cognitive linguistics of timbre. In E. I. Dolan & A. Rehding (Eds.), *The Oxford Handbook of Timbre*. Oxford University Press.

Zacharakis, A., Pastiadis, K., & Reiss, J. D. (2014). An interlanguage study of musical timbre semantic dimensions and their acoustic correlates. *Music Perception*, *31*(4), 339–358.

Which Timbral Features Granger-Cause Colour Associations to Music?

PerMagnus Lindborg
School of Creative Media, City University of Hong Kong, HKSAR
pm.lindborg@cityu.edu.hk

Introduction

Sensory information processing is inherently multimodal. An organism normally perceives the environment using all its senses simultaneously. Crossmodal correspondence might take place at any stage of neural processing (Spence 2011; Deroy & Spence 2016), and studies have provided evidence that many non-arbitrary correspondences exist between auditory and visual stimulus features (Martino & Marks 2001; Bresin 2005; Palmer et al. 2013; Whiteford et al. 2018). However, few studies take electroacoustic music compositions as substrate for stimuli, despite the great variations of timbre within and across such works. This paper outlines an ongoing study of audiovisual correspondences through time series analysis. We investigate Granger-causality and other measures of association from time series analysis of multivariable acoustic features (describing nine music excerpts via timbral and dynamic features) and multivariate visual features (size and colour), collected in a perceptual experiment (N = 21) with a continuous response interface.

In previous work we charted colour-to-sound associations in film music (Lindborg & Friberg 2015) and in electroacoustic music (Lindborg, 2019a) with audio excerpts of around 15 seconds. Since the stimuli were fairly short they could be treated as independent data points in a random distributed variable. The present study extends these studies to longer music excerpts through time series analysis techniques. Listening to a recorded piece does not alter the audio file and in this sense the information flow is one-directional. Hence for our purposes the acoustic features may be treated as independent variables and colour responses as dependent variables in a time series regression analysis. A paper detailing experiments on continuous responses to music was published by Emery Schubert (1999) who then extended the analysis to time series (2001). Recent developments are in no small measure due to Roger Dean and collaborators (e.g. Dean & Bailes, 2010; 2011; Pearce ,2011; Bailes & Dean, 2012; Dean & Dunsmuir, 2016).

Materials

Audio excerpts of approximately three minutes duration were selected from nine electroacoustic pieces: well-known works by Chowning, Harvey, Risset, and Wishart, as well as recent pieces by Winderen, Martin, and the author. After normalisation by loudness (Nygren, 2009), they were presented in randomised order in an experiment (N = 21; eight females, median age 30, all right-handed, no reported colour vision deficiency or hearing impairment) following the same procedure as in (Lindborg, 2019a). While listening, participants manipulated two interfaces with the hands to control the size and colour of a visual object presented on a screen. Their task was to continuously match this object to the music.

Responses were sampled at 10 Hz, and colour was represented in *CIELab* which closely matches human perception (Hoffman, 2003; Shaw & Fairchild, 2002). It has three orthogonal dimensions that correspond to lightness (L), green-to-red (a), and blue-to-yellow (b). See example in Figure 1, left panel. Specifying colours within a perceptual scheme has advantages over parametric schemes (such as RGB or HSL) in terms of replicability and relevance to visual perception. Colour spaces and the design of the experimental response interface are discussed in (Lindborg & Friberg, 2015).

A large number of acoustic features were extracted computationally using the *MIR Toolbox* (Lartillot, 2013). Note that most are highly inter-correlated, due to the way the algorithms are structured. A selection of around 20 was made based on previously reported results, and to these we joined psychoacoustic descriptors extracted with *PsySound* (Cabrera et al., 2008). Time series were cleaned by imputing missing values (0.03% of responses in total) and handling outliers (altering 0.3% of the most extreme values two-

tailed, corresponding to trimming at ±3 SD in a normal distribution), and all series were down-sampled to a common rate of 4 Hz.

Methods

In empirical time series the elements almost always display some form of serial dependency. In our experimental setup, it is clear that if the music changes the visual response object is likely to be changed as well, but it will do so gradually, since the new position of the interfaces depend on their previous positions. Each new data point is to some degree correlated with the preceding ones: the time series is autocorrelated. In order to use parametric statistics to evaluate the degree of association between two time series, the values must be independent and identically distributed within each.

We conducted analysis in *R* (R Core Team, 2020) following the approach outlined in (Dean & Dunsmuir, 2016), with each music excerpt considered as a separate case study. For details on the statistical methods mentioned below see e.g. (Hyndman & Athanasopoulos, 2018) especially chapter 8, and (Box et al., 2015), especially chapters 4–5. See also (Jebb et al., 2015) for applications in psychology, and (Pfaff, 2008) for econometrics.

Time series were reshaped to achieve 'weak stationarity' by differencing. This removes trends in the data. In nearly all cases, d = 1 was an adequate degree, as judged by the KPSS and Augmented Dickey-Fuller tests. We then performed initial modelling tests including Granger causality (Granger, 1969) as an exploratory tool to assess whether the relationship between two stationarized series contains a causal element, at some lag. See Figure 1, right panel, for an example. However, it does not inform us about the strength of the predictive causation nor yield a transfer function that allows us to model the relationship.

In predictive regression modelling we need to be able to evaluate the cross-correlations between series, at a range of lags. Before significance levels of predictors can be correctly assessed the autocorrelation needs to be removed from at least one of the series being compared. This 'prewhitening' process involves modelling the autocorrelation structure so that the residuals display desired statistical properties, including serial independence, normal distribution, and heteroskedasticity. For each series, we estimated the autoregressive components with an ARIMA model. After obtaining reasonable maximum parameters (for p and q, since d was previously determined) from the autocorrelation and partial autocorrelation functions, we searched from one degree higher (p_{max}+1, d+1, q_{ma}x+1) down to (0, 0, 0) and then selected an optimal solution, as indicated by BIC, among those where the residuals passed the portmanteau Ljung-Box test on a range of lags. However, a fully automatic process might lead to overfit i.e. models that do not generalise well. Ultimately, our primary interest is mechanistic, seeking to identify psychological mechanisms by which crossmodal association processes might be explained. Robust predictive modelling is an important step towards this goal.

Therefore we are currently investigating the *auto.arima()* function (Hyndman et al., 2018) which implements a more powerful search method and an interface that is flexible when multiple exogenous predictors are included, i.e. ARIMAX. Since several acoustic features can potentially be influencing the colour response, a parsimonious set of predictors can be found by a systematic process of stepwise reduction, where predictor coefficient significance, error variance, and BIC are used as guides as to which predictors to include or exclude. For each case under study, the resulting model informs us of the predictive influence onto a dependent response variable from three sources: its own autoregressive function, a transfer function of the optimal set of acoustic predictor variables at different lags, and a white-noise error term. The transfer function is our focus of interest. The explanatory strength of the model is estimated from the cross-correlation series over a range of lags.

In our data the response is a multivariate time series {Size, L, a, b}. We are currently investigating univariate series separately, i.e. Size and a Change variable derived from the four, both for individual participants and for a group average (see example in Fig. 1). As the *CIELab* variables display multicollinearity, a full analysis requires a multivariate approach.

Discussion

The modelling approach outlined above assumes that the character of crossmodal associations at play are stable over time. It is an acceptable simplification given the experiment at hand, but the analysis of a natural situation calls for a dynamic approach. We have assumed that in the course of the visual association task, the listener "locks in" on something that s/he hears, and chooses a strategy, most likely intuitively, to match the colour response. As shown in (Lindborg & Friberg, 2015), emotion can be a strongly mediating factor. When presented with another stimulus, it might be that another acoustic feature takes on greater salience and a therefore a different matching strategy emerges. During the time the response is "locked in", a higher degree of influence of one or more acoustic features onto one or more of the response parameters will be observed.

The continuous response method supports situations where expert subjects are closely tracking their perception of spectro-morphological features of music. Using colour is a way to side-step semantic cognitive processing and can potentially reflect lower-level crossmodal processing mechanisms. Such methods can provide advantageous experimental tasks with non-expert subjects, or for those unable to use words (such as small children or stroke patients), or in experimental situations where the cognitive load of having to translate perceptions into semantic labels might distract from the task or from the act of listening itself (Lindborg, 2019b; cf. Saitis et al., 2019).

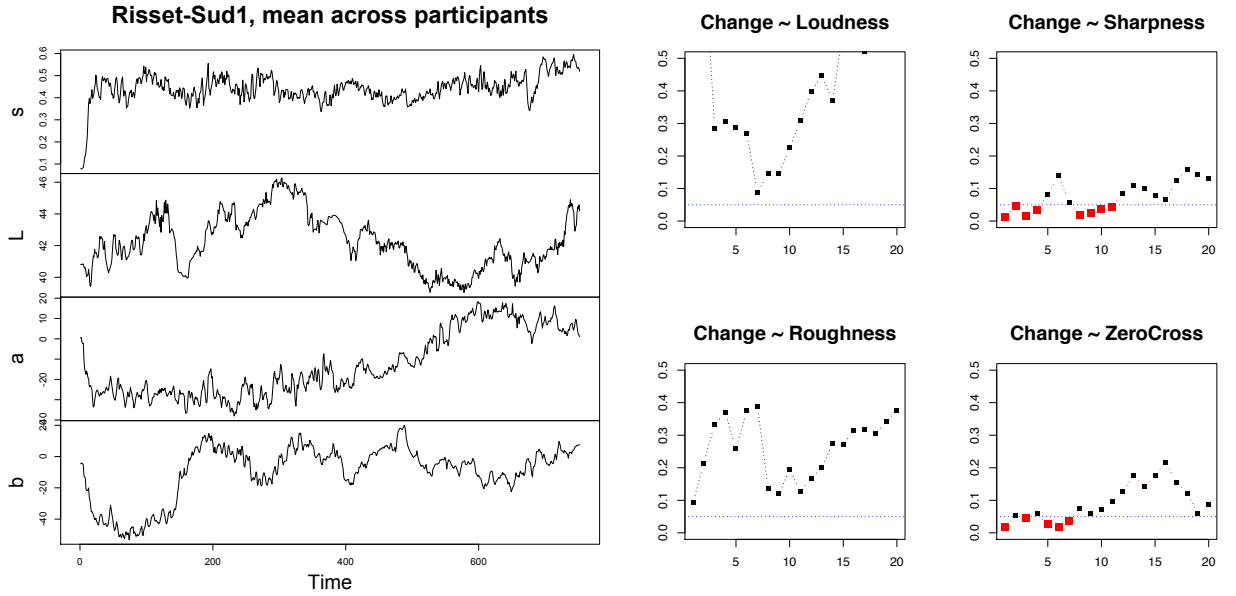


Figure 1: Plots for J.-C. Risset's Sud (three-minute excerpt from the beginning of the first part). Left: Size, L, a, b time series averaged across participants. Right: Change response Granger-caused by respectively Loudness, Sharpness, Roughness, and ZeroCross for a range of lags. Squares (red colour) below the dotted line indicate lags with a predictive-causal relationship significant at α = 0.05.

References

Bailes, F., & Dean, R. T. (2012). Comparative time series analysis of perceptual responses to electroacoustic music. *Music Perception: An Interdisciplinary Journal, 29*(4), 359-375.

Box, G. E., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). *Time series analysis: forecasting and control*: John Wiley & Sons.

Bresin, R. (2005). What is the color of that music performance? *Proceedings of the International Computer Music Conference*

Cabrera, D., Ferguson, S., Rizwi, F., & Schubert, E. (2008). PsySound3: a program for the analysis of sound recordings. *Journal of the Acoustical Society of America, 123*(5), 3247.

Dean, R. T., & Bailes, F. (2010, Oct.). Time series analysis as a method to examine acoustical influences on real-time perception of music. *Empirical Musicology Review, 5:4*, 152-175.

Dean, R. T., & Bailes, F. (2011, Apr.). Modelling perception of structure and affect in music: Spectral centroid and Wishart's Red Bird. *Empirical Musicology Review, 6:2*, 131-137.

Dean, R. T., & Dunsmuir, W. T. (2016). Dangers and uses of cross-correlation in analyzing time series in perception, performance, movement, and neuroscience: The importance of constructing transfer function autoregressive models. *Behavior research methods, 48*(2), 783-802.

Deroy, O., & Spence, C. (2016). Crossmodal correspondences: Four challenges. *Multisensory research, 29*(1-3), 29-48.

Granger, C. W. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: journal of the Econometric Society*, 424-438.

Hoffmann, G. (2003). *Cielab color space*. Retrieved from http://docs-hoffmann.de/cielab03022003.pdf

Hyndman, R., Athanasopoulos, G., Bergmeir, C., Caceres, G., Chhay, L., O'Hara-Wild, M., . . . Yasmeen, F. (2018). forecast: Forecasting functions for time series and linear models. R package version 8.4. *URL: https://CRAN. R-project. org/package= forecast*.

Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting: principles and practice*: OTexts.

Jebb, A. T., Tay, L., Wang, W., & Huang, Q. (2015). Time series analysis for psychological research: examining and forecasting change. *Frontiers in psychology, 6*, 727.

Lartillot, O. (2013). MIRtoolbox 1.5, User's Manual. *Finnish Centre of Excellence in Interdisciplinary Music Research*.

Lindborg, P. (2019a.). What is the Color of that Electroacoustic Music? *Proceedings of the International Computer Music Conference j.w. New York City Electroacoustic Music Festival*, New York, NY, USA.

Lindborg, P. (2019b). How do we listen? 에밀레 *Emille Journal of the Korean Electro-Acoustic Society, 16*, 43-49.

Lindborg, P., & Friberg, A. K. (2015). Colour association with music is mediated by emotion: Evidence from an experiment using a CIE Lab interface and interviews. *PloS One, 10*(12).

Martino, G., & Marks, L. E. (2001). Synesthesia: Strong and weak. *Current Directions in Psychological Science, 10*(2), 61-65.

Nygren, P. (2009). *Matlab code for the ITU-R BS. 1770–1 implementation. Appendix E to Master Thesis: Achieving Equal Loudness between Audio Files.* (MSc). KTH Royal Institute of Technology,

Palmer, S. E., Schloss, K. B., Xu, Z., & Prado-León, L. R. (2013). Music–color associations are mediated by emotion. *Proceedings of the National Academy of Sciences, 110*(22), 8836-8841.

Pearce, M. T. (2011, Apr.). Time-series analysis of music: Perceptual and information dynamics. *Empirical Musicology Review, 6:2*, 125-130.

Pfaff, B. (2008). *Analysis of integrated and cointegrated time series with R*: Springer Science & Business Media.

R Core Team. (2020). R: A language and environment for statistical computing. In. Vienna, Austria: R Foundation for Statistical Computing.

Saitis, C., & Weinzierl, S. (2019). The semantics of timbre. In *Timbre: Acoustics, perception, and cognition* (pp. 119-149): Springer.

Schubert, E. (2001). Continuous measurement of self-report emotional response to music. In P. N. Juslin & J. A. Sloboda (Eds.), *Series in affective science. Music and emotion: Theory and research* (pp. 393–414): Oxford University Press.

Schubert, E. (1999). Measuring emotion continuously: Validity and reliability of the two-dimensional emotion-space. *Australian Journal of Psychology, 51*(3), 154-165.

Shaw, M., & Fairchild, M. (2002). Evaluating the 1931 CIE color-matching functions. *Color Research & Application: Endorsed by Inter-Society Color Council, The Colour Group (Great Britain), Canadian Society for Color, Color Science Association of Japan, Dutch Society for the Study of Color, The Swedish Colour Centre Foundation, Colour Society of Australia, Centre Français de la Couleur, 27*(5), 316-329.

Spence, C. (2011). Crossmodal correspondences: A tutorial review. *Attention, Perception, & Psychophysics, 73*(4), 971-995.

Whiteford, K. L., Schloss, K. B., Helwig, N. E., & Palmer, S. E. (2018). Color, music, and emotion: Bach to the blues. *i-Perception, 9*(6), 2041669518808535.

Characterizing Subtle Timbre Effects of Drum Strokes Played with Different Technique

Francesco Bigoni[1], Michael Grossbach[2], and Sofia Dahl[1†]

[1] Department of Architecture, Design and Media Technology, Aalborg University, Copenhagen, Denmark
[2] Institute of Music Physiology and Musicians' Medicine, Hanover University of Music, Drama and Media, Hannover, Germany
[†] Corresponding author: sof@create.aau.dk

Introduction

In instrumental playing, musicians control sound characteristics such as loudness and timbre. While the many years of training for professional musicians ensure that most changes in timbre are intentional, others arise from changes in playing position, technique applied, or timing (Danielsen et. al 2015). Although subtle, such changes in timbre may still be heard and add to the quality of the performance, also for brief, percussive sounds. Drum strokes where the drumstick is allowed to freely rebound from the drum head – "normal" strokes – appear to have different audible quality compared to "controlled" strokes, where a player restrains the stick from freely moving up after the hit (Dahl & Altenmüller, 2008). Although audible, these differences in timbre are not always captured by features traditionally used such as log attack time and temporal centroid. An additional problem is the brevity of percussive sounds, making the use of those descriptors that require frame-based processing (e.g. spectral flux, spectral contrast) difficult (Bigoni & Dahl, 2018). In this context, it is crucial to identify the signal phases (e.g. attack vs. decay or transient vs. steady-state) in a way that is perceptually relevant and meaningful for drum sounds. The goal of this study is to evaluate whether audio descriptors that capture characteristics of the transient part of the waveform differ between Normal and Controlled strokes. Similar to Danielsen et al. (2015), we define this transient part to occur between the onset and the temporal centroid of each stroke, but we also separate the initial attack.

Method

The data consisted of a set of 1102 drum strokes played on a 14-inch rototom with instructions that determined the grip of the drumstick for each stroke, as described by Dahl & Altenmüller (2008). Eight professional players were instructed to play *Normal* (N) strokes, where the stick was allowed to freely rebound, whereas for *Controlled* (C) strokes, the player was instructed to control the ending position of the drumstick, stopping it as close as possible to the drum head after a stroke. A trial typically consisted of 10-13 strokes, where each stroke was allowed to ring out before the next one. The striking area of the drumhead was defined by a circle, 5 cm in diameter. An omnidirectional condenser microphone was mounted at a distance of 50 cm and angled 45 degrees with respect to the drumhead surface.

We separated each drum stroke into separate files and extracted audio descriptors using MIRtoolbox 1.7.2 (Lartillot, Toiviainen & Eerola, 2008). In order to separate the early energy part of the waveform, which we believe to be perceptually relevant, from the later tonal and "ringing" phase of each waveform, we defined algorithms for detection of onsets and offsets of each stroke. Rather than relying on amplitude envelopes for detection of onset and offset time (c.f. Nymoen, Danielsen & London, 2017) we used a threshold peak-picking algorithm directly on the waveform. To avoid spurious peaks in the waveform, the maximum peak time was estimated from a smoothed signal envelope instead. These time events, together with the temporal centroid (see Figure 1), were used to define four signal phases: 1) attack (max peak time - onset time), 2) early decay (temporal centroid - max peak time), 3) late decay (offset time - temporal centroid), 4) total (offset time - onset time). In all, 25 audio descriptors were extracted using the MIRtoolbox algorithms: duration (x 4 phases), sound pressure level (SPL, x 4 phases), spectral centroid (x 4 phases), temporal centroid, temporal flatness (i.e. geometric mean / arithmetic mean, calculated on the amplitude envelope values, x 4 phases), spectral flatness (i.e. geometric mean / arithmetic mean, calculated on the

spectral bin values, x 4 phases), and crest factor (i.e. max peak value / rms value, calculated on the waveform, x 4 phases).

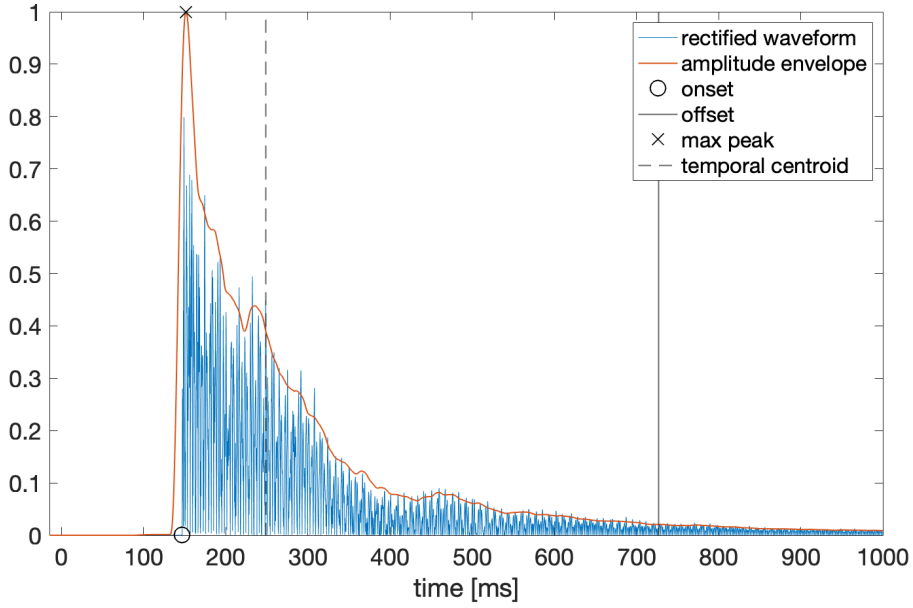


Figure 1: Rectified waveform of a stroke with the smoothed envelope. The markings show extracted onset (circle); maximum peak (cross); temporal centroid (dashed vertical line); and offset (full vertical line). For the modelling, we used descriptors based on the early decay phase (between max peak and temporal centroid).

Results

As expected, our descriptors differ between players in terms of range, and some are highly correlated. Judging from the density plots of the descriptors, we could not discern substantial differences between N and C strokes: that indicates that differences between players and playing arm influenced our data to a high degree.

Departing from our initial visual and auditory exploration of the data, as well as the reported results from Danielsen et al (2015), we started by modelling the spectral centroid of the early decay phase (earlyDecSC) in a multilevel Bayesian regression with a skewed normal link function and informative priors, using the package brms (Bürkner, 2017) in R (R Core Team, 2020). The best model according to a leave-one-out cross-validation (Vehtari et al., 2017) showed a clear difference in the estimated mean of earlyDecSC conditional on instruction (N or C), with the expected value for the dependent variable for N-strokes estimated to be 47.65 (Credible Interval [CI]: 58.31to 37.59) lower compared to that of C-strokes (Mean [CI]: 773.45 [759.67 to 788.26]). There was no interaction with the playing arm. Furthermore, the model suggests that the playing instruction had an effect on skewness as well as on spread, both increasing for C-strokes. Playing with the non-dominant arm also increased skew and spread, but to a lesser degree.

Discussion

Based on the model, we can infer that the playing instructions cause differences in the mean, the skewness, and spread of spectral centroid calculated across the early decay phase (between max peak and temporal centroid). The results show a substantially lower mean spectral centroid for normal strokes, with no effect from the playing arm. Since spectral centroid is commonly considered as a perceptual correlate of

brightness, our model agrees with our informal listening explorations, where we found controlled strokes to sound slightly harsher/brighter around the hit point and getting faster to the tonal, ringing phase compared to normal strokes. The listening test in Dahl & Altenmüller (2008), using a subset of the strokes from one of the players, reported the controlled strokes being rated as more flat or dull compared to normal strokes despite having higher peak force and shorter contact durations. Intuitively, one would think that stopping the drum stick as soon as possible after the impact would prolong the contact time, but the opposite was found for this player. Rather, a firm, "cramped" grip around the drum stick could alter the vibrational modes that influence the force pulse during contact with the drum head, thus altering the spectral centroid and perceived brightness.

The increase in skewness and spread for controlled strokes seem reasonable in that both inter- and intra-player variability would be more likely to increase in their execution of these strokes, but also that some players may have enforced the stopping of the drumstick to a higher degree than others. A rather high intra-class correlation coefficient for a null model, including only the subjects as grouping structure, indicates that more than 70 percent of the variation can be explained by the individual participants, most likely an effect of the small sample size in combination with high inter-individual variation.

Although collected from a small sample, our results provide further insights to how subtle timbre changes can be described for brief, percussive sounds. We will proceed to include temporal flatness and crest factor into a multivariate model over the early decay phase for normal and controlled strokes. A later listening tests will aim to verify that the investigated descriptors are useful as perceptual predictors. Although subtle, we argue that the changes are perceivable and that investigating suitable descriptors provide important knowledge on the link between action and perception in the control of musical instrument playing. For example, percussionists often train to produce even loudness and timbre across many repeated strokes, a skill very much needed in performance of pieces such as Ravel's *Bolero*.

Acknowledgments

The authors would like to thank Olivier Lartillot for sharing the updated version of *mirevents* before the release of MIRtoolbox 2.0. The original data was collected during a MOBILITY-2.1 Marie Curie Intra-European Fellowship (EIF) awarded to Dahl. The contribution by Dahl is partially funded by NordForsk's Nordic University Hub Nordic Sound and Music Computing Network (NordicSMC), project number 86892.

Author SD conceived and designed the study; SD & MG collected the data; FB preprocessed the data, performed feature extraction and initial analysis; MG made the statistical modelling; all authors participated in the interpretation of results and writing of the final manuscript.

References

Bigoni, F., & Dahl, S. (2018). Timbre Discrimination for Brief Instrument Sounds. In E. Gómez, X. Hu, E. Humphrey, & E. Benetos (eds.), *Proceedings of the 19th International Society for Music Information Retrieval Conference (ISMIR 2018),* (pp. 128–134). Paris, France.

Bürkner, P.-C. (2017). brms: An R Package for Bayesian Multilevel Models Using Stan, *Journal of Statistical Software,* Foundation for Open Access Statistic, 80.

Dahl, S., & Altenmüller, E. (2008). Motor control in drumming: Influence of movement pattern on contact force and sound characteristics. In *Proceedings of Acoustics'08, Intl. Meeting Acoustical Society of America, ASA, the European Acoustics Association, EAA, and the Société Francaise d'Acoustique, SFA.*, (pp. 1489–1494). Miami, Florida.

Danielsen, A., Waadeland, C. H., Sundt, H. G., & Witek, M. A. (2015). Effects of instructed timing and tempo on snare drum sound in drum kit performance. *The Journal of the Acoustical Society of America*, *138*(4), 2301–2316.

Lartillot, O., Toiviainen, P., & Eerola, T. (2008). A matlab toolbox for music information retrieval. In: Preisach, C., Burkhardt, H., Schmidt-Thieme, B., Decker, R. (eds), *Data analysis, machine learning and applications* (pp. 261-268). Springer, Berlin, Heidelberg.

Nymoen, K., Danielsen, A., & London, J. (2017). Validating attack phase descriptors obtained by the Timbre Toolbox and MIRtoolbox. In: Tapio Lokki, Jukka Pätynen, and Vesa Välimäki (eds), *Proceedings of the SMC Conferences* (pp. 214-219). Aalto University, Finland.

R Core Team (2020). R: A Language and Environment for Statistical Computing.

Vehtari, A., Gelman, A. & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC, *Statistics and Computing*, *27***,** 1413–1432.

On the difficulty to relate the timbral qualities of a bowed-string instrument with its acoustic properties and construction parameters

Claudia Fritz
Institut Jean le Rond d'Alembert, Sorbonne Université / CNRS, Paris, France
claudia.fritz@sorbonne-universite.fr

Introduction

A long-standing goal in musical acoustics is to identify relations between the perceived qualities (timbre, playability, …) of a given individual musical instrument (as evaluated by players and listeners) compared to other instruments of the same family and the description of its structural and acoustic properties. Here, we are not comparing the timbre of different instruments as explored through the evaluation and measurement of a single note of for example a clarinet versus a single note of a trumpet, but the timbral qualities of one violin (across the whole register) with the qualities of another. Our aim is to identify the acoustic phenomena that can account for these perceptual timbre differences, which could provide insight to violin makers about how to empirically modify some construction parameters in order to reach some desired sound qualities.

Method

While better understanding the perceived qualities of instruments from the violin family and linking them to acoustic properties has been the core of the author's research for over a decade [e.g. references below], the task has proven really challenging and there are, in the end, few clear-cut results. One of the reasons is that players and listeners hardly agree among themselves. Another reason is that any two given instruments differ by so many parameters that it is hard to know what perceived quality (when there is some agreement) is due to what acoustical parameter. Two recent projects, one on the viola (Obialto project), one on the violin (Bilbao project), have been partially designed to address this issue and to explore links between perceptual properties with construction characteristics (and not only acoustical parameters).

While the outline geometry differs slightly between violins, it can vary considerably between violas, which are less standardized. During the 2016 Oberlin workshop (organized by the Violin Society of America), a group of instrument makers have collectively designed the so-called Obialto outline. 25 violas were then built following this model (but without any other constraint except for the set of strings) and brought to the 2017 workshop during which two short excerpts (one in the low register, one in the high) were recorded by a professional player in a recording studio. The recordings were used in a series of listening tests, based on a free categorization task. The data were analysed in terms of statistics (leading to hierarchical trees) and linguistics (analysis of the verbal descriptions of the different classes). In addition, various audio descriptors were calculated on the recordings and vibro-acoustical measurements made on the instruments were processed to characterize their main low frequency resonance modes. Relationships were then searched between the set of perceptual data, the acoustical descriptors and measurements, and the constructional parameters.

The Bilbao project aims at relating intrinsic characteristics of the materials (wood density and stiffness) and some geometric characteristics of the violin's constituent parts (thicknesses of the plates) with the tonal qualities of the complete violins. To this end, six instruments were carefully built at Bilbao's violin making school (BELE): three violins with normal backs, each paired with a thin (pliant), normal or thick (resistant) top; similarly, three with normal tops, each paired with a thin, normal or thick back. The two examples of normal top paired with normal backs serve as a control. Wood for tops and backs were closely matched in density and sound speeds – all tops and backs from the same trees. Greater control was achieved by having all plates and scrolls cut by Computer Numerical Control routers. The outside surface was not modified, as the graduation was performed entirely on the inside surface. In addition, structural measurements were taken at many steps during the building process and radiation measurements were taken at the end. Another

set of six violins was built by six "external" makers, with slightly less constraints in order to have a wider range of plate thicknesses.

This set of violins has been used in a series of experiments; playing tests (a free categorization task involving 20 players in Bilbao and a preference sorting experiment involving 18 players and makers in Oberlin) as well as listening tests in two concert halls, in Bilbao (about 60 listeners) and Oberlin (about 50 listeners).

Results and discussion

In the light of some results of these perceptual tests conducted on violins and violas, we will show that no strict mapping between the perceived timbre and audio descriptors as well as vibro-acoustical measurements could be found, due to the complexity of the concept of timbre that induce a large variability between musicians: the lack of consensus between the musicians' evaluation criteria actually results from the diversity of interpretations when evaluating the timbre of an instrument globally (across the whole register, for different dynamics and playing techniques, …) and assessing the subtle acoustic differences between instruments of the same family. In addition, the multiplicity of the parameters during the building process allow instrument makers to obtain a certain set of perceptual qualities with very different strategies, which makes it difficult to find relationships between perceived qualities and construction parameters if strict constraints have not been imposed on the latter (like in the Bilbao project).

Acknowledgments

The author would like to thank her collaborators (D. Dubois, F. Krafft, G. Stoppani, U. Igartua, R. Jardon Rico,) and students (V. Fraisse, P. Cerântola, B. Souchu, A. Verrier and R. Montero Murillo), as well as all the makers, players and listeners who took part in these projects and experiments.

References

Fritz C., Curtin J., Poitevineau J. & Tao F.-C. (2017). Listener evaluations of new and Old Italian violins. In: Purves D. (ed), *Proceedings of the National Academy of Sciences,* 114, 5395-5400.

Saitis C., Fritz C., Scavone G.P., Guastavino C. & Danièle Dubois (2017) Perceptual evaluation of violins: A psycholinguistic analysis of preference verbal descriptions by experienced musicians. *Journal of Acoustic Society of America,* 141, 2746-2757.

Wollman I., Fritz C., Poitevineau J. & McAdams S. (2014) Investigating the Role of Auditory and Tactile Modalities in Violin Quality Evaluation. *PlosOne 9*(12): e112552.

Fritz C., Curtin J., Poitevineau J., Borsarello H., Wollman I., Tao F.-C. & Ghasarossian T. (2014) Soloist evaluations of six Old Italian and six new violins. In: Purves D. (ed), *Proceedings of the National Academy of Sciences,* 111, 7224-7229.

Saitis C., Giordano B.L., Fritz C. & Scavone G.P. (2012) Perceptual evaluation of violins: A quantitative analysis of preference judgments by experienced players. *Journal of Acoustic Society of America,* 132, 4002-4012.

Fritz C., Curtin J., Poitevineau J., Morrel-Samuels P. & Tao F.-C. (2012). Players preferences among new and old violins. In: Purves D. (ed), *Proceedings of the National Academy of Sciences,* 109, 760-763.

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

One singer, many voices: Distinctive within-singer groupings in Tom Waits

Joshua Albrecht
Hugh A. Glauser School of Music, Kent State University, Kent, OH, USA
jalbrec6@kent.edu

Introduction

Timbre plays an essential role in popular music (Fink *et al.,* 2018, p. 2), in which a unique sound can craft a distinctive identity for a band or artist even as other musical features show less variability, such as stock chord progressions and more limited melodies. A singer's vocal quality alone can create a recognizable persona (Tagg, 2012, p. 350) and become associated with the singer's brand. Moreover, the human voice is one of the most sophisticated and musically relevant sources of timbral variety in music, and humans are remarkably adept at deciphering timbral cues in voices. Hughes *et al.* (2004) found that listeners are consistently accurate at inferring speakers' age, height, weight, socio-economic status, personality traits, and emotional and mental states from listening to recordings alone. These results suggest that one may be able to define "sonic fingerprints" for individual singers by examining distinctive acoustic characteristics of their voice, at least in theory. However, pinpointing the sonic markers that distinguish one singer from another by examining only the audio source can be difficult, and the acoustic parameters associated with recordings of different singers are likely confounded by many other recording artifacts.

An alternative approach would control for *singer* and focus on different timbral approaches the same singer uses to influence voice quality. Several singers are known for mastery over a wide range of vocal timbres, such as Billy Joel (Duchan, 2016) and Bob Dylan (Rings, 2013). One of the more extreme cases is that of Tom Waits. One of the most immediately recognizable aspects of his music is his distinctly rough vocal timbre(s). Solis (2007) argues that Waits actually uses many different voices, each distinct and recognizable. While rock music since the 1960's typically claims some form of implicit autobiography, Waits' songs by contrast are (often overtly) inhabited by fictional personas who speak in distinct voices, both figuratively (lyrical meaning) and literally (unique vocal timbres). However, many of his songs share similar vocal timbres, suggesting links or shared meaning between them. The music of Tom Waits provides an interesting case study to compare groups of recordings that are perceived to have similar vocal timbres against other groups that are perceived as having different timbres, but yet controlling for possible confounds associated with different singers.

Method

The ideal approach for a study like this would be to locate masters of recordings and isolate the voice. Vocal tracks could then be subjected to automatic feature extraction and/or a perceptual study. Unfortunately, all of my attempts to contact Waits's studios have been unsuccessful, and so all tracks examined consisted of Waits's voice along with instrumental accompaniment. Without being able to isolate the voice, automatic feature extraction would be problematic and similarities between recordings would be likely to be strongly influenced by instrumentation rather than vocal quality alone. However, human listeners are exceptionally skilled at auditory scene analysis and can sift timbral properties of voices out of complex acoustical environments (Bregman, 1990). Consequently, this study used a perceptual method with human participants who listened to excerpts and sorted them by hand. Due to the important role that semantic description plays in capturing perceptually relevant timbral properties (Saitis & Weinzierl, 2019), participants also provided descriptive labels for their timbral categories

Sample:
Tom Waits's entire output, consisting of 255 vocal tracks over 19 studio albums, was too large for the present study. However, his career is often divided into two phases. The second phase, beginning after Waits married his wife Kathleen Brennan who encouraged him to experiment with a more adventurous

range of vocal timbres and instruments, is more relevant to the current study. Beginning with the album *Swordfishtrombones* (1983), Waits produced 146 tracks with voice over ten albums. For this study, the first five seconds that included at least 4 seconds of vocal sound from each of these 146 tracks was sampled.

Participants:

A total of 134 undergraduate music majors participated, 73 from the University of Mary Hardin-Baylor and 61 from Kent State University. 84 of the participants were female, and 50 were male, with a mean age of 20.7 (sd = 5.5). 101 participants reported Rock music as the musical genre they primarily listened to, with 21 reporting classical music, 11 jazz, and 1 not reporting. Discerning timbral differences is a challenging skill, and so musical sophistication was an important consideration for participant selection. The majority of participants were music major undergraduates, and participants averaged a Goldsmith's Musical Sophistication Index for general musical sophistication of 97.3 out of 126 total (sd = 11.7). Average scores of subsets of the Gold-MSI measure included 47.8 out of 63 for active engagement (sd = 6.9), 51.0 out of 63 for perceptual abilities (sd = 6.1), 36.7 out of 49 for musical training (sd = 6.8), 34.9 out of 42 for emotions (sd = 4.7), and 35.5 out of 49 for singing abilities (sd = 6.4).

Procedure:

Participants were provided 40 randomly-selected five second excerpts from the full set, represented as boxes on a computer screen (see Figure 1). The interface was run on a private Amazon Web Services website through the Google Chrome browser. UMHB participants were seated in groups of 8 in a computer lab using headphones. Due to the COVID-19 pandemic, KSU participants used their personal computers, but were instructed to complete the experiment in one sitting in a distraction-free environment, using headphones. After listening to their excerpts in numerical order, participants could drag the boxes into as many as eight groups or categories and re-listen as many times as needed by clicking on the box's play button. Participants were told to sort the excerpts according to vocal timbre and to ignore instrumentation, texture, meaning of the text, and genre. After the initial sort, participants were presented with each of their assembled groups in the second phase, listening to all excerpts in the group in random order. They were then asked to provide the best description of the vocal timbre of the excerpts in the group. In the third phase, participants progressively merged groups until there were only two left.

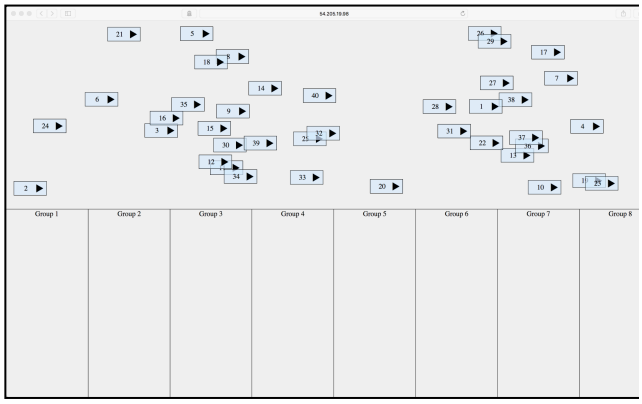


Figure 1: The interface. 40 randomly-selected excerpts scattered across the top of the screen and participants freely dragged them into as many as eight categories.

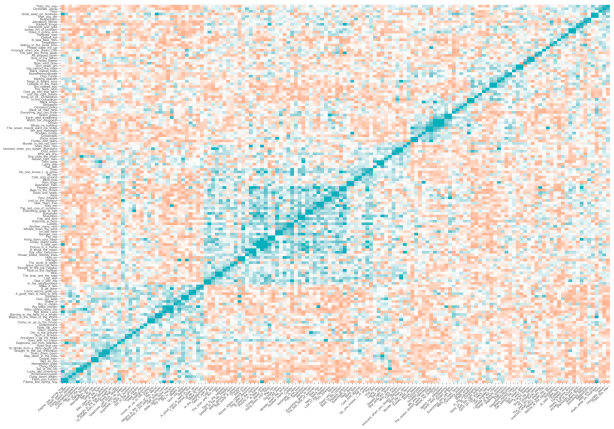


Figure 2: Dissimilarity matrix of the participant grouping responses for the entire dataset. Max dissimilarity is in red and minimum in blue.

Results

Participant data were used to estimate the timbral similarity between each pair of excerpts. Similarity measures were obtained by calculating the number of times two excerpts were grouped together at any stage of the grouping procedure for any participant by the number of times they *could have been* grouped

together. For example, if two excerpts were always grouped together every time they were co-present, they would have a similarity of 1, and if they were never grouped together in any condition, they would have a similarity of 0. Every possible two-excerpt pair were presented together in the same experiment at least four times. The grouping data were subjected to cluster analysis, in which the proportion of excerpt grouping was treated as the distance measure. Cluster analysis require *dis*similarity data, so the similarity scores were subtracted from 1. In other words, if two excerpts were never grouped together, their dissimilarity was 1. The dissimilarity matrix for preliminary results on data from the 72 UMHB participants for all 146 excerpts appears in Figure 2. Maximum dissimilarity is shown in orange and minimum dissimilarity is shown in blue.

It is not straightforward to determine the optimal number of clusters in a cluster analysis. A number of metrics have been proposed to provide an empirical means of determining the optimal number of clusters, though the intuition of the research remains an important consideration. The gap statistic is a test of how many clusters are most appropriate in a given dataset by comparing the total within-cluster variation for different numbers of clusters against the expected values under null reference distribution of the data (Tibshirani *et. al* 2001). The gap statistic for the preliminary dataset suggests an optimal number of seven clusters (see Figure 3). Hierarchical clustering dendrograms are also useful to visualize how many clusters appear appropriate for the data. A dendrogram showing Ward's method presented in circular form for reasons of space appears in Figure 4 with the 7-cluster solution shown. This solution appears robust, with only a 3-cluster solution appearing to have more distance between cluster heights. After verifying the appropriateness of 7 clusters, a separate k=7 k-means cluster analysis was conducted. A 3D multidimensional scaling of the results are provided in Figure 5. Clustering is apparent, but weak.

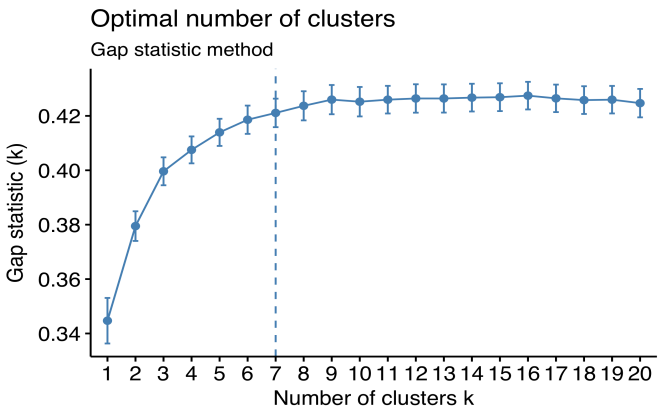


Figure 3: Dissimilarity matrix for the participant grouping responses. Maximum dissimilarity is in red.

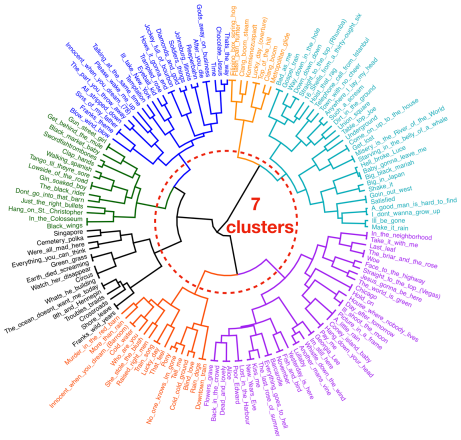


Figure 4: Ward's hierarchical cluster results for seven clusters

Once initial groups were established, each participant provided a descriptive label of the timbral quality of each excerpt as a free response. These responses were cleaned by removing all words that were not descriptors (generally, adjectives were kept and other words removed). When appropriate, nouns were turned into adjectives, and adjective modifiers like "really" or "very" were removed. When there were two adjectives in different forms, the simpler was retained, so "stranger" was changed to "strange." After cleaning the timbral descriptors, there remained 6,124 total terms and 407 unique terms describing Waits's vocal timbres of the preliminary dataset. The seven most common terms for each of the seven k=7 k-means clusters are displayed in Table 1. Of note is that "raspy" is the most common descriptor for all seven clusters, appearing a remarkable 9.5% of the total dataset. But, the rank-ordering of the next most common terms for each cluster reveals interesting differences between patterns of perceptions of vocal timbre for each cluster. For example, after raspy, C1 is described as deep, smooth, and breathy, whereas C6 is rough, growly, screaming, and harsh, and C4 is speech-like, deep, and low. The rank-ordered differences are consistent with different timbral characteristics for each cluster.

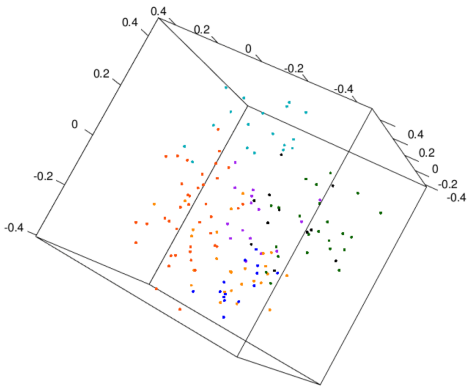


Figure 6: 3D multi-dimensional scaling for k=7 k-means clustered data.

C1 (33)	C2 (22)	C3 (15)	C4 (21)	C5 (16)	C6 (16)	C7 (23)
raspy (150)	raspy (83)	raspy (48)	raspy (93)	raspy (58)	raspy (52)	raspy (97)
deep (76)	nasal (45)	smooth (35)	speech-like (42)	rough (18)	rough (31)	growly (38)
smooth (74)	rough (29)	breathy (31)	deep (33)	deep (15)	growly (25)	rough (38)
breathy (63)	harsh (19)	soft (21)	low (29)	smooth (15)	scream-ing (21)	deep (27)
soft (45)	speech-like(19)	relaxed (16)	soft (27)	breathy (14)	harsh (15)	yelling (22)
low (38)	jazzy (18)	light (15)	breathy (24)	growly (14)	nasal (14)	gravelly (20)
scratchy (30)	light/ growly(13)	airy (11)	smooth (24)	heavy (13)	scratchy (13)	husky (17)

Table 1: The most common timbral description terms by cluster.

Discussion

Participant grouping data for the corpus of Tom Waits's studio songs is consistent with the existence of a small number of discernible groups based on the vocal timbre used, suggesting that this dataset could be a useful tool for examining inter-singer timbral differences. The qualitative descriptions provided by participants reveals striking differences between groups from the same singer. While the preliminary data suggests that seven groups is an appropriate number of song subsets in the Waits corpus, more refined subsets could be possible when the additional 61 participant dataset is added to the analysis. A close examination of Figure 2 reveals the presence of dozens of small clusters of 3-5 songs that are highly similar.

References

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

Duchan, J.S. (2016). Depicting the working class in the music of Billy Joel. In K. Williams & J. A. Williams (eds.), *The Cambridge Companion to the singer-songwriter*. Cambridge: Cambridge University Press.

Fink, R., Latour, M., & Wallmark, Z. (2018). *Timbre in popular music: The relentless pursuit of tone*. Oxford: Oxford University Press.

Hughes, S.M., Dispenz, F., Gallup, G.G. (2004). "Ratings of voice attractiveness predict sexual behavior." *Evolution and Human Behavior*, *25*, 295-304.

Montandon, M. (ed.) (2005). *Innocent when you dream: The Tom Waits Reader.* New York: Carroll & Graf.

Rings, S. (2013). A foreign sound to your ear: Bob Dylan performs "It's alright, ma (I'm only bleeding)," 1964-2009. *Music Theory Online*, *19* (4).

Saitis, C., & Weinzierl, S. (2019). The semantics of timbre. In K. Seidenberg, C. Saitis, S. McAdams, A. Popper, & R. Fay (eds.), *Timbre: Acoustics, Perception, and Cognition*. (pp. 119–149). Springer Handbook of Auditory Research. Springer, Cham.

Solis, G. (2007). "Workin' hard, hardy workin'/ Hey Man, you know me": Tom Waits, sound, and the theatrics of masculinity. *Journal of Popular Music Studies, 19* (1), 26-58.

Tagg, P. (2012). *Music's meanings: A modern musicology for non-musos.* New York: The Mass Media Music Scholars' Press.

Tibshirani, R., Walther, G., & Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B, 63* (2), 411-423.

Neural Mechanisms for Timbre: Spectral-Centroid Discrimination based on a Model of Midbrain Neurons

Braden N. Maxwell[1, 2, 3], Johanna B. Fritzinger[2], and Laurel H. Carney[2,3†]
[1] Department of Music Theory, Eastman School of Music, University of Rochester, Rochester, NY, USA
[2]Department of Neuroscience, University of Rochester, Rochester, NY, USA
[3]Department of Biomedical Engineering, University of Rochester, Rochester, NY, USA
[†] Corresponding author: Laurel.Carney@Rochester.edu

Introduction

The spectral centroid, or center of mass of the spectrum, has been shown to play a substantial role in the perception of timbre (McAdams, 2019); however, the neural code for the spectral centroid is not fully understood. Here we used computational models to test spectral-centroid encoding by auditory neurons. Amplitudes of slow fluctuations (slow temporal changes in firing rate) in auditory-nerve (AN) responses vary systematically across neurons tuned to frequencies near the spectral centroid. These changes in fluctuation amplitudes are reflected in the average discharge rates of midbrain neurons that are sensitive to fluctuations (Carney, 2018). We show that the spectral-centroid discrimination thresholds from Allen and Oxenham (2014) can be estimated based on population responses of model midbrain neurons (Carney & McDonough, 2019). Additionally, model midbrain representations of the spectral centroid are influenced by changes in the fundamental frequency (F0) of the stimulus, suggesting a sub-cortical basis for the interaction between pitch and the perception of timbre (Krumhansl & Iverson, 1992; Marozeau & de Cheveigné, 2007, Allen & Oxenham, 2014).

Method

Spectral-centroid stimuli matched those of Allen and Oxenham (2014), experiments 1 and 2, including roving of stimulus parameters. Average rates of tonotopic populations at two stages of the auditory system in response to two-interval stimuli were simulated using computational models for the AN (Zilany, Bruce, & Carney, 2014) and midbrain (including cochlear nucleus stage; Carney & McDonough, 2019). Model populations included logarithmically spaced characteristic frequencies (CF, the frequency to which a neuron is most sensitive) over ranges specified below. Ten independent, high-spontaneous-rate, AN fibers and one midbrain neuron were simulated for each frequency channel. Midbrain simulations were based on band-suppressed neurons, a common midbrain cell type that is excited by unmodulated sounds and suppressed over a range of amplitude-modulation frequencies. For these simulations, the range of suppression was centered near 100 Hz, as commonly observed in the midbrain. The spectral centroid was generally reflected in increased activity of model neurons tuned near the peak of the spectral envelope (Fig. 1). Midbrain model responses were affected by the AN average rates and by the amplitudes of slow fluctuations in the AN responses.

An estimate of the spectral centroid as represented in the model neural responses was based on the weighted average frequency of the population response. Weights were the average driven rates for each frequency channel, and log(CF) was used to calculate the weighted average. Driven rates were computed by subtracting an estimate of an upper bound on the spontaneous rate (mean + 4 × standard deviation of spontaneous rate; Fig. 1, red lines: 156 sp/s for AN, 28 sp/s for midbrain) from the average response rate for each frequency channel. This calculation was performed for model responses to the two intervals in each experimental trial, and the interval with the higher weighted average frequency was selected as the interval with the higher spectral centroid. The AN model included random internal noise; thus, the percent of correct model responses was computed based on 80 trials for each spectral centroid difference. A logistic function fit to the model %-correct results was used to estimate the difference limen (DL); a 70.7%-correct criterion for the DL was chosen to match the procedure in Allen and Oxenham (2014).

Results

Figure 1 shows an example spectral-centroid stimulus (Allen & Oxenham, 2014) with AN and midbrain model population responses to stimuli at three sound levels. Centroid estimates based on each model response are shown in corresponding colors. Some aspects of the AN model are similar to other techniques for acquiring physiologically-based estimates of spectral centroid (e.g., Marozeau, de Cheveigné, McAdams, & Winsberg, 2003); for example, the AN model channels sample the stimulus spectrum on a logarithmic scale and filter widths change with frequency. However, the AN response used here was also shaped by several nonlinearities implemented in the Zilany et al. (2014) model. These nonlinearities realistically prevent the more densely packed components at high frequencies from resulting in higher average rates and cause the filter bandwidths (and response) to change with sound level (multi-colored curves, Fig. 1B), among other effects. The general flattening of the stimulus representation ultimately makes pitch interference more possible by emphasizing components farther from the spectral peak relative to the peak itself.

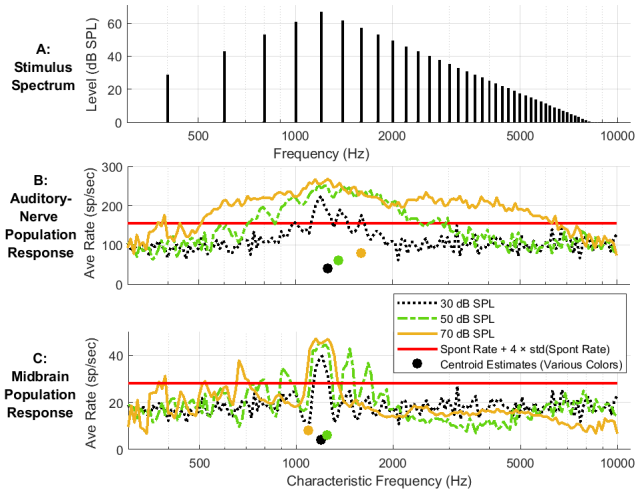


Figure 1: Spectral-Centroid Encoding in Model Populations. (A) Example stimulus: spectral centroid at 1200 Hz, F0=200 Hz, 70 dB SPL (Allen and Oxenham, 2014). (B,C) Average rates across tonotopic auditory-nerve (AN) and midbrain populations, respectively, and centroid estimates (circles) for responses to three sound levels. Only response rates above thresholds based on spontaneous rate (horizontal red lines) were included in centroid estimates. Here 160 CFs were logarithmically spaced from 0.3-10 kHz, to illustrate the broad AN population response. Note: Due to spread of excitation, centroids based on model AN responses were more level-dependent than those based on midbrain responses.

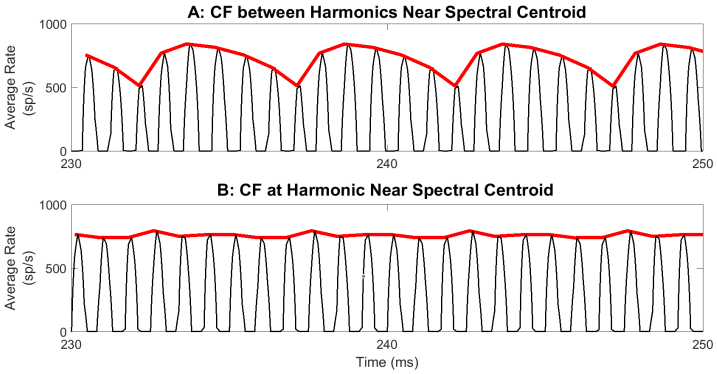


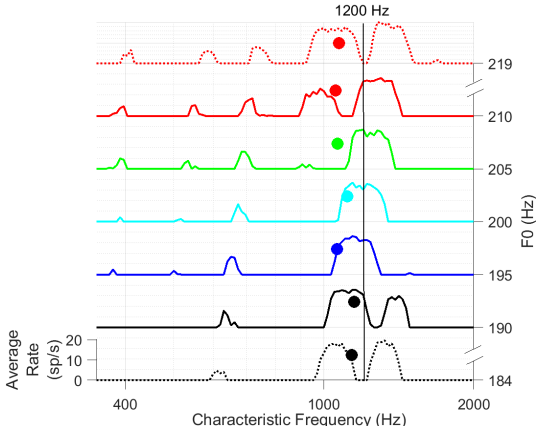
Figure 2: Contrast in Fluctuations. (A) F0-fluctuations (red) in response of AN neuron with CF near centroid, but between harmonics (CF=1075 Hz). (B) Non-fluctuating (flat) response of AN neuron with CF at highest-magnitude harmonic (CF=1200 Hz). Stimulus is same as shown in Fig. 1A. Note: In Fig. 1C, band-suppressed midbrain neurons at CFs with fluctuating responses are suppressed, whereas those tuned to non-fluctuating channels have higher rate responses.

AN rates alone do not provide a suitable code for the spectral centroid. As shown in Fig. 1B, the sharp AN representation of the stimulus peak at a low sound level becomes broad at 70 dB SPL, a level in the range used for music listening (Epstein, Marozeau, & Cleveland, 2010) or conversational speech (Carney, 2018) and used in Allen and Oxenham (2014). This representation remains broad even if all spontaneous activity of the neurons is subtracted (average + 4 standard deviations, red line in Fig. 1B). The centroid estimate based on the AN response changes dramatically with increasing level, contrary to the general impression that the brightness of a sound is stable when the sound is rescaled over a wide range of levels. The midbrain population response offers a sharper, more precise spectral-centroid code that is consistent across a wide range of sound levels (Fig. 1C).

The enhanced code in the midbrain response reflects the influence of temporal patterns in AN responses. Figure 2 demonstrates these different temporal patterns, called neural fluctuations (Carney, 2018). Model AN fibers with CFs far from the centroid or *between* harmonics near the spectral centroid (Fig. 2A) have amplitude fluctuations at F0 due to beating between multiple harmonics. In contrast, AN CFs *at* harmonics near the spectral centroid (Fig. 2B) have minimal amplitude fluctuations at F0 because their response is dominated by a single harmonic, due to both peripheral filtering and to nonlinear 'capture' of the response (Carney, 2018). The band-suppressed midbrain model (Fig. 1C) is excited at CFs with minimal fluctuations (Fig. 2B) and inhibited at CFs with strong fluctuations (Fig. 2A). Thus, for this midbrain cell type, maximal responses occur for CFs at harmonics near the centroid, and the maxima are sharply delineated by midbrain suppression of responses at higher and lower frequencies. This sharp midbrain representation leads to a model threshold for centroid discrimination that is very close to that of humans (Fig. 4, leftmost points). Note that smaller peaks in the midbrain population response are often shifted away from harmonic frequencies (e.g., 900-Hz peak in Fig. 1C, green curve, corresponds to 1000-Hz harmonic). The response profile is strongly shaped by beating between harmonics, which is optimized when two components have equal magnitudes in a filter output. On the slopes of the spectral envelope, filters with equal magnitude components in their outputs are asymmetrically positioned between components (see Henry et al., 2017, for similar phenomenon).

The sharpened representation of harmonics in the midbrain response provides a mechanism for interaction between F0 and spectral centroid perception. Figure 3 shows driven responses for 1200-Hz centroids with varied F0s (increasing from bottom to top). As F0 increases over a moderate range, the major peak in the driven response shifts to higher frequencies, consistent with Allen and Oxenham (2014) experiment 3, in which spectral-centroid discrimination was easier for congruent trials (for which F0 moved in the same direction as the centroid) than for incongruent trials. However, peaks in response to lower harmonics pull the centroid estimate used here to lower frequencies, complicating the influence of F0 variation.

Figure 3: Effect of Shifting F0 on Centroid Estimation. Peak of spectral envelope remained constant at 1200 Hz while F0 shifted (increasing, bottom to top). Driven model responses are shown (e.g. response minus an upper bound on spontaneous rate). Centroid estimates are indicated by circles. Double peaks for F0s of 219 and 184 Hz indicate response to harmonics numbered [5,6] and [6,7], respectively. 180 CFs were logarithmically spaced from 0.125-2 kHz.



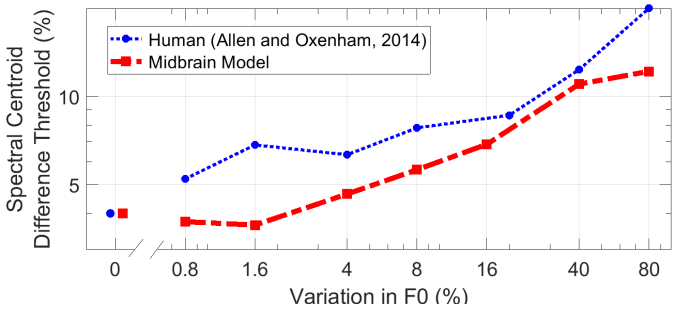


Figure 4: Midbrain model thresholds (dashed red curve) for spectral-centroid discrimination with increasing random F0 variation, compared to human (musician) thresholds (replotted from Allen and Oxenham, 2014; dotted blue curve). Model population responses were for 186 CFs logarithmically spaced from 0.125-2.4 kHz. Simulation for 0% variation used slightly fewer CFs: 180, over the same range.

Additional complications of F0 variation include: (1) large F0 variations may result in two harmonics that are equal in magnitude, and a shift in the ranks of harmonics near the centroid (dotted lines, Fig. 3); (2) the difference in the spectral centroid between intervals affects the prominence of different harmonics; (3)

between-trial shifts of F0 and spectral centroid for both intervals, called rove, further complicate the comparison. Nevertheless, when all of these factors from the original experiment were included, simulations of human thresholds using the midbrain population responses showed a pattern of F0 interference in spectral centroid discrimination that was generally similar to Allen and Oxenham's result (Fig. 4). More remains to be understood regarding the mechanisms driving this interference in the midbrain model.

Discussion

These modeling results demonstrate a possible role in timbre perception for known response properties of sub-cortical auditory neurons. The basis for these model results is the sensitivity of the midbrain to slow fluctuations in AN responses. The plausibility of this mechanism is supported by the elevation in model thresholds for spectral centroid discrimination when the F0 was made variable across intervals (Fig. 4), as observed by Allen and Oxenham (2014). These model results are consistent with the hypothesis that the midbrain may play an important role in spectral-centroid encoding, and that pitch and timbre interactions may have origins in the early stages of the auditory system. Future tests of this model should include spectral centroids that are not aligned with the spectral peak and spectra with multiple spectral peaks. As long as one or more relatively high-magnitude harmonics influence the perceptual spectral centroid, the proposed neural code may pertain. It should be emphasized that the sharpness of the midbrain rate code results from sharp changes in temporal fluctuations in AN responses. Even if information is not extracted by the midbrain in the way described by these models, such timing information may be extracted by similar processes at other stages in the auditory pathway.

Acknowledgments

This work was supported by NIH-NIDCD Grant No. DC010813.

References

Allen, E. J., & Oxenham, A. J. (2014). Symmetric interactions and interference between pitch and timbre. *JASA*, *135*(3), 1371-1379.

Carney, L. H., & McDonough, J. M. (2019). Nonlinear auditory models yield new insights into representations of vowels. *Attention, Perception, & Psychophysics*, *81*(4), 1034-1046.

Carney, L. H. (2018). Supra-threshold hearing and fluctuation profiles: Implications for sensorineural and hidden hearing loss. *Journal of the Association for Research in Otolaryngology*, *19*(4), 331-352.

Epstein, M., Marozeau, J., & Cleveland, S. (2010). Listening habits of iPod users. *Journal of Speech, Language, and Hearing Research*, *53*, 1472-1477

Henry, K. S., Abrams, K. S., Forst, J., Mender, M. J., Neilans, E. G., Idrobo, F., & Carney, L. H. (2017). Midbrain synchrony to envelope structure supports behavioral sensitivity to single-formant vowel-like sounds in noise. *Journal of the Association for Research in Otolaryngology*, *18*(1), 165-181.

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

Marozeau, J., & de Cheveigné, A. (2007). The effect of fundamental frequency on the brightness dimension of timbre. *JASA*, *121*(1), 383-387.

Marozeau, J., de Cheveigné, A., McAdams, S., & Winsberg, S. (2003). The dependency of timbre on fundamental frequency. *JASA*, *114*(5), 2946-2957.

McAdams, S. (2019). The perceptual representation of timbre. In: Siedenburg K., Saitis C., McAdams S., Popper A., Fay R. (eds), *Timbre: Acoustics, Perception, and Cognition* (pp. 23-57). Springer Handbook of Auditory Research, vol 69. Springer, Cham.

Zilany, M. S., Bruce, I. C., & Carney, L. H. (2014). Updated parameters and expanded simulation options for a model of the auditory periphery. *JASA*, *135*(1), 283-286.

Age and musicianship-related use of timbral auditory streaming cues

Sarah A. Sauvé[1†], Jeremy Marozeau[2] and Benjamin Rich Zendel[13]
[1] Division of Community Health and Humanities, Faculty of Medicine, Memorial University of Newfoundland, St. John's, Newfoundland and Labrador, Canada
[2] Department of Health Technology, Technical University of Denmark, Lyngby, Denmark
[3]Aging Research Centre – Newfoundland and Labrador, Grenfell Campus, Memorial University
[†] Corresponding author: sarah.a.sauve@gmail.com

Introduction

Understanding speech in noisy environments becomes increasingly difficult with age, and is the most commonly reported hearing issue in older adults (Pichora-Fuller et al., 2016). Auditory stream segregation is crucial to understanding speech in noise, with a growing literature investigating how it is affected by aging. Current evidence suggests that concurrent stream segregation – the segregation of simultaneous sounds – suffers (Alain et al., 2001) while sequential stream segregation – the segregation of sounds over time – remains intact when frequency is the primary auditory cue for segregation of two auditory streams (Snyder & Alain, 2006). Musical training has also been linked to better auditory stream segregation, where less difference between streams is needed for successful segregation (François et al., 2014; Marozeau et al., 2013). We present two studies that investigate how musical training interacts with aging to impact the relative salience of intensity, spectral envelope and temporal envelope as auditory streaming cues.

Method

In Study 1, a repeating four-note *target* melody was interleaved with semi-random *distractor* tones that were manipulated in terms of three *features*, intensity, spectral envelope and temporal envelope, over 20 equally spaced *levels* of increasing dissimilarity to the target (see Marozeau et al., 2013 for details). Two of the target melody notes were inverted 25% of the time to create a *deviant* melody, which participants identified by pressing the space bar on a keyboard. This task is only possible when the target is segregated from the distractor, which is easiest when the distractors are less similar to the target. This type of task generates *hits* and *false alarms*, allowing the calculation of *d' score*, a common measure of sensitivity. 54 participants took part in Study 1, 28 younger (< 38 years; 16 female) and 26 older (> 60 years; 9 female) and 12 in Study 2, (6 younger, 6 older; 10 from Study 1, 2 from lab). Study 2 was a dissimilarity rating paradigm with 15 four-note target melodies with combinations of intensity, spectral envelope and temporal envelope levels. This will allow direct comparison of *d'* scores between the three features by generating a common perceptual dissimilarity scale (see Marozeau et al., 2013 for details).

Results

Mixed effects linear modelling was used to measure effects of *age*, *musicianship* as measured by the Gold-MSI musical training sub-section (Müllensiefen et al., 2014), *dissimilarity*, based on the MDS solution generated by Study 2, *feature,* and their interactions on *d'* scores. There was a significant main effect of musicianship but not of age. The interaction between age and musicianship was not significant; all other interactions were. Figure 1 illustrates *d'* scores for each participant group and each feature along the common perceptual dissimilarity scale.

Discussion

Our results provide evidence that sequential auditory streaming, when cued by intensity, spectral envelope and temporal envelope is similar in older and younger adults. With respect to timbre, poor performance by all participant groups when the temporal envelope was manipulated suggests that greater

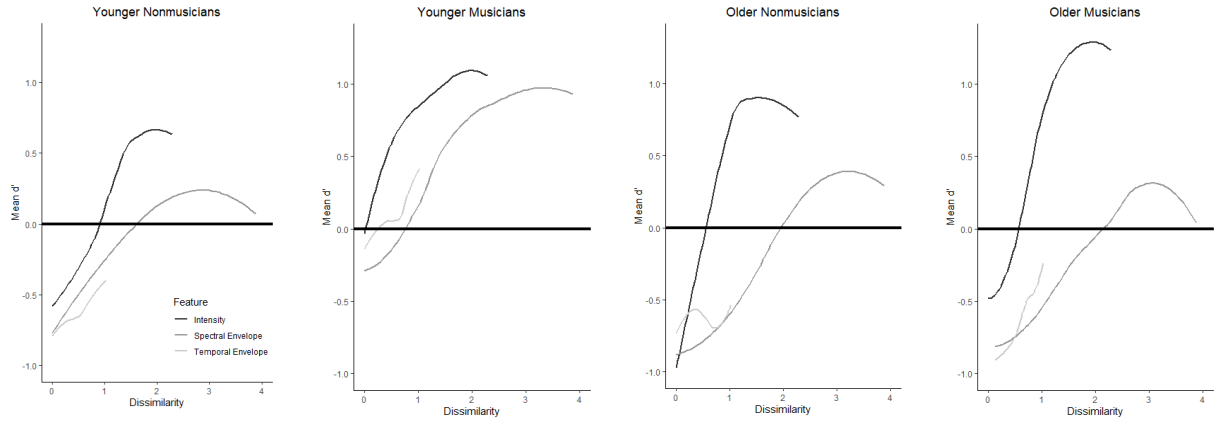


Figure 1: Mean d' scores for each participant group and feature plotted on the perceptual dissimilarity scale established by Study 2. For visualization purposes, musicians are defined as having scored > 50% on the Gold-MSI musical training sub-scale.

differentiation was needed for the temporal envelope to be a useful auditory streaming cue. Furthermore, the interaction between age and feature suggests that older adults used spectral envelope less than younger adults, relying more strongly on intensity as a streaming cue. This may be due to high frequency hearing loss, where older adults may be less able to detect changes in spectral envelope. However, older adults performed similarly to younger adults with low Gold-MSI scores for both timbre features, suggesting that timbre perception is unchanged by aging, though it may become a less reliable auditory streaming cue.

Acknowledgments

This research was funded by B.R. Zendel's Canada Research Chair. Thank you to Liam Foley and Alex Cho for assistance with data collection.

References

Alain, C., Arnott, S. R., & Picton, T. W. (2001). Bottom–up and top–down influences on auditory scene analysis: Evidence from event-related brain potentials. *Journal of Experimental Psychology: Human Perception and Performance*, *27*(5), 1072–1089.

François, C., Jaillet, F., Takerkart, S., & Schön, D. (2014). Faster Sound Stream Segmentation in Musicians than in Nonmusicians. *PLoS ONE*, *9*(7), e101340.

Marozeau, J., Innes-Brown, H., & Blamey, P. J. (2013). The Effect of Timbre and Loudness on Melody Segregation. *Music Perception: An Interdisciplinary Journal*, *30*(3), 259–274.

Müllensiefen, D., Gingras, B., Musil, J., & Stewart, L. (2014). The Musicality of Non-Musicians: An Index for Assessing Musical Sophistication in the General Population. *PLoS ONE*, *9*(2), e89642.

Pichora-Fuller, M. K., Kramer, S. E., Eckert, M. A., Edwards, B., Hornsby, B. W., Humes, L. E., Lemke, U., Lunner, T., Matthen, M., Mackersie, C. L., & others. (2016). Hearing impairment and cognitive energy: The framework for understanding effortful listening (FUEL). *Ear and Hearing*, *37*, 5S–27S.

Snyder, J. S., & Alain, C. (2006). Sequential auditory scene analysis is preserved in normal aging adults. *Cerebral Cortex*, *17*(3), 501–512.

Perceptual ratio scales of timbre-related audio descriptors

Savvas Kazazis[1†], Philippe Depalle[1], and Stephen McAdams[1]
[1] Schulich School of Music, McGill University, Montreal, Quebec, Canada
[†] Corresponding author: savvas.kazazis@mail.mcgill.ca

Introduction

Most of the past research on timbre psychophysics has focused on determining acoustic correlates of perceptual dimensions derived from multidimensional scaling of dissimilarity ratings, in order to quantify the ways in which we perceive sounds to differ. However, there is little empirical evidence to date demonstrating that acoustic features derived from correlational analysis causally correspond to psychological dimensions. Most importantly, even for cases in which the causality has been verified, there is almost no research on understanding how the sensation magnitudes of such acoustic features are apprehended. To the best of our knowledge, there have been no previous attempts in psychophysical scaling of timbre-related audio descriptors other than perhaps the preliminary results of Almeida et al. (2017) who attempted to derive a ratio scale of timbral brightness as a function of spectral centroid.

In order to investigate whether listeners perceive timbre-related descriptors on perceptual ratio scales, we conducted a ratio scaling experiment in which we tested the following descriptors: spectral centroid, spectral spread, spectral skewness, odd-to-even harmonic ratio, spectral deviation, and spectral slope (measured in dB/octave) (Peeters et al., 2011).

Method

Twenty participants, 6 female and 14 male, with a median age of 25 years (range: 18–41) were recruited from the Schulich School of Music, McGill University. All of them were self-reported amateur or professional musicians with formal training in various disciplines such as performance, composition, music theory, and sound engineering. Participants were compensated for their time.

The stimulus sets of each audio feature were synthesized and (wherever possible) independently controlled through additive synthesis with appropriate spectral amplitude distributions. For spectral centroid, spread, and skewness, the stimuli had an f0 of 120 Hz and the initial spectrum (prior to filtering according to Gaussian distributions) contained harmonics up to Nyquist frequency (i.e., 22.05 kHz). In addition, for spectral spread and skewness, three different stimulus sets were constructed with centroids at 1640, 5600, and 7800 Hz. For odd-to-even ratio, deviation, and slope, three different sound sets were constructed with f0's at 120, 300, and 720 Hz and with 9 harmonics except for the sets of spectral deviation in which 16 harmonics were used.

The listener's task was to equisect a continuum of a particular audio feature. Each equisection was performed using the *progressive solution* according to which listeners progressively partition the continuum formed by the stimuli into a number of equal-sounding intervals. In order to create a continuum within a range of a particular feature, several stimuli were constructed with multiple imperceptible successive differences. The total number of sounds used for each stimulus set and the ranges of feature values for a particular set are indicated in Table 1. In a first step, listeners bisected the continuum of an audio feature into two equal-sounding intervals, by triggering each stimulus with a cursor along a horizontal bar that contained the stimuli, and by placing a marker over the stimulus-bar. Each resulting section was then bisected in the next step. In total there were three bisections: the first one was made between the stimuli of the total range, and the other two between the lower and upper bisected ranges. In a final step, listeners were presented with all their bisections and were instructed to make further fine adjustments so that all four intervals they had created in the previous steps sounded equal. The equality of sensory intervals implies that the intervals themselves have ratio properties (Marks & Gescheider, 2002) and thus, the results of this experiment led to ratio scale measurements.

The equisection scales were then derived by fitting well-behaving functions to the listeners' ratings (Figure 1, left panel)[1]. The criteria used for choosing the form of the function were monotonicity, maximum explained variance, and good continuation (i.e., no oscillations) outside the tested range, which was useful for extrapolating the fitting function (Figure 1, right panel). The reliability of the derived scales across listeners was evaluated according to *Cronbach's alpha* (α).

Table 1: The ranges of feature values within designated stimulus set are shown in bold. The number of sounds on which the feature values were computed are shown in parentheses. The reported ranges for the spectral spread and skewness stimulus sets were computed on stimuli with 5600-Hz spectral centroid. Linear regression over normally distributed spectral amplitudes is futile.

Stimulus Sets (# sounds)	Feature Ranges					
	Centroid (Hz)	Spread (Hz)	Skewness	Odd-to-Even Ratio	Deviation	Slope (dB/octave)
Centroid (505)	**[1642, 9560]**	[479, 480]	[0.00, 0.02]	[1.00, 1.00]	[0.00, 0.00]	-
Spread (100)	[5600, 5600]	**[181, 1439]**	[0.00, 0.00]	[1.00, 1.00]	[0.00, 0.00]	-
Skewness (97)	[5600, 5600]	[1079, 1080]	**[–0.88, 0.96]**	[1.00, 1.00]	[0.00, 0.00]	-
Odd-to-Even Ratio (349)	[1260, 1500]	[768, 848]	[0.00, 0.21]	**[0.25, 1250.00]**	[0.00, 0.11]	[–11.67, –2.62]
Deviation (265)	[1723, 2550]	[1292, 1396]	[0.00, 0.28]	[1.00, 1.19]	**[0.00, 0.06]**	[0.00, –5.04]
Slope (349)	[332, 2082]	[134, 785]	[–1.04, 6.68]	[1.25, 15.05]	[0.00, 0.03]	**[–24.00, 5.44]**

Results

The results indicate that listeners can produce ratios of descriptor values, which in turn enabled the construction of perceptual ratio scales of each descriptor tested. As evidenced by *Cronbach's alpha*, the reliabilities of the scales were overall excellent, with the scales of spectral centroid and spectral skewness having the highest reliability ($\alpha = 0.96$). The lowest reliability was observed for the equisections of spectral spread ($\alpha = 0.89$), followed by the odd-to-even ratio ($\alpha = 0.78$). With the exception of spectral skewness, for which the best fitting function on the median ratings was a third-order polynomial, the best fitting functions for the rest of the descriptors were all power functions albeit exhibiting significantly different shapes, which indicates that each descriptor is perceived on a different psychophysical scale. After identifying the form of the function fitted on listeners' equisections, the psychophysical scales were constructed by defining a *unit* for each scale and its *zero point*. With the exception of spectral centroid, for which the zero point of the scale was assigned to 20 Hz, which marks the lower limit of pitch perception, the rest of the scales were assigned a zero point that has a physical meaning (e.g., zero skewness). The units of the scales were defined by empirically assigning specific numerals to the points of the equisection scale (e.g., by assigning the numeral 10 to the 1-kHz spectral centroid), so as to facilitate comparisons between the derived perceptual ratio scales of all descriptors (Figure 1).

Discussion

The stimuli used in each of the presented experiments were constructed through specifically designed additive-synthesis algorithms that enabled control of each audio descriptor independently of the rest and that therefore isolated as much as was feasible the effect of each descriptor on listeners' perceptions. However, spectral slope is the descriptor that is the most difficult to control independently of centroid, spread and skewness descriptors, because these two sets of descriptors are physically intercorrelated (Table 1).

The construction of psychophysical scales based on such univariate stimuli allowed for the establishment of *cause and effect* relations between audio features and perceptual dimensions, contrary to past research that has relied on multivariate stimuli and has only examined the correlations between the two. The derived

[1] The upper limit of the centroid range is high (10 kHz), essentially to serve as an anchor point, and accommodate specific high-pass sound signals such as a hi-hat.

scales along with their respective units designate a *perceptual coordinate system of audio features* in which sounds can be grouped and ordered according to their perceived sound qualities that relate to each descriptor. The perceptual coordinate system of descriptor values could potentially be used to define a "control surface" for applications that include computer-aided orchestration, perceptually motivated sound effects, and synthesis algorithms.

In addition, audio descriptors have been widely used as predictor variables in statistical regression models for interpreting and predicting listeners' responses on a variety of tasks that relate to timbre. However, the physical values of these predictors may lead to false-positive interpretations about their perceptual significance on a particular task. The derived scales allow timbre researchers to use perceptually informed values of spectral descriptors as predictors in their statistical models that may lead to more sustainable conclusions and accurate interpretations in terms of perception.

Nonetheless, it should be mentioned that this study does not imply that all the descriptors tested here constitute perceptual dimensions because, it only provides evidence that individual descriptors can be perceived on perceptual ratio scales when the rest of them remain relatively constant. However, this study does not test the extent to which each descriptor is independently perceived when multiple descriptors covary. Verifying or rejecting that hypothesis is left for future work.

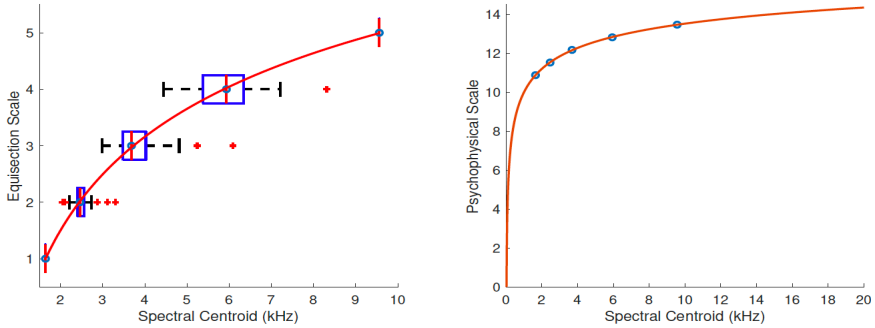


Figure 1: Equisection and psychophysical scales of spectral centroid. Whiskers extend to 2.7 the Standard Deviation.

Acknowledgments

We would like to thank Bennett K. Smith for programming the user interface of the experiment, and Erica Huynh for her help in recruiting and running participants.

References

Almeida, A., Schubert, E., Smith, J., & Wolfe, J. (2017). Brightness scaling of periodic tones. *Attention, Perception & Psychophysics*, *79*, 1892–1896.

Marks, L. E., & Gescheider, G. (2002). Psychophysical scaling. In H. Pashler & J. Wixted (eds.), *Stevens' handbook of experimental psychology: Perception and motivation; learning and cognition* (pp. 91–138). Hoboken, NJ, US: John Wiley & Sons Inc.

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

Spectral and Temporal Timbral Cues of Vocal Imitations of Drum Sounds

Alejandro Delgado[1,2†], Charalampos Saitis[1], and Mark Sandler[1]
[1] Centre for Digital Music, Queen Mary University of London, London, United Kingdom
[2] Research and Development Team, Roli Ltd., London, United Kingdom
[†] Corresponding author: a.delgadoluezas@qmul.ac.uk

Introduction

The imitation of non-vocal sounds using the human voice is a resource we sometimes rely on when communicating sound concepts to other people. Query by Vocal Percussion (QVP) is a subfield in Music Information Retrieval (MIR) that explores techniques to query percussive sounds using vocal imitations as input, usually plosive consonant sounds. The goal of this work was to investigate timbral relationships between real drum sounds and their vocal imitations. We believe these insights could shed light on how to select timbre descriptors for extraction when designing offline and online QVP systems. In particular, we studied a dataset composed of 30 acoustic and electronic drum sound recordings and vocal imitations of each sound performed by 14 musicians [1]. Our approach was to study the correlation of audio content descriptors of timbre [2] extracted from the drum samples with the same descriptors taken from vocal imitations. Three timbral descriptors were selected: the *Log Attack Time* (LAT), the *Spectral Centroid* (SC), and the *Derivative After Maximum* of the sound envelope (DAM). LAT and SC have been shown to represent salient dimensions of timbre across different types of sounds including percussion [2]. In this sense, one intriguing question would be to what extent listeners can communicate these salient timbral cues in vocal imitations. The third descriptor, DAM, was selected for its role in describing the sound's tail, which we considered to be a relevant part of percussive utterances.

Method

After computing the three descriptors using the Essentia library, we constructed two distance matrices for each of them: one for the acoustic space of drum samples and another one for the acoustic space of vocal imitations. The first one was built by taking the euclidean distance between a single descriptor taken from all drum samples. It was, therefore, a symmetric matrix of dimensions 30x30. For the second distance matrix, we first measured the euclidean distances between the descriptors taken from the vocal imitations of individual participants and we then averaged the 14 resulting matrices into a single one of size 30x30. Once the two distance matrices were built, we ran the Mantel test, which measures the degree of statistical correlation between two symmetric matrices. Lastly, we picked the two most correlated descriptors and applied a hierarchical clustering algorithm to group imitators based on the distances between their Mantel test scores (Fig. 1b). This was done to see if there were any interpersonal differences in the way these descriptors were imitated. For reference, we also plotted the values of these two descriptors against each other for both drum sounds and vocal imitations, using different colors for the five imitated instruments (Fig. 1a). We did this to assess whether the descriptors from same-category instruments were close to their imitations.

Results

The Mantel test scores of the DAM descriptor for all 14 imitators gave a mean result of $\bar{r}=.561$, a standard deviation of $\sigma_r=.114$, and a maximum p-value of $p<.001$. For the SC, the results were $\bar{r}=.428$,$\sigma_r=.153$, $p<.024$ for a subset of 13 participants, and $\bar{r}=.102$, $p=.411$ for one participant alone. For the LAT descriptor, the scores were $\bar{r}=.011$,$\sigma_r=.130$, $p<.997$.

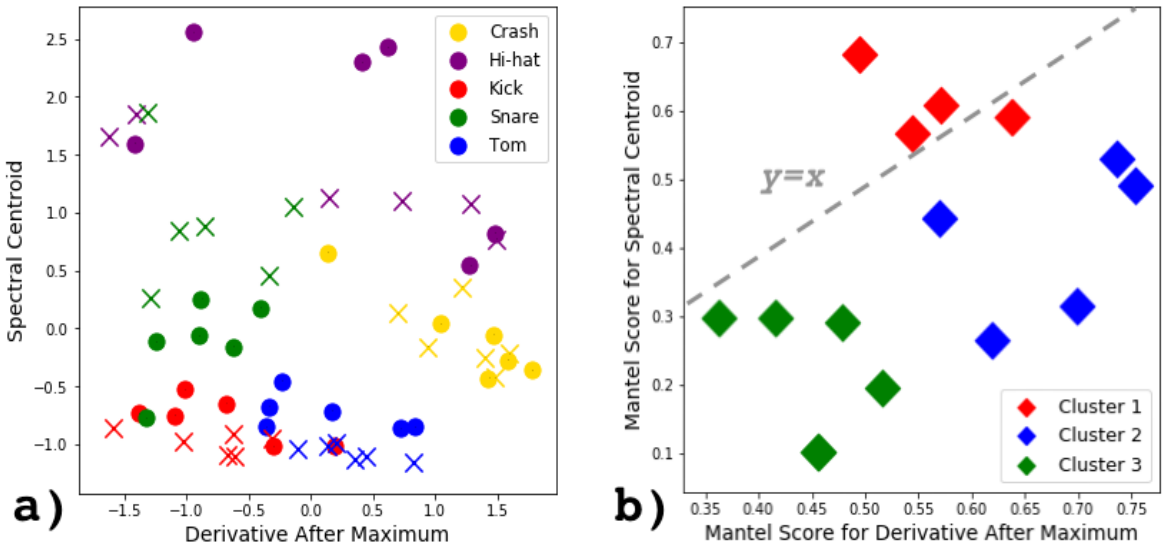


Figure 1: (1a) Normalised mean values of SC plotted against DAM for all sounds (circle markers for drum sounds and crosses for vocal imitations averaged across imitators). (1b) Mantel scores of SC plotted against those of DAM for all imitators (applied hierarchical clustering to group participants).

Discussion

The fact that the LAT descriptor failed to be a good predictor for one acoustic space given the other indicates that, despite playing an important role in timbre perception, LAT might be difficult to reproduce vocally with enough precision, at least in the case of percussive utterances (quick attack). Instead, it appears that listeners can better imitate a temporal envelope cue related to the length of the sound's tail (DAM), an observation that inspires further research from a timbre perception perspective. Interestingly, participants appear to imitate the SC cue dexterously, which seems to reinforce its role in describing one of the most salient dimensions of timbre. Considering QVP systems, the irrelevance of the LAT could make online QVP more challenging, needing more complex descriptors to quickly classify utterances. Meanwhile, offline QVP systems can safely rely on the SC and the DAM to link imitations and samples. Fig. 1a shows how these two descriptors can help cluster the different instruments and their imitations. The emergence of imitator clusters in Fig. 1b warrant further investigation into interpersonal differences in vocal imitation of non-vocal sounds, for example, differences between expert and naive listeners.

Acknowledgments

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 765068

References

Mehrabi, A., Choi, K., Dixon, S., & Sandler, M. (2018, April). Similarity measures for vocal-based drum sample retrieval using deep convolutional auto-encoders. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 356-360). Calgary, Canada.

Caetano, M., Saitis, C., & Siedenburg, K. (2019). Audio content descriptors of timbre. In K. Siedenburg, C. Saitis, S. McAdams, A. Popper and R. Fay (eds) *Timbre: Acoustics, Perception, and Cognition* (pp. 297-333). Springer Handbook of Auditory Research, vol 69. Springer, Cham.

Player recognition for traditional Irish flute recordings using K-nearest neighbour classification

Islah Ali-MacLachlan[1†], Edmund Hunt[2] and Alastair Jamieson[1],
[1] DMTLab, Birmingham City University, Birmingham B4 7XG, UK
[2] Royal Birmingham Conservatoire, Birmingham B4 7XR, UK
[†] Corresponding author: islah.ali-maclachlan@bcu.ac.uk

Introduction

Irish traditional music (ITM) is central to the idea of Irish cultural identity. The wooden traverse flute is one of several instruments used in ITM alongside the fiddle (violin), uilleann pipes and tin whistle. Mastery in musicianship is displayed by a player's translation of an unadorned traditional melody into a personalised rendition containing stylistic traits such as ornamentation, dynamics and timbre (Breathnach, 1996). In ITM, stylistic differences between flute players have been attributed to many influences, including regional playing styles, *sean nós* singing, uilleann pipes technique, and players' personal preferences (Johnston, 1995). Within the ITM community, the importance of timbre is illustrated by the frequent use of adjectives such as 'reedy', 'earthy' or 'warm' to describe an individual flute player's tone. However, these descriptions of timbre are highly subjective and difficult to analyse.

The aim of this work is to identify individual traditional Irish flute players from recordings, and to understand the influence of two types of features, harmonic magnitudes and and mel-frequency cepstral coefficients (MFCCs) in attaining an overall classification accuracy. A number of experiments have used differences between magnitudes of a range of harmonics along with changes in amplitude envelope to indicate timbral variances (Grey, 1977; Iverson & Krumhansl, 1993). Harmonic magnitudes can be used to identify individual flute players of different ability levels (Ali-MacLachlan et al., 2013). MFCCs are used to discriminate between sonorant and non-sonorant speech (Dumpala et al., 2015) and have been used successfully in flute player identification (Ali-MacLachlan et al., 2018). Studies have shown that it is difficult to identify professional flute players from recordings by using only harmonic magnitudes, due to experienced players having more breath and embouchure control (Ali-MacLachlan et al., 2015). Studies often use the steady-state central part of notes and in this work we also investigate a comparison between steady-state, attack, release and whole event.

Method

The recordings chosen for analysis were part of the ITM-Flute-99 dataset[1]. We used the dataset described in (Ali-MacLachlan et al., 2015) containing five players, each playing four solo unaccompanied traditional Irish flute tunes. All of these players are regarded as having a distinctive musical signature. A typical playing style includes the ornamentation of a traditional melody with *cuts* and *taps* – very short pitch deflections up and down caused by quick finger movements over the holes above and below. These ornaments are usually in the range of 0.02-0.08 sec. whereas most melody notes are in the range of 0.1-0.3 sec. (Köküer et al., 2014).

The audio for each event is extracted by using ground truth timing data and in the case of shorter events, zero padded to 2048 samples. The Fourier transform is applied and the absolute values are taken to provide a spectrum. Magnitudes are then compressed by applying logarithm. H1 – H5 harmonic magnitudes are identified semi-automatically by calculating the localised maxima around the multiples of the event frequency from 1 to 5. 13 MFCC coefficients are also extracted and the first one is discarded as it contains a constant offset relating to the average log-energy of the input signal.

A study was conducted using a KNN classifier with a range of k=1-10 for a dataset using all notes (see Table 1). Based on this, k=1 was found to return the highest accuracy.

[1] www.github.com/izzymaclachlan/datasets

Table 1. Comparison of identification accuracy (%) for the Notes dataset using different k values with HMAG and MFCC features

k	1	2	3	4	5	6	7	8	9	10
Notes	91.2	88.9	90.3	88.5	88.5	86.9	87.3	86.2	86.3	85.8

The KNN classifier (k=1) was then trained with harmonic magnitude (HMAG) and mel frequency cepstral coefficient (MFCC) features derived from attack, sustain or release portions of events. 4-fold cross-validation was used, where 75% of the data was combined to train the model, which was tested on the remaining 25%. The process was repeated four times and overall performance aggregated across the four folds.

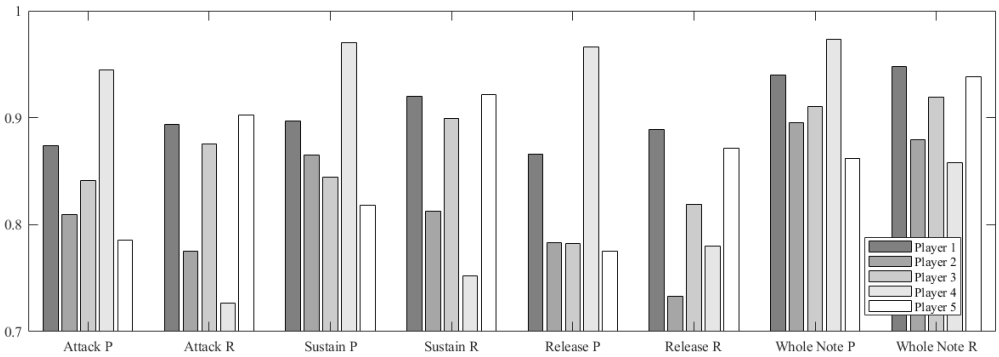


Figure 1- Comparison of precision (P) and recall (R) for attack, sustain, release and whole notes, showing results per class. All HMAG and MFCC features are used on audio of notes only.

Results

Figure 1 shows precision and recall results across 5 players when the classifier is trained using data from either attack, sustain (steady state) or release sections, or by using the whole note. The average precision and recall for the sustain section is 0.8148 and 0.7667 in comparison to whole notes at 0.9162 and 0.9087. The precision and recall levels for different players are variable showing that different styles contribute to individuality in different sections of the note.

Figure 2 shows that harmonic magnitudes make very little contribution towards precision and recall. Average precision and recall across 5 players is 0.9164 and 0.9073 when using MFCC only, and 0.9162 and 0.9087 when using both MFCC and HMAG. Using higher order harmonics H4 and H5 alongside H1, H2 and H3 increase classification accuracy when HMAG is used alone.

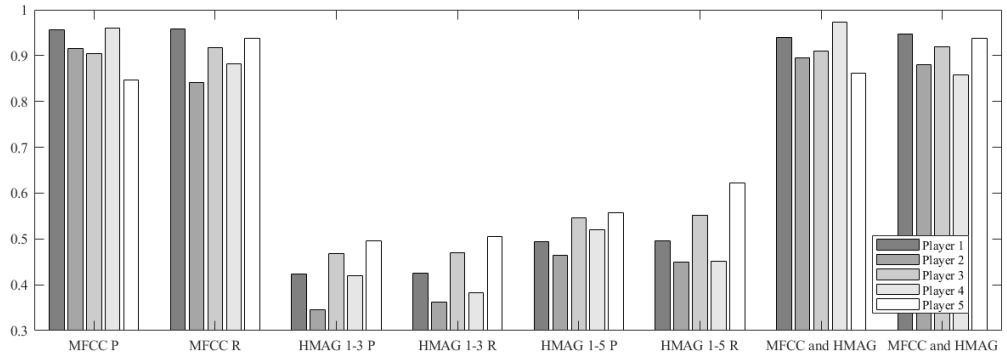


Figure 2- Comparison of precision (P) and recall (R) for MFCC only, HMAG H1-H3, HMAG H1-H5 and MFCC with HMAG H1-H5. All HMAG and MFCC features are used on audio of notes only.

Figure 3 shows a differences in precision and recall when training a model with notes or ornaments individually or together (all events). Average precision and recall across 5 players is 0.9164 and 0.9073 with notes only in comparison to 0.8569 and 0.8519 when training with all events.

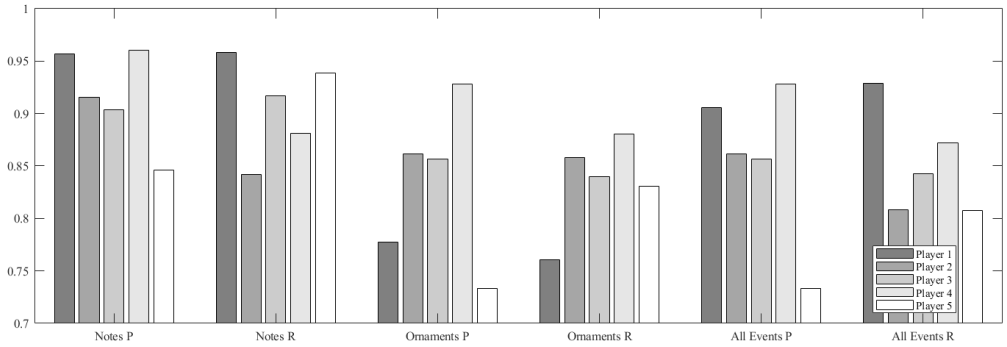


Figure 3- Comparison of precision (P) and recall (R) for notes, ornaments and all events (notes + ornaments). MFCC features are used in all cases.

Conclusions

The findings contribute to a better understanding of timbre in traditional flute playing by allowing us to quantify the difference in results that are achieved by analysis of different note sections and use of harmonic magnitudes in comparison to MFCCs. The results show that MFCCs return substantially better results and that using harmonic magnitudes alongside make little difference.

Our previous studies have concentrated on using the steady state central section of notes but this study shows that the average precision and recall across the 5 player classes for whole notes is higher than for the central section only. We also found that using longer melodic notes without shorter ornaments give better results.

In future research, we hope to compare these results to listening tests in order to explore the correlation between computational analysis and a flute player's comprehension of style and tonal quality.

References

Ali-MacLachlan, I., Köküer, M., Athwal, C., & Jancovic, P. (2015). Towards the identification of Irish traditional flute players from commercial recordings. In: Association Dirac (ed), *Proceedings of the 5th International Workshop on Folk Music Analysis (FMA) (pp.13-17)*. Paris, France.

Ali-MacLachlan, I., Köküer, M., Jancovic, P., Williams, I., & Athwal, C. (2013). Quantifying Timbral Variations in Traditional Irish Flute Playing. In: Kranenburg P., Anagnostopoulou C., Volk A. (eds),

Proceedings of the 3rd International Workshop on Folk Music Analysis (FMA). (pp. 7–14). Amsterdam, Netherlands.

Ali-MacLachlan, I., Southall, C., Tomczak, M., & Hockman, J. (2018). Player recognition for traditional Irish flute recordings. In: Holzapfel A., Pikrakis A. (eds), *Proceedings of the 8th International Workshop on Folk Music Analysis (FMA)* (pp.3-7). Thessaloniki, Greece.

Breathnach, B. (1996). *Folk music and dances of Ireland*. London: Ossian Publications Ltd.

Cowdery, J. R. (1990). *The melodic tradition of Ireland*. Ohio: Kent State University Press.

Dumpala, S. H., Nellore, B. T., Nevali, R. R., Gangashetty, S. V., & Yegnanarayana, B. (2015). Robust features for sonorant segmentation in continuous speech. In: Möller S., Ney H., Möbius B., Nöth E., Steidl S. (eds), *Sixteenth Annual Conference of the International Speech Communication Association*. (pp.1987–1991). Dresden, Germany.

Grey, J. M. (1977). Multidimensional perceptual scaling of musical timbres. *The Journal of the Acoustical Society of America*, *61*(5), 1270–1277.

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

Johnston, T. F. (1995). The Social Context of Irish Folk Instruments. *International Review of the Aesthetics and Sociology of Music*, *26*(1), 35–59. https://doi.org/10.2307/836964

Köküer, M., Ali-MacLachlan, I., Jancovic, P., & Athwal, C. (2014). Automated Detection of Single-Note Ornaments in Irish Traditional flute Playing. *Proceedings of the 4th International Workshop on Folk Music Analysis (FMA)*. Istanbul, Turkey.

Widholm, G., Linortner, R., Kausel, W., & Bertsch, M. (2001). Silver, gold, platinum-and the sound of the flute. In: Bonsi D. (ed), *Proceedings of the International Symposium on Musical Acoustics*, *1*, 277–280.

Perceptual characteristics of spaces of music performance and listening

Thomas Chalkias[1†] and Konstantinos Pastiadis[1]
[1]School of Music Studies, Aristotle University of Thessaloniki, Thessaloniki, Greece
[†]Corresponding author: thomashalkias1993@gmail.com

Introduction

Four major auditory percepts (loudness, pitch, duration and timbre) are considered of interest in Music and Acoustics. Compared to the other three ones, timbre is the most complex and least understood characteristic of sound (Zacharakis & Pastiadis, 2014). Given the multidimensional nature of timbre on its own, the combination of timbre and reverberation creates a large number of additional variables that affect the perception of timbre.

In room acoustical quality research history, the main goal has been to investigate what is considered a "good listening environment" (Kuusinen, 2016). Today, concert hall design aims at specific acoustic properties that can be reproduced using acoustic simulation and acoustic measurements. (Kahle, 2013). Several researchers attempted to describe the features of room acoustic halls, with as few as possible number of verbal descriptors (e.g. Kuusinen, 2016,).

Weinzierl, Lepa, & Ackermann, (2018), created a measuring instrument for the auditory perception of rooms. Three types of stimuli (orchestra, solo, speech) were used under 35 x 2 (rooms x listener positions) different room acoustics instances. The participants were asked to interpret the stimuli with 46 different room acoustical quality descriptors. Hence, the Room Acoustical Quality Inventory (RAQI) was produced.

In this paper we attempt to investigate aspects of the interaction between musical timbre and reverberation using semantic descriptors of room acoustics with the use of the 9-factor RAQI (Quality, Strength, Reverberance, Brilliance, Irregular Decay, Coloration, Clarity, Liveliness, Intimacy) by Weinzierl, Lepa, & Ackermann, (2018) as the verbal description mechanism.

Method

Our research is based on two verbal characterization experiments. Both of these experiments used the 9-factor RAQI as verbal descriptors. In both experiments, the participants were asked to quantify each RAQI factor for each stimulus sound (orchestra, solo trumpet, and speech in the 1st experiment and various instruments producing a single tone in the 2nd experiment). The RAQI terms appeared both in English and Greek. Three room acoustics conditions were considered, e.g. anechoic, medium reverberation (DTU), high reverberation (Hagia Sophia). In the second experiment a set of individual instruments' tones (similarly to Zacharakis, Pastiadis, & Reiss, 2014) were quantified using the RAQI, under the 3 aforementioned acoustical conditions, attempting to identify interactions of different instrument families/timbres with reverberation conditions. The participants were highly trained musicians.

Results

In Experiment 1, in most cases, in the orchestral stimuli (Polyphonic timbre) there was an augmentation in RAQI score in moderate reverberation and a decrease in high reverberation. In the solo instrument the relationship between the RAQI scores and reverberation is not monotonic. In Experiment 2 lower reverberation enhanced RAQI scores with percussions or keyboards, while winds, brass and strings led to higher RAQI scores in conditions with higher reverberation. The RAQI scores show consistency between different types of sounds in low and high reverberation but tend to be instrument dependent in medium reverberant environments.

Discussion

Timbre's interaction with reverberation (in terms of room acoustics characterizations) varies among instrument families. A significant differentiation lies in the type of timbre, namely polyphonic or monophonic. Although the RAQI score shows consistent quantifications in low-high reverberation, the variability in particular factors across different timbres and reverberation conditions is important for the evaluation of both the statistical properties of RAQI per se and the investigation of the relation between timbre and room acoustics.

References

Kahle, E. (2013, 12). Room Acoustical Quality of Concert Halls: Perceptual Factors and Acoustic Criteria -Return from Experience. *Building Acoustics, 20*(4), 265-282. doi:10.1260/1351-010X.20.4.265

Kuusinen, A. (2016). *Multidimensional perception of concert hall acoustics - Studies with the loudspeaker orchestra.* (phD Thesis) Aalto University, Finland.

Weinzierl, S., Lepa, S., & Ackermann, D. (2018). A measuring instrument for the auditory perception of rooms: The Room Acoustical Quality Inventory (RAQI). *Journal of the Acoustical Society of America, 144*(3), 1245-1257.

Zacharakis, A., & Pastiadis, K. (2016). Revisiting the Luminance-Texture-Mass Model for Musical Timbre Semantics: A Confirmatory Approach and Perspectives of Extension. *Journal of The Audio Engineering Society, 64*(9), 636-645.

Zacharakis, A., Pastiadis, K., & Reiss, J. (2012). An interlanguage study of musical timbre semantic dimensions and their acoustic correlates. *Music Perception, 31*(4), 339-358.

Perception of Action and Object Categories in Typical and Atypical Excitation-Resonator Interactions of Musical Instruments

Erica Y. Huynh[1,2†], Joël Bensoam[3], and Stephen McAdams[1,2]
[1] Music Technology Area, McGill University, Montréal, Québec, Canada
[2] Center for Interdisciplinary Research in Music Media and Technology, McGill University, Montréal, Québec, Canada
[3] Instrumental Acoustics Team, Institut de recherche et coordination acoustique/musique, Paris, France
[†] Corresponding author: erica.huynh@mail.mcgill.ca

Introduction

Our ability to recognize sound sources is automatic, yet very little is known about how this process works. Physical sources generate sounds that carry information about the object and material of the source and the action required to produce the sound. Listeners perceive impacted materials produced by the same action or material as more similar (Hjortkjær & McAdams, 2016) and can accurately identify the actions and materials of sound sources across broad categories (Lemaitre & Heller, 2012). A recent review suggests that listeners classify tones under the same category and rate them as more similar if they are played by similar excitation methods or instruments of the same family (Giordano & McAdams, 2010). These findings are often based on sound sets comprising tones from recorded orchestral instruments or synthesized versions of them, which are highly familiar in everyday listening. Siedenburg et al. (2016) collected dissimilarity ratings based on sounds that were familiar (i.e., recorded orchestral instrument tones) and unfamiliar (i.e., timbral transformations of recorded instrument tones that preserve acoustic properties). Perceived dissimilarity depended on an interplay of categorical (e.g., instrument family, excitation method) and acoustic (e.g., brightness) information. The goal of the current study is to directly examine how actions and objects are identified when they are combined in ways that are either typical or atypical of acoustic musical instruments. We used a synthesis paradigm that allowed for direct application of isolated actions to isolated objects. We assessed identification with two experimental methods: rating the resemblance of each stimulus to different actions (Experiment 1a) and objects (Experiment 1b), and explicit categorization of their actions and objects (Experiment 2).

Method

Stimuli. We used Modalys, a digital physical modelling platform, to synthesize stimuli that simulated three actions (bowing, blowing, and striking) and three objects (string, air column, and plate). Modalys allows for independent control of these actions and objects, so they can be freely associated, such that physically impossible sounds become possible with physically inspired modelling (Dudas, 2014). Thus, we combined each action with each object, creating nine classes of action-object interactions. Some interactions were typical of acoustic musical instruments (bowed string, blown air column, struck plate, and struck string), whereas others were atypical (bowed air column, bowed plate, blown string, blown plate, and struck air column). Interaction typicality was associated with the limitations of sound production in the physical world. For example, air columns and plates can be blown and struck, respectively; however, not many other actions set these objects into vibration in the physical world on a daily basis. Strings can be bowed and struck; but very rarely are they blown. Three exemplars for each action-object interaction were chosen through an exploratory approach to demonstrate the variability in their timbres (see Huynh, 2019).

Apparatus. Experiment 1 was conducted in the Perceptual Testing Lab at the Center for Interdisciplinary Research in Music Media and Technology (CIRMMT) at McGill University. It ran on a Mac Pro computer running OS 10.7 (Apple Computer, Inc, Cupertino, CA) and was displayed on an Apple Display 23-inch screen. We ran Experiment 2 in the Music Perception and Cognition Lab at McGill University in an IAC model 120act-3 double-walled audiometric booth (IAC Acoustics, Bronx, NY). The experiment ran on a Mac Pro computer running OSX (Apple Computer, Inc., Cupertino). In both Experiments 1 and 2, stimuli were presented over Seinnheiser HD280 Pro headphones (Sennheiser Electronic GmbH, Wedemark,

Germany) and were amplified through a Grace Design m904 monitor (Grace Digital Audio, San Diego, CA). The experiments were programmed in the PsiExp computer environment (Smith, 1995).

Procedure. Experiments 1a (N=41, 22 female, 19 male) and 1b (N=41, 24 female, 17 male) each comprised three blocks. Ratings in each block concerned only one type of action (1a) or object (1b). The order of the blocks was randomized for each participant. Each block had 27 trials—one for each stimulus (3 actions × 3 objects × 3 exemplars). In each trial, participants played a stimulus and rated its resemblance to the target action or object on a continuous scale from "not at all" to "completely". The order of stimulus presentation within each block was pseudo-randomized, such that two stimuli produced by the same action-object interaction were not presented in successive trials.

Experiment 2 (N=47, 35 female, 11 male, 1 other) involved a categorization task with two blocks. Each block concerned either action categorization or object categorization of the stimuli. The order of the blocks was randomized for each participant. Within each block, there were 27 trials. Depending on whether participants were categorizing the actions or objects of the stimuli, there were three boxes presented on the screen representing each category. Positions of the three boxes were randomized for each participant. For a given trial, participants played a stimulus and clicked the box corresponding to the action or object they thought produced it. The order of stimulus presentation within each block was pseudo-randomized in the same manner as in Experiment 1.

Results

Experiment 1. We averaged ratings for each action and object across the three exemplars of each action-object pair for each participant. We conducted two 3×3 repeated-measures Multivariate Analyses of Variance (MANOVAs), one for the action-resemblance ratings (Experiment 1a) and one for the object-resemblance ratings (Experiment 1b). For both MANOVAs, the independent variables were the action properties and object properties. We used Pillai's Trace, V, as the multivariate test statistic to accompany the F statistic, since it has been reported to be more robust to violations of assumptions (Olson, 1974).
For the action-resemblance ratings, we found significant within-groups effects of both the action properties (V=1.42, F(6,158)=65.11, p<.001, η_p^2=.71) and object properties (V=1.54, F(6,158)=87.20, p<.001, η_p^2=.77), as well as a significant interaction between actions and objects (V=0.99, F(12,480)=19.69, p<.001, η_p^2=.33). The different actions, objects, and their combinations influenced how listeners perceived the three actions. We conducted univariate analyses to test for the effects of the action-object interactions on each of the three action-resemblance ratings, since those interactions represent our stimuli. A conservative adjustment of the F statistic (i.e., epsilon) is reported to account for violation of the sphericity assumption. Significant interactions between actions and objects were revealed for: bowing ratings, F(3.68,147.12)=38.01, p<.001, ε=.92, η_p^2=.49; blowing ratings, F(3.77,150.76)=23.60, p<.001, ε=.94, η_p^2=.37; and striking ratings, F(2.58,103.20)=24.93, p<.001, ε=.65, η_p^2=.38.

For the object-resemblance ratings, the MANOVA revealed significant within-groups effects of: objects, V=1.56, F(6,158)=28.68, p<.001, η_p^2=.78; actions, V=1.04, F(6,158)=28.68, p<.001, η_p^2=.52; and their interaction, V=1.06, F(12,480)=21.82, p<.001, η_p^2=.35. Objects and actions that produced the sounds, as well as their interactions influenced listeners' perceptions of the objects. We further assessed this with univariate analyses to examine whether the different action-object interactions influenced resemblance ratings for each object. Accounting for the violation of the sphericity assumption, we report a conservative adjustment of the F statistic where appropriate. There was a significant interaction between action and object properties for: string ratings, F(3.32,132.88)=42.07, p<.001, ε=.83, η_p^2=.51; air column ratings, F(4,160)=34.72, p<.001, η_p^2=.47; and plate ratings, F(3.59,143.68)=33.62, p<.001, ε=.90, η_p^2=.46.

The significant interaction between action and objects for each of the action- and object-resemblance ratings are summarized in Figure 1. For stimuli representing typical action-object combinations, participants assigned the highest resemblance ratings to the actions and objects that actually produced the sounds. For example, bowed strings had the highest bowing and string ratings; similar patterns were also observed for blown air columns, struck strings, and struck plates. For the atypical interactions, listeners confused

different actions or objects for one another. Bowed air columns and blown plates had the highest blowing and air column ratings. Additionally, bowed plates were perceived as struck plates and struck air columns were perceived as struck plates and struck strings. In these cases, listeners perceived either the correct action or object, seldom both. The correct property they perceived biased perception toward the complementary property that it most typically interacts with in acoustic musical instruments. An interesting case is the blown string, for which bowing, blowing, string, and air column ratings were highest. This may have to do with the fact that both bowed and blown sounds result from continuous excitations. As bowing is typically applied to strings and blowing is typically applied to air columns, participants confused these objects for one another.

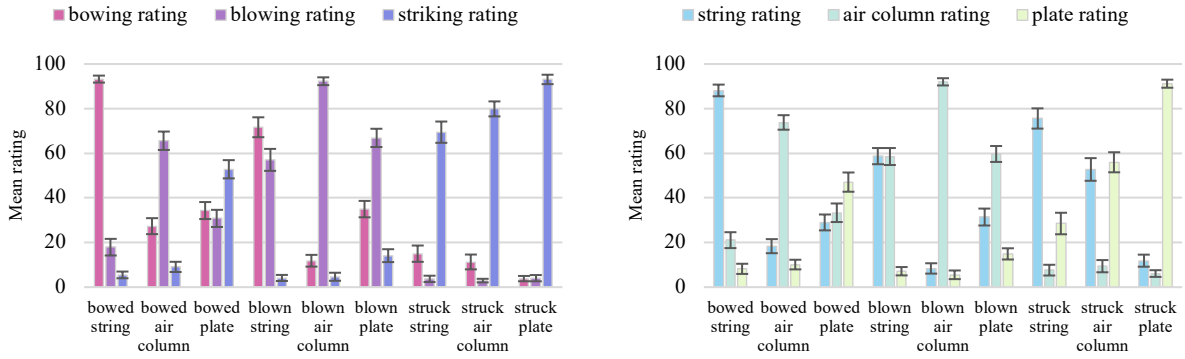


Figure 1: Mean action (left) and object (right) resemblance ratings for each action-object interaction. Error bars represent standard error of the mean.

Table 1: Categorization confusion matrix for the nine classes of action-object interactions. Stimuli are represented in the rows and response categories are represented in the columns. Typical stimuli are displayed in red font. Correct responses are indicated in bold.

	Action identification			**Object identification**		
Stimulus	**Bow**	**Blow**	**Strike**	**String**	**Air column**	**Plate**
bowed string	***131***	*6*	*4*	***134***	*5*	*2*
bowed air column	***17***	*123*	*1*	*9*	***123***	*9*
bowed plate	***59***	*23*	*59*	*16*	*26*	***99***
blown string	*91*	***48***	*2*	***85***	*55*	*1*
blown air column	*4*	***137***	*0*	*2*	***138***	*1*
blown plate	*41*	***98***	*2*	*19*	*93*	***29***
struck string	*24*	*0*	***117***	***129***	*0*	*12*
struck air column	*11*	*0*	***130***	*62*	***18***	*61*
struck plate	*4*	*0*	***137***	*3*	*4*	***134***

Experiment 2. Since this experiment involved explicit categorization of the actions and objects of the stimuli, we were able to further assess the ambiguities resulting from the resemblance ratings in Experiment 1. We computed confusion matrices for the categorization of actions and objects (Table 1). The values represent the number of times a stimulus was chosen as being produced by a certain action or object. Similar to Experiment 1, listeners correctly identified the actions and objects of stimuli produced by typical action-object interactions and confused different actions or objects for one another when they categorized atypical interactions. As in Experiment 1, bowed air columns and blown plates were categorized as blown air columns, and struck air columns were categorized as struck strings and struck plates. Listeners correctly categorized either the action or object and consequently perceived the complementary property it most typically interacts with. Listeners correctly categorized bowed plates as plates, but were confused between

bowing and striking. This may be because there was an impulsive sound arising from the bow's contact with the plate before the bowing occurs, which could be heard as a strike. For blown strings, which were rated as sounding like bowing, blowing, strings, and air columns in Experiment 1, listeners more decisively categorized them as bowed strings. Although they correctly identified the string, they were biased into perceiving bowing, the action most commonly applied to it.

Discussion

The current study demonstrates that action and object identification of nine types of action-object interactions depended on: (1) the familiarity with the interaction and (2) its perceived mechanical plausibility. Typical action-object interactions represented musical instrument families that listeners are familiar with: string instruments (bowed and struck strings), wind instruments (blown air columns), and percussive instruments (struck plates). Consequently, listeners are frequently exposed to the sounds these instrument families produce. Moreover, listeners were sensitive to the differences between impulsive (striking) and continuous (bowing, blowing) excitations, revealing that struck sounds were easily identified, but continuously excited sounds were often confused. For the atypical sounds, there was a general trend that listeners identified either the correct action or object and were consequently biased toward identifying the complementary property that most commonly interacts with it. This suggests that listeners seemed to interpret atypical interactions as conforming to a sound for which they already have mental models. Our findings indicate that it was either difficult for participants to perceive the actions or objects that interact atypically or that physically inspired modeling approaches cannot entirely convey what listeners simply do not have the mental models for. Although considering the timbre of a sound is essential for identifying its source properties, our findings suggest that our perceptions of sounds are not always accurate. Thus, novel sounds for which the actions or objects are difficult to identify are perceived as belonging to a category of sounds that has developed through a lifetime of exposure.

Acknowledgments

This research was financially supported through grants to Professor Stephen McAdams from the Natural Sciences and Engineering Research Council of Canada (NSERC). We would like to thank the Centre of Interdisciplinary Research in Music Media and Technology (CIRMMT) for providing us with the space to run participants.

References

Dudas, R. (2014). Modalys, Version 3.4.1 [computer software]. Paris: Institut de recherche et Coordination acoustique/musique.

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

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

Huynh, E. (2019). *Bowed plates and blown strings: Odd combinations of excitation methods and resonance structures impact perception* (Unpublished master's thesis). McGill University, Montréal, Québec.

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

Olson, C. L. (1974). Comparative robustness of six tests in multivariate analysis of variance. *Journal of the American Statistical Association*, *69*(348), 894–908.

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

Smith, B. K. (1995). PsiExp: An environment for psychoacoustic experimentation using the IRCAM musical workstation. In: Wessel D. (ed), *Society for Music Perception and Cognition Conference*, (pp.83-84). Berkeley, University of California.

New materials, new sounds: how metamaterials can change the timbre of musical instruments

Carolina Espinoza[1,2†], Alonso Arancibia[2], Gabriel Cartes[2] and Claudio Falcón[1]

[1] Departamento de Física, Facultad de Ciencias Físicas y Matemáticas, Universidad de Chile, Santiago, Chile
2 Departamento de Sonido, Facultad de Artes, Universidad de Chile, Santiago, Chile
[†] Corresponding author: carolinaaespinozao@uchile.cl

Introduction

The main goal of this research is to seek the expansion of the sound properties of existing traditional musical instruments, which through the time have evolved their forms and materialities from the technological advances of the cultures where they have been developed. The final purpose of this work is to contribute to this enrichment, applying the new possibilities that come from the development of metamaterials, trying to permeate into the culture and given to interpreters new ways of musical expression, in the same way that digital technologies, such as augmented reality, machine learning and others, are trying to do in the field of digital musical instruments (Tan & Lim, 2016; Bovermann et al., 2017).

Traditional musical instruments are complex acoustic systems. In most of them, the production of sound depends on the collective behavior of several vibrators, which can be made of different materials, weak or strongly coupled to each other (Fletcher & Rossing, 1998). If we couple an absorbent material to the sound box of a musical instrument, its sound changes: some harmonics are dampen or all the components of the sound are attenuated. In this research, the relevant question is: what happens if we couple to the sound box of a musical instrument a material with the capability of to absorb specifics and tunable ranges of frequencies? If the spectral content of a sound is modified perceptibly while the amplitude and fundamental frequency remain constant, we say that the timbre changes. The main goal of this work is to characterize acoustically synthetic materials, called tunable mechanical metamaterials, and to explore the opportunities of sound manipulation that they bring for the modification of the tonal qualities of musical instruments.

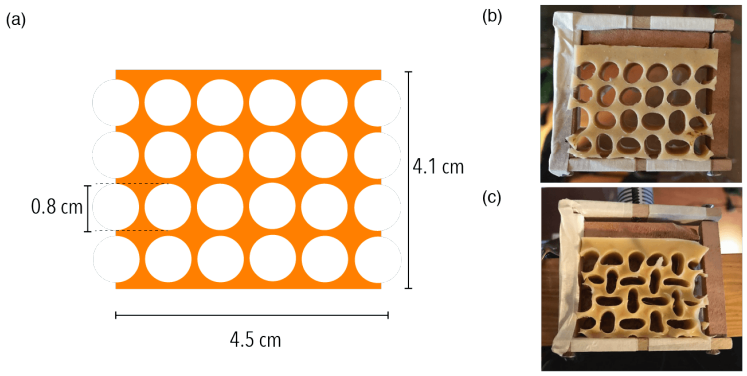


Figure 1: (a) Metamaterials: squared arrays of circular holes in a silicone elastomer matrix. (b) Metamaterial with a deformation of 1%. (c) Metamaterial with a deformation of 11%.

Metamaterials are rationally designed composites aiming at effective material parameters that go beyond those of their ingredients. Some mechanical metamaterials are artificially structured composite materials that enable manipulation of the dispersive properties of vibrational waves. We use mechanical metamaterials with the structure presented in figure 1 (a), i.e. arrays of circular holes in a elastomer matrix. These metamaterials exhibit auxetic behavior with a negative Poisson's ratio. It means that when the structure is stretched in the axial direction, it expands in the transversal direction, in contrast with typical materials (Bertoldi & Boyce, 2008; He & Huang, 2017). Furthermore, they have frequency band gaps

(ranges of frequencies with no vibrational transmission), which can be tuned by adjusting their geometry and/or stiffness by different levels of mechanical deformation (see figure 1 (b) and (c)), giving us a simple way of making tunable mechanical passband filters.

Experimental method

We make a metamaterial, labeled as M1, with the geometry shown in figure 1: an array of 6x4 circular holes of 8 mm of diameter in a elastomer matrix made of silicone Elite Double 22 Fast, with a thickness of 5 mm. In order to obtain its frequency response, it is compressed at different levels while a mechanical excitation is performed, using the experimental setup described in figure 2. A spectrum analyzer (Stanford System SR780) sends a sweep sine signal from 300 Hz to 1000 Hz with constant amplitude. The signal is amplified by a power audio amplifier (Gemini XGA5000) and a mechanical vibrator (SF 9324) is activated, which excites the metamaterial. An accelerometer (PCB 356A14), connected to a signal conditioner (PCB 408E09), receives the response of the metamaterial, and the power spectrum is obtained by the spectrum analyzer. The specimen is compressed by a mechanical press and the measurement is repeated. The frequency spectra are stored in a computer for further analysis.

In order to measure the effect of applying metamaterials to a vibrant system, we attached the characterized metamaterial M1 to the sound box of an acoustic guitar using a coupling gel, to obtain and analyze its acoustic behavior using the setup described in figure 3. A string tuned to a convenient frequency is plucked, with and without the M1 sample coupled to the sound box at different positions, and the sound is recorded for frequency analysis. A pencil condenser microphone Samson C02 and a Steinberg UR44 interface were used to perform the experiment. The metamaterial was coupled in positions A and B (see figure 3 (a)). Each measurement was performed five times in equal conditions.

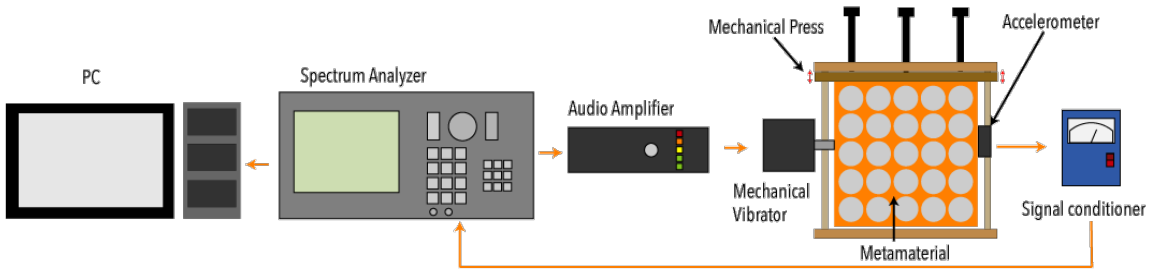


Figure 2: Experimental setup used for acoustic characterization of metamaterials: a spectrum analyzer, audio amplifier and a mechanical vibrator were used for excitation. An accelerometer, signal conditioner, and a computer were used for receiving and storing the material response. The deformation was performed using a mechanical press.

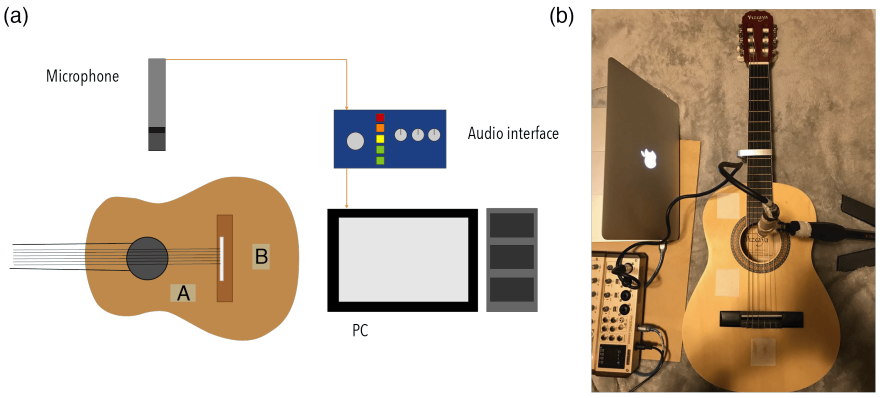


Figure 3: (a) Schematic of the experimental setup used for acoustic characterization of the influence of the metamaterial coupled to the sound box of an acoustic guitar. A pencil condenser microphone Samson C02 and a Steinberg UR44 interface were used to perform the experiment. The metamaterial was coupled in positions A and B. (b) Setup photography.

Results

Figure 4 (a) shows the power spectrum of the M1 sample in two states: the metamaterial without deformation (S1 spectrum, black dashed line), which presents one band gap at 400 Hz, and the metamaterial with a deformation of 11% (S2 spectrum, pink line), with a band gap around 530 Hz. It means that the components around 400 Hz of a mechanical vibration could be absorbed by the metamaterial without deformation, while if it is tuned with a compression of 11%, components around 530 Hz will be attenuated. The insert figure present the ratio S1/S2.

In figure 4 (b) we observed the average FFT of the sound signals produced by the acoustic guitar. The plucked string was tuned in 500 Hz. Black lines correspond to the results without coupling the M1 sample. Red lines are the FFT of the signals produced with M1 coupled in A position, with a deformation of 11%. Finally, blue lines are the same case than red ones, but with M1 positioned in B. Each measurement was performed five times. No changes in sounds were perceived in any case, but we observe a mean attenuation of 2.3 dB between the amplitude of the fundamental frequency, measured at 504 Hz, without and with M1. Although the sounds produced by the guitar are indistinguishable, we can see these changes in the frequency spectra. These preliminary results seem promising for future explorations that will allow us to change the timbre of musical instruments using metamaterials.

Conclusion

Our main result is that we can measure band gaps, between 300 Hz and 1000 Hz (audible regime), in soft, 2D periodic structures called mechanical metamaterials, which depend on their deformation level. In addition, we present preliminary measurements of the effect of attaching a metamaterial to the sound box of an acoustic guitar. Through a frequency analysis we show that although we cannot hear the effects of band attenuation, we can see them. Tunable mechanical metamaterials can be used as a tool to create mechanical filters that allow us to expand the acoustical properties of musical instruments, modifying their tonal qualities, in a dynamical and reversible way.

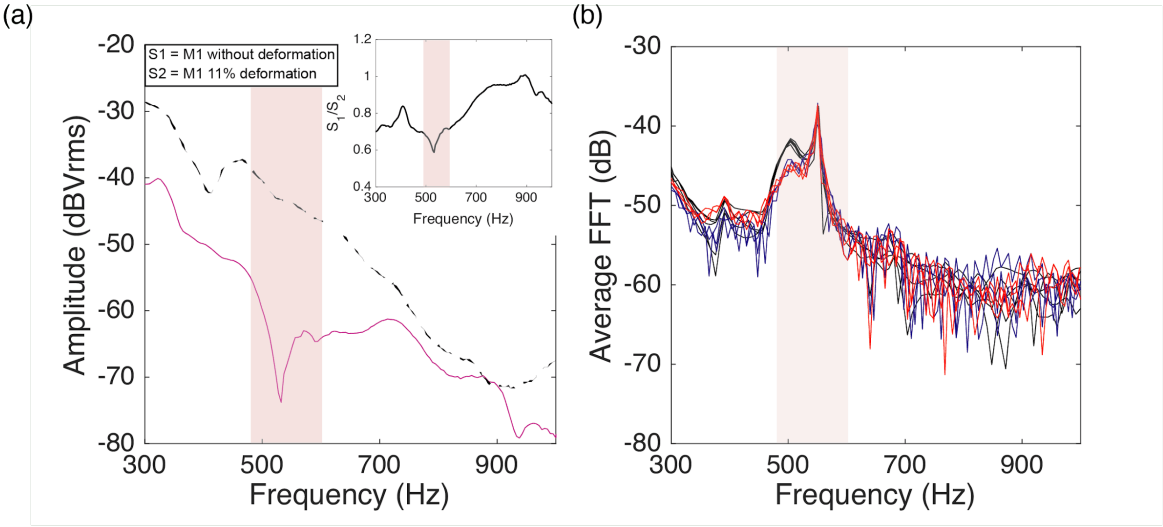


Figure 4: (a) Power spectrum for M1 without deformation (S1, black dashed line) and with deformation of 11% (S2, pink line). Frequency response presents a band gap around 530 Hz. (b) Fast Fourier transform average of of the sound signal produced by the acoustic guitar. A string tuned in 500 Hz was plucked: (black lines) without M1 coupled to its resonance box, (red lines) with 11% deformed M1 coupled in A position, (blue lines) with 11% deformed M1 coupled in B position. Each measurement was performed five times.

Acknowledgments

We acknowledge the support of Fondecyt Postdoctoral Grant #3200239, Fondecyt Grant #1190005 and Iniciativa Científica Milenio Grant #NM_CS_18_01.44.

References

Bertoldi, K., & Boyce, M. C. (2008). Mechanically triggered transformations of phononic band gaps in periodic elastomeric structures. *Physical Review B,* 77, 052105.

Bovermann, T., de Campo, A., Egermann, H., Hardjowirogo, S.-I. and Weinzierl, S. (2017). *Musical Instruments in the 21st Century. Identities, Configurations, Practices.* Singapore: Springer.

Fletcher, N. H., & Rossing, T D. (1998). *The physics of musical instruments*. New York, USA: Springer.

He, H. & Huang, H. (2018). Tunable Acoustic Wave Propagation Through Planar Auxetic Metamaterial. *Journal of Mechanics,* 34, 113-122.

Tan, K. & Lim, C. (2016). Development of traditional musical instruments using augmented reality (AR) through mobile learning. In: Faizatul A., Nifa A., Khai Lin C., Hussain A. (eds), *Proceedings of the 3rd International Conference on Applied Science and Technology ICAST'18* (pp. 020140-1–020140-6). Penang, Malaysia.

Timbre Latent Space: Exploration and creative aspects

Antoine Caillon[1†], Adrien Bitton[1], Brice Gatinet[1], Philippe Esling[1]
[1]Institut de Recherche et Coordination Acoustique Musique (IRCAM)
UPMC - CNRS UMR 9912 - 1, Place Igor Stravinsky, F-75004 Paris
[†]Corresponding author: caillon@ircam.fr

Introduction

Recent studies show the ability of unsupervised models to learn invertible audio representations using Auto-Encoders (Engel et al., 2017). While they allow high quality sound synthesis and high-level representation learning, the dimensionality of the latent space and the lack of interpretability of each dimension preclude their intuitive use. The emergence of disentangled representations was studied in Variational Auto-Encoders (VAEs) (Kingma et al., 2014, Higgins et al., 2017) and has been applied to audio. Using an additional perceptual regularization (Esling et al., 2018) can align such latent representation with the previously established multi-dimensional timbre spaces, while allowing continuous inference and synthesis. Alternatively, some specific sound attributes can be learned as control variables (Bitton et al., 2019) while unsupervised dimensions account for the remaining features. In this paper, we propose two models and suited interfaces that were developed in collaboration with music composers in order to explore the potential of VAEs for creative sound manipulations[2]. Besides sharing a common analysis and synthesis structure, one has a continuous latent representation and another has a discrete representation, which are applied to learning and controlling loudness invariant sound features.

Models

We consider a dataset of audio samples, such as performance recordings of an instrument. A variable-length audio x can be processed by analyzing series $\{x_0, \dots, x_L\}$ of signal windows $x_i \in R^{d_x}$ with an encoder E_ϕ mapping each frame into a latent code as $E_\phi : x_i \mapsto z_i \in R^{d_z}$. This encoder is paired with a decoder D_θ that inverts these features as $D_\theta : z_i \mapsto \hat{x}_i$. The vanilla auto-encoder optimizes its parameters $\{\theta, \phi\}$ on a reconstruction objective such that $\hat{x} \approx x$ (Figure 1).

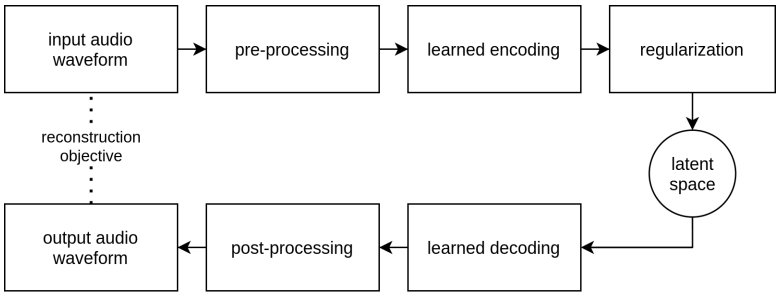


Figure 1. Block diagram of a VAE with optional pre and post audio processing.

Usually, we choose $d_z \ll d_x$ so that the latent variables embed a compressed representation of the data from which we can synthesize new samples. However, this continuous representation often remains highly-dimensional and does not disentangle data properties on separate latent dimensions. The usability of such representation and its quality for sampling or interpolation are thus limited. These considerations highlight the need for additional training objectives that enforce useful properties in the latent representation. We

[1] These authors contributed equally

[2] See https://acids-ircam.github.io/timbre_exploration/ for additional information about models, interfaces and sound examples.

consider two separate models, comparable in their overall encoder-decoder structure[3], but different in how the representation is regularized during training.

Continuous model. The first model aims to construct a latent space that is invariant to loudness in order to embed features that mainly account for the instrument timbre. It is achieved with an adversarial domain adaptation, where a latent regressor is trained at predicting loudness, and a gradient reversal optimization (Ganin et al., 2015) leads to a loudness-invariant encoder representation. Besides this objective, the VAE latent space is regularized on a Gaussian prior distribution n $N(0,1)$ which ensures local smoothness and favors independence between latent variables.

Discrete model. The second model is based on the Vector-Quantized VAE (VQ-VAE) proposed in van den Oord et al. (2017). It optimizes a discrete set of latent features q^j. Each encoder output is matched to its nearest codebook element $q_i^* \in \{q^0, \ldots, q^k\}$, before being decoded. This latent space is disentangled from a gain applied to the decoder output, which produces short-term features that are invariant to audio levels. Given that the set of latent features $q^j$is finite, we can analyze and map this codebook with acoustic descriptors.

Both models are intended to learn latent audio features that are invariant to loudness. The continuous model offers unconstrained and smooth feature manipulations. The discrete model can be analyzed in order to predict the output acoustic features embedded in the representation.

Experiments

Descriptor-based synthesis. Each vector of the discrete representation is individually decoded and the output signal is analyzed with a descriptor. It is thus possible to compute the mapping between a descriptor curve and the series of nearest latent features (details in Bitton et al., 2020). Latent synthesis can be directly controlled by following a user-defined descriptor target, as shown in figure 2. The codebook can be ordered and traversed according to different properties, such as centroid or fundamental frequency.

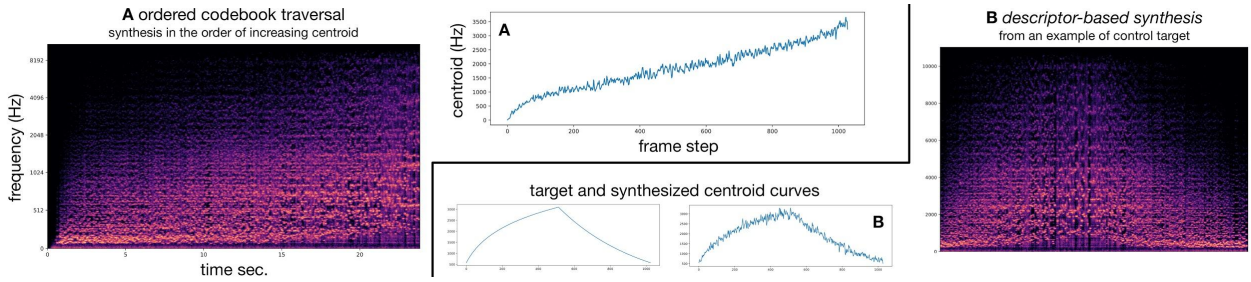


*Figure 2. The discrete representation can be analyzed with the spectral centroid and traversed in the increasing order (**A**). A control target can be synthesized by selecting the nearest latent features, the decoded audio approximately follows the curve provided (**B**).*

Continuous latent interpolations. In order to display the local smoothness of the continuous model, we consider the time variant linear interpolation z_{interp} between two latent series z_a *and* z_b of the same size inferred from two audio samples A and B. Decoding z_{interp} results in an audio sample smoothly interpolating between sample A and sample B, as shown in figure 3.

[3] Architectural differences are not detailed in this paper since we focus on discussing the representation properties.

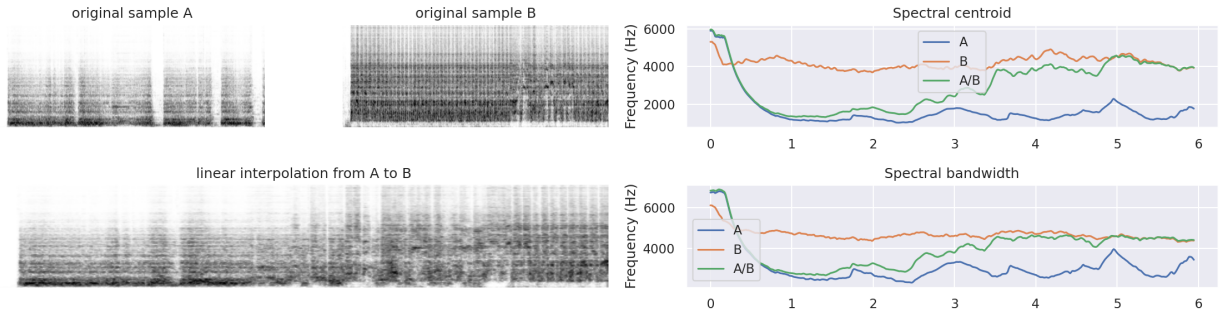


Figure 3. Linear interpolation in the latent space between two audio samples. We can see that the centroid and the bandwidth of the interpolated audio sample performs a smooth transition between those of the two original audios.

In order to facilitate a creative use of this model, we present two interfaces designed to circumvent the problem of identifying latent dimensions by facilitating their exploration.

Continuous model interfaces

The first interface is a Max/MSP application that is a graphical equivalent[4] to the command line tools we usually have to test the model. It features several high-level interactors such as mathematical operators on the latent series, manual editing, and an interpolation plane. We have built this application in collaboration with A. Schubert[5], aligning with his remarks on how to improve visualization and control over the generation. This interface is intended to be used in order to grasp the main characteristics of a model trained on a specific dataset.

This stand-alone interface has built-in interactions but a limited integration and restrictions in the possible operations. We have thus developed a second interface built in collaboration with B. Gatinet, implementing the encoder and the decoder as PureData abstractions that can be combined with any other regular objects. New aspects of the continuous model emerge from this interface, as it allows uninterrupted exploration with realtime rendering, enabling the use of complex signal processing techniques on both the audio and latent series. As this interface can be integrated in real time inside a digital audio workstation, it is more suited for composition workflows. It is furthermore a strict superset of the first interface in terms of functionalities.

The use of these interfaces has brought to light new ways of generating audio signals, whether by explicit control of an audio descriptor, or by morphing between different existing sounds. Training a model on an audio domain and using it to resynthesize an audio sample from a different domain can also lead to an implicit synthesis method. Additional results on audio conversion of instrument sounds can be found in Bitton et al., (2020).

Conclusion

This research has studied VAEs with continuous and discrete latent sound representations as creative tools to explore timbre synthesis. The discrete model allows the generation of a new audio signal by directly controlling acoustic descriptors. Manipulations of the continuous model are eased by developing specific interfaces and real-time rendering, which greatly enrich composition and sound design possibilities. And in turn, it gives further insights on the generative qualities found in the learned representations, as well as the relevance of their different parameters and controls with respect to the new timbres that are synthesized.

[4] See our website for a screenshot of the interface

[5] See http://www.alexanderschubert.net/

Acknowledgment

This work is supported by the ANR:17-CE38-0015-01 MAKIMOno project, the SSHRC:895-2018-1023 ACTOR Partnership and Emergence(s) ACIDITEAM project from Ville de Paris and ACIMO projet from Sorbonne Université. The authors would also like to thank Alexander Schubert for his creative inputs.

References

Bitton, A., Esling, P., Caillon, A., & Fouilleul, M. (2019). Assisted Sound Sample Generation with Musical Conditioning in Adversarial Auto-Encoders. In: Stables R., Hockman, J., Välimäki, V., Fontana F. (eds), *Proceedings of the 22nd International Conference on Digital Audio Effects*. Birmingham, UK.

Bitton, A., Esling, P. & Harada, T. (2020). Vector-Quantized Timbre Representation. In *Arxiv:2007.06349*.

Engel, J., Resnick, C., Roberts, A., Dieleman, S., Norouzi, M., Eck, D., & Simonyan, K. (2017). Neural Audio Synthesis of Musical Notes with WaveNet Autoencoders. In *Proceedings of the 34th International Conference on Machine Learning, PMLR 70*. http://arxiv.org/abs/1704.01279.

Esling, P., Chemla, A., & Bitton, A. (2018). Bridging Audio Analysis, Perception and Synthesis with Perceptually-regularized Variational Timbre Spaces. In: Gomez E., Hu X., Humphrey E., Benetos E. (eds), *Proceedings of the 19th International Society for Music Information Retrieval Conference*. (pp.175-181). Paris, France.

Ganin, Y., Lempitsky, V. (2015). Unsupervised Domain Adaptation by Backpropagation. In: Bach F., Blei D. (eds), *Proceedings of the* 32nd *International Conference on Machine Learning* (pp.1180-1189). Lille, France.

Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S. & Lerchner, A. (2017). beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. *Proceedings of the 5th International Conference on Learning Representations (ICLR)*. Toulon, France

Kingma, D. P., & Welling, M. (2014). Auto-Encoding Variational Bayes. In Bengio Y., Lecun Y. (eds), *Proceedings of the 2nd International Conference on Learning Representations (ICLR)*. Banff, Canada.

Van den Oord, A., Vinyals, O., & Kavukcuoglu, K. (2017). Neural Discrete Representation Learning. In: Guyon I., U.V. Luxburg U.V., Bengio S., Wallach H., Fergus R., and S. Vishwanathan S., Garnett R. (eds), *Advances in Neural Information Processing Systems 30* (pp. 6306–6315). Long Beach, CA.

Uncovering the meaning of four semantic attributes of sound: Bright, Rough, Round and Warm

Victor Rosi[1†], Olivier Houix[1], Nicolas Misdariis[1] and Patrick Susini[1]
[1] Sound perception and sound design group, STMS Lab (IRCAM-CNRS-SU), Paris, France
[†]Corresponding author: rosi@ircam.fr

Introduction

Several studies discuss the semantics of words used by sound professionals or musicians to describe timbre in particular situations, such as sound engineering or instrument playing. The present study aims to understand the use and the definition of four terms selected from the sound lexicon developed by Carron et al. (2017) as they are cited in numerous studies for sound description. Bright (*brillant*), round (*rond*), warm (*chaud*) and rough (*rugueux*) are four terms vastly used in the French language for sound description in sound creation processes such as music performance, orchestration, sound engineering or sound design, yet they lack formal, standardized definitions. This work is based on interviews with sound professionals from these different fields. The goal is to get definitions, or semantic portraits, for each word with corresponding sound samples from a musical instrument dataset.

Method

We organized individual interviews with 32 French-fluent sound professionals (musicians, composers, sound designers, acousticians...), during which the four terms were discussed sequentially with the participants. The study of one term had two main parts, during the first part, the interviewees were asked to give a definition of the studied term, then they selected sound samples from a musical instrument dataset that match their perception of the term. During the second part, the interviewees first chose sound samples that were opposed to the studied term and then tried to define the opposite concept. The sound dataset was mainly composed of the Ircam Studio-Online Library (SOL) mixed with parts of the Vienna Symphonic Library (VSL) for additional instruments. The dataset was presented to the participants through a Max/MSP interface they could manipulate.

After the transcription of the interviews, the definitions of the studied terms were processed with basic NLP (Natural Language Processing) steps: tokenization, lemmatization and filtering of the stop words. We assessed the lemma/interviewee frequency for each term (i.e. the number of interviewees using one lemma for each definition).

Results

Informed by the literature on semantic analysis of timbre (Wallmark, 2018; Carron et al., 2016; Porcello, 2004; Faure, 2000), and the experimental raw data, we proposed 10 categories in order to structure the data and to better compare the description strategies for the four terms. These categories are organized in three greater categories. The first one groups all the acoustic, the second one groups all the information on the source, and finally the third category gathers the metaphorical descriptions of sound. The categories were validated with an inter-rater agreement measure with the four authors (Fleiss' kappa $\kappa = 0.69$, $p<0{,}001$). The categories along with verbal examples used for the description of the terms extracted from the corpus are presented in Table 1.

Table 1 – Categories of description strategies with verbal examples translated in English along with the original verbatims in French.

Acoustic	
Spectral	high-pitch (*aigu*), harmonics (*harmoniques*), medium (*medium*)
Temporal	attack (*attaque*), release (*décroissance*), steady (*stable*)
Dynamic	*forte, piano, crescendo*
Sound specific semantic	nasal (*nasal*), resonant (*résonnant*), noisy (*bruité*)
Source related	
Source	trumpet (*trompette*), voice (*voix*), orchestra (*orchestre*)
Excitation mode	rub (*frotter*), *vibrato*, breathing (*souffler*)
Metaphoric	
Crossmodal correspondance (CMC)	warm (*chaud*), harsh (*dur*), clear (*clair*)
Matter (shape, density, material)	round (*rond*), full (*plein*), organic (*organique*)
Effect	enveloping (*enveloppant*), scratching (*qui gratte*), straightforward (*franc*)
Affect	pleasant (*agréable*), aggressive (*agressif*), comforting (*réconfortant*)

The description strategies most used for each term helped to shape the definitions. For instance, we noted that all the participants defined *Bright* through spectral descriptions, while *Rough* was more often explained metaphorically through analogies to the sense of touch, with temporal aspects or with definitions of an excitation mode. Finally, *Round* and *Warm* shared many similarities in their metaphorical and spectral descriptions, although, *Round* seems to be substantially more described by temporal aspects compared to *Warm*.

From these results, by parsing the context of the most frequently occurring lemmas in each definition, and with a comparison with the most elected sound samples we were able to summarize semantic portraits for each term:

- A *bright* sound has most of the spectral energy in the high frequencies. It is often a high-pitched sound that can be composed with a sharp attack.
- A *warm* sound seems to be a low-pitched or mid-low-pitched sound. It gives a feeling of spectral richness in the mid-low frequencies. A *warm* sound has a rather soft attack. It is a fairly pleasant sound that gives a sensation of envelopment.
- A *round* sound has a soft attack and is temporally stable. It tends to also have a soft release or a long resonance. A *round* sound is spectrally perceived as full with a spectral balance located in the mid-low frequencies.
- A *rough* sound is temporally unstable; it presents fast temporal variations that can bring some sort of noise. It gives a rubbing/scratching sensation.

While these free verbalizations allowed us to dig deep into the nature of the four types of sounds, it makes the work of the researcher tedious and conducive to interpretation as there are many syntactic elements, negations or quantifiers to take into account in the definitions. Because of the diverse and sometimes conflicting nature of the definitions given by the experts, it is necessary to homogenize, reduce and hierarchize the information gathered in order to formulate both accurate and detailed definitions.

In a second study, a corpus of phrases was extracted from the verbatims and related the most occurring lemmas selected from the definitions of each term to their oppositions gathered in the interviews. As part of an online experiment currently in progress, we want to ask a bigger population of sound professionals the degree of familiarity that they associate with the phrases of the corpus and the relevance of this phrase to the definition of the associated term. There are other subsequent goals to this study: first, clustering some presumably similar descriptions (e.g. *"a sound with a soft attack", "a sound with a slow attack", "a sound*

without an attack"), enabling a disambiguation and reduction of the information. secondly, we expect to better understand the use of certain metaphorical description such as *rich*, or *full.*

Discussion

This two-part study allows us to report the variety of description strategies employed by sound experts in the French language. The methodology employed could be used for the study of other semantically ambiguous terms related to sound. Finally, the definitions formulated and the sound samples will be incorporated in the sound lexicon (*SpeaK*) currently in development, following Carron's work. One of the purposes of this lexicon is to enable better communication about timbre descriptions of sound in a sound design process.

The next step of our study is to connect the definitions of the four terms, *bright, warm, round* and *rough,* with their acoustic characterizations. Firstly, we chose to annotate the sound dataset used during the interviews (~600 sounds) with the four terms. To that end, we are currently adapting a novel annotation experiment called *Best-Worst Scaling* (BWS). During a BWS procedure, at each trial, participants are asked to elect the best and the worst items along a latent subjective dimension in a tuple of *N* items. At the end of the procedure, scores are computed for each item using for instance a simple counting method that results in a ranking of all the items. For example with the term *Bright,* we could gather scores from the most to the least *bright* sound of our dataset.

Previous studies (Hollis & Westbury, 2018; Kiritchenko & Mohammad, 2017) have adapted this method, originally designed for small datasets, to semantic research with many-item datasets. The comparison with procedures using rating scales showed that BWS gives better consistency. BWS seems to have the perks of pairwise comparison without its time consuming downfall that prevents from using such methodology in many-item paradigms.

Following this step, we imagine a feature extraction procedure in the form of a machine learning experiment whose purpose will be to obtain the salient acoustic correlates responsible for the ranking of the sounds along each studied term. Ultimately, we wish to create a validation experiment that will confront the previously obtained definitions and the sound samples elected by the experts with designed sounds depending on the result of the machine learning experiment. These results aim to propose a methodology in order to define terms frequently used for timbre description. This approach could be used for other terms and languages.

References

Carron, M., Rotureau, T., Dubois, F., Misdariis, N., & Susini, P. (2016). Speaking about sounds: A tool for communication on sound features. *J Design Res 15*(2), 85–109.

Faure, A. (2000). *Des sons aux mots, comment parle-t-on du timbre musical?* (PhD Thesis). Ecole des Hautes Etudes en Sciences Sociales (EHESS), Paris.

Porcello, T. (2004). Speaking of Sound: Language and the Professionalization of Sound-Recording Engineers. *Social Studies of Science*, *34*(5), 733–758. https://doi.org/10.1177/0306312704047328

Wallmark, Z. (2018). A corpus analysis of timbre semantics in orchestration treatises. *Psychology of Music*, *21, 1-21*. https://doi.org/10.1177/0305735618768102

Hollis, G., & Westbury, C. (2018). When is best-worst best? A comparison of best-worst scaling, numeric estimation, and rating scales for collection of semantic norms. *Behavior Research Methods*, *50*(1), 115-133. https://doi.org/10.3758/s13428-017-1009-0

Kiritchenko, S., & Mohammad, S. M. (2017). Best-Worst Scaling More Reliable than Rating Scales : A Case Study on Sentiment Intensity Annotation. *ArXiv:1712.01765 [Cs]*. http://arxiv.org/abs/1712.01765

Louviere, J. J., Flynn, T. N., & Marley, A. A. J. (2015). *Best-worst scaling : Theory, methods and applications*. Cambridge: Cambridge University Press.

The difference between shrieks and shrugs: Spectral envelope correlates with changes in pitch and loudness

K. Jake Patten[1†] and Michael K. McBeath[2]
[1] College of Health Solutions, Arizona State University, Tempe, Arizona, USA
[2] Department of Psychology, Arizona State University, Tempe, Arizona, USA
[†] Corresponding author: kjp@asu.edu

Introduction

Stimulus intensity can impact the perception of pitch. Stevens (1935) and, later, Gulick (1971) demonstrated that, for non-dynamic tones, pitch and loudness are positively correlated for frequencies above 2 kHz but become negatively correlated below 250 Hz. More ecologically valid, dynamic sounds, however, function differently. Pitch and loudness are strongly correlated throughout the range of audible frequencies. Participants experience an illusory increase in pitch as auditory objects draw near and a similar illusory decrease in pitch as objects recede (Neuhoff & McBeath, 1996). The experienced change can be as much as 8 semitones when the physical change amounts to only 2; a full half-octave greater than the actual change (McBeath & Neuhoff, 2002). This bias may have arisen because natural sounds (up to roughly 1 kHz), musical instruments, and human and animal vocalizations naturally exhibit simultaneous increases and decreases in f_0 and intensity (Johnston, 2009; McBeath, 2014; Wang, Astfalck, & Lai, 2002). In particular, when moving from whispering to normal speech to shouting over noise to yelling in anger, frequency consistently rose along with intensity (Scharine & McBeath, 2018). Past research has also shown that shouting is often correlated with a consistent shallowing of the spectral envelope and a reduction in harmonicity (Raitio et al., 2013; Wallmark & Allen, 2020), as well as being identified as one of two primary organizational non-dynamic dimensions of timbre (Patten, McBeath, & Baxter, 2018). The current study investigates the existence of a reliable correlation and natural regularity between spectral envelope, fundamental frequency, and intensity.

To test this, the current study uses North American vowel phonemes, which – when controlled for fundamental frequency and intensity – are changes in timbre that are well-known to all participants. One well-known phenomenon pertaining to vowel sounds is intrinsic fundamental frequency (If_0); the finding that some North American vowels are often voiced higher than others (Crandell, 1925). This production bias may be partly explained by anatomical constraints, as some researchers have found evidence of this bias persisting across cultures (Whalen & Levitt, 1994). Furthermore, high vowels are far more susceptible to f_0 changes when a speaker's larynx is raised than low vowels (Sundberg & Nordstrom, 1976). However, other researchers find evidence of If_0 in some registers of tonal languages and not others or of If_0 not existing at all (Connell, 2002; Zee, 1980). It is possible that If_0 is not solely determined by biophysical constraints, but by the perception of the pitch of vowels and other natural sounds.

Method

Participants in all experiments were undergraduate students enrolled in introductory psychology and speech and hearing science classes at Arizona State University. The average age in all three experiments ranged from 19.4 to 22.6 years.

Experiment 1A. Participants listened to 10 North American monophthongs as well as the steady [a] and [o] portions of [aI] and [ou] spoken by a professional voice actor and recorded in a B_T environment. All vowels were digitally altered to hold f_0 and intensity constant. Participants were first asked to rate the similarity (1 – 10, latter being most similar) of all paired vowels. These ratings were used to derive the multidimensional scaling (MDS) solution in Figure 1. Next, participants were asked to order the vowels from highest to lowest pitch, allowing for a correlation to be computed between the MDS solution and tone height.

Experiment 1B. The MDS solution from 1A informed the choice of phoneme in this and subsequent experiments. Participants were asked to move a mouse on their screen to indicate perceived changes in pitch and loudness (all participants indicated both in two separate, counterbalanced blocks) for vowels that changed in timbre ([i] to [ʌ]), f_0, and/or intensity.

Experiment 2. This experiment consisted of an analysis of different song types to understand how the perceptual and production bias of the pitch of [i] and [ʌ] is manifest in different song types. Scat, a vocalization style where the mouth is used to replicate instruments, was hypothesized to exhibit the largest production bias as the singer would use all techniques to increase their vocal range. Conversational interviews from podcasts were hypothesized to exhibit a small, though significant difference. Finally, lyrical songs – with constraints at the word, phrase, or musical score level – were hypothesized to show no difference. The f_0 of the first 30 instances of both [i] and [ʌ] were observed for six instances of each song type.

Experiment 3. To further demonstrate the bias in using [i] to produce high f_0 sounds, participants were tasked with recreating sounds, both high and low, outside their vocal range (60 Hz and 8 kHz sine waves). Participants' were unconstrained in the vocalizations they could make, though all used vowels.

Results

Experiment 1A. Participants' ratings were used to construct the MDS solution in Figure 1. The *y*-dimension of the solution correlates with rated pitch (though all phonemes were presented at a constant f_0 and intensity), r (10) = .90, $p < .001$. This is greater than the correlation of the *y*-dimension – ostensibly, tone height or experienced pitch – and the second formant, r (10) = .76, $p < .05$. The *x*-dimension of the solution correlates with harmonicity, r (10) = .60, $p < .05$.

Experiment 1B. Psychophysical functions derived from the results reveal that moving from [i] to [ʌ] is equivalent to a .38 semitone decrease in f_0 and a .75 dB decrease in intensity.

Experiment 2. A repeated measures ANOVA reveals an overall mean difference of fundamental frequency between [ʌ] (M = 209.95 Hz), [I] (M = 250.90 Hz), and [i] (M = 261.45 Hz) phonemes [main effect of phoneme, F (2, 438) = 27.02, $p < .001$, $\eta^2 = .04$] and between interviews (M = 147.53 Hz), songs (M = 251.48 Hz), and scat (M = 323.29 Hz) [main effect of type of recording, F (2, 438) = 90.95, $p < .001$, $\eta^2 = .42$]. Importantly, all three hypotheses regarding the extent of the production bias across song type were borne out by a significant interaction between phoneme and recording type, F (4, 438) = 17.15, $p < .001$, $\eta^2 = .05$. This can be seen in Figure 2.

Experiment 3. The [i] phoneme is chosen much more often to produce a high-pitched sound than other phonemes and the [ʌ] phoneme is selected more often for replicating low-pitched sounds, χ^2 (5, 68) = 64.57, $p < .001$, Cramer's V = .44. This is shown in Figure 3.

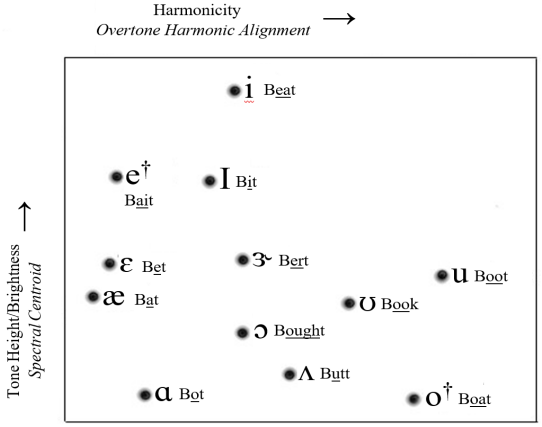


Figure 1. The MDS solution from Experiment 1A. Axes are labeled by the timbre dimension they represent and how that dimension was calculated.

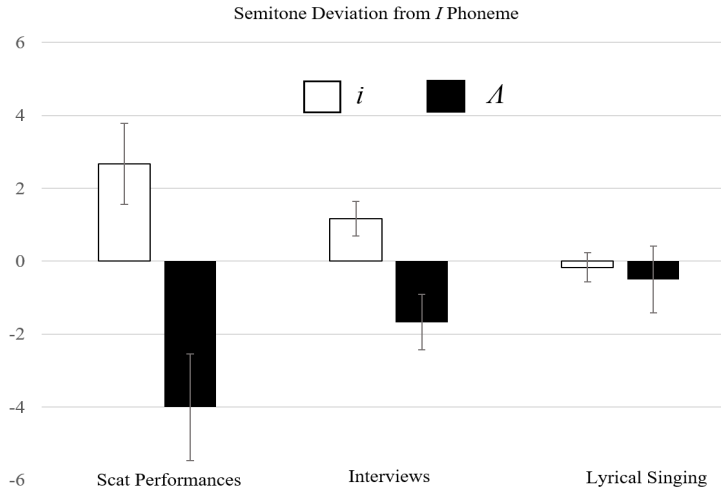


Figure 2: Semitone deviations from [I] for [i] and [ʌ]for scat songs, interviews, and lyrical singing. There was a significant difference for scat songs and natural conversation, though the latter exhibited the bias to a lesser degree. There was no difference for lyrical singing.

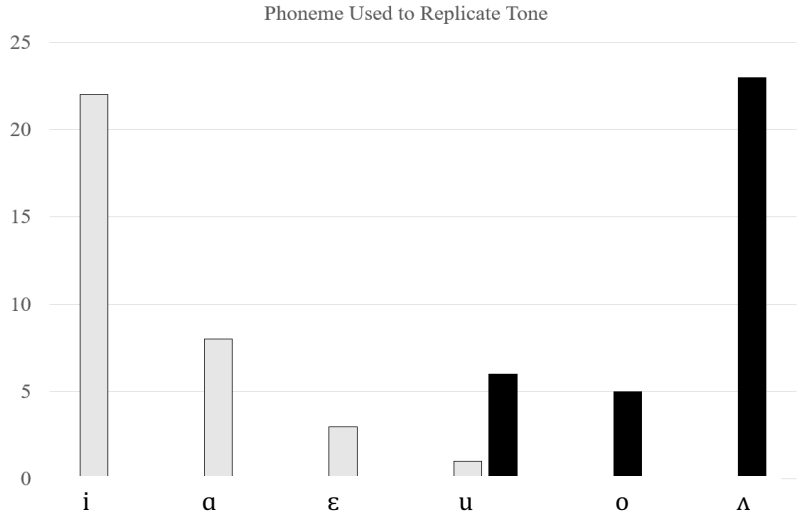


Figure 3. Phonemes used to replicate high (8 kHz) and low (60 Hz) tones in Experiment 3.

Discussion

Experiment 1A illustrated the way vowel phonemes are organized in cognitive space when f_0 and intensity are held constant. This experiment also demonstrated that the two most salient dimensions of vowel phoneme organization are tone height and harmonicity. The extreme vowels of [i] and [ʌ] informed the construction of Experiment 1B, which quantified the extent to which timbre impacts perceptions of frequency and intensity. This finding also suggests the existence of a natural regularity between spectral envelope, f_0, and intensity. Experiment 2 tested the prevalence of this natural regularity in a condition where singers would want to exploit it (scat singing), in a natural setting (interviews), and a condition where the regularity would be suppressed (lyrical singing). Finally, Experiment 3 pushes the findings of Experiment

2 further by tasking non-singers with replicating sounds outside their range. On average, participants used the [i] phoneme to replicate a very high tone and the [ʌ] phoneme to replicate a very low sound. Overall, these three experiments demonstrate a high degree of correlation between spectral envelope, f_0, and intensity. This correlation can be used to enhance the function of synthetic speech, voice recognition, hearing aids, and communication compression algorithms. These findings also support the use of the natural regularities framework to investigate timbre through illusions and discover new truths about perception and hearing.

References

Connell, B. A. (2002). Tone languages and the universality of intrinsic f_0: Evidence from Africa. *Journal of Phonetics*, *30*, 101-129.

Crandall, I. B. (1925). The sounds of speech. *The Bell System Technical Journal, 4*(4), 586-639.

Gulick, W. L. (1971). *Hearing: Physiology and Psychophysics*. New York: Wiley.

Johnston, I. (2009). *Measured Tones. Boca Raton*, FL: CRC Press.

McBeath, M. K. (2014). The Fundamental Illusion. *Paper presented at the 55th annual meeting of the Psychonomic Society*, Long Beach, California.

McBeath, M. K., & Neuhoff, J. G. (2002). The Doppler effect is not what you think it is: Dramatic pitch change due to dynamic intensithy change. *Psychonomic Bulletin and Review, 9*(2), 306-313.

Neuhoff, J. G., & McBeath, M. K. (1996). The Doppler illusion: The influence of dynamic intensity change on perceived pitch. *Journal of Experimental Psychology: Human Perception and Performance, 22*(4), 970-985.

Patten, K. J., McBeath, M. K., Baxter, L. C. (2018). Harmonicity: Behavioral and neural evidence for functionality in auditory scene analysis. *Auditory Perception and Cognition, 1*(3-4), 150-172.

Raitio, T., Suni, A., Pohjalainen, J., Airaksinen, M., Vainio, M., & Alku, P. (2013). Analysis and synthesis of shouted speech. *INTERSPEECH*, 1544-1548.

Scharine, A. A. & McBeath, M. K. (2018). Natural regularity of correlated acoustic frequency and intensity in music and speech: Auditory scene analysis mechanisms account for integrality of pitch and loudness. *Auditory Perception & Cognition*, *1*(3-4), 205-228.

Stevens, S. S. (1935). The relation of pitch to intensity. *Journal of the Acoustical Society of America, 6*(3), 150-154.

Sundberg, J. & Nordström, P-E. (1976). Raised and lowered larynx – The effect on vowel formant frequencies. *STL-QPSR, 17*(2-3), 035-039.

Wallmark, Z. & Allen, S. E. (2020). Preschoolers‘crossmodal mappings of timbre. *Attention, Perception, and Psychophysics, 82*, 2230-2236.

Wang, C., Astfalck, A., & Lai, J. C. S. (2002). Sound power radiated from an inverter-driven induction motor: Experimental investigation. *IEE Practical Electrical Power Applications, 149*(1), 46-52.

Whalen, D. H. & Levitt, A. G. (1994). The universality of intrinsic f_0 of vowels. *Haskins Laboratories Status Report on Speech Reseach SR-117/118*, 1-14.

Zee, E. (1980). Tone and vowel quality. *Journal of Phonetics, 8*(3), 247-258.

Distorted Pieces of Something: A Compositional Approach to Luminance as a Timbral Dimension

Ivonne Michele Abondano Flórez
School of Music (PGR), University of Leeds, Leeds, West Yorkshire, UK
michele.abondano@gmail.com

Introduction

Generally, there is a need to describe timbre in relation to physical features of the possible sound source or the type of experience involved in its perception. Wallmark has provided a typology of timbre descriptors in seven basic categories: affect, matter, cross-modal correspondence, mimesis, action, acoustics, and onomatopoeia (2019). Zacharakis, Pastiadis & Reiss have studied English and Greek words that describe timbre as well as the correlations between them and the acoustic conditions that determine timbral perceptual qualities to propose three semantic dimensions: luminance, mass, and texture (2014). Here, I approach *luminance* as the semantic dimension that accounts for the amount and intensity of light perceived in timbre, through the compositional process of the piece *Distorted Pieces of Something. Study on Light (when it rains)* (2019), for soprano saxophone and viola.

Luminance is the dimension that describes timbre in terms of how *brilliant* it is, as a metaphor associated with visual perception: major inharmonicity and a stronger spectral centroid fluctuation seem to reduce the brilliance perceived in timbre, while the presence of a fundamental frequency tends to make timbre brilliant (Zacharakis, Pastiadis & Reiss, 2014). Lochhead develops the concept of 'radiance' as a formal property that emerges from the interaction of three types of musical phenomena: moments of sonic *luminance*, moments of '*flickering*', and moments of *intensity*, so the 'events' of luminance include sounds with prominent upper partials, higher pitch, and a louder dynamic (2016). Liza Lim approaches the concept of 'shimmer' in her compositions inspired by a sacred painting technique called *bir' yun*, developed by the ancestral culture Yolngu in Australia, in which fine cross-hatching drawn in high-contrast colours over the surface projects a shimmering brightness that is read as a representation of ancestral power as well as felt as a direct manifestation of it (Rutherford-Johnson, 2011). These ideas of 'flickering' and 'shimmer' stress the importance of movement and contrast to understand that the perception of light in timbre is neither a steady nor fixed experience.

The musical analysis of new music repertoire has enabled me to recognise how musical descriptors for luminance seem to rely on metaphors or poetical expressions that evidence a desire to distinguish the identity of timbre clearly, a kind of effort to discover its purity or describe the obstacles to bringing it about. In the piece *eyam I (it takes an ocean not to)*, Ann Cleare uses several expressions to accompany specific instrumental techniques. For instance, for the dyad multiphonics (alternating with aeolian tones or just air) she writes: 'gently and fluently: like a whispering voice in the fog', or 'New element III: tiny piercing lights' (2009-14). Such evocative experiences could be central to explorations of nuances in the perception of light in timbre. Indeed, Amato writes: "in illuminating the desirable, lighting exposes the undesirable" (2001). The experience of light could transgress the idealisation of timbre; the presence of light could make evident conditions that were covered by other factors, but an excess of light could exaggerate or distort some characteristics. On the other hand, the absence of sound has been usually associated with the experience of darkness. The perception of different levels of darkness in timbre can be approached from the impossibility of listening as a consequence of very soft dynamics, but also as an intentional covering effect when a timbral quality is behind a kind of screen that hides it. Rebecca Saunders explores the phenomenon of silence in her work *vermilion* (2003) inquiring about its inner qualities and function (Saunders & musikFabrik, 2008). She uses the following fragment from Samuel Beckett's *Company* as an epigraph for the score: "By the voice a faint light is shed. / Dark lightens while it sounds. / Deepens when it ebbs. / Is whole again when it ceases." (1980).

These references exemplify the strong influence of the visual experience of light in the perception of luminance in timbre. Consequently, this research seeks the development of compositional strategies that allow to approach this timbral dimension from a creative and technical perspective.

Method

I propose a scale of two phases that measures the *amount* of light going from dark to clear, as well as in terms of the *intensity* of light for qualities from clear to distorted. The terms used to mark each level reunite the most common descriptors for the perception of luminance in timbre that have been recognised through the literature review and analysis of instrumental repertoire for this research, as well as my own experience and compositional criteria.

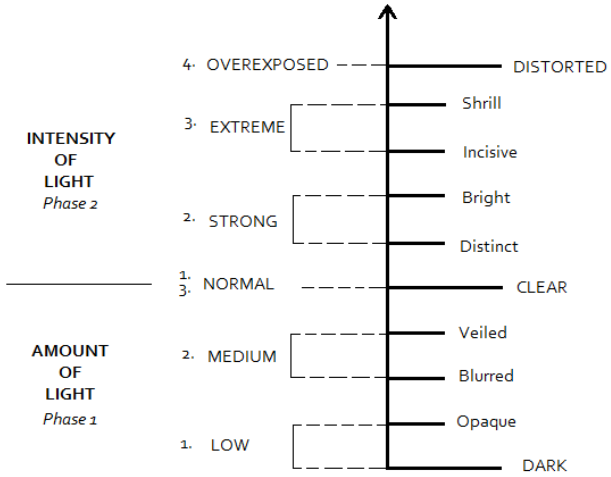


Figure 1: Luminance Scale for the measurement of intensity and amount of light perceived in timbre.

The levels of the scale function as structural points for five uninterrupted micro-sections: *a* (bright-incisive), *b* (shrill), *c* (distorted), *d* (normal-distinct), *e* (blurred-veiled). Saxophone and viola techniques are developed and categorised according to the specific timbral experience pursued in each section. The composition is approached as a process of interconnections of temporal luminance experiences in which the instrumental techniques are continuously altered through changes in the execution that directly affect the levels of luminance perceived in the timbral experience. This piece was workshopped with Wojciech Psiuk (saxophone) and Aleksandra Demowska-Madejska (viola) during the SYNTHETIS International Summer Course for Composers (July 15-27, 2019). Recordings of these sessions have been spectrally analysed to recognise the acoustic characteristics and the conditions of the interaction between the parameters of sound that determine the timbral behaviour and lead the resulting timbre to be measured according to the luminance scale.

Results

Results are not separable from the specific compositional experience of the piece *Distorted Pieces of Something. Study on Light (when it rains)*; therefore, it is recommended to listen to the work for a clearer understanding of this information. Thus, two parameters of sound are recognised as fundamental in the perception of luminance from a compositional perspective: dynamic envelope and spectral content.

Dynamic envelope: a soft attack can make timbres unclear or undefined mostly because parameters like pitch and harmonic content cannot be easily discriminated, which leads listeners to perceive qualities like those in the first phase of the scale while, with a medium attack these parameters become more precise, making timbres clearer or even distinct and bright. Very strong attacks can easily lead to the perception of

the inharmonicity of timbre making it complex or unstable and more likely to be in the third and fourth stages of the intensity phase of the scale. Moreover, the amount or intensity of light perceived through the development of a timbre can be affected by dynamics. The increase or reduction of loudness has a direct consequence for the spectral centroid: that is, constant medium dynamics can make timbres steady or stable during the course of their development and likely to be perceived as clear. Consequently, loud dynamics stimulate large variations and fluctuations in the spectral centroid, which results in timbres that are more likely to be in the top range of the intensity of light. Very low levels of loudness can make timbres unstable producing unclear fundamental pitches and increasing inharmonicity, which tends to be perceived in the first phase of the scale.

Spectral content: the presence of high partials and the distribution conditions that produce harmonic dissonance can lead to the perception of both extremes of the scale according to the level of loudness. Then, the perception of high intensity of light in timbre can be a consequence of high loudness while soft dynamics can lead to perceive low amount of light. In addition, timbres with a strongly present fundamental frequency and low perception of higher partials seem to be perceived as clear. Particularly, low pitch fundamentals are more likely to be perceived in the range of the first phase of the scale.

The following graphs allow the analysis of timbre from the relation between the instrumental techniques and the compositional approach.

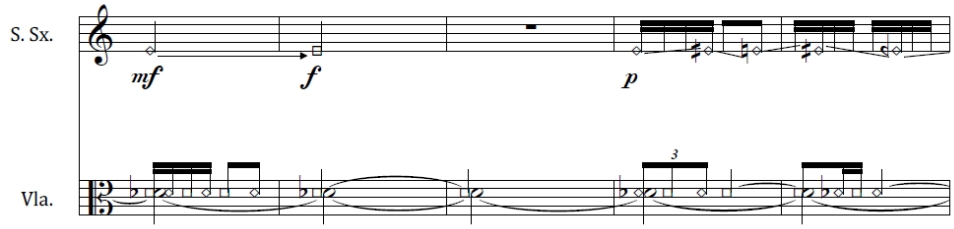


Figure 2: Score excerpt (mm. 118-122)

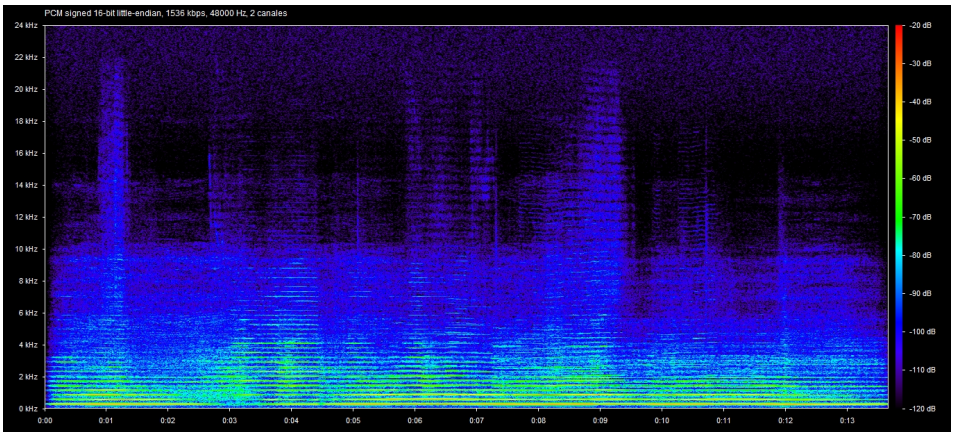


Figure 3: Spectrum corresponding to the excerpt showed in Fig. 2 (premiere recording).

This timbre, composed for *section e*, can be perceived as *veiled-blurred* because of the lack of higher partials. The fundamental frequency is weak due to the instrumental techniques approached: saxophone's aeolian sounds and viola's very low finger pressure. The inharmonicity produced by these techniques as well as the fluctuation generated by the proximity of frequencies played in both instruments at the same time, act as a filter for light that prevents the perception of a complete revealed timbre. This condition is reinforced by general low levels of loudness.

Discussion

This practice-led research allows the understanding of luminance in terms of the *amount* and *intensity* of 'light' perceived in timbre, from the correlations between specific timbral conditions and the words that describe them. In the compositional practice, it is possible to recognise that *dynamic envelope* and *spectral content* exert a strong influence in the perception of luminance. The attempt to measure light in a two-phase scale is a productive strategy for the study and organisation of instrumental resources, as well as for attributing specific qualities to the timbral experiences intended in each composition. Working directly with performers is fundamental for the development of specific instrumental techniques that respond to the different levels of luminance perception in the composition. Nonetheless, this research is still searching for new structural approaches that reflect more accurately the complexity of the dynamic and multidimensional condition of timbre.

Acknowledgments

School of Music, Faculty of Arts Humanities and Cultures, University of Leeds (UK).
SYNTHETIS International Summer Course for Composers (Poland).

References

Amato, J. A. (2001). *Dust: a history of the small and the invisible*. Berkeley & Los Angeles: University of California Press.

Beckett, S. (1980). *Company*. London: John Calder.

Cleare, A. (2009-14). *eyam I (it takes an ocean not to)*.

Lochhead, J. (2016). *Reconceiving structure in contemporary music: New tools in music theory and analysis*. New York, USA: Routledge.

Rutherford-Johnson, T. (2011). Patterns of Shimmer: Liza Lim's Compositional Ethnography. *Tempo Vol. 65 (258)*, 2-9.

Saunders, R. & musikFabrik (2008). *Stirring Still* [CD]. Mainz, Germany: Wergo.

Wallmark, Z. (2019). A corpus analysis of timbre semantics in orchestration treatises. *Psychology of Music, 47(4)*, 585-605.

Zacharakis, A., Pastiadis K., & Reiss, J. (2014). An Interlanguage Study of Musical Timbre Semantic Dimensions and Their Acoustic Correlates. *Music Perception: An Interdisciplinary Journal, 31(4)*, 339-358.

Evidence for timbre space robustness to an uncontrolled online stimulus presentation

Asterios Zacharakis[1†], Ben Hayes[2], Charalampos Saitis[2] and Konstantinos Pastiadis[1]
[1] School of Music Studies, Aristotle University of Thessaloniki, Thessaloniki, Greece
[2] Centre for Digital Music, Queen Mary University of London, London, UK
[†] Corresponding author: aszachar@mus.auth.gr

Introduction

Research on timbre perception is typically conducted under controlled laboratory conditions where every effort is made to maintain stimulus presentation conditions fixed (McAdams, 2019). This conforms with the ANSI (1973) definition of timbre suggesting that in order to judge the timbre differences between a pair of sounds the rest perceptual attributes (i.e., pitch, duration and loudness) should remain unchanged. Therefore, especially in pairwise dissimilarity studies, particular care is taken to ensure that loudness is not used by participants as a criterion for judgements by equalising it across experimental stimuli. On the other hand, conducting online experiments is an increasingly favoured practice in the music perception and cognition field as targeting relevant communities can potentially provide a large number of suitable participants with relatively little time investment from the side of the experimenters (e.g., Woods et al., 2015). However, the strict requirements for stimuli preparation and presentation prevents timbre studies from conducting online experimentation. Despite the obvious difficulties in imposing equal loudness on online experiments, the different playback equipment chain (DACs, pre-amplifiers, headphones) will also almost inevitably 'colour' the sonic outcome in a different way. Despite the above limitations, in a social distancing time like this, it would be of major importance to be able to lift some of the physical requirements in order to carry on conducting behavioural research on timbre perception. Therefore, this study aims to investigate the extent to which an uncontrolled online replication of a past laboratory-conducted pairwise dissimilarity task will distort the findings.

Method

A pairwise dissimilarity study presented in Zacharakis et al. (2015) was replicated in an online experiment. Sixteen musically trained listeners with normal hearing took part in the experiment (12 male, 4 female, average age: 30.7, average years of musical practice: 16.4, std of years of musical practice: 8.1). Their task was to rate the pairwise differences among 24 musical tones (300 pairs overall) –consisting of acoustic, electric and synthetic instruments– using the free magnitude estimation method. That is, they rated the perceptual distances of 300 pairs (same pairs included) by freely typing in a number of their choice to represent dissimilarity of each pair (i.e., an unbounded scale) with 0 indicating a same pair. Prior to the main listening test, a headphone screening test similarly to Woods et al. (2017) along with a familiarisation phase and a short training phase took place to make sure participants used headphones and understood the required task adequately enough. The pairs of the main experiment were presented in random order and the presentation order within each pair was also randomised. In the beginning of the experiment the participants were asked to set a comfortable playback level and keep it constant throughout the process.

Results

The Cronbach's Alpha for this set of responses was .9, indicating a strong internal consistency of responses, albeit a little lower compared to the .94 and .96 identified for English and Greek speaking participants for the controlled experiment. Comparison of the average raw dissimilarities between the two experiments (English speaking participants) showed a very strong correlation (Pearson's $\boldsymbol{r = .9}$, $p<.001$). This fact already indicates that the uncontrolled online experiment did not result in a substantial alteration of the acquired ratings overall.

A subsequent non-metric Multidimensional Scaling Analysis with dimension weighting (INDSCAL within SPSS PROXSCAL algorithm) was applied to the obtained dissimilarities. An examination of the

measures-of-fit that is presented in Table 1–in comparison with the respective metrics for the controlled experiment– indicates that the data are optimally represented by a 3-dimensional model since the improvement of the measures diminishes when a fourth dimension is added.

Table 1: Measures-of-fit and their improvement for different MDS dimensionalities between the laboratory and online experiments. S-Stress is a measure of misfit. The lower the value (to a minimum of 0) the better the fit. D.A.F.: Dispersion Accounted For is a measure of fit. The higher the value (to a maximum of 1) the better the fit.

	Online				**Laboratory**			
Dimensionality	**S-Stress**	**Improv.**	**D.A.F.**	**Improv.**	**S-Stress**	**Improv.**	**D.A.F.**	**Improv.**
1D	*.32*	*-*	*.81*	*-*	*.36*	*-*	*.81*	*-*
2D	*.18*	*.14*	*.92*	*.11*	*.19*	*.17*	*.92*	*.11*
3D	*.13*	*.05*	*.95*	*.03*	*.13*	*.06*	*.95*	*.03*
4D	*.09*	*.04*	*.97*	*.02*	*.10*	*.03*	*.97*	*.02*

The comparison between the two 3-dimensional timbre spaces that resulted from the laboratory and the online replication of the experiment was based on one index for configurational similarity, namely the Tucker's congruence coefficient and a second one for assessing dimensional similarity (i.e., the direct relationships between the dimensions of the two timbre spaces), namely the modified RV coefficient. A more detailed explanation regarding the use of these metrics for timbre space comparison can be found at Zacharakis & Pastiadis (2016) and Zacharakis et al. (2017). In general, values of the Tucker's congruence coefficient greater than .92 are considered fair and values larger than .95 practically imply perfect equivalence between the compared configurations (Lorenzo-Seva & Ten Berge, 2006). The statistical significance of the congruence coefficient between the two configurations was tested using a bootstrap analysis method (Monte Carlo estimate of its expected value under chance conditions). The modified RV coefficient for overall dimensional similarity between matrices varies between 0 and 1 and should be interpreted in a similar manner to the correlation coefficient between two unidimensional variables (Abdi, 2007).

*Table 2: Metrics for configurational and dimensional similarity between the two spatial configurations. The expected value and the standard deviation of the congruence coefficient were estimated through bootstrapping with 10000 runs (**: significance at the .01 level).*

	Congruence coefficient (expected value, standard deviation)	**RV-mod**
Laboratory timbre space vs. online timbre space	.97 (.90, .008)	.71**

A comparative assessment of the two metrics that are presented in Table 2 revealed a strong similarity both at configurational and at dimensional level. This can be also confirmed by inspection of the timbre spaces themselves shown in Figure 1. Although some specific differences may exist between the two configurations, the main characteristics such as the distinct major clusters between the impulsive and the continuous sounds, notable outliers (e.g., the double bass, the bowedpad, the marimba or the saxophone) and certain smaller clusters (e.g., Acid-Moog-Saxophone, Violin-Cello) are preserved.

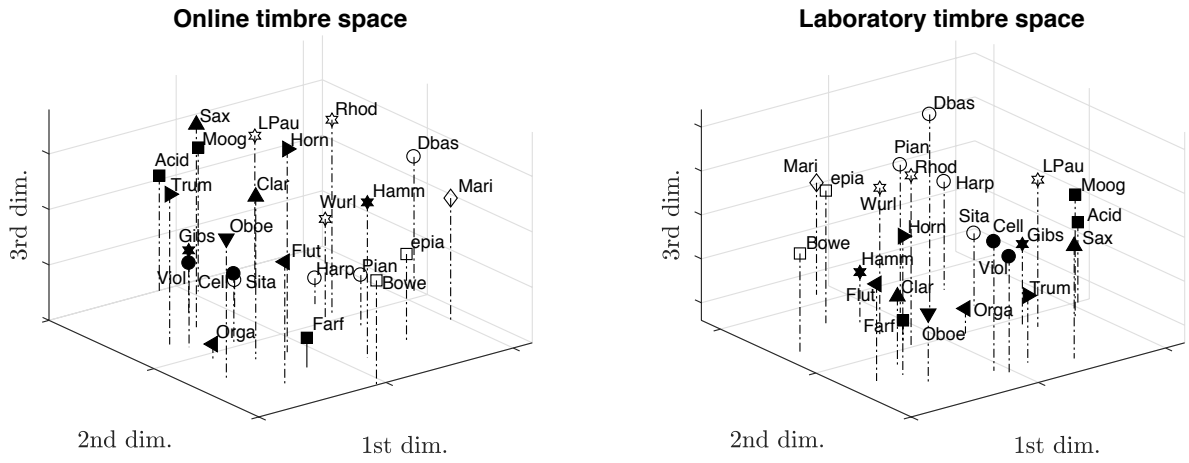


Figure 1: The two 3-dimensional timbre spaces for each experiment. The uncontrolled timbre space that came from the online experiment is shown on the left and the controlled that came from the laboratory experiment on the right. Black symbols: continuant, white symbols: impulsive, ▲: Single reed, ▼: Double reed, ◄: Aerophone, ►: Lip reed, ●: Chordophone, ♦: Idiophone, ⋆: Electrophone , ■: Synthesizer. Abbreviations of instrument names, Acid: Acid, Bowe: Bowedpad, Clar: clarinet, DBas: double bass pizzicato, epia: electric piano (rhodes), Farf: Farfisa, Flut: flute, Gibs: Gibson guitar, Hamm: Hammond, Horn: French horn, Harp: Harpsichord, LPau: Les Paul Gibson guitar, Mari: marimba, Moog: Moog, Oboe: oboe, Orga: pipe organ, Pian: piano, Rhod: Rhodes piano, Sax: saxophone, Sita: sitar, Trum: trumpet, Cell: cello, Viol: violin, Wurl: Wurlitzer.

Discussion

This study investigated whether or not controlling for playback level and frequency colouration due to differences in playback equipment significantly affects the perceptual timbre space of a certain set of stimuli. It was motivated by the difficulty in having physical access to participants in order to conduct controlled laboratory experiments on timbre perception due to the outbreak of the COVID-19 global pandemic. Our analysis showed that the dissimilarity ratings obtained through the online test featured high internal consistency and, most importantly, were robust to the uncontrolled experimental conditions, at least for familiar instrumental tones and musically trained participants with no hearing impairments. This naturally led to a timbre space that exhibited strong similarity with the one resulted from the laboratory conditions. To put this in perspective, the configurational similarity quantified by the Tucker's congruence coefficient between the online and laboratory spaces was .97 when the respective value between the same laboratory condition and one other identical laboratory condition featuring participants that simply spoke a different native language was found to be .98 (Zacharakis et al. 2015).

Since this experiment concerns a more or less familiar set of sounds, it could be argued that the higher level cognitive mechanism for timbre recognition that has been shown to play a role in dissimilarity judgments (Siedenburg et al., 2016) and which seems to be also evident in the current timbre spaces, could account for the robustness of comparative timbre judgements to presentation conditions. Thus, if the categorical information conveyed by familiar stimuli was able to counterbalance the lack of controlled presentation conditions the next step would be to examine whether diminishing such information through the modification of known stimuli or the presentation of synthetic unfamiliar stimuli altogether would result in timbre space variability for an uncontrolled experimental condition. Indeed, there is already some evidence that presentation conditions (i.e., the existence of background noise) rearranges the timbre space of unfamiliar synthetic tones. In addition, there is also some evidence that the degree of exploitation of categorical information may be mediated by musical expertise (Siedenburg & McAdams, 2017). Therefore, despite this first positive step, future work needs to examine variables such as the balance

between acoustic and categorical information present in the stimulus set as well as the level of musical expertise in order to come up with a comprehensive online testing protocol for the typically highly controlled task of pairwise dissimilarity rating. Although in this case the mean raw dissimilarities were already strongly correlated, an additional point of interest would be to test the suggested remedial properties of the non-metric MDS algorithm on noisy data (Shepard, 1966; Young, 1970).

Acknowledgments

This study was supported by a post-doctoral scholarship issued by the Greek State Scholarships Foundation (grant title: "Subsidy for post-doctoral researchers", contract number: 2016-050-050-3-8116), which was co-funded by the European Social Fund and the Greek State. The authors wish to thank the participants of the online experiment.

References

Abdi, H. (2007). The RV coefficient and congruence coefficient. In N. Salkind (ed.), *Encyclopedia of measurement and statistics*, (pp. *849*-853). Thousand Oaks: CA, Sage.

ANSI. (1973). *Psychoacoustical Terminology*, S3.20-1973, New York: American National Standards Institute.

Lorenzo-Seva, U., & Ten Berge, J. M. (2006). Tucker's congruence coefficient as a meaningful index of factor similarity. *Methodology*, *2*(2), 57-64.

McAdams S. (2019). The Perceptual Representation of Timbre. In: Siedenburg K., Saitis C., McAdams S., Popper A., Fay R. (eds), *Timbre: Acoustics, Perception, and Cognition* (pp. 23-57). Springer Handbook of Auditory Research, vol 69. Springer, Cham.

Shepard, R. N. (1966). Metric structures in ordinal data. *Journal of Mathematical Psychology*, *3*(2), 287-315.

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

Siedenburg, K., & McAdams, S. (2017). The role of long-term familiarity and attentional maintenance in short-term memory for timbre. *Memory*, *25*(4), 550-564.

Woods, K. J., Siegel, M. H., Traer, J., & McDermott, J. H. (2017). Headphone screening to facilitate web-based auditory experiments. *Attention, Perception, & Psychophysics*, *79*(7), 2064-2072.

Woods, A. T., Velasco, C., Levitan, C. A., Wan, X., & Spence, C. (2015). Conducting perception research over the internet: a tutorial review. *PeerJ*, *3*, e1058.

Young, F. W. (1970). Nonmetric multidimensional scaling: Recovery of metric information. *Psychometrika*, *35*(4), 455-473.

Zacharakis, A., Pastiadis, K., & Reiss, J. D. (2015). An interlanguage unification of musical timbre: Bridging semantic, perceptual, and acoustic dimensions. *Music Perception*, *32*(4), 394-412.

Zacharakis, A., & Pastiadis, K. (2016). Revisiting the luminance-texture-mass model for musical timbre semantics: A confirmatory approach and perspectives of extension. *Journal of the Audio Engineering Society*, *64*(9), 636-645.

Zacharakis, A., Terrell, M. J., Simpson, A. J., Pastiadis, K., & Reiss, J. D. (2017). Rearrangement of Timbre Space Due To Background Noise: Behavioural Evidence and Acoustic Correlates. *Acta Acustica united with Acustica*, *103*(2), 288-298.

Questioning the Fundamental Problem-Definition of Mridangam Transcription

Kaustuv Kanti Ganguli[1†], Akshay Anantapadmanabhan[2] and Carlos Guedes[1]
[1] Music and Sound Cultures research group, New York University Abu Dhabi, UAE
[2] Freelance Musician, India
[†] Corresponding author: kaustuvkanti@nyu.edu

Introduction

There have been several attempts to analyze and characterize percussion instruments using computational methods, both in the context of Western (Sandvold et al., 2004; Tindale et al., 2004) and non-western percussion, specifically on automatic transcription of tabla (Chordia, 2005; Gillet & Richard, 2003) and mridangam strokes (Anantapadmanabhan et al., 2013; 2014). Although Anantapadmanabhan et al. (2013) provide greater insight into the mridangam strokes and their relation to the modes of the drumhead, the transcription approach is limited by its dependency on prior knowledge about the specific modes of the instrument. This puts a constraint for the method to be generalized to other instruments or different tonics (Anantapadmanabhan et al., 2014). Another concern is the unavailability of a unique mapping between the acoustic properties of a segmented stroke and its nomenclature in the vocabulary. It is often observed that the same stroke is uttered differently in the konakkol vocalization (the art of performing percussion syllables vocally in Indian art music), based on contextual variations or grammatical impositions. Most notably, even an expert musician is often unable to resolve such ambiguities on isolated presentation of a stroke. In this paper, we attempt to address this problem by proposing a combination of acoustic and semantic approaches for the contextual transcription of mridangam strokes.

We address the problem in an analysis-by-synthesis framework. First, a corpus of mridangam compositions is constructed and annotated. The annotations include both syntactic (i.e. an expert musician adhering to the lexicon without a reference to the acoustic properties of the audio) and listening-based (i.e. perceptual classification by a musician having no exposure to the mridangam repertoire). This facilitates modeling the task from both top-down and bottom-up approaches. In this work, we address the problem of mridangam stroke transcription at the intersection of these two approaches. The rest of the paper is structured in terms of description of the methodology, experimental results, and finally, discussion of the obtained insights.

Method

A corpus of mridangam compositions (both pre-composed traditional ones and spontaneous free-flowing) as well as konakkol was recorded by virtuoso percussionist Akshay Anantapadmanabhan (also a co-author). We employ traditional spectral methods (Bello et al. 2005; Peeters, 2010; Tindale et al., 2004) for transcription of the mridangam strokes detected by a spectral-flux based onset detection algorithm. To avoid the lexicon-based ambiguity, we initially performed a mid-level (manual) annotation based on 5 frequency bands: low, mid1, mid2, mid3, and hi. This was a reduction to five essential stroke types that can still represent the sequences according to Anantapadmanabhan. Figure 1 shows the ground truth onsets for the case-study excerpt. Comparing with the transcription in the form of konakkol, we see that there is a reduction from the set of unique konakkol syllables to the mridangam strokes. This leads to investigating the alias components in the transcription. This phenomenon occurs for two reasons: (i) to ensure fluency of reciting the konakkol at higher rhythmic densities, this is linked with physiological constraints of speech production; and (ii) to ensure the defined solkattu (rhythmic solfege for different subdivisions of the beat) to be recited in its integral form, this is linked to the language model.

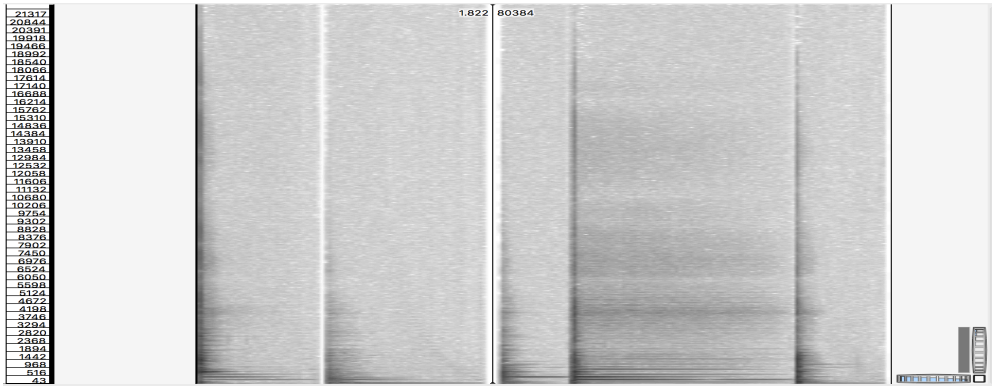


Figure 1: Spectrogram of typical strokes from the frequency-based classes [hi, low, mid1, mid2, mid3*] (left to right sequence) where the energy density is observed to vary along the frequency axis*.

Table 1 presents four typical phrases where the acoustically equivalent strokes are placed at different placeholders based on the grammar — mandated by ease of production for the konakkol. According to the mapping of konakkol-to-mridangam sounds, essentially Dhi ≡ Ka ≡ Ki — they are the same strokes in principle and are extensively used in the pedagogy to alert the pupils to realize the disambiguation process.

Table 1: Example phrases showing the acoustically equivalent strokes. The last two columns indicate the proportion of unique strokes to total stroke count without and with the acoustically equivalent strokes.

Index	***Phrases***	***Without equivalent***	***With equivalent***
1	*Ta Tha Cha Tha Ki Ta Tha Ka*	*5/8*	*4/8*
2	*Ki Ta Ki Ta Dhi Dhom Dhom Ka*	*5/8*	*3/8*
3	*Ki Ta . Ki Num . Ki Ta Ki Ta Dhom Dhom Ka*	*5/11*	*4/11*
4	*Tha . Dhi . Dhi . Thom Thom Ka . Dhi . Dhi Dhom Dhom Ka*	*5/11*	*4/11*

Thus we see that there is a further reduction in the set of (acoustically) unique[1] strokes after the introduction of acoustically equivalent strokes. This indicates that it is impossible to replicate the top-down (syntactic) transcription without the knowledge of this language model (mapping of the equivalents) into the algorithm.

Results

Given the 5 frequency-band based classes defined on the set of strokes are available as ground truth, the first approach we try is to use unsupervised clustering with the given number of clusters using a combination of frequency- and energy-based features. Figure 2 (left) shows the scatter plot of the detected onsets for a case-study excerpt with [Zero Crossing Rate (ZCR), Spectral Centroid] as the feature vector. The K-Means clustering with K=5 yields 5 clusters, denoted by different colors. All features are scaled between [-1,1]. For the second approach, we employ a simplistic supervised machine learning model K-Nearest Neighbor (KNN) classifier. We train the KNN with 30 strokes for each of the 5 classes (to ensure data balance) with a frequency-based (ZCR) and an energy-based (Spectral Centroid) feature. Figure 2 (right) shows the scatter plot of the ground truth onsets for the case-study excerpt. The predicted class labels are denoted by 5 different colors. Even though the homogeneity of the obtained clusters (left) are supported by a high cluster purity (0.97), there is a considerable overlap among the classified instances (right) that attracts further investigation on the mapping between the ground truth labels and the corresponding acoustic features.

[1] The definition of unique does not means identical in the signal domain, but similar within a small threshold. The standardization of similarity computation is an independent topic beyond the scope of this work. Without loss of generality, we can assume that the manner / place of articulation and hand gesture is identical.

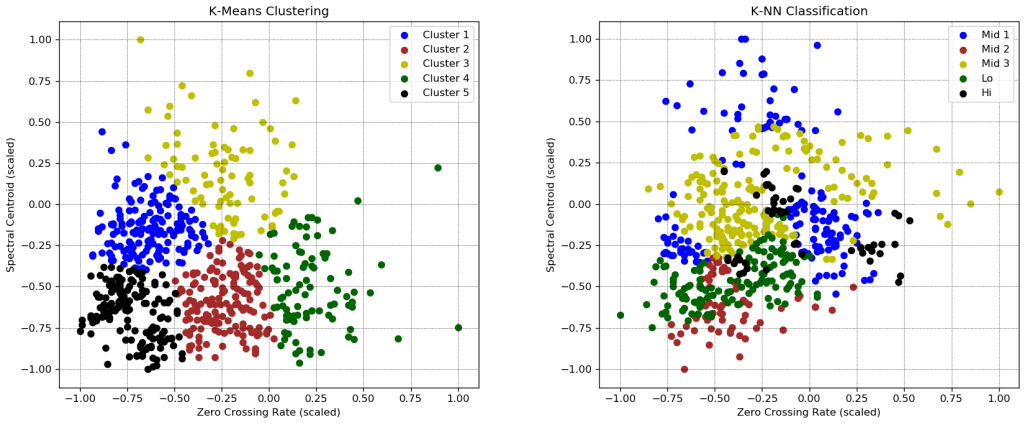


Figure 2: Scatter plots for K-Means clustering (left) and KNN classification (right) on the detected and ground truth onsets respectively, for a case-study excerpt.

The bottom-up computations show that a combination of machine learning approaches, i.e. unsupervised clustering (e.g. K-Means) and supervised classification (e.g. KNN), can lead to abstraction of frequency- and energy-based discrimination and labeling of mridangam strokes. However, the actual transcription labels (musicological) do not bear an intuitive mapping with the acoustic properties of the segmented mridangam stroke. This makes a two-pass system — where the knowledge constraints are to be incorporated in the post-processing stage — particularly relevant.

Discussion

We propose a 2-stage process for the assignment of the grammatically-accurate label for the transcribed mridangam stroke. In the first stage, we employed timbre-modeling techniques to assign a generic label for the stroke instance. In the second stage, the acoustically equivalent strokes are renamed according to the n-gram models learnt from the grammatical-annotation of the corpus as knowledge-constraints (see Ganguli & Guedes (2019)). This enables the mapping between konakkol-to-mridangam sounds adhering to the grammar prescriptions. Hence we contest the fundamental definition of the stroke-transcription problem. Even though the existing methods can reliably transcribe isolated strokes, the contextual transcription is still ill-defined and becomes particularly challenging. Moreover, there is no clear theory whether in the top-down transcription, konakkol syllables bear any correspondence with the frequency- or energy-characteristics of the corresponding strokes. Ganguli & Guedes (2019) showed that in the case of higher rhythmic density (tested on a phrase *Tha Ri Ki Ta Thom*, originally recorded at 90 bpm*)*, a time-compressed version of the reference phrase played in 4x speed is perceivably different from the same reference phrase played at 4x speed. This indicates that there is a gestural difference in articulating the same phrase at different speeds. Co-articulation effect mandates the performer to perceive the whole phrase as a gestalt (as opposed to a sequence of individual strokes) to modify the hand gestures which changes the acoustic properties of the realized stroke whereas the transcription label remains unaffected.

As future work, we propose word-vector (Mass et al., 2011) and graph-community (Blondel et al., 2008) based approaches to learn the context-dependence (e.g. language model) of the konakkol i.e. the symbolic subsequences. This will facilitate disambiguating the acoustically equivalent strokes from the detected (acoustically accurate) stokes from the first stage. Additionally, the network visualizations can in turn be fed back into the pedagogy as an interactive interface for engagement-learning.

Acknowledgments

This research is part of the project “Computationally engaged approaches to rhythm and musical heritage: Generation, analysis, and performance practice," funded through a grant from the Research Enhancement Fund at the New York University Abu Dhabi.

References

Anantapadmanabhan, A., Bello, J., Krishnan, R., & Murthy, H. (2014). Tonic-independent stroke transcription of the mridangam. In *Proceedings of Audio Engineering Society Conference: 53rd International Conference: Semantic Audio. Audio Engineering Society*. [Online]. Available: http://www.aes.org/e- lib/browse.cfm?elib=17102

Bello, J. P., Daudet, L., Abdallah, S., Duxbury, C., Davies, M., & Sandler, M. B. (2005). A tutorial on onset detection in music signals. In *IEEE Transactions on Speech and Audio Processing, 13*(5), 1035-1047.

Blondel, V. D., Guillaume, J. L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. In *Journal of Statistical Mechanics: Theory and Experiment*, *10*, P10008.

Chordia, P. (2005). Segmentation and Recognition of Tabla Strokes. *In Proceedings of the International Society for Music Informational Retrieval (ISMIR) conference*, (pp. 107-114), London, UK.

Ganguli, K., & Guedes, C. (2019). An approach to adding knowledge constraints to a data-driven generative model for Carnatic rhythm sequence. In *Trends in Electrical Engineering, 9*(3), 11-17.

Gillet, O., & Richard, G. (2003). Automatic labelling of tabla signals, In *Proceedings of the International Society for Music Informational Retrieval (ISMIR) conference*, Maryland.

Guedes, C., Trochidis, K., & Anantapadmanabhan, A. (2017). CAMeL: Carnatic percussion music generation using N-gram and clustering approaches. In *Proceedings of the 16th Rhythm Production and Perception Workshop,* Birmingham, UK.

Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., & Potts, C. (2011). Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the Association for Computational Linguistics: Human Language Technologies, volume 1*, (pp. 142-150). Association for Computational Linguistics.

Peeters, G. (2010). Spectral and temporal periodicity representations of rhythm for the automatic classification of music audio signal. *IEEE Transactions on Audio, Speech, and Language Processing, 19*(5), 1242-1252.

Sandvold, V., Gouyon, F., & Herrera, P. (2004). Percussion classification in polyphonic audio recordings using localized sound models. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR) conference,* (pp. 537-540), Barcelona, Spain.

Tindale, A. R., Kapur, A., Tzanetakis, G., & Fujinaga, I. (2004). Retrieval of percussion gestures using timbre classification techniques. In *Proceedings of the International Society for Music Informational Retrieval (ISMIR) conference,* Barcelona, Spain.

The medium is the message: Questioning the necessity of a syntax for timbre

Moe Touizrar[1 †], Kai Siedenburg[2]
[1] Schulich School of Music, McGill University, Montreal, Canada
[2] Department of Medical Physics and Acoustics and Cluster of Excellence *Hearing4all*, University of Oldenburg, Oldenburg, Germany
[†] Corresponding author: moe.touizrar@mail.mcgill.ca

Introduction

The notion of a timbral syntax has been of persistent appeal to composers, music theorists, and music psychologists alike. Following a discussion by Arnold Schoenberg (1911) as part of his theory of harmony, several conceptions of timbral syntax have been outlined (Lerdahl, 1987; McAdams, 1989; Bigand et al., 1998; Murail, 2005; Nattiez, 2007). The discourse surrounding these proposals is speculative: authors suggest ways in which timbre could function syntactically in a yet-to-be materialized musical system but give little consideration to how timbre actually functions in music. Of course, music neither possess nor displays a true syntax in the normative linguistic sense, nevertheless, beginning with the writings of Hugo Riemann in the late 19th century, music theorists have proposed a syntax-like organization for music. According to Swain (1995) syntaxes, both musical and linguistic, have two basic functions: first, to control information load, and secondly to mediate expressed relationships. More specifically, Meyer (1989) proposes that to qualify as syntax "successive stimuli must be related to one another in such a way that specific criteria for mobility and closure are established" (p. 14). By mediating tension with closure, music is thus segmented into loosely recursive units, which in turn make possible more complex hierarchical structures and organization, such as musical form. Importantly, Meyer denied timbre the ability to contribute meaningfully to musical syntax (and therefore to formal organization) due to its secondary nature with respect to more syntactically salient parameters such as pitch, harmony, and rhythm. In our presentation, we revisit this theme and argue that syntax is a sufficient, but not a necessary condition for timbre to contribute meaningfully to an unfolding musical form. Moreover, we propose that contrary to any syntactical underpinning, the various levels of timbral organization that span a musical work can be apprehended by a process of mnemonic agglutination. In doing so, we bridge recent empirical research on memory for timbre, together with score analysis and a consideration for the listening act itself—for what it means to attend to and remember music (and thus timbral characteristics) over time.

Method

Two questions frame the presentation: (1) How does memory for timbre play out in ecologically valid contexts? (2) If an underlying syntax for timbre is indeed not necessary to the organization and apprehension of a piece of music, how might composers structure the deployment of timbre such that it serves a form-bearing function (McAdams, 1989)? First, we review contemporary theories of auditory memory (Siedenburg & Müllensiefen, 2019) that demonstrate how fine-grained timbral properties in music may be extracted, memorized, and stored over long time-spans, hence providing the basis for timbre to act as a bearer of musical form. We propose that timbre does not need syntax in order to carry forward musical information over time, but that its aesthetic qualities—themselves traditionally only considered as a medium to other musical information—can constitute the musical message. Next, drawing on theories of the experience of musical form (Cook, 1990; Levinson, 1997; London, Cox, Morrison, Maus & Repp, 1999; Zbikowski, 1999; Huovinen, 2013), we argue that one prominent way by which timbre bears form is via the recognition and agglutination of discrete timbral-mnemonic units into contours across large spans of time. We call this process *apperception* (Touizrar, 2019). Finally, drawing on analyses of orchestral music of the 19th and 20th centuries, we demonstrate how composers construct and modify large-scale apperceptive contours, thus forging orchestrational form using timbre as a primary mnemonic construct.

Results

The result of our theoretical proposition, when applied to score analysis, demonstrates that composers exhibit in their works an explicit concern for the mnemonic structuring of timbre across large spans of musical time. Composers such as Nikolai Rimsky-Korsakov and Arnold Schoenberg construct what we term *iconic motives* whose primary constituents include a prominent timbral profile in addition to other parameters such as pitch and rhythm. Developments to the various parameters of these recurring motives—and especially to timbre—across the work allow listeners to apperceive large-scale contour spanning the entire work whose principal grouping involves the agglutination of several instrumentally-varied iterations over time. We demonstrate how these large-scale grouping structures depend on the sonic impression of a memorable psychological present, followed by dependency on episodic and working memory that all together contribute to both implicit and explicit retention and recognition of form-generating timbral information.

Discussion

Our presentation addresses central questions faced by research on musical timbre: how, and to what degree might timbre participate in the construction of musical form? Moreover, if timbre does indeed function in a form-bearing capacity, how can we identify and demonstrate its explicit and implicit contributions to a musical form both in terms of its construction by composers as well as its reception by listeners? In tonal music, syntax serves as a building block of large-scale formal organization via the recursive harmonic progressions and their inevitable cadential closures that make up complete statements such as theme, periods, and sentences. These medium-scale units are grouped together into larger functional sections that progress through harmonic key areas, which are themselves structured according to the same tension-release principles that govern small-scale harmonic progressions. If timbre does indeed play a syntactic role in the unfolding of musical form, we should expect to find a palpable and recursively structured scheme for tension and release that is independent yet co-occurrent with harmonic, melodic, and rhythmic elaborations. Without denying the possibility of such structuring for timbre in other musical traditions, the battle to force timbre into a syntax seems to have been lost. Since no compelling evidence yet exists that demonstrates a coherent timbral syntax, let alone one that is easily apprehended, we propose that a change in approach to timbre as a form-bearing element is required. Instead, the relationship between syntax and form should be newly examined from an experiential vantage point; one that takes into account the imaginative experience of form-building undertaken by listeners (a group that includes composers, performers, and theorists). Memory is a key component in the experience of musical form, and understood experientially, timbre's contribution to the temporal elaboration of musical form is at the very least palpable. We propose that memory for timbre, and more specifically diachronic mnemonic grouping of timbre can play an active role in the unfolding of musical form. By merging existing evidence and theories together with score and phenomenal analysis, we wish to make a case for timbre's form-bearing capacity that is to be examined in an interdisciplinary and ecologically valid context.

Acknowledgments

KS is supported by a Freigeist Fellowship of the Volkswagen Stiftung.

References

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

Cook, N. (1990). *Music, imagination, and culture*. New York: Oxford University Press.

Huovinen, E. (2013). Concatenationism and anti-architectonicism in musical understanding. *The Journal of Aesthetics and Art Criticism*, 71(3), 247–260.

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

Levinson, J. (1997). *Music in the moment*. Ithaca NY: Cornell University Press.
London, J., Cox, A., Morrison, C., Maus, F., & Repp, B. (1999). "Music in the moment": A discussion. *Music Perception*, *16*(4), 463–494.
McAdams, S. (1989). Psychological constraints on form-bearing dimensions in music. *Contemporary Music Review*, *4*(1), 181–198.
McAdams, S. (2019). Timbre as a structuring force in music. In K. Siedenburg, C. Saitis, S. McAdams, A. Popper and R. Fay (eds), *Timbre: Acoustics, Perception and Cognition* (pp. 211–244). Springer Handbook of Auditory Research, vol 69. Springer, Cham.
Meyer, L. B. (1989). *Style and music: Theory, history, and ideology*. Chicago: The University of Chicago Press.
Murail, T. (2005). The revolution of complex sounds. *Contemporary Music Review*, *24*(2/3), 121–135.
Nattiez, J-J. (2007). Le timbre est-il un paramètre secondaire? [Is timbre a secondary parameter?]. *Les Cahiers de la Société Québécoise de Recherche en Musique*, *9*(1–2), 13–24.
Schoenberg, A. (1911/1983). *Theory of harmony*. Translated by Roy E. Carter. Berkeley: University of California Press.
Siedenburg, K., & Müllensiefen, D. (2019). Memory for timbre. In K. Siedenburg, C. Saitis, S. McAdams, A. Popper and R. Fay (eds), *Timbre: Acoustics, Perception and Cognition* (pp. 59–86). Springer Handbook of Auditory Research, vol 69. Springer, Cham.
Swain, J.P. (1995). The concept of musical syntax. *The Musical Quarterly*, *79*(2), 281–308.
Touizrar, Moe. (2019). *From ekphrasis to apperception: The sunrise topic in orchestral music*. (PhD Dissertation). McGill University, Montreal.
Zbikowski, L. (1999). Musical coherence, motive, and categorization. *Music Perception*, *17*(1), 5–42.

Orchestration and Drama in J.-P. Rameau's *Les Boréades*

Didier Guigue[1†] and Charles de Paiva Santana[2]
[1] Dept. de Música, Universidade Federal da Paraíba, Brazil
[2] Interdisciplinary Center for Sound Communication, University of Campinas, Brazil
[†] Corresponding author: didierguigue@gmail.com

Introduction

In this proposal, we discuss how the French composer Jean-Philippe Rameau (1683-1764) uses the tonal palette and sonic qualities of the baroque orchestra as a powerful device towards the construction of meaning and expression of a dramatic subject. Despite the consensus on the great originality of the composer regarding the orchestration of his operatic works, systematic examination of its structural function is yet to be done, as the majority of systematic musicology studies about Rameau are concerned with his harmonic language.

To illustrate the composer's procedures, we will take a closer look at the instrumental introduction of the *ariette Un horizon serein*, extracted from the last of his five *tragédies lyriques*, *Les Boréades* (1763). In this opera, the librettist, believed to be the composer's faithful collaborator and encyclopedist Louis de Cahusac (1706 – 1759), develops naturalistic metaphors to convey the conflicting and unsettling feelings faced by the young Queen Alphise. For instance, the *ariette* portrays a quiet seascape under a serene blue sky being thereafter disrupted by a ravaging tempest. That setting symbolizes the tragic future that threatens the enamored queen, her people, and her country, against her dream of peaceful love and reign.

From the poem, Rameau selects three motives based on nature: the 'calm' or serene horizon; the 'rumbling' wind; and the wind-raised sea waves. Each such image is given a particular musical topos, that is, a musical figure which is transformed through a series of variations following the dramatic developments. We will describe how we categorize each topos family. What is striking here is that the composer avoids using the same orchestral configurations for the different topoi instances. That means that the variations are also, and sometimes exclusively, timbre-based variations.

In this presentation, we will focus on the introduction's orchestration and textural design, that is, the way Rameau blends, highlights, and coordinates the different instruments of the orchestra to display the psychological conflict set forth by the poem. For this task, we developed an experimental analytic model for the classification and estimation of complexity and classification of texture and orchestration, grounded on the mathematical *Theory of Partitions* (Andrews, 1998), Pauxy Gentil-Nunes (2008), and Wallace Berry (1987).

Method

The method puts forward a general numerical representation that allows the abstraction and subsequent computational manipulation of textural configurations. It also proposes a hierarchy of criteria of dispersion, or textural situations, that allow the stratification of the musical surface into different real components. It defines a measure that allows the quantification of heterogeneity relationships in textural configurations and a measure that estimates how diversely sound resources are used on the realization of textural configurations. Finally, it puts forward and a model for relative texture complexity based on how diverse is the orchestration and how intricate is the allocation of real components in a given musical work.

The numerical representation involves the use of nested lists of integer numbers. For example, the list [2, 1, 2] may refer to a musical segment where two voices are coordinated in a single layer or textural part, one textural layer consists of a single voice, and two other voices are coordinated in a last textural stratum.

From the numerical representation, the level of agglomeration can be calculated through the ratio between the sum of the real component's pairwise combinations and the total, and a dispersion index is obtained through the complement. Depending on the number of criteria and which ones are used, agglomeration or

dispersion can be weighted. The ensuing measure combined with the logarithmic coefficient of used sound resources per setup provides the model for textural complexity.

Results

After describing this model, we show its application to our *ariette* excerpt, which resulted in the discovery of a textural and timbre-based arch-form. The symmetrical structure is set up by going from a simple pairing of two instruments, to represent a single topos, to a complex stacked structure of six independent threads to express up to four topoi, and back again to a unique pairing. Note that this textural behavior does not converge with the periodic, harmonic infrastructure. It consists, instead, of an independent, additional level of the musical structure.

Discussion

Based on our analysis, we believe that Rameau is on the way towards the modern concept of orchestration, of music as being “organized sound”. We situate his orchestral techniques within a decidedly progressive perspective. Although his influence will not be immediately felt during the subsequent *Style Galant*, it will be essential to the advent of modernist thinking in music composition, leading to the concept of "pure music", abstracting its dependence to an external narrative. The writer of the *Treatise of Harmony* is also a forerunner of the perspective where sound itself, the timbre, has a prominent structural role, resurfaced during the romantic and the later modern period.

Acknowledgments

This research is supported by the São Paulo Research Foundation under grant 2018/04229-6 and by the Brazilian National Council for Scientific and Technological Development (CNPq).

References

Andrews, G. E. (1998). *The theory of partitions* (No. 2). Cambridge: Cambridge University Press.

Berry, W. (1987). *Structural functions in music*. New York: Dover Publication.

Gentil-Nunes, P. (2009). *Análise particional: uma mediação entre composição musical e a teoria das partições*. (PhD thesis). Rio de Janeiro, Universidade Federal do Estado do Rio de Janeiro.

Kintzler, C. (1983). Jean-Philippe Rameau. *Splendeur et naufrage de l’esthétique du plaisir à l’âge classique*. Paris: Éditions Le Sycomore.

The Semantics of Orchestration: A Corpus Analysis

Jason Noble[1†], Kit Soden[1], and Zachary Wallmark[2]
[1] Schulich School of Music, McGill University, Montréal, Québec, Canada
[2] School of Music and Dance, University of Oregon, Eugene, Oregon, USA
[†] Corresponding author: jason.noble@mail.mcgill.ca

Introduction

The burgeoning field of timbre semantics has revealed many insights into cognitive-linguistic and cross-modal bases for ways we experience and describe timbre (e.g., Saitis, 2019; Wallmark, 2019; Zacharakis & Pastiadis, 2016), but has usually focused on the semantics of *individual* timbres, especially the timbres of musical instruments. There remain many questions about the semantics of timbres and textures arising from orchestral *combinations*: for example, what leads orchestration pedagogue Samuel Adler (2002) to describe some orchestral textures as "flickering" while others are "noble," "bombastic," or "muddy"? Semantics of combinations of timbres and textures are germane to ecologically valid experiences of music, especially ensemble music. They also complement the much-discussed gray area between timbre and harmony (e.g., Hasegawa 2009; Harvey 2000) with another equally fascinating liminality: between timbre and texture.

A rich source of information about the semantics of orchestral combinations comes from the many published orchestration treatises. Wallmark (2019) analyzed a corpus of eleven treatises and manuals, focusing on descriptions of individual instrumental timbres. He organized the resulting sample of descriptors into seven basic categories and analyzed the frequency with which particular terms and categories appeared in the corpus as indices of timbre conceptualization and cognition. Building upon this precedent, we analyze a corpus of six orchestration treatises published over the last century (Table 1) for semantic descriptions of orchestral combinations.

Table 1: Orchestration treatises used in our corpus study.

Author	Title	Publication year
Adler, Samuel	*The Study of Orchestration*	1982/2002
Blatter, Alfred	*Instrumentation and Orchestration*	1997
Jacob, Gordon	*Orchestral Technique: A Manual for Students*	1982
Read, Gardner	*Style and Orchestration*	1979
Piston, Walter	*Orchestration*	1955
Forsyth, Cecil	*Orchestration*	1935

Method

The authors and their research assistants thoroughly reviewed the treatises and catalogued all semantic descriptions of timbres and textures arising from combinations of instruments (i.e., two or more instruments sounding concurrently). Information captured for each entry included the descriptive terms used, the instruments involved, the numbers of instruments, and the number of instrument families. Where possible, each entry was catalogued according to the orchestral effects taxonomy developed by Goodchild, Soden, and McAdams (in prep), using categories such as blend, surface texture, stratification, and timbral contrasts. Information about register, dynamics, articulations and so forth was also captured, and will be analyzed in future stages of this research, along with the inclusion of descriptors mined from additional treatises.

Here we present the first stage of analysis, based primarily on conventional corpus linguistic measures and preliminary qualitative descriptions of the contents of our corpus. A more detailed statistical analysis will be the subject of a future paper.

Results

From these six books, we extracted a total of 1288 instances (tokens) of semantic descriptions of timbres and textures arising from instrumental combinations, of which 545 were unique (types), resulting in a Type/Token Ratio of .42 (CTTR; a simple index of lexical diversity on 0–1 scale). Table 2 shows the top 43 semantic types in descending order by frequency (*f*), each of which had at least six tokens in this corpus. An additional 502 types appeared with five tokens or fewer.

Table 2: Top 43 most frequently occurring semantic descriptors for instrumental combinations.

descriptor	*f*	descriptor	*f*	descriptor	*f*	descriptor	*f*
reinforcing	40	warm	13	fresh	9	exciting	7
soft	38	bright	11	full	9	expressive	7
rich	37	colorful	11	heavy	9	penetrating	7
powerful	25	emphatic	11	sharp	9	subdued	7
brilliant	23	homogeneous	11	sonorous	9	intense	6
light	18	interesting	11	strengthening	9	interlocking	6
dark	17	loud	11	blending	8	nasal	6
beautiful	16	smooth	10	dominant	8	pulsating	6
prominent	16	balanced	9	ethereal	8	thickening	6
contrasting	15	clear	9	supportive	8	transparent	6
strong	13	deep	9	background	7		

A thorough comparison with the findings of Wallmark (2019) is beyond the scope of this short paper, but even at a glance, it is clear that the treatise authors do not simply apply the same descriptive vocabulary to timbral combinations as to individual timbres. Comparing the top 43 types from each study, there are 18 terms in common (*bright, brilliant, clear, dark, deep, expressive, full, heavy, intense, nasal, penetrating, powerful, rich, smooth, soft, sonorous, strong, warm*). Of the 25 remaining terms in the top 40 from our corpus, 14 specifically invoke combination, comparison, or multiplicity (*background, balanced, blending, colorful, contrasting, dominant, homogeneous, interlocking, prominent, reinforcing, strengthening, supportive, thickening, transparent*).

About half of the semantic descriptions were applied to generic accounts of multiple instruments with no number specified (e.g., "violins," "strings"; see Table 3a). Those that did specify numbers of instruments followed a classic long-tail distribution, with descriptions of smaller combinations being generally higher in tokens and types. A similar pattern obtains for family representation (Table 3b), with 45% of descriptions referring to only one of the four families of standard orchestral instruments (strings, woodwinds, brass, percussion), and only 4% referring to all four. In both cases, lexical diversity increased as a function of the number of instruments/families involved in the orchestral combination.

Table 3a: Breakdown of corpus by # instruments.

# instruments	%	tokens	types	TTR
2	17.24	222	150	.67
3	10.02	129	88	.68
4	6.13	79	67	.85
5	5.59	72	54	.75
6	2.25	29	23	.79
7	1.40	18	16	.89
8	1.32	17	16	.94
9	2.02	26	25	.96
10	1.16	15	14	.93
11+	1.79	23	20	.87
not specified	51.09	658	352	.53

Table 3b: Breakdown of corpus by # families.

# families	%	tokens	types	TTR
1	44.71	494	271	.55
2	32.40	358	209	.58
3	18.91	209	130	.62
4	3.98	44	36	.82

Table 4 shows the types with five or more tokens for semantic descriptions involving 1–3 families (there were no types for 4 families with more than 3 tokens).

Table 4: Most frequently occurring descriptors per number of families

1 family		2 families		3 families	
descriptor	*f*	descriptor	*f*	descriptor	*f*
rich	17	reinforcing	14	reinforcing	9
soft	16	rich	14	powerful	7
reinforcing	13	soft	12	soft	6
powerful	12	light	8	beautiful	5
brilliant	8	brilliant	7	colorful	5
homogeneous	8	strengthening	6	dark	5
strong	8	warm	6	emphatic	5
beautiful	7	dark	5		
bright	7	dramatic	5		
expressive	7				
penetrating	7				
balanced	6				
sonorous	6				
contrasting	5				
dark	5				
dominant	5				
full	5				
interesting	5				
loud	5				
smooth	5				
warm	5				

Analyzing these descriptions in light of the orchestral effects taxonomy reveals interesting asymmetries of distribution. Although many perceptual effects can arise from combinations of orchestral instruments, a few seem to be disproportionately represented in this corpus (Table 5), which may indicate aesthetic priorities of the authors and/or the composers about whose works they were writing. *Blend*, defined as "the fusion of different sources of acoustic information into a more or less unified auditory event," appears to be of singular importance. Although blend is only the third-most represented category, *timbral augmentation* and *timbral emergence* are sub-categories of blend, and *timbral heterogeneity* is the negation of blend, so 649 tokens (50.4% of the corpus) are blend-related. Also emphasized are musical layers and musical lines. *Stratification*, "groupings of events into strata of different prominence (e.g., foreground, middleground, background)," accounts for 20.4% of the corpus, and stream, stream segregation, and stream integration, which deal with groupings of sequential events into unitary musical patterns, account for 6.3% of the corpus.

Table 5: Breakdown of corpus by orchestral effects taxonomy

taxonomic category	*f*	%	taxonomic category	*f*	%	taxonomic category	*f*	%
timbral augmentation	284	22.0	surface texture	24	1.9	general contrast	8	0.6
stratification	263	20.4	stream integration	22	1.7	gradual reduction	5	0.4
blend	248	19.3	stream	16	1.2	sudden addition	4	0.3
timbral heterogeneity	77	6.0	gradual addition	14	1.1	sectional contrast	3	0.2
stream segregation	43	3.3	sudden reduction	13	1.0	timbral shifts	3	0.2
timbral emergence	40	3.1	antiphonal contrast	9	0.7	timbral echo	1	0.1

Table 6 shows the types with six or more tokens for the most represented taxonomic categories. Note the high count for *reinforcing* and the moderately high count for *supportive*, which are functional paraphrases of "timbral augmentation."

Table 6: Most frequently occurring types per taxonomic category

timbral augmentation		stratification		blend	
reinforcing	29	light	13	rich	10
brilliant	8	soft	11	reinforcing	8
soft	8	prominent	8	powerful	7
emphatic	7	brilliant	6	soft	7
rich	7	beautiful	5	blending	6
supportive	6	dominant	5		

Discussion

Before we even begin to analyze the specific instrumentations denoted in these semantic descriptions and the acoustical attributes to which they may correspond, a number of interesting observations may be made about the semantic conventions of instrumental combinations. There is substantial overlap between the terminology used by authors of orchestration treatises to describe individual instruments and their terminology for combinations, but there also appear to be substantial differences. Many new terms appear, often reflecting an intuitive awareness of perceptual grouping principles (Goodchild & McAdams, 2018; McAdams et al., in prep). Smaller combinations, along with those that do not specify a number of instruments, tend to receive more discussion than larger combinations; this holds true for both instruments and families. Of all the different perceptual effects that may arise from combining orchestral timbres, these authors seem to focus on blends, layers, and lines.

This study begins to address questions of great importance to musical experience that have received relatively little attention in the scholarship. We anticipate that a more thorough statistical analysis will yield many greater insights, and we hope that this corpus analysis will lay a foundation for future empirical studies.

Acknowledgments

The authors thank Stephen McAdams, Bob Hasegawa, the ACTOR project, and research assistants Michelle Nadon-Belanger and Rachel Hottle.

References

Adler, S. (2002). *The study of orchestration,* Third ed. New York: W.W. Norton and Company.

Goodchild, M., & McAdams, S. (2018). Perceptual processes in orchestration. In A. Rehding & E. I. Dolan (eds.), *Oxford Handbook of Timbre.* New York: Oxford University Press.

Harvey, J. (2000). Spectralism. *Contemporary Music Review*, *19*(3), 11–14.

Hasegawa, R. (2009). Gérard Grisey and the 'nature' of harmony. *Music Analysis*, *28*(2-3), 349–71.

McAdams, S., Goodchild, M., & Soden, K. (manuscript in preparation). *A taxonomy of perceptual effects of orchestration related to auditory grouping principles.*

Saitis, C., & Weinzierl, S. (2019). The semantics of timbre. In K. Siedenburg, C. Saitis, S. McAdams, A. Popper, & R. Fay (eds.), *Timbre: Acoustics, Perception, and Cognition* (pp. 119–149). Springer Handbook of Auditory Research, vol 69. Springer, Cham.

Wallmark, Z. (2019). A corpus analysis of timbre semantics in orchestration treatises. *Psychology of Music*, *47*(4), 585–605.

Zacharakis, A., & Pastiadis, K. (2016). Revisiting the luminance-texture-mass model for musical timbre semantics: A confirmatory approach and perspectives of extension. *Journal of the Audio Engineering Society*, *64*(9), 636-645.

Towards a Theory and Analysis of Timbrebased on Auditory Scene Analysis Principles: A Case Study of Beethoven's Piano Sonata Op. 106, Third Movement

Nathalie Hérold
GRÉAM and ACCRA (UR 3402), University of Strasbourg, Strasbourg, France
nathalieherold@hotmail.com

Introduction

The field of music theory and analysis has been traditionally and primarily focused on pitch organization systems, in particular through the development of tonal and post-tonal theories, including also duration aspects of music as is the case in theories of musical forms. However, theory and analysis based on the timbral dimension of music have become a developing area since the end of the 1970s. In studies like Erickson (1975), Cogan (1984), or Guigue (2009), specific attention is given to sound aspects of music, timbre no longer being considered as a secondary parameter (Nattiez, 2002), but rather as a multidimensional meta-parameter. In these studies, new analytical methods were developed, including in particular new approaches to musical scores as well as the use of visual representations in the form of graphs or spectrograms. Perceptual aspects – yet fundamental – are indeed discussed, but generally with few references to the psychological area. Only more recently has some research attempted to draw concrete propositions for the integration of psychological models such as the auditory scene analysis (Bregman, 1990) in the field of music theory and analysis, considering in particular orchestral repertoire (Touizrar & McAdams, 2019; Lalitte, 2019).

It is the aim of the present study to gain insight into the possibilities of application of auditory scene analysis principles for the study of timbre in a theoretical and analytical perspective, and to discuss their relevancy in the case of the third movement of Beethoven's Piano Sonata Op. 106. Composed in 1817–19 and titled "*Hammerklavier*", this sonata is fully part of the last Beethoven style, known for its new conception of musical forms as well as its structural use of musical timbre (Boucourechlivev, 1991). Indeed, the third movement of Op. 106, organized as a wide sonata form in F sharp minor, is also characterized by its formal use of the *una corda*/*tre corde* sonority resulting in an overall *ABABA* multipartite timbre structure (Hérold, 2011, 414–17). This piece is therefore worth considering as a case study for a timbre analysis based on auditory scene analysis principles.

Method

This research draws on the score of the Beethoven Adagio, as well as its audio recording (by pianist Alfred Brendel) – as written and sonic representations of the piece, respectively. After an intensive and detailed listening phase, the score is annotated in a systematic manner by means of colored frames, following the taxonomy and annotation system developed in the context of the Orchestration Analysis & Research Database (OrchARD, http://orchard.actor-project.org), as part of the ACTOR (Analysis, Creation and Teaching of Orchestration) partnership. Based on auditory scene analysis principles, this annotation system refers to the identification of different types of timbre and orchestration effects, as a result of concurrent, sequential and segmental auditory groupings (Goodchild & McAdams, 2018). As this method was originally developed for the analysis of symphonic music, its application to piano music requires some adjustments as regards the annotation tool itself as well as the underlying taxonomy.

Results

Regarding the taxonomy and annotation system, the analysis of the Beethoven Adagio shows the possibility to describe a whole piano sonata movement in a systematic manner with analytical categories pertaining to auditory grouping principles. Originally designed for the analysis of orchestral music, their level of generality allows them to apply as well to other musical repertoires. Most of the categories are relevant in Beethoven's movement, which also accounts for the diversity of timbre and orchestration situations that can be encountered within this piece. It is also interesting to note that these timbre and

orchestration effects are not always independent from each other and that some of these can occur simultaneously, giving rise to superimposed analytical annotations on the score and, consequently, to different layers and levels of timbre and orchestration effects.

Furthermore, the analysis of the Beethoven Adagio sheds light on some aspects of the formal organization of the movement. With regards to the relation between the timbre and orchestration effects and the tonal-thematic dimension of the movement, it is worth noting that some effects have a greater weight in specific sections of the musical form. This is particularly obvious when considering the first thematic group in F sharp minor (mm. 1–26), which is mainly characterized by a perceptual fusion effect – including timbral augmentation but also timbral emergence effects, according to the taxonomy used – that predominates in duration in comparison with other secondary effects. In this passage of the movement, the fusion effect can thus be considered as a characteristic of a higher structural level – a kind of "dominant" timbre effect –, the other secondary effects having much more an ornamental function. The same kind of observation is relevant in the case of the transition section (mm. 27–44), based on stratification as a dominant effect, as well as the second thematic group in D major (mm. 45–68), provided with a coupling of stratification and timbral echoes effects.

Finally, the analysis of Beethoven's Op. 106, third movement, questions, beyond auditory grouping effects as determinant of some aspects of timbre – but the latter being non-reducible to the former –, the general timbre dimension of the piece, which also involves piano specific features related to idiomatic writing configurations. This is particularly evident when considering effects pertaining to pedaling – including sustaining as well as *una corda* pedal aspects –, as well as effects related to a pianistic use of registers and doublings. These atypical situations are worth discussing with regard to the timbre and orchestration effects taxonomy that underlies the present Beethoven analysis.

Discussion

This research leads to a better understanding of the Beethoven Adagio itself in terms of timbre organization, which is likely to converge with or diverge from the sonata form as a structural model. It brings some insight into the use of timbre and orchestration strategies in Beethoven's late style and their role in order to renew the classical forms. Beyond this particular Beethoven case study, this research also brings to light concrete tools for the analysis of timbre in piano music and emphasizes the idea of orchestral piano (Hering, 1974) from an analytical perspective. Finally, this study contributes to the development of music theoretical aspects related to timbre and its structural conception, in particular through the idea of timbral effects on different hierarchical structural levels.

Acknowledgments

This research is made possible by support from the GRÉAM research centre (*Groupe de Recherches Expérimentales sur l'Acte Musical*) and the ACCRA research unit (*Approches Contemporaines de la Création et de la Réflexion Artistiques*, UR 3402) from the University of Strasbourg, as well as the ACTOR partnership (Analysis, Creation and Teaching of Orchestration).

References

Boucourechliev, A. (1991). *Essai sur Beethoven*. Arles: Actes Sud.

Bregman, A. S. (1990). *Auditory Scene Analysis: The Perceptual Organization of Sound*. Cambridge (Mass.): The MIT Press.

Cogan, R. (1984). *New Images of Musical Sound*. Cambridge (Mass.) : Publication Contact International.

Erickson, R. (1975). *Sound Structure in Music*. Berkeley, Los Angeles, London: University of California Press.

Goodchild, M., & McAdams, S. (2018). Perceptual Processes in Orchestration. In E. Dolan & A. Rehding (eds), *The Oxford Handbook of Timbre*. New York: Oxford University Press. http://doi.org/10.1093/oxfordhb/9780190637224.013.10

Guigue, D. (2009). *Esthétique de la sonorité: l'héritage debussyste dans la musique pour piano du XXe*

siècle. Paris: L'Harmattan.

Hering, H. (1974). Orchestrale Klaviermusik. *Acta musicologica*, XLVI/1, 76–91.

Hérold, N. (2011). *Timbre et forme : la dimension timbrique de la forme dans la musique pour piano de la première moitié du dix-neuvième siècle* (PhD thesis), University of Strasbourg, Strasbourg. https://publication-theses.unistra.fr/public/theses_doctorat/2011/HEROLD_Nathalie_2011.pdf

Lalitte, P. (2019). Vers une analyse texturale de la performance fondée sur les principes de l'analyse de scène auditive. In: P. Lalitte (ed), *Musique et cognition: perspectives pour l'analyse et la performance musicales* (pp. 231–252). Éditions Universitaires de Dijon, Dijon, France.

Nattiez, J.-J. (2007). Le timbre est-il un paramètre secondaire ? *Les cahiers de la SQRM*, 9/1–2, 13–24.

Touizrar, M., & McAdams, S. (2019). Aspects perceptifs de l'orchestration dans *The Angel of Death* de Roger Reynolds: timbre et groupement auditif. In P. Lalitte (ed), *Musique et cognition: perspectives pour l'analyse et la performance musicales* (pp. 55–88). Éditions Universitaires de Dijon, Dijon, France.

Klangfarbenmelodie in 1911: Anton Webern's Opp. 9 and 10

Matthew Zeller
McGill University, Schulich School of Music
matthew.j.zeller@gmail.com

Introduction

In 1911, Arnold Schoenberg theorized *Klangfarbenmelodie* (timbre-melody) in his treatise *Harmonielehre* (Theory of Harmony). That same year, Anton Webern showed works to Schoenberg hoping his teacher and friend would find *Klangfarbenmelodie* in his new compositions. This paper examines Webern's works of 1911 in the context of Schoenberg's formulation of *Klangfarbenmelodie* and shows that Webern uses timbre as a primary compositional parameter in the manner Schoenberg theorized.

A critical reexamination of Schoenberg's music and theoretical writings reveals two definitions of *Klangfarbenmelodie*: (1) a timbre-melody, that is, the directed process of a timbral progression; and (2) a textural style of presentation akin to homophony or polyphony—a type of chromaphony (timbre music).[1] Timbral lines are similar to melodic lines; they are cohesive, autonomous constructive unities connected by their intrinsic values that move forward through the music. Once these timbral lines are arranged with a logic that satisfies, there is a new textural style of presenting music.

The intellectual history of *Klangfarbenmelodie* has been shaped (or misshaped) by critical reception that created two opposing definitions: (1) a quasi-static pitch with morphing timbres, and (2) the fragmentary, pointillistic distribution of linear pitch material among different timbres. Respectively, these notions are often characterized as the composition *of* timbres (*Komposition der Klangfarben*) and composition *with* timbres (*Komposition mit Klangfarben*), and associated with Schoenberg's "Farben" and Webern's works (Ligeti, 2007).[2] I refer to these ideas as static and dynamic *Klangfarbenmelodie*.

The static and dynamic notions of *Klangfarbenmelodie* may accurately reflect how some twentieth-century musical thinkers approached the concept, but they were not the views of Schoenberg or Webern (Zeller, 2020). Both notions are based on the critical response that characterized the two composers' works by their outward manifestations rather than the technique's foundations in musical logic. Static *Klangfarbenmelodie* stems from 1919 when Arnold Schering writes of Schoenberg's *Five Orchestral Pieces*, Op. 16, No. 3, "A certain chord remains immobile for a long time in *pp*, but receives an ever-changing color gradation from half-measure to half-measure" (Schering, 1919, p. 153). And the idea of dynamic *Klangfarbenmelodie* in Webern's music comes from Erwin Stein when he writes about the *Six Bagatelles for String Quartet*, Op. 9, in 1923: "…in the melodies, almost every tone is apportioned to a different instrument, almost every one in a different timbre (harmonics, pizzicato, col legno, etc.). … Schoenberg's idea of timbre-melodies may have been influential" (Stein, 1923, p. 15). Even in this early period of reception, however, some scholars saw through the differences in surface features to grasp the underlying structural aspects of Schoenberg's and Webern's *Klangfarbenmelodie*. Discussing the technique in 1919, Alfredo Casella notes that timbre acts in both the vertical and horizontal dimensions, something melody, harmony, and rhythm cannot do (Casella, 1924). Not only does Casella approach discerning Schoenberg's goal of combining the horizontal and vertical, he foresees a musical twentieth century guided by the beacon of timbre. And in 1924 Paul A. Pisk hints at understanding the concept as a textural principle when he writes, "The juxtaposition of different lines results in stratifications" (Pisk, 1924, p. 1023). Timbral lines creating textural stratification, if it becomes an organizational principle, is *Klangfarbenmelodie*. Unfortunately, historical precedent was set, and both the single-pitch and pointillistic conceptions have been largely attributed as the term's original meaning.

[1] For a detailed discussion of *Klangfarbenmelodie* see Zeller, 2020, pp. 71–245; for discussion of chromaphony see pp. 4–6.
[2] Also discussed in Iverson, 2009.

The static and dynamic types of timbre-melodies can be techniques of creating a timbral progression, but they are just two of many possibilities. And if present, they must be an organizational principle in the musical logic of the work to create *Klangfarbenmelodie* the stylistic principle. In Schoenberg's estimation, new forms are needed. He writes, "…progressions of tone-colors would certainly demand constructions different from those required by progressions of tones, or of harmonies.… Quite different forms had to be produced by homophony and the art of counterpoint" (Schoenberg, 1975, p. 485). In actuality, the old forms remained, but timbre provided new ways of making them comprehensible.

Method

This presentation uses planal analysis to elucidate musical connections and draw out textural streams and timbral lines. The planes of planal analysis are analytical planes, though they often align with musical foreground and background in textural analyses.[3] Building upon Kathryn Bailey's work with tone-rows (Bailey, 1991), this paper uses musical block topography (Zeller, 2020) defined by both score-based analysis and auditory "chunking" (Goodchild and McAdams, 2018). The musical units I call blocks are one or more textural streams chunked into musically coherent parcels. Musical blocks are organized into textures with varying block topographies: monophonic, homophonic, or polyphonic, and combinations thereof. Block topographies describe the relations of the blocks themselves, not necessarily the internal content that forms each block. Planal analysis is then employed to illustrate the textural relationships of the blocks in however many analytical planes are dictated by the music. Since one of the Second Viennese School's goals was to combine the principles of homophony and polyphony, block topography becomes a powerful tool for analyzing this repertory. Another style of planal analysis employed in this presentation is timbral analysis, where each timbral line is rendered in its own plane.

Results and Discussion

Webern's timbral language: In his *Sechs Bagatellen für Streichquartett*, Op. 9, Webern removes the Bagatelles from the timbral identity of a string quartet, creating a new sound for "the new music" (Webern, 1975). Throughout all six of the Bagatelles there is a pervasive de-emphasis of pitch through playing technique. Above all, Webern's extensive use of *am Steg* (at the bridge or *sul ponticello*) is a concrete, physical move away from pitch primacy. Bowing at the bridge actually reduces the sound level of the fundamental frequency in comparison to its overtones.[4] Artificial harmonics also reduce the fundamental in favor of the much more prominent overtone of the fingered node. Webern regularly employs artificial harmonics with resultant tones two octaves above the stopped fundamental. They are still tones of definite pitch, but compositional weight is placed on the timbre over the pitch. If the pitch was all that was important there would be no need for the harmonics; the instruments could play the same absolute pitches in pure tones if Webern would have wanted that. Tremolo de-emphasizes pitch by creating a constant state of acoustic attack, eliminating the more stable sustain portion of the tone's ADSR spectrum. In combination with *am Steg*, tremolo at the bridge heavily masks the fundamental pitches sounded by the technique. Furthermore, Webern uses mutes extensively. Other playing techniques also work to elevate timbre over pitch in a more understated way. Playing at the fingerboard (*am Griffbrett*) changes the spectral characteristics of the tone. It does not reduce the fundamental in the same way playing at the bridge does, and correspondingly, is not employed with the same regularity. When used, it moves the sound away from prototypical unmodified arco tones, yet it allows a certain degree of continuity with the "normal" arco tones (muted) that Webern employs. Finally, Webern's instruction *an der Spitze*—at the tip [of the bow]—shows the familiarity he had with string instruments and his incredible insight into timbral control. By playing at the tip of the bow, its weakest point, additional bow pressure may be required from the performer. This increase in pressure also increases the amount of bow

[3] For more on planal analysis, see Zeller, 2020.

[4] Physicist Joe Wolfe has shown that the second, third, fourth, and sixth harmonics are much more prominent than the fundamental in at the bridge bowing (Wolfe, 2020).

noise present at the beginning of each stroke. Playing at the tip is yet another, more subtle way to understate pitch.

Op. 9/5: There are two levels of structural organization in the Fifth Bagatelle—timbre and pitch. A monophonic chain of nine sequentially presented musical blocks comprises the movement (Example 1). Within each block there is a homophonic texture presenting a complete aphoristic musical phrase consisting of a timbral idea, dynamic swell, and unique tetrachord—except Block 8 which is polyphonic. The result is a series of clearly audible, distinct phrases, usually separated by rests. Locally, each block contains an aphoristic *Klangfarbenmelodie* statement, each phrase is a microcosm of a larger musical universe. The movement's form is binary, defined by two complete aggregates of chromatic saturation. In addition, the structure has two concise closing statements: a reflecting statement (x) and a coda (c), resulting in an ABxc form.[5] Speaking specifically about Op. 9 in his 1932–33 lectures, Webern said, "The most important thing is that each 'run' of twelve notes marked a division within the piece, idea, or theme" (Webern, 1975, 51). In the Fifth Bagatelle, each phrase block contributes to the aggregate, and when the chromatic "run" is complete, so too is the large-scale formal unit. The chromatic aggregates provide the skeletal structure, but not the substance of Webern's music. They are the blank canvas stretched across a frame, waiting for the artist's paint. Rendered upon that structure are tetrachords and timbres, and these swells of sound make the aggregates comprehensible as a form.

The Fifth Bagatelle expands timbral-registral space with arco timbres escaping from their encapsulation within those of *am Steg* (at the bridge), to the sounds of pizzicato's exodus from its containment between arco tones (Example 2). Globally, the nested wedges created by the timbral lines delineate the formal divisions of the pitch aggregates. The *am Steg* line matches the aggregates' compositional pacing and divides the binary form. Spanning across the work, a wedge of registral expansion echoes the formal units outlined by the timbral wedges (Example 3). Unity throughout the pitch and timbre domains reinforces the work's cohesion.

Op. 10/1: Symmetry is an important aspect of Webern's musical language, and his *Klangfarbenmelodie* works are no exception. In the First of his *5 Stücke für Orchester* (Op. 10), Webern composes the work's symmetrical form through a continuously unfolding timbral process. Rather than a pitch construction, in this case, the axis of the symmetrical form is the timbre of the brass choir in mm. 6–7.[6] Expanding outward, the structural timbres on either side of the axis are: flutter-tongue flute in m. 8–9 mirroring the flute over celesta trill in mm. 4–5; glockenspiel in mm. 9 and 2; harp in mm. 9 and 1; celesta colored with bowed string harmonics in mm. 10 and 1; and the harp in mm. 10 along with the trumpet in m. 12 mirroring the trumpet combined and harp in the anacrusis (Example 4). Webern reinforces the axial timbre by highlighting muted trumpet at the beginning, mid-point, and end of the form, providing an anchor for the symmetrical form's timbral progression.

A loose symmetrical pitch process also exists, but it is not nearly as well-defined as the strict timbre process, and it lacks the structural vigor to be convincing as a form-bearing element. There is one complete aggregate of chromatic saturation in the work; however, it does not provide structure as it did in the Bagatelles. Particularly noteworthy, however, is that the timbre blocks that make up the body of the work are composed of sets of nine of the twelve tones. As with the tetrachords in Op. 9/5, the nonachords are not structural in this movement, but they are a way of organizing pitch for Webern, and they do have structural implications for the other movements. Op. 10/1 radically minimizes pitch-structure to the point of not allowing it to create the architecture of the work. Timbre's rigorous organization, on the other hand, clearly indicates that it was Webern's primary organizational parameter in the composition of this movement.

[5] For a discussion of reflecting statements in Webern's music see Zeller, 2020, pp. 198–221.

[6] "Axis of symmetry" usually refers to the midpoint of vertical pitch distribution in post-tonal theory. Here, the brass choir is not an axis of hierarchical timbre space, it is the axis around which a symmetrical form unfolds.

In the First Orchestral Piece, Webern actively engages with the orchestral tradition from which he is in the process of breaking away. The homophonic texture created by sections of instruments in Blocks 2 and 3 becomes a polyphonic voice as a unit that is then entwined in polyphony with the independent main line in Block 1 (Example 5). Webern finds another way to create the happy mixture of presentation styles. The rigid timbral structure of this movement is consistent, well-formed, and logical.

Representational of Webern's works of 1911, Op. 9/5 and 10/1 show two compositional strategies employing timbre in directed compositional processes—timbre-melodies—that convey clear musical ideas. *Klangfarbenmelodie* is the stylistic presentation of a musical idea by conveying its musical logic through timbre. Schoenberg's concept, and Webern's application of it, nourished the nascent chromaphony of the twentieth century. Schoenberg and Webern did not have diverging notions of *Klangfarbenmelodie*, that belief was created by critical reception after the historical schism of World War I. Webern's aphoristic works display an unprecedented adherence to timbre-based composition. Cohesive timbral statements are defined by strict musical logic developed within each work. In the historical moment of pre-War expressionism, *Klangfarbenmelodie* was one path to the new music the Second Viennese School explored.

Musical examples discussed in this presentation can be found at www.matthewzeller.com/timbre-2020.

References

Bailey, K. (1991). *The Twelve-Tone Music of Anton Webern: Old Forms in a New Language*. Cambridge: Cambridge University Press.

Bregman, A. S. (1990). *Auditory Scene Analysis: The Perceptual Organization of Sound*. Cambridge: MIT Press.

Casella, A. (1924). *L'Evoluzione della musica* (Anonymous, trans.). London: J. & W. Chester. (Original work published 1919).

Cramer, A. (2002) Schoenberg's *Klangfarbenmelodie*: A Principle of Early Atonal Harmony. *Music Theory Spectrum* 24(1), 1–24.

Goodchild, M. & McAdams, S. (2018) Perceptual Processes in Orchestration. In E. Dolan & A. Rehding (eds), *The Oxford Handbook of Timbre*. Oxford: Oxford University Press. DOI: 10.1093/oxfordhb/9780190637224.013.10.

Iverson, J. (2009). *Historical Memory and György Ligeti's Sound-Mass Music 1958–1968*. (Doctoral diss.). University of Texas, Texas.

Ligeti, G. (2007). Eine Neue Wege im Kompositionsunterricht. In M. Lichtenfeld, (ed), *Gesammelte Schriften I* (131–56). Mainz: Schott. (Original work published 1968).

Pisk, P. (1924). Die Moderne seit 1880: Deutsche. In G. Adler (ed), *Handbuch der Musikgeschichte* (pp. 1002–38). Berlin.

Schering, A. (1919). Die expressionistische Bewegung in der Musik. In *Einführung in die Kunst der Gegenwart* (pp. 139–61). Leipzig: E. A. Seemann.

Schoenberg, A. (1975). *Style and Idea: Selected Writings of Arnold Schoenberg* (L. Stein, Ed., L. Black, Trans.). London: Faber and Faber.

Schoenberg, A. (1978). *Theory of Harmony* (R. E. Carter, trans.). Berkeley: University of California Press. (Original work published 1911).

Schönberg, A. (1911) *Harmonielehre*. Vienna: Universal Edition.

Stein, E. (1923). Alban Berg-Anton v. Webern. *Musikblätter der Anbruch* V, 13–16.

Webern, A. (1975). *The Path to the New Music*. (Willi Reich, ed., Leo Black, trans.). New York: Universal Edition. (Original work published 1960).

Wolfe, J. (2020). *Articulation and vibrato on the violin.* University of New South Wales, Sydney. https://newt.phys.unsw.edu.au/jw/violinarticulation.html#sulponti.

Zeller, M. (2020). *Planal Analysis and the Emancipation of Timbre: Klangfarbenmelodie and Functional Orchestration in Mahler, Schoenberg, and Webern*. (Doctoral diss.). Duke University, Durham.

Musical OOPArts: early emergences of timbral objects

Felipe Pinto-d'Aguiar
Escuela de Artes Musicales y Sonoras, Universidad Austral de Chile, Valdivia, Los Ríos, Chile
felipe.pintodaguiar@uach.cl

Introduction

In his article 'Tempus ex Machina…', composer Gérard Grisey discusses a number of examples in classical music, where the traditional flow of time is interrupted to explore timbre. I present and analyze further examples, including cases which are not characterized by 'temporal suspensions', but other parameters relevant for generating timbre. Then, I introduce the notion Musical OOPArt in order to organize evidence that shows a progressive focalization on timbre during the romantic period and its impact in the development of future spectral and timbre-based musical styles.

Method

This research was conducted by listening and analyzing musical examples (recordings and scores) from the classical, romantic, and early modern repertoire, forming a collection of timbral-relevant objects to investigate.

Results

I expect to show how a progressive interest in timbre accentuated during the romantic period and lead to the development of spectral music and other timbre-based aesthetics.

Discussion

Starting Point

The spectral movement, which took sound as a self-referential model, has been identified for its key developments in the domains of harmony, texture, and timbre, and has had a significant impact on a variety of contemporary musical styles. Previous examples to the timbral music from the seventies, emphasizing timbre over conventional musical parameters associated to a traditional syntax such as melody, rhythm, and formulaic structures, are found in the compositions of G. Ligeti, G. Scelsi, O. Messiaen, and others. Even earlier examples can be observed within the classical and romantic periods. In the article 'Tempus ex Machina…', Gérard Grisey provides us with some notable examples from W. A. Mozart (Symphony No 40, 1st movement, bars: 58-62; 241-245), J. Brahms (Piano Concerto No 2, 1st movement, bars: 238-244; 245-260), A. Bruckner (Symphony No 9, 1st movement, bars 539-549, and 3rd movement, bars 21-29; 121-129), and R. Wagner (Das Rheingold, Beginning of the Prelude), where the standard musical development is temporarily suspended and replaced by iterations or extreme prolongations of sonic material, which produce a focus on the vertical axis of sound, almost as if composers were attempting to scrutiny the musical fabric through a microscope, although only momentarily.

One can understand timbre in two ways: as local color (micro time) or as overall color (like the characteristic sound of a rock band). The first relates to immediate perception, and the second to memory. Both approaches are relevant for timbre-centered music, but as I will show later, there are other aspects than time to consider.

Not Just Time

Additional examples in the same vein explored by Grisey are found in works by L. V. Beethoven (Symphony No 6, 2nd movement 'Nachtigall', eleven bars before the end of the movement, where time is suspended due to a *senza tempo* bird-like gesture. There is also a break in the linear narrative of Symphony No 9 with what seems to be an early example of a musical loop, thirteen bars before the end).

Something similar happens with a passage of F. Liszt in the Three Concert Études, 'Un Sospiro' (bars 46-51 '*leggierissimo volante*'), which presages the circular obsessiveness of Vortex Temporum.

In the precedent examples the 'temporal suspensions' (Grisey, 1987) allow to focus on the quality of sound, however I found several other passages in the musical literature, which are especially relevant from a timbral perspective, and that are related to other aspects distinct from the flow of time, including elements such as harmony, texture, gesture, tone color, and space. I will come back to this point later.

Musical OOPArts

To continue digging into notable musical peculiarities from the past, I would like to introduce the notion of Musical OPPArt. OOPArts or 'out of place artifacts' are archeological and puzzling, presumably human-made objects, which do not belong to the time period where found. For instance, there is an OPPArt called the 'Antikythera Mechanism' (an archaic form of analog computer created in Greece around 100 BCE), which forced to re-consider the estimated degree of advancement of the civilization where it was discovered.

In the case of music, when we encounter Musical OOPArts, we have to reconsider historical assumptions and the place of timbre in music from the past. While most of the archeological objects have been discredited or proved hoaxes by researchers, I find the concept useful in music to describe intriguing sound objects which question conventional narratives of the linearity of history and of musical progress, particularly in relation to timbre, showing that certain musical situations were pushing, perhaps always but more and more frequently since the romantic period, towards a new notion of time, texture, and timbre. Musical OOPArts are non-motivic or thematic materials. Marilyn Nonken argues that composers like Brahms and Liszt explored and revitalized neutral (non-melodic or thematically relevant) materials (Nonken, 2014), which transitioned from being mere connectors or accompaniment to become foreground. Precisely some of those innocuous or isolated musical materials have the potential to become Musical OOPArts, breaking the traditional musical syntax to open timbral explorations.

I would like to stress that these contextual anomalies are in fact that: unique events. In precedent centuries to the 20th, the timbral-textural-non-developmental objects appeared more as an exception than a rule, and apparently composers just let them occur, with no attempt to capture or routinize them. In fact, they usually come unexpectedly and leave inadvertently, followed by some conventional melodic formula perhaps dictated by the *status quo* of their time, although for contemporary listeners they strike as meaningful, discarding the possibility of being purely ornamental or digressional.

Types of Musical OOParts

In the following lines, I offer some examples of Musical OOPArts, sketching a taxonomy of their types. I should first warn that this classification is an exercise and that some elements that I will discuss could easily fit into more than one category due to the holistic nature of timbre.

1st Musical OOPArt (rhythm).

This is the previously examined type, including the examples by G. Grisey where temporal suspensions direct the attention of the listener to immediate color, anticipating timbral-based music.

2nd Musical OOPArt (harmony).

Perhaps the most obvious parameter moving from the tonal-system era to what could be called the century of timbre (the 20th and beginning of the 21st) is harmony. Harmony in relation to timbre is important because, as Tristan Murail states, these two elements work as a continuum (Murail, 2005). Harmony understood as a collection of frequencies plays a key role in defining the overall color of a sound, not just from a pitch perspective, but impacting more nuanced perceptual qualities such as brightness or mellowness, and even creating emotional associations. For instance, harmony can make us consider a sound as aggressive, sweet or mysterious. In the 3rd movement of Beethoven's Waldstein Sonata (bars 255-256; 261-262; 267-268) there is a harmonic progression in which three chords call my attention. The three suspects are harmless dominant 7th chords, although colored with sharp 9ths (C7[#9], F7[#9], B♭7[#9]).

To my jazzy ears, we are transported all of a sudden to the world of Jimmy Hendrix, Funk, and Fusion Music. Even if in the context of Beethoven these oddities are only expressive additions, they are opening the door to working with harmony as color and paving the way to what will come later with C. Debussy, A. Scriabin, and others.

3rd Musical OOPArt (texture).

Musical texture can be defined by the number of layers and the hierarchies or relationships between the component sounds. According to Panayiotis Kokoras, the paradigmatic texture of our time is what he calls *holophony*, which is a combination or synthesis of different sounds into a whole (Kokoras, 2007). The emergence of this kind of texture is partly explained by the development of electroacoustic music where the traditional categories of texture (monophony, polyphony, homophony) are replaced by open and mixed textures, resulting in a continuum of sound. If we listen to the opening (bars 1-2) of the 6th movement of G. Mahler's Das Lied von der Erde 'Der Abschield', we encounter a musical material that cannot be classified under the traditional textural types. It is really a sound object, which could have time-traveled from the middle of the 20th century to create this introductory atmosphere that is not melody, neither harmony nor any other conventional musical constituent. By building a refined orchestration, Mahler is anticipating the notion of 'composing the sound itself' (Bauer, 2001) instead of composing with a predetermined palette of sounds. As usual with Musical OOPArts, this anomalous moment is quickly abandoned, and then it turns into a conventional —fragmented however— accompanied melody.

4th Musical OOPArt (gesture).

Gesture is perhaps one of the most recent interests of musical scrutiny (along with timbre). Musical gestures are linked with physical actions literally or figuratively. Considering gesture in its metaphorical aspect, that is as a musical material characterized by active movement, I would like to concentrate on three examples from R. Strauss. Two of them appear in the Alpine Symphony. The first is found three bars before rehearsal number 43 and marked 'At the Waterfall'. Extremely fast descending note-streams not only effectively achieve the association with a cascade, but also surprise because of the novelty of the gesture and the brightness of the orchestration. Since the music is fast, it leaves a memorable sound signature, thus gesture becomes timbre. Here, Strauss anticipates the material of page 31 in Grisey's Partiels. At rehearsal number 43 'Apparitions' the same material of the precedent waterfall gesture is stretched and its energy de-radicalized to transform the material into a stable texture, which preserves the overall timbre. In rehearsal number 5 of Josephslegende, we find a violin gesture which displays physical action, and also stands out for its originality as sonic substance. The virtuosity here is not related to the craft of N. Paganini, but more directly to the one of S. Sciarrino. Although the passage requires a mastery level of the instrument, it seems more a virtuosity of sound than acrobatics of the hands.

In bars 106-110 of B. Bartók's Concerto for Orchestra (one of his latest works) one can appreciate a music that breaks the linearity of narrative to focus on 'a single body of sound' (Johnsons 2015). Here the gesture is stretched to become texture, one that does not progress, but that is self-referential, announcing some of the sonorities of the music from the seventies. Bartók is perhaps the 'missing link' between late romanticism, and timbre-based music of the 20th and 21st centuries —probably even more than I. Stravinsky and A. Schönberg were, and whose rivalry seems more related to the music of the first half of the 20th century.

5th Musical OOPArt (tone color).

It may seem tautological to discuss a timbre-defining object in terms of its tone color, but I wish to refer to a very specific case, which is when special attention is given for creating a new sound in a more deliberate way. There are several examples in the literature when sound is altered to produce an 'effect', but there are cases where an effect acquires transcendence. If we pay attention to bars 84-85 of R. Wagner's Das Rheingold, 'Entrance of the Gods into Valhalla' we will encounter sound blocks, which could perfectly belong to a passage orchestrated by O. Messiaen, who would later introduce the notion of 'harmonic-timbral complex'. This example is specially interesting since we have a reference in the same movement to compare to. Bars 84-85 are a sort of zoom-in or distorted version of the precedent bars 68-

69, which are indeed more conventional in their sonority. It seems that Wagner is opening the possibility of incorporating timbre as a procedure for musical development.

Another example of utilizing color to create an evocative passage appears in the Rite of Spring (last bar of 'The Sage'), where Stravinsky only needs two seconds to capture our attention and transports us into the domain of the most adventurous works of G. Ligeti.

The previously mentioned example of Mahler could fit into this type of Musical OPPArt as well, although since it is an object which expands time itself, I consider it more related to textural development than to immediate color.

6th Musical OOPArt (space).

The final type of Musical OOPArt I have reflected upon carries us into a more distant past than the romantic era. Longtime before the days of amplification, effect-pedals, and recording studio techniques —and the later assimilation of these resources by purely acoustic music, which mimics delay, chorus or reverb effects—, spatial considerations were relevant. Space dramatically affects timbre (in the micro and the macro levels). Take for instance bars 10-15 of G. Gabrieli's Canzon in Echo a 10. The written echo effects —and panning effects if performed in two opposing choruses— in addition to the resonances of an enclosed space, provide the brass instruments with a timbre completely different to the one achieved in open spaces. By writing echoes in the score, Gabrieli enhances an acoustic phenomenon, and in this way becomes one the precursors of turning form and material into one thing. One could also discuss the associated practice of stereophonic choruses, no doubt a visionary Out of Place Practice or 'OOPPra'…

Conclusions

Several taxonomies of sound objects or musical types have been attempted since Pierre Schaffer. I am particularly attracted to the thoughts of Helmut Lachenmann on this respect (Lachenmann, 1970) and to the more experimental approach proposed by Xavier Hautbois with *Les UST* (Temporal Semiotic Units) (Frey et al., 2014). These new systems of classification have emerged as a necessity to organize contemporary musical entities, which do not conform to traditional grammars of music, and that are heavily structured on timbre. Some of these sonic entities, born long ago but described only in recent decades, have existed as 'out of time' musical fragments in precedent centuries, perhaps ignored *'objets à découvrir'* of their days, which acted as windows to the music of the future. The Musical OOPArts previously discussed are a testimony of the weirdness of history, and perhaps the ubiquity of certain musical aspects beyond time. They also suggest the possibility that being attuned to current singularities (contemporary Musical OOPArts) could influence, in a more conscious manner, the music to become.

References

Bauer, A. (2001). Composing the Sound Itself: Secondary Parameters and Structure in the Music of Ligeti. *Indiana Theory Review*, 22(1), 37-64.

Frey, A., Hautbois, X., Bootz, P., & Tijus, C. (2014). An experimental validation of Temporal Semiotic Units and Parameterized Time Motifs. *Musicae Scientiae*, 18(1), 98–123.

Grisey, G. (1987). Tempus ex Machina: A composer's reflections on musical time. *Contemporary Music Review*, 2(1), 239-275.

Johnson, J. (2015). Out of time: Music and the making of modernity. New York: Oxford University Press.

Kokoras, P. (2007). Towards a holophonic musical texture. *The Journal of Music and Meaning, 4,* 1–7.

Lachenmann, H. (1970). *Klangtypen der neuen Musik: Klangbeispiele*. Stuttgart: Ichthys Verlag.

Murail, T. (2005). After-thoughts. *Contemporary Music Review*, 24, 269-272.

Nonken, M., & Dufourt, H. (2016). *The spectral piano: From Liszt, Scriabin, and Debussy to the digital age.* Cambridge: Cambridge University Press.


TIMBRE