

UiO : **University of Oslo**

Tejaswinee Kelkar

Computational Analysis of Melodic Contour and Body Movement

Thesis submitted for the degree of Philosophiae Doctor

Department of Musicology
Faculty of Humanities

RITMO Center for Interdisciplinary Studies in Rhythm Time and
Motion



2019

© Tejaswinee Kelkar, 2019

Faculty of Humanities, University of Oslo

All rights reserved. No part of this publication may be reproduced or transmitted, in any form or by any means, without permission.

Cover: Hanne Baadsgaard Utigard.

Print production: 07-Media Oslo.

to ajji

Acknowledgements

Writing this thesis has been an unbelievable challenge, and a great source of enjoyment for me. I would like to take this opportunity to thank everyone who enabled me to spend three years thinking about melodic contour and their relation to the body - this act itself is an immense privilege.

Thank you Alexander for letting me do this project, and always showing the light in all of the different projects that I was able to get involved in. Reading your work itself is always revealing, clarifying, and inspires a lot of traction and hope for me to engage with music in deeper, and more meaningful ways.

Thank you so much to Rolf Inge for motivating this work philosophically, and for teaching me the ways to approach the cores of research problems. Every conversation with you has always been extremely motivating and calming, and full of parables and things to remember for life.

Thank you very much Anne for being the director of the center of dreams, and such an inspiration. Thank you also for all your kindness. Thank you Peter, Målfrid, and Ancha for all your positivity and making all of this work possible. Thank you Stan for your encouragement. Thank you Bipin sir for teaching me some valuable lessons early on during my masters that have guided me throughout. Thank you Bruno for the encouragement throughout the projects.

I would like to thank all the participants of these experiments for their contributions, insights, and thoughtful reflections on the experiments. Last but not least, I would like to thank each and every one of my colleagues both at IMV and RITMO for building a research environment where we thrive on discussion, learning from each other, and leaning on each other.

Thank you to my parents. It is comical to thank you because I owe everything to the way you brought me up. Thank you especially Aai for talking to me every single day and always being there. Nachi and Kadambari, I get motivated for anything at all thinking about you. Thank you Udit for being such a pillar of strength and support. Rajvi my love you are always here and to Malathi for being the biggest remote support. Thank you very much to Chitrakleha for the excellent work with copy-editing, and Dayita for the detailed feedback.

Ragnhild, my darling, for your presence in my life. I don't remember you not being there in it at all. Victoria for teaching me a way to live here, to speak, I am really thankful for being able to share so much with you! Tore, for your insight, your nuanced and calm way of being a friend. Aine and Derek, thank you for adopting me and always being there. Mari, my first ever friend here, and a mentor for so much else. Sanskriti, thank you for all of your music, and your friendship, it has been invaluable to me and will always be. Ulf, for your humor and presence and *hjørnekontoret* and chicken. Ingrid, thanks for being the *stjerna*. Thank you Kayla for babysitting me this year, and to Lucas. Thank

Acknowledgements

you very much to Charles, Victor, and Olivier for guidance and help. Thanks to my compatriots co-warriors of thesising: Emil, Gui, Stephane, Bjørnar, Kjell-Andreas, Marek, Benedikte, Agata, and Merve. Ingeborg, for your friendship, and my Marius. Thanks to dear friends from the music / kunst miljø Andreas, Petrine, Karoline, Kjetil, Åsmund, and Martin. I would like to thank Deepak dada, Wadegaokar guruji, Manas Vishwaroop, Kamod Arbedwar, Achal Yadav, Suyash Medh, Kelcey Gavar, and James Bunch, I will always owe you a lot.

• **Tejaswinee Kelkar**

Oslo, November 2019

List of Papers

Paper I

Kelkar, T., & Jensenius, A. R. (2017). Exploring melody and motion features in “sound-tracings”. In Proceedings of the 14th *Sound and Music Computing Conference* (pp. 98-103). Aalto University.

Paper II

Kelkar, T., & Jensenius, A. R. (2017, June). Representation strategies in two-handed melodic sound-tracing. In Proceedings of the 4th International Conference on Movement Computing (p. 11). ACM.

Paper III

Kelkar, T., & Jensenius, A. (2018). Analyzing free-hand sound-tracings of melodic phrases. *Applied Sciences*, 8(1), 135.

Paper IV

Kelkar, T., Roy, U., & Jensenius, A. R. (2018). Evaluating a collection of Sound-Tracing Data of Melodic Phrases. In Proceedings of the 19th *International Society for Music Information Retrieval Conference*, Paris, France. (pp. 74-81)

Contents

Acknowledgements	iii
List of Papers	v
Contents	vii
List of Figures	xi
List of Tables	xiii
1 Introduction	1
1.1 Introduction	1
1.2 Motivation	2
1.3 Research Objective and Research Questions	3
1.4 Approach	5
1.4.1 Open Science	7
1.5 Thesis Outline	7
2 Melody	9
2.1 Introduction	9
2.2 Speech Melody	10
2.2.1 Prosody	11
2.2.2 Intonation	11
2.2.3 Model for a Speech–Song Spectrum	12
2.3 Musical Melody	13
2.3.1 Pitch	14
2.3.2 Organization	16
2.3.3 Recognizability	19
2.3.4 Succession	23
2.3.5 Shape	24
2.4 Other Aspects of Melody	26
2.4.1 Voice and Melody	26
2.5 Discussion	29
2.6 Summary	29
3 Auditory and Motor Imagery	31
3.1 Introduction	31
3.1.1 Terminology	31
3.2 Auditory Imagery	32
3.2.1 Auditory Scene Analysis	32

	3.2.2	Ecological Pitch Perception	33
	3.2.3	Conceptual Shapes and Music Theory	33
3.3		Gestalt and Shape	34
	3.3.1	Melodic Gestalt	34
	3.3.2	Contour Continuity and Gestalt	35
3.4		Motor Imagery	36
	3.4.1	Phenomenology of Shape	37
	3.4.2	Neural Correlates of Phenomenology of Shape	38
	3.4.3	Lines and Movement	38
3.5		Summary	39
4		Body Movement	41
	4.1	Introduction	41
	4.2	Sound-Tracing	41
	4.3	Embodied Music Cognition	43
	4.3.1	Early Research on Music Embodiment	43
	4.3.2	Recent Research on Music Embodiment	44
	4.3.3	Auditory Working Memory	45
	4.3.4	Multimodality	46
	4.3.5	Ecological Psychology	46
	4.3.6	Conceptual Metaphors	47
4.4		Music Related Movement	48
	4.4.1	Gestural Imagery	48
	4.4.2	Gesture and Musical Gestures	50
4.5		Action–Sound Perception	52
	4.5.1	Ideomotor Theory	53
	4.5.2	Motor Control	54
4.6		Summary	55
5		Data Sets and Experiments	57
	5.1	Introduction	57
	5.2	Stimulus Set	57
	5.2.1	Descriptions of the Musical Styles	58
	5.3	Experiments	60
	5.3.1	Experiment 1: Melodic Tracings	60
	5.3.2	Experiment 2: Melodic Tracings, Tracing Repetitions, Singing Melodies	61
	5.4	Data Sets Used	62
	5.5	Summary	63
6		Methods	65
	6.1	Introduction	65
	6.2	Sound Analysis	65
	6.2.1	Computational Representation and Features of Melody	65
	6.2.2	Symbolic Approaches	67

6.2.3	Music Information Retrieval	69
6.2.4	Extraction of Pitch Contours and Contour Analysis	70
6.2.5	Pitch Detection Algorithms	71
6.2.6	Contour Analysis Methods	72
6.3	Motion Analysis	73
6.3.1	Tracking Data through Infrared Motion Capture Systems	74
6.3.2	Details for the Experiments in the thesis	75
6.3.3	Post Processing	75
6.3.4	File Formats for Motion Capture Files	76
6.3.5	Feature Extraction	78
6.3.6	Toolboxes for MoCap Data	79
6.4	Motion-Sound Analysis	80
6.4.1	Visual Inspection and Visualization	81
6.4.2	Statistical Testing	81
6.4.3	Machine Learning and Artificial Intelligence	82
6.4.4	Time Series Distance Metrics	84
6.4.5	Canonical Correlation Analysis (CCA)	85
6.4.6	Deep CCA	87
6.4.7	Template Matching	87
6.5	Summary	87
7	Conclusions	89
7.1	Research Summary	89
7.1.1	Paper I	89
7.1.2	Paper II	90
7.1.3	Paper III	91
7.1.4	Paper IV	92
7.2	Discussion	92
7.3	General Discussion	95
7.3.1	Verticality	95
7.3.2	Imagery	96
7.3.3	Voice	97
7.3.4	Body	98
7.3.5	Cultural Considerations	99
7.4	Impact and Future Work	100
7.4.1	Research in Melody and Prosody Perception	100
7.4.2	Search and Retrieval Systems	100
7.4.3	Interactive Music Generation	101
7.4.4	Future Work	102
	Bibliography	103
	Papers	120
I	Exploring melody and motion features in “sound-tracings”	121

II	Representation Strategies in Two-handed Melodic Sound-Tracing	129
III	Analyzing free-hand sound-tracings of melodic phrases	135
IV	Evaluating a collection of Sound-Tracing Data of Melodic Phrases	159
	Appendices	169
A	Details of the Experiments	171
	A.1 Marker Labeling	171
	A.2 List of Melodic Stimuli	172
	A.3 File Formats for Symbolic Music Notation	173
B	List of implemented functions and features	175
	B.1 Dependencies	175
	B.2 Data Types	175
	B.3 Functions	175
	B.3.1 Motion Features	175

List of Figures

- 2.1 Speech–Song spectrum. In this figure, I have tried to represent the many forms between speech and song. Most categories and their place on this spectrum are not assigned without debate; this figure is only indicative. 13
- 2.2 A melodic entity might be stable, as discussed above, to several kinds of variation. Which kinds of variations are most relevant depends upon the musical culture within which a similarity judgment is to be made. 20
- 2.3 A melodic entity can be recognized as belonging to one of several melodic frameworks. A framework might mean different things according to the musical context, as illustrated. 21
- 2.4 Range of melodic and sound contours from previous studies from Seeger (1960); Schaeffer et al. (1967a); Śarmā (2006); Hood (1982); Adams (1976) 24
- 2.5 On the left, we see Palestrina’s Iubilate deo universa terra" psalm verses in neumes first published in 1593 in Offertoria totius anni, no. 14. It is argued that neumatic notation is derived from cheironomic hand gestures indicating changes in pitch. On the right, is an illustration of how neumes evolved into mensuration notation, and finally as modern notation. 26
- 2.6 Tibetan Yang notation from silkscreen prints, ca. nineteenth century. This form of notation goes back to the sixth century (Collection, 2019) 27

- 3.1 A and B represent the property of isotropy in this illusion; the triangle is perceived despite scaling and rotation. In C and D, with the change of an angle, we perceive the third edge of the triangle to be slightly curved. This alludes to the properties of smoothness and locality. The lines at the corners are still straight lines, but the effect is of a curved edge instead. 36

- 4.1 Solfa gestures are used to help memorize intervals in a major scale. This method has been used quite often to teach students of singing. 51
- 4.2 Similarities between the action–fidgeting model and the posture-based motion planning theory. 53

List of Figures

5.1	A stimulus set containing 16 melodies was used for two motion capture experiments, resulting in three data sets: one for melody–motion pairs, the second for repetitions in sound-tracings, and the third for singing back melodies after one hearing.	57
5.2	The 16 melodies used as the stimulus set for all the experiments in the thesis come from four different music cultures and contain no words. The X-axis represents time, and Y-axis represents pitch height in MIDI notation.	58
5.3	Experimental flow for experiment 1.	60
5.4	Flow of Experiment 2. The repetitions and sung sections are included in this experiment.	61
6.1	An illustration of different levels of data handling and analyses in the experiments. The three types of data each have their own signal	66
6.2	Pictures from the FourMs MoCap Lab where the experiments are conducted. On the left is the lab with the cameras and speakers mounted as shown, on the right is a participant wearing reflective markers.	74
6.3	An example of a post-processed motion capture stick figure. A detailed list of marker labels can be found in Appendix B. . . .	76
6.4	An illustration of the flow of mocap data. The mocap files are exported from QTM, imported into python. Normalized files are re-exported, and analyzed in python for obtaining features. . . .	77
6.5	Visualizations of quantities of motion. Visual inspection	82
6.6	A representation of the CCA algorithm used; fa and fb represent the two different data sets—melodic and motion features, respectively.	86
A.1	The 16 melodies used as the stimulus set for all the experiments in the thesis come from four different music cultures and contain no words.	172

List of Tables

- 6.1 The features extracted from the motion capture data to describe hand movements. 79
- 6.2 Quantitative motion capture features that match the qualitatively observed strategies. QoM refers to *quantity of motion*. 80

Chapter 1

Introduction

*Melody, moving downstream.
A string of barges
Just lit against the blue evening
The fog - giving each light - a halo
Moving with the river but not the drift*

*a little faster perhaps,
or is it slower?
A singing
Sung if it is sung quietly
within the scored crashing and the
almost inaudible hum impinging
upon the river's
seawardness.
- Denise Levertov (Levertov, 1983)*

1.1 Introduction

Many people think of melodies as having contours. The association between musical melodies and the visual representation of a contour—an arch, a rainbow, a zigzag line, a circle, and so on—appears to be united in our minds, and represents an essential quality of melodic identity. Lessons in songwriting often teach students how to diversify their thinking by creating contrasts with contours, although there are few formal methods to analyze contours, and they are mostly based on the analysis of sheet music.

‘Contour’ also has signification beyond the visual representation of music. Be it the evocative nature of optimistic arch-like melodies of theme songs of Disney princesses, the use of large arch-like leaps of a sixth or above in the ‘princess’ songs, ranging from *Somewhere Over the Rainbow* from the Wizard of Oz to Mulan’s *Reflection*, or the linear contours in Alban Berg’s compositions, as discussed by Perle, p. 86, and the descending contours of many lament songforms; melodic contours have signification for us beyond just being musical artifacts. Representations of contours on paper range from neumes and squiggly lines drawn above musical notations to remind the reader of the contours of a musical phrase, to representing these phrases with hand movements during improvisation, as is done in, for example, the Hindustani musical tradition.

In this thesis, I investigate *why* we think of melodies as contours, and *how* this differs from symbolic or score-based representations of melodies. A primary objective of this thesis is trying to understand how people move to melodies. I

1. Introduction

come to this from the perspective of a vocalist, which is my entry point to these research questions.

1.2 Motivation

I started thinking about melodic contours and the embodiment of music when I studied Hindustani vocal music as a child. While improvising, musicians would use elaborate hand gestures to accompany melodic phrases. When I was new to this musical genre, it seemed as though only experts were allowed to use gestural improvisation upon mastering the style. However, regardless of their expertise, different singers use various styles of gestural elaboration that communicate improvised melodic phrases to the audience, through their hands, heads, and facial expressions. These movements not only serve a communicative function, but they also assist the process of singing through the manipulation of resonance centers, and facilitating breath supply. During lessons, my teachers would illustrate nuances in the melody with hand movements, which helped me understand the phrasing faster than in their absence, an experience that is well documented by several learners of this style (Pearson, 2016). Over time, movement-based representation seemed the only way to understand the intricacies of melodic phrasing.

When I learned operatic music, however, this particular method of using visual shape to understand melody was completely absent. I also noticed that the body was used differently in this cultural context. The idea of cultural body use has been explained in research (Kimmel, 2008,p.77). The vocabulary of gestures and visual metaphors for melodic shapes that are integral to Indian music are simply not the ones used in operatic singing. Instead, melodies were visualized using shapes that related to the resonance centers in the body activated in different vocal transitions. Students like me, in trying to replicate the body movements from classical Indian music to operatic singing, ended up influencing the articulation of the voice in this new style. I wondered why this might be—what is the relationship between the singer’s body and the melody? How is melodic imagery related to abstract shapes? What are the different ways in which melodic contour is enumerated, understood, and used? Furthermore, I was interested in the intersubjectivity in the experience of these shapes. How do people differ in their conceptualizations of melodic shape, and why?

Experiences like this with my training as a vocalist led me to read research on the use of hand movements in Hindustani music, and their semiotic and cognitive implications. While researching this topic for my master’s thesis, I came across the semiotics of these gestures and their pedagogical function. I analyzed video recordings of performers singing the same set of ragas, to understand their specific use of hand gestures to represent rhythm, vowels, and phrase termination.

Although my masters’s research project on this subject was restricted to Indian music, the concept of melodic shape is not. Visual representations of melodies and melodic contours are found in places ranging from ancient notation forms, such as neumes, to modern music visualizers; melodic shape appears to

be a robust concept that requires further investigation in parallel with research on how the body is used to express this motor imagery. I have investigated these shapes in this thesis, using a set of vocal melodic stimuli, and have asked people to respond to the music physically, using actions. Inevitably, this subject deals with how we remember and imagine melody, which in turn, is also influenced by the affective content of melodic inflection. I have brought together the aspects of contour dealing with motor imagery, melodic memory, and melodic affect in this thesis by presenting articles that deal with each in different ways.

I have used real sound recordings as stimulus material. Much research done in the area of melodic contour perception and analysis involves the use of isochronous melodies using MIDI and symbolic representation. However, by using isochronous and symbolic melodies, we lose out on a range of information that contributes to melodic contour perception. In order to avoid this, I have used signal processing methods and continuous motion capture to reflect on the multimodal nature of melodic contours. Ultimately, the goal is to learn more about melodic, and by extension, pitch perception.

1.3 Research Objective and Research Questions

The primary research objective of this thesis is to:

Understand the role of embodiment in melodic contour perception using the sound-tracing experimental methodology.

The following postulates form the basis of the questions explored in this thesis:

1. Melodies have contours or lines.
2. Melodies and lines are thought of as having movement.
3. This movement is imagined and can be represented physically.
4. We learn about melodic perception by studying these movements.

From this primary objective, the following research questions emerge:

RQ 1: How do listeners represent melodic motion through body movement, and how can we analyze motion representations of melodic contours?

The primary objective of this thesis is to understand melody through motion. I explore the idea that the vocabulary for melodic undulation is tied to the language of describing melodies. When we ask people to represent these contours without the use of language, what do we find, and what do these findings suggest about contour perception?

To fulfil this objective, I asked people to “draw” these contours in air, and recorded them using motion capture technology. Motion traces were then analyzed and cross-compared to reveal patterns in the movements.

1. Introduction

In this thesis, I have referred to people's intentional body movements as *movement*, and the data captured from these movements as *motion*, from their use in systems for *motion capture*. I believe that the term movement intrinsically references intentionality rather than motion. As such, I have been consistent to this distinction between movement, as it is performed, and motion, as data gathered from body movements, which could pertain to motion with or without intentionality. The word *trajectories* is used to represent the motion traces in motion data, pertaining to body parts that are tracked using markers.

RQ 2: What are the characteristics and applications of motion representations of melodic contours?

Since contour representations rely on imagining contour movements as shapes, it is only possible to reflect upon phrasal shapes in prospectively or retrospectively, trying to anticipate or remember the memory of such a shape. Melodic shapes may also be context dependent, and differences in vocal styles and genres can influence tracings. This information is typically lost when we transcribe melodies down to discrete pitches, and play them back from, for example, a MIDI sequence, which is a common way of conducting contour-related experiments.

RQ 3: How can we test if motion related to melodic contours is consistent: a) within participants, and b) across participants?

Comparing the motion representations of contours of several people to find commonalities between these contour representations, I analyzed if the motion representations of participants can be modeled individually, representing consistency of a mental model of melodic contour.

Another sub-objective is to develop technology for the analysis of sound and movement pairs with each other. This involves creating toolboxes and libraries that facilitate the analysis of sound and motion. The work done to achieve this objective includes the creation of a stand-alone library to analyze motion capture data from sound-tracings and various features from those data.

RQ 4: Can we build a system to retrieve specific melodies based on sound-tracings?

Could our understanding of sound-tracings and melodic embodiment be used to build systems that would be able to retrieve specific or similar melodies? This problem might help add to both the gamut of interactive interface literature as well as music information retrieval applications.

Can the answers to research questions from 1-3 can provide data that can help build a retrieval system to explore this question. Such a system would require learning between two different paired modalities. For this, it would be necessary to model tracing-based representations and variations well.

1.4 Approach

Interdisciplinarity

I draw on research in music cognition to investigate the multi-modal interactions between music and movement, and contribute to the computational methods for analysis. The aim has been to combine these approaches, which are already interdisciplinary, to deal with music information using state-of-the-art algorithms and tools, and to investigate both their fit for the human transliteration of data, and also the other way around. Melodic grammar and perception has been modelled in a number of ways algorithmically to perform tasks that are quite simple for humans to perform, such as identifying melodic similarity.

Owing to its interdisciplinary approach, this thesis handles terminology from three domains: auditory perception, motor action, and abstract imagery and its geometry. The central idea is that melodic contour perception gives rise to shape imagery, which is realizable through motor action. In other words, there is something movement-like about melodies, and we can think of them as having a geometric structure.

There are three levels of sound and movement representations handled in this thesis: *physical*, *digital*, and *perceptual*. Physical representations of sound and movement are included in the data from direct recordings. Digital representations of sound include transcriptions of melodies and transformation of contour data into the symbolic domain. Perceptual descriptors of sound and motion include computational models that mimic perceptual qualities, such as ‘smoothness’ of movement and ‘loudness’ of sound.

More specifically, for auditory descriptors, I refer to three different levels of concepts: the *acoustic*, *psychoacoustic*, and *musical*. Acoustic analysis includes features that are calculated mathematically; for example, the energy of a sound signal or its spectral features. Psychoacoustic features can also be approximated through computational methods, such as perceptual loudness of a sound signal. Musical features of a sound stimulus are different from these, and may be embedded within a musical culture; for instance, cadence or specific intervals. They may also be psychoacoustic approximations that are understood as having a ‘musical’ quality, such as melody or intervals.

This research deals with human body motion, or the ‘response’ material for the perceived qualities in melody, as data that could be purely *physical* or *perceptual*. Analysis of these data involves, for example, the notion of an *effector*, which refers to any body part that might be involved carrying out a movement (hands, legs, the head), or an object held by or attached to the body, such as a wand. Whether body movement is measured using cameras, motion capture, and so on, also dictates how the data are obtained, and what these data can show us. In this thesis, I mainly present motion capture data, which gives precise 3D positions using infra-red markers. Some measures calculated from these data are physical, such as ‘Quantity of Motion’, whereas others are based on modeling movement perception, like smoothness. As such, this is comparable to acoustic and psychoacoustic features. The details of motion features are explained in

Chapter 6.

Despite how far research on melodic modeling and pitch perception has progressed in recent decades, there is a lot more about melodic contour perception, and its entanglements with speech, that we do not yet comprehend or have been still trying to model (Schmuckler, 2004). The key tenets of embodied music cognition explain how we might understand nuances of sound perception when we understand how the body reacts to sound and music in the environment, as several studies about beat perception and bodily entrainment have shown.

Through this thesis, I have approached melody in a similar way. What can our knowledge of embodiment add to our understanding of melodic perception, and how can we use it to inform music information related systems? How can we improve interactive interfaces for the creation of music using this knowledge?

Nymoen et al. (2013) have explored the ideas of “active listening”, where we control listening to musical stimuli using our bodies, by using devices to track body motion. This can help us actively control for example, the speed of playback, triggering samples, and so on using movement. In research involving both movement and music, these elements have often been treated separately (Müller, 2007). Even if we do understand music-related movement well, the question of how it informs music analysis could still be answered and understood in ways that allow us to explore the embodied nature of melodic perception. For instance, by designing experiments that pay attention to our natural instincts for representing music in the context in which it is heard with our bodies. Or in other words, to incorporate embodied listening into the practice of music analysis is a goal of this work.

Limitations and Scope

I have focused on analyzing melody motion data pairs in various different ways that best illustrate spontaneous melody-motion associations. The broad results of this study imply that *metaphoric thinking* is natural to most participants, regardless of their explicit experience with movement and melodic motion. I explain the details of movement metaphors in Chapter 4.

Even though this thesis touches on theoretical perspectives in speech-prosody, and contour perception in speech, I am unable to get into the experimental analysis of melodies across speech and music due to time constraints. Still, I find it important to mention a range of studies in this domain, because exaggerated contours of speech-melody form an important part of our early experiences with melodic contour. But the extent to which these experiences contribute to cognition of ‘musical’ melodies is widely debated (Patel, 2010; Zatorre and Baum, 2012).

The melodic stimuli used in this thesis are from four different music-cultures, being classical vocalise, scat singing, Hindustani classical singing, and the Sami joik. Despite this, this work does not involve cross comparisons of melodic grammars within these music-cultures. However, I do discuss implications of the participants having prior knowledge about the use of the body in some of these genres. But the experiments are not intended to highlight ‘cross-

cultural' differences; they are not designed to be able to comment on them, and socioeconomic and geocultural factors are outside the scope of this thesis.

Some findings of this thesis are also relevant to understanding gender differences in a movement analysis context. While it could have been interesting to comment on gender differences in music related movement, it is outside the scope of this thesis to discuss these findings in light of gender theory. I will focus instead on the findings from the data analysis that are directly connected to the research questions.

1.4.1 Open Science

All the articles published in this thesis are open access. In addition, the data sets and codes for running all experiments have been documented and released online. In this way, This makes the work comply to the principles of open access, and open data. The papers, code, data, and descriptions have been released on my website at

<http://tejaswineek.github.io>

1.5 Thesis Outline

The thesis consists of two parts. The first part is an introduction to the theoretical frameworks, research motivations, methods, and experiments conducted. The second part is a collection of papers published in various peer-reviewed journals and conference proceedings. In the summary section of the thesis, I introduce the articles and elaborate the key findings of the research.

The main problems posed in this thesis relate to body movement and melody. Chapters 2, 3, and 4 offer an overview of the theoretical motivations and key concepts surrounding the interaction of music and motion, specifically melodic contour. Chapter 5 describes the data sets and experiments conducted. Chapter 6 presents the main frameworks or the disciplinary areas that inform the work in this thesis, including analysis methods and technologies used. Chapter 7 provides a summary of key results in each of the appended articles, and additional results that were not included in the articles for various reasons. In addition, Chapter 7 offers a discussion of how the research questions raised in this introduction have been answered, and presents applications of the research done, elaborating on future work.

All papers published in this thesis are openly available, not just on the university website; they are written as open access articles. The data sets are also publicly available and released at links mentioned in Chapter 5, and on my personal website. Appendix 1 describes details of the experiments including the stimuli, and details of motion capture. The code written for the analysis is also open source and accessible on my personal website, and described in Appendix 2.

Chapter 2

Melody

*Actually, almost any note
can be played if
there is a melodic shape to the line.*
- Bob Mintzer (2004,p. 24)

2.1 Introduction

The above quotation summarizes how many musicians think about improvisation: music as melodic shapes rather than as notes. Jazz improvisation is often a combination of several practiced ‘licks’ and phrases in the repertoire of an improvising musician, which are played in different combinations, and over different scales. I find this interesting because melodies are considered both, a cluster of intervals as well as a contour unfolding over time, as if the contour properties overrule the effects of the intervals. But do they?

Think about a familiar melody, say, *Twinkle Twinkle Little Star*. When remembering a melody, many people recall it ‘as a whole’, at a faster pace than when they would actually sing it. However while speeding through the melody, they do not distort the durations of the notes in the melody (Andrews et al., 1998). People also recall the melody by imaginizing singing it under their breath, but the speeding up aspect is most interesting to me. In mental recall of melody, people are less likely to have a clear image of the actual intervals, and may be more invested in the act of singing through the contour. To me, this demonstrates many properties of melodies in our imagination: they are compressible and expandable, and they are transposable to any key and octave. A melody is accessible to us as one holistic *object*, with its contour, form, and rhythm embedded in the melodic *entity* or *phrase*. This property of resilience is reflected in our ability to tolerate badly sung renditions of known melodies. We are capable of smoothing over details in a melody when we are listening to, for instance, children trying to repeat a melody that they do not know well. As such, Mintzer’s quotation rings true—the melody is what it is, as long as the line holds its shape. So, either the melodies themselves are robust, or we are forgiving of melodic distortions, or both.

Exaggerated contour explorations are particularly common in infant babble—a slow exploration of the apparatus of enunciation, from vocables and vowels to non-speech sounds. Children repeat *melodic contours* and *melodic phrases* over and over during play. These infant melodies are essential to the development of speech and hearing. Specifically, contour acquisition is equally important to understand emotional nuances of speech as it is for remembering musical melodies. Whether our nuanced understanding of contours is developed *for*

2. Melody

either speech or music is not a question this thesis explores; throughout history, though, researchers have wondered about the crossovers between speech melodies and song, song-like speech, and speech-like song in various musical cultures.

Classic examples of the speech–song illusion include *Sometimes Behave So Strangely* by Diana Deutsch, where a broken replay of these words somehow makes the phrase sound like a tonal melody, to the point that if the phrase is encountered within a speech excerpt, it automatically sticks out as a ‘song’ (Deutsch et al., 2008, 2011).

The idea of speech–song is not new, especially to the internet generation, as speech and interview excerpts can be easily transformed into catchy songs by mobile applications. Every time a fragment (such as the ‘So Strangely’ one) that we learn to hear as a tonal melody plays, we switch to the song mode of listening on hearing the first syllable, even before we hear the whole melody.

This property of recall for familiar melodies makes them a lot like *objects* or *icons*—identifiable from the onset and resilient to variability. This holistic perception of a melody is what I have chosen to work on as the theoretical goal of this thesis. Why is it that melodies are understood as a whole, and we are able to think of them as stable shapes, while simultaneously understanding how these melodies unfold in time. These properties are similar to the phenomenological understanding of geometry of shapes, about which I go into detail in the next chapter.

In this chapter, I introduce the central theme of this thesis: melody. In Section 2.2, I show the interconnections between speech melody and melodic cognition, and how we sometimes hear speech melodies as having musical qualities. In Section 2.3, I explain five essential characteristics of melodies and how they relate to melodic contour. I also discuss how melodic *entities* are established in various music cultures, the peculiarity of vocal melodies, and the connection between verticality and melody. I detail what specifically about contour is of interest thereafter, in Section 2.5.

2.2 Speech Melody

Is there melody in speech? The ups and downs in speech intonation across different languages are studied in detail in linguistics as *prosody* and *intonation*. *Prosody* refers to the suprasegmental properties of speech, such as the modulation of voice pitch, the durations and stresses of syllables, and fluctuations of loudness; the pitch–curve–like properties are studied more frequently as *intonation*. The connection between speech melody and musical melody has always been at the forefront of discussions on the definitive aspects of melodies. To quote Bolinger, “Since intonation is synonymous with speech melody, and melody is a term borrowed from music, it is natural to wonder what connection there may be between music and intonation.” (Bolinger and Bolinger, 1986, p.28)

Speech is often also accompanied by gestures: by movements of the hands, head, or body (McNeill, 1992). The relationship between speech melody and

co-speech gestures will be discussed in detail in Chapter 4. Here, I will discuss contours of speech melodies in more detail.

The systematic study of speech accents and speech contours across different languages may not directly be related to the study of contours in musical melodies. However, intonation contours are studied using fundamental pitch extraction and by annotating high and low points in order to discuss contour families in different languages (Ladd, 2008; Bolinger and Bolinger, 1986; Wittmann, 1980). I would like to draw upon this approach for discussing musical melodies. But musical melodies are mostly studied in the form of pitch transcriptions and symbolic notation, and traditionally, contour analysis is not usually researched unless the research is explicitly about melodic contour.

2.2.1 Prosody

The word ‘prosody’ can be traced back to ancient Greek. A combination of two words that meant ‘towards song’, it was used to mean ‘song sung to music’ (Nooteboom, 1997). Prosody generally refers more to the timing-related elements of speech, rather than pitch levels. For example, in a rhyming poem, the timing of a metrical foot is used to give rhythmic sense to spoken words, as seen in poetry from Shakespeare and Kabir, to Matsuo Basho and Kendrick Lamar. In clever poems, such as *Jabberwocky*, Lewis Carroll plays with pseudo-words to make poetic sense, so long as the prosodic context is maintained. Using event-related-potentials (ERPs) in the brain, Pannekamp et al. (2005) found that prosodic, and not segmental cues are responsible for phrase-boundary detection in language.

Another interesting example is that of auctioneers’ speech. Studies done on auctioneers and their rapid speech reveal that they use a programmatic language subset, that requires rehearsal of fast utterances in order to be able to reproduce that speech quickly. Vowels and syllables fuse with each other, in a phenomenon referred to as *coarticulation* (Kent, 1977), which gives rise to a new speech that almost sounds intelligible but requires familiarity and training to comprehend. Even though the contents of their speech are rooted in the basic principles of the operating language (Kuiper and Haggio, 1984).

The above examples show how our perception of an utterance as ‘melody’ does not *require* that it belongs to a scale or tonality framework. We are capable of understanding and appreciating speech utterances as melodies, using contour profiles to guide us.

2.2.2 Intonation

In linguistics, *intonation* refers to three different levels of understanding phonological organization (Ladd, 2008, p. 1-6): *suprasegmental*, referring to pitch, stress, and quantity; *post-lexical*, which refers to pitches, whole phrases, or sentences; and *linguistically structured*, referring to how sentence- and phrase-level intonational features interact with the variable states of the speaker (for example, degree of arousal and so on). A four-level structure containing linguistic

segmental, linguistic suprasegmental, paralinguistic, and kinesic features is described in spoken communication (Wittmann, 1980). This is split into proximal and distal attributes. The study of ‘distal’ attributes of language, such as intonation, is often considered to be suprasegmental and above. This means that intonation conveys the meaning in a certain set of contexts, rather than words. For example, the way in which we understand that a statement is a question, even if we do not hear the words, is because of the intonation contours codified in the spoken contour even in the absence of a question word.

There is usually a consensus on intonation curves across native speakers of a language. Despite this, large variations in intonation are understood as dialects of any given language across its geographical spread. We even perceive pseudo-language, or imitations of language, as passable based on intonation contours. In a research article, Mehler and Dupoux (1992) found that babies as young as four days were able to distinguish between intonation patterns in their mother tongue and a foreign language. Mora suggests that “discourse intonation, the ordering of pitched sounds made by a human voice, is the first thing we learn when we are acquiring a language.” (Mora, 2000, p.149). Intonation, however, is not explicitly musical or a speech melody with a musical purpose. In essence, three properties are said to separate speech melodies from musical melodies (Patel, 2010):

1. Declination: The presence of fixed sentence-level contour structures for spoken languages, which are different for different languages.
2. Tonality: We perceive tonal relationships in most musical melodies, but not in speech melodies.
3. Diversity of linguistic intonation: Speech melodies may contain a larger number of ‘intervals’ than musical melodies.

Micro-inflections in intonation are essential to understanding emotional affect. Picking up on a friend’s mental state even before they have articulated it for themselves, questioning someone’s enthusiasm based on the tone of their affirmation, and so on, are just some examples. This topic has gained much traction lately, especially in the era of ‘smart’ voice assistants such as Amazon Echo, Google Home, and others. Computational analysis of affect perception in speech melodies will probably become more important in the future, in applications such as robotic caregiving.

2.2.3 Model for a Speech–Song Spectrum

I would like to propose a model for analysing a range of genres, and musical and speech forms, to understand the many forms of melodic and poetic utterances on a continuum between speech and melody in Figure 2.1. If we consider the extremes of the spectrum to be ‘full speech’ and ‘full song’, then several forms lie in between. I have tried to classify the forms more related to rhythmicity and prosody, such as poetic forms with a fixed number of syllables, several forms of

chanting that contain rules for syllable pronunciations, and perhaps an overall contour framework.

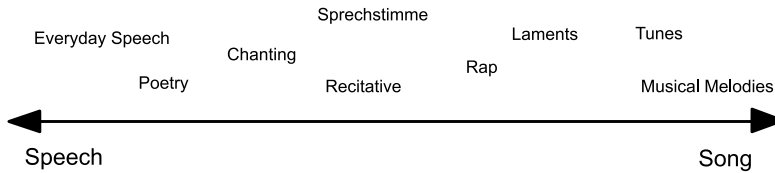


Figure 2.1: Speech–Song spectrum. In this figure, I have tried to represent the many forms between speech and song. Most categories and their place on this spectrum are not assigned without debate; this figure is only indicative.

Somewhere in the middle are the operatic forms, such as recitative, which is written as musical scores, but are dialogues by the characters, propelling the story, and Schönberg’s Sprechstimme. Then there are the more song-like rap melodies, with contour directions that must be obeyed. Rhythm alone does not propel rap music, but the contours of the phrases are also important to the style. Laments in some traditions are based on singing through pitch contours over and above ‘hitting pitches’ (Tolbert, 1990). To the very right are musical melodies. Melodies without text specially use the voice idiomatically as a musical instrument, are placed to the very right. The proposed model is not perfect, but it is the start of what we could think of as a spectrum from speech to song, and it is interesting to discuss how some phrases, upon repetition, move to the extreme ends of this spectrum.

2.3 Musical Melody

Melody is often described as the ‘salient’ or ‘hummable’ musical line, often found in higher pitch registers than the ‘rest’ of the music. Despite this operational definition, we hear melody in many sounds around us, such as birdsong, poetry, and repeated fragments of speech. It is interesting how melodies, with their pitched and rhythmic identities, are often described as lines; in turn, these lines are described as having contours, which means that they have an outline, or contour, that unfolds over time. I would like to stress, in particular, that I am mainly interested in the holistic nature of melodic perception. In discussions about melodic typologies, especially contour typologies, musicologists often focus on pitch classes and pitch relationships. However, linguists generally use fewer categories of pitch levels to describe linguistic intonation. So, how is musical melody studied?

In a chapter titled ‘Pitch and Pitch Structures’, in the book *Ecological Psychoacoustics*, Schmuckler (2004) endeavors to provide a framework for pitch

2. Melody

perception from an ecological perspective. This approach draws primarily from the work of J. J. Gibson from the 1950s, and has found multiple applications, particularly in explaining visual perception and movement. An approach with ecological psychoacoustics encourages us to study perceptual properties at the behavioral level. Schmuckler states that in order to adapt this approach into a meaningful study on pitch perception, we might focus our attention instead on the *apprehension of pitch objects*—perhaps study *melodic objects* and how we hear them. In this thesis, I find it important to study melodies ‘as they are heard’ from different cultures, and as they are vocalized or sung, instead of trying to reinterpret them as isochronous sequences.

Additionally, it is important to locate the experiments in this thesis at the right level in the hierarchy of musical importance. Many experiments on pitch perception take a bottom-up approach, studying the perception and cognition of single tones and the organization of scales in experimental conditions. However, as Schmuckler points out, “The alphabet (an alphabet of pitch materials to which the rules for creating well-formed patterns are applied), however, is often defined in terms of tonal sets, implicitly building tonality into the serial patterns” (Schmuckler, 2004, p.282). Although tonality determines the well-formedness of melodies, it certainly does not reflect the ways in which, for example, children play with melodies.

In general, melodies are understood as the salient, linear, and hummable monophonic abstractions of musical expression. In polyphonic and homophonic music, melody is often written as the topmost voice. The *Concise Oxford Dictionary of Music* defines melody as “A succession of notes, varying in pitch, which has an organized and recognizable shape.” (Kennedy et al., 2013). The key takeaways from this definition are the following words: 1. *pitch*, 2. *organized*, 3. *successive*, 4. *recognizable*, and 5. *shape*.

2.3.1 Pitch

Melody, in its everyday definition, is said to comprise pitches. Pitch has an everyday description—the “stuff music is made up of”—and is defined as being that attribute of music which can be ordered on a scale from low to high (Fuchs, 2010, p.71). Pitch and melody rely on each other for their definitions circularly. Using the words ‘low’ and ‘high’ to describe pitch sets the stage for considering pitch perception as directional, and moreover, as having spatial orientation. Many languages describe pitch in terms of dullness and brightness, suggesting that this perception of brightness corresponds to the periodicity, and by extension, the frequency of sounds (Shayan et al., 2011). Modeling pitch successfully requires an understanding of how pitch is perceived, and how it behaves in different musical contexts.

Frequency and Pitch Models

Frequency—mathematically defined as the number of repeating cycles of a regular signal; and pitch, are related but different. Pitch is a psychoacoustic component,

while frequency is a physical measure. Psychoacoustic components rely on perception to be realized. This means that pitch does not exist without a listener. Studies have revealed much about pitch perception, including the auditory perceptual scale for pitch discrimination, or just-noticeable-difference, which varies in the range of human hearing.

In and of itself, a pitch model can be a mathematical abstraction of pitch processing, a physical model of the hearing apparatus, or the description of neural firing in response to pitch stimuli. What we get from a model depends on what it is built for, and what we wish to obtain from it. De Cheveigne describes models and what they are useful for:

A very broad definition [of a model] is: a thing that represents another thing in some way that is useful. This definition also fits other words such as theory, map, analogue, metaphor, law, etc., ... “Useful” implies that the model represents its object faithfully, and yet is somehow easier to handle and thus distinct from its object. Norbert Wiener is quoted as saying: “The best material model of a cat is another, or preferably the same, cat.” I disagree: a cat is no easier to handle than itself, and thus not a useful model. Model and world must differ. (de Cheveigne, 2005, p.3)

Since pitch is a psychoacoustic (not a physical) phenomenon, some have argued that it is impossible to model pitch without a mind. However, models that closely approximate how pitch is abstracted are used for various applications. Most models for pitch estimation annotate absolute pitch, but as humans, we seem much better at approximating relative pitch.

Computational models of pitch perception rely on an understanding of the apparatus of pitch perception, which is curious in the case of auditory perception. Pitch perception remains stable despite missing fundamentals, or even when the bottom-most fundamental partial is missing from the spectral analysis. Pitch is perceived as stable over varying factors, such as amplitude, duration, spectra, and duration of stimulus. Melodies are also stable over a range of factors. Transpositions do not, for example, throw us off—we can identify melodic phrases in a wide range of transpositions. Moreover, melodic phrases remain unchanged when played on a range of instruments, and sometimes, distortions of scale and intonation do not disrupt the identification of melody. Lastly, we are able to recognize a melody as ‘the same’ across a large range of time variations. This means that we are able to perceive structural embellishments as external to melodic identity. Thus, we are seemingly able to construct a skeletal schema for melodic identity that is extremely robust.

Pitch Perception

We understand speech intonation, and melody through variations in pitch. The study of pitch perception, in trying to understand the local effects that melodic contexts have on pitch, incorporates a wide range of questions that encompass the breadth of our hearing spectrum. Being able to abstract a fundamental

2. Melody

pitch or an approximation of a fundamental frequency from a wide range of timbres and spectral shapes seems to be a unique property of human hearing. *Physiological models* of pitch deal with the shape and biological properties of the cochlea, and the coding of tonotopy in the cortex. *Algorithmic models* try to compute pitch using time- or spectrum-based signal processing methods. Pitch perception and melodic contour are closely related, but some differences are significant.

Melodic contour is clearly present in both music and language perception, but it is hard to find an inclusive definition of melody that applies to both language and music. Broadly, contour is defined in the same work as “a melody’s pattern of ups and downs of pitch over time without regard to exact interval size” (Patel, 2010, p.99). Experimental research has suggested that contours are a lower-level perceptual feature, in that we acquire it in early childhood (Trehub et al., 1984). This research also shows that infants are sensitive to directional changes in melodies. In 1994, Dowling et al. experimented with identification of unfamiliar or unknown melodies, to understand the role that contours play in their recognition. Participants in Dowling’s study also used contour and intonation distractors, which were similar stimuli to the target melody. It was reported that contour-distractors were more often confused to be the target melody than intonation-distractors. Early ethnographic research on melodic contour types focused on identifying contours in different musical cultures (Boer and Fischer, 2011), and mapping the frequency of contours in contour typologies.

Mysteries of melodic contours are relevant to this discussion. Contours are a coarse-level feature that we acquire very early in childhood is well known. Infant directed (ID) speech contains highly exaggerated contour profiles compared to adult directed (AD) speech. These experiments show that while ID and AD speech do not differ in prosodic shape, the contour profiles themselves contain a high level of emotional exaggeration in ID speech (Trainor et al., 2000). Acquiring coarse categories for prosodic meaning is important for the verbal, and by extension melodic, development of children. Melodic contours also play an important role in emotion detection in speech. In his research, Ross studied a case with clinical difficulty in processing affective speech prosody (Bell et al., 1990; Ross, 1993). Huron also argues for, on the one hand the co-occurrence of musical acuity and social development in genetic disorders such as Wilson’s disease; and on the other end, the connection between autism and proclivity to absolute pitch perception, and difficulty in ‘getting into’ music, is also observed. Melodic perception is essential, thus, to understanding affect in speech and music.

2.3.2 Organization

The organization of pitch and pitch structures is a key factor in determining the ‘musicality’ of pitched material. The organization of pitch structures includes, broadly, the following components: key, tonality, temperament, scale, and grammar.

Key and Tonality

A musical *key* refers to the perception of adherence of a piece of music to a single tone, around which the scale and other notes in the music seem to revolve. In tonal music, this single tone is the tonic, and the property of adherence is described as tonality. The tonic, or the key center also appears to be the stable pitch level; when there is a sense of ‘resolution’ in the music.

Tonality refers to the arrangement of pitches in a hierarchical order of perceived relationships and stabilities. It also refers to an understanding of a stable ‘key’. When a melody is constructed in a tonal framework, tonality dictates how and where a melody rests, and how its constituent notes are interrelated. A large amount of research done in this area is on the framework of phrasal grammars makes up. Many experiments on octave perception have tried to understand the perceptual organization of pitch classes.

Temperament

The intonational relationships between different tones in a melodic scale are referred to as temperament. Intonation deals with the ratios between different notes in an octave. In equal temperament for example, octaves are divided into 12 equal parts, so that every note is more or less equally out of tune. Cuddy (1982) designed experiments to identify interference of logarithmic and linear temperaments on contour perception in absolute and non-absolute pitch listeners. She found that contour and temperament affected the recognition of melodies, but that listeners apprehend contour even when they encounter unexpected intervals.

Scale

Scale refers to the arrangement of intervals in tonal melodies. In western music, major and minor scales are most commonly used; however, this is not representative of most music in the world, which features a large diversity of intonations and scales.

If the mode of presentation of a melody is changed—for example, *Happy Birthday* is played in a minor key—we can discern it as the same melodic sequence but with a different ‘flavor’. An iconic melody changes its identity more if the contour is dramatically different than if the scale or mode of presentation is altered. In the latter case, the melody retains its contours. Studies by Dowling (1978, 1972) show that contour is the principle factor for identifying melodies. Dowling also shows that inversions, retrogrades, and other melodic operations make the same melody harder to recognize, indicating that contour, or the time-unfolding properties of pitch, take precedence over scale in the recognition of melodies.

Grammar

Expectations of tonality rely on our exposure to certain musical grammars. ‘Probe tone’ experiments are often used to test tonal expectations; for instance, an incomplete melody is completed using a probe or question tone, and participant ratings help us understand how the question tone is perceived in the context of the melody (Tillmann et al., 2000; Tillmann and Bigand, 2010). Cross-cultural probe-tone experiments have helped reveal the extent of cultural learning for phrase completion (Curtis and Bharucha, 2009), while neurological activation studies involving probe tone experiments have also identified brain-areas tracking tonality (Janata et al., 2002). Some studies have also found that the tonic is not uniquely stable in major-mode melodies (Curtis and Bharucha, 2009; West and Fryer, 1990), and the dominant and subdominant were also rated equally high in some cases. Eerola et al. (2002) tested two sets of melodies with a total of 40 from two sources, with 27 isochronous sequences (Eerola et al., 2002). They selected factors from probe tone experiments that are known to influence melodic expectations. Isochronous melodies created using a generative model based on ‘typical transition probabilities’ were tested in the experiments asking participants for continuous ratings of the predictability of melodies.

To model phrasal grammars of western tonal music, several important ideas have been proposed, but most of these models are for a particular musical style, culture, or time period. For example, *Generative Theory of Tonal Music* (Lerdahl and Jackendoff, 1987) deals with a linguistic analysis approach to western tonal composition. Melodic grammars are also proposed for understanding specific composers or their work, such as for Bach chorales (Baroni and Jacoboni, 1978). Schenkerian analysis, originally to analyze tonal music has been modeled computationally modeled as context free grammar (Temperley, 2011). An influential model for melodic grammar that also incorporates ideas from music cognition as a whole is Narmour’s model of melodic expectancy (Narmour, 1992), also called the Implication Realization or the IR model. The core fundamentals of this model are that any melodic interval that is not perceived as closed, is an *implicative interval*, (Schellenberg et al., 2000, p.296), while between the following tone and the second tone after the implicative interval is the *realized interval*. The theory claims that these implications result from five perceptual predispositions that we learn from exposure to music: *registral direction*, *intervallic difference*, *registral return*, *proximity* and *closure*. As such, the Narmour (1992)’s IR model is the most generalized model of melodic grammar, that has been used to investigate cross-cultural melodic expectancy (Krumhansl et al., 2000; Pearce and Wiggins, 2006).

The aforementioned features: key, tonality, temperament, scale, and grammar are organizational features of melodies. The recall and recognizability of melodies is often studied through experiments in music psychology.

2.3.3 Recognizability

We recognize melodies that have a wide range of variabilities, but what is the basis of our recognition? It has been shown through research in psychology that scale and contour influence our memory of melodies. Explicitly tonal melodies are generally easier to remember than atonal melodies (Dowling, 1978; Vuvan and Schmuckler, 2011; Bod, 2002). Research also supports the enhancement of melodic memory when the stimuli are vocal (Weiss et al., 2012).

Dowling (1978) proposed a model to understand how melodies are stored in long- and short-term memory, using stimuli with scale and contour variations. The first component is the perceptual–motor schema of the musical scale. The second component, melodic contour, is shown to function independent of pitch interval sequences in memory. Dowling also underlines the importance of contour while repeating melodies from unfamiliar scales.

Melodic Identity

What do we call a recognizable melody? In the preceding sections, I have presented research studies on melodic contour that have primarily been conducted in the West, with western classical music as the main source material. While discussing these studies, I have explained how contour identity helps us recognize melodies, despite variations. But how much do melodic properties have to vary before the melody becomes unrecognizable as the original? It turns out that this depends upon musical style or musical culture. Cambouropoulos (2001) discusses this in relation to Quine’s observation about the identity of any object in a discourse. Quine states that objects that are indistinguishable from each other in a given discourse are identical for that discourse (Quine, 1950). Cambouropoulos, extending this discussion to melody states that a melodic phrase might be identical only to itself if pitch is most important in the musical context, if we imagine a theoretical context where no variation is accepted in melodic identity. But if instrumentation is most important, then the same pitches played on different instruments might not be recognized as the same melody.

I would like to define a *melodic entity* here as a melodic phrase that we can identify in repeated hearings, and recognize its belonging to a larger melodic framework; for instance, a song. I define *melodic framework* as a collection of melodies, including composition rules in some cases, that represent an identifiable style. A symphonic piece that features a thematic *melodic entity* might represent a framework. Other examples of melodic frameworks include grammatical arrangements of melodies, such as in a *raga* or *makam*; a tune family, such as those found in Irish folk music; or a style of improvisation, such as in an era of jazz. Elements of personal style can also be recognized as melodic frameworks, while other symbolic references may be attributed to melodic frameworks; for example, a general descending pattern might represent sadness in some contexts. A melodic entity could belong to one of many melodic frameworks, as I have illustrated in Figure 2.3.

I make a distinction here between melodic phrase and melodic entity. A

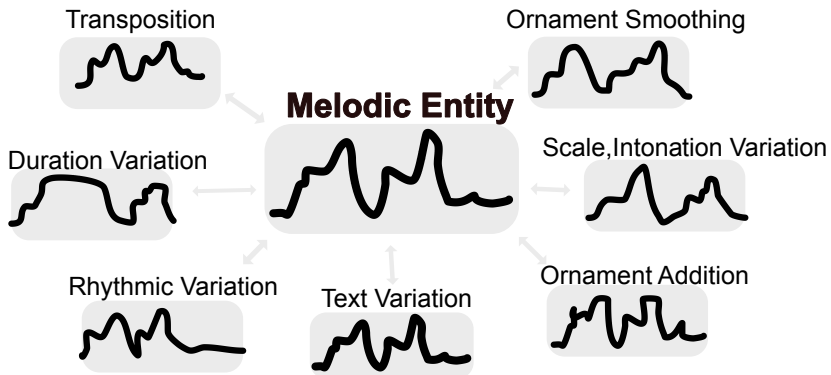


Figure 2.2: A melodic entity might be stable, as discussed above, to several kinds of variation. Which kinds of variations are most relevant depends upon the musical culture within which a similarity judgment is to be made.

melodic phrase can be defined as a melodic unit that has a self-contained quality. That is, a melodic phrase can be understood as independent. Every melodic phrase may or may not be a melodic entity, as I have described above, but this purely is related to the use of a phrase in a particular context, and not its structural properties.

In Figure 2.2, I elaborated upon variations that are tolerable in most musical cultures across repeated hearings of the same melodic entity. Most often, variations in the transposition of melodies and in the simplification of ornaments do not affect the recognition of a melodic entity. In some cases, melodies with lyrical variations might be treated as essentially the same. Melodic entities with changes in rhythm and duration may be treated as versions of the same. Finally, melodic entities may be treated as the same even with the addition of ornaments and embellishments.

Based on the assumption that melodic entities have acceptable variations, as described in Figure 2.2, we can understand which cultural or functional contexts tolerate which of these variations. I present some examples below. Even though practitioners often study the melodic contours of music through annotated notes on paper, the majority of musical cultures in the world rely on learning ‘tunes’ as a key part of the tradition. This means that melodic entities are subject to a large number of variations, and their integrity is largely determined by whether practicing musicians decide is or is not the same melody. James Cowdery, writing about Irish folk tunes, puts it succinctly:

“How should we characterize this entity? The problem is academic and not practical: a folk musician is content to call “The Blackbird” a tune. The scholar, however, realizing that all musicians play it

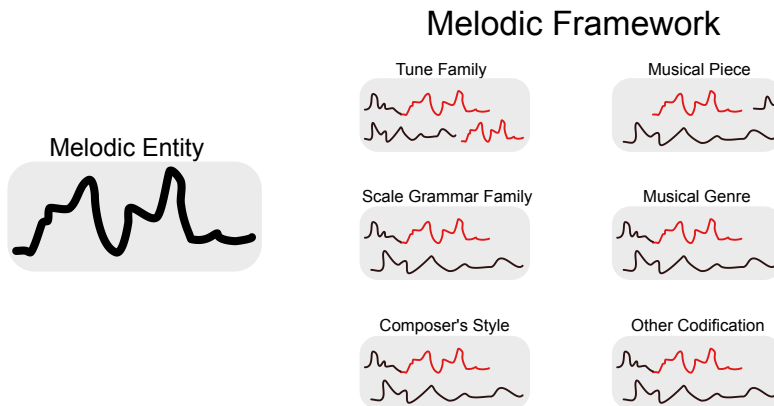


Figure 2.3: A melodic entity can be recognized as belonging to one of several melodic frameworks. A framework might mean different things according to the musical context, as illustrated.

slightly differently and that many never even play it identically twice, eventually comes to the same impasse that Gavin Greig described. Like him, we must ask, “Then where is the tune?”” (Cowdery, 1990, p.44)

While discussing tunes that are orally transmitted, Bronson (1951) notes variations in American folk music traditions. Other more computational models have also tried to describe variations in melodic entity, such as Bohak and Marolt (2009) calculating folk song variations, and Volk et al. (2007) using rhythms to identify various folk songs. Savage and Atkinson (2015) also analyze tune families using sequence alignment methods.

While a tune in orally transmitted folk melodies can have these types of variations on melodic entity, what happens if the melodic variation is restricted by the lyrical style? An example of this can be found in Cantonese Opera. Cantonese is a tonal language, which means that all vowels have pitch levels, or contours, that distinguish them from each other. Cantonese has six different tone shapes, that are categorized into two main families: ‘light’ and ‘dark’. Within these categories, there are three shapes for rising, falling, and flat tones. This affects the melodic formation of sung poetry in Cantonese, and has a direct effect on melodic contour (Yung, 1991).

Compositionally, Cantonese opera uses a limited number of ‘tunes’ to accompany different texts. The structural identities of these tunes are discussed in terms of melodic contour in books on Cantonese opera (Yung, 1989); Yung compares melodic and vowel contours, remarking that melodic contours in these arias often follow the contours of the vowels.

Contours are an important part of learning improvisation in the two largest

classical music traditions in India—Hindustani and Carnatic music. A framework called the *raga* framework in these traditions, provides scale material as well as melodic grammar to an improvising musician. There are hundreds of ragas that have individual scale and grammar combinations. Establishing a raga’s identity relies on the use of scales, which often have different notes in ascending and descending contours. Rules and ornamental features, that can be loosely described as a compositional grammar, must be remembered while presenting each individual raga. There are many instances of different ragas accommodated in the same scale, where the distinguishing factor is the unique melodic grammar.

Within the Hindustani tradition, there are four recognizable contour types mentioned in theory. Together, they are called *varnas*. The four contour types are ascending, descending, flat, and varying. Although descriptions of these contour types exist, they are usually included in the study of ragas as general guidelines rather than as formal analytical methods.

The accompanying hand gestures in improvised singing have been studied using different approaches, from anthropological to cognitive (Clayton and Leante, 2013; Pearson, 2016; Paschalidou et al., 2016; Kelkar, 2015). Studies reveal that contour properties and contour metaphors play a role in our cognition of melodic nuance, and in communicating this nuance during improvisation.

Melodic identity can be defined in various ways as described, and computational applications in which people evaluate melodic similarity usually are within one of the above contexts. For example, we might have a task to look for variations of folk tunes, the ground-truth data is a tune-label, while another task might be related to identifying a raga based on pitch contours.

Melodic Similarity

In their 2005 paper, Hoffman et al. review a large number of music information retrieval (MIR) related papers on melodic similarity. They classify the approaches into: 1. contrast models, 2. distance models, 3. dynamic programming, and 4. transition matrices.

A more recent paper by Cheng et al. (2018) evaluates parameters in a recurrent neural network (RNN) for melodic similarity analysis. The authors compare string edit distance, n-gram based measures, alignment-based methods, and RNN features using cosine distances for RNN parameters. The authors also note that changes in scale and duration do not impact melodic similarity. Instead, similarity measures appear to be invariant to musical transformations such as changes in pitch and tempo (Müllensiefen et al., 2004; Urbano et al., 2012). For this, the authors state that using symbolic music as opposed to audio signal might be simpler.

Dalla Bella et al. (2003) found that people are generally good at recognizing familiar melodies on hearing only a few notes. This happens regularly while listening to the radio or music from any unfamiliar source—we do not wait for the whole melody to unfold to identify it. The authors found that with just five to seven notes, both musicians and non-musicians are generally able to identify the correct melody.

2.3.4 Succession

How are we able to separate melodies into different phrases, and what defines a phrase? Several models have been proposed to automatically segment long melodic sequences into phrases. The idea of melodic succession relies on the separation of a stream of continuous melodies into constituent phrases. However, all these phrases are not remembered equally, nor are equally important, and we have several concepts to define and determine the phrases that matter more to us, such as a theme, a motif, an idea, a leitmotif, and so on.

White (1960) applied 12 types of distortions on familiar melodies to find out which of these distortions was most effective in obscuring melodic identity. Many of the distortions are common compositional tools used to expand and grow thematic material. White found that linear transformations were least likely to disrupt identification, while nonlinear transformations and temporal reversal were most disruptive. This indicates that succession is a core element of melodic cognition. He also found that melodies were just as easily identified from the first six notes as they were from the first 24 notes.

Frankland et al. (2004) present experiments that compare the empirical parsing of melodies predictions, derived from the four grouping preference rules of GTTM (generative theory of tonal music) (Lerdahl and Jackendoff, 1987). The experiment was to ask participants to annotate melodic segments, and they found within-subject consistency of boundary placements across three repetitions.

Pearce et al. (2010) present a model for a phrase segmentation task, and summarize the musicological and psychological background to the task, and review existing computational models. They propose a new model for phrase segmentation called IDyOM, based on statistical learning and information dynamics analysis. In another paper, Pearce et al. (2008) introduce a melodic segmentation model based on an information dynamics analysis of melodic structure. The performance of the model is compared to several existing algorithms that predict annotated phrase boundaries in large corpuses of folk music.

Other phrase segmentation models include: GTTM (Lerdahl and Jackendoff, 1987), the local boundary detection model (LBDM) (Cambouropoulos, 2001), Temperley's grouper model (Temperley, 1999), phrase structure preference rules (PSPRs) (Temperley, 2011), Tenney and Polansky's temporal Gestalt model (Tenney and Polansky, 1980), melodic density model by Ferrand et al. (2003), and Bod's supervised learning approach to analyze grouping in melodic boundary detection (Bod, 2002). IDyOM or Information Dynamics of Music model (Cambouropoulos, 2006) accounts for musical segmentation from a given melodic 'surface'. The beginning and ending points of prominent repeating musical patterns influence the segmentation of a musical surface; the discovered patterns are used to determine probable segmentation points in the melody.

2.3.5 Shape

The fifth defining attribute is melodic shape. I will discuss the phenomenological understanding of a shape later in Chapter 3. Melodic shape is often initially understood in the context of pitch direction. Henceforth, I will use shape to mean melodic contour. Melodic contour is described as the general shape of a melody or a melodic phrase (Kennedy et al., 2013). Some words that are used interchangeably with contour are *shape*, *configuration*, *outline*, or simply up–down movement. For instance, Adams (1976) uses melodic *configuration* and *outline* to mean melodic contour. In this thesis, I will refer to the shape properties of melodies as *melodic shape* and *melodic contour*. Some contour typologies defined in this work are compiled in Figure 2.4, based on previous research (Hood, 1982; Adams, 1976; Seeger, 1960; Schaeffer et al., 1967b; Vasant, 1998). It is evident that all these contour shapes are represented as lines that move primarily from left to right, and the direction of pitch is represented vertically.

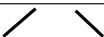

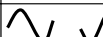
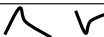


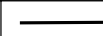







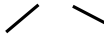


Seeger	xx	xy	xyy	xyx
				
Schaeffer	Impulsive	Iterative	Sustained	
				
Varna	Ascending	Descending	Stationary	Varying
				
Hood	Arch	Bow	Tooth	Diagonal
				
Adams	Repetition	Recurrence		
				

Figure 2.4: Range of melodic and sound contours from previous studies from Seeger (1960); Schaeffer et al. (1967a); Śarmā (2006); Hood (1982); Adams (1976)

Pitch and Verticality

Pitch is often discussed as rising and falling. However, the up–down metaphor for pitch is not present in every language. In fact, several languages use other binaries to represent pitch—bright–dark, thin–thick, small–large, light–heavy, and so on—as described by Shayan et al. (2011). Some studies have investigated cross-modal pitch correspondences by studying pre-verbal infants who tended to look longer at visual stimuli in which an object moving up corresponded with a sound that rose frequency (Walker et al., 2014).

Rusconi et al. (2006) proposed the SMARC effect (Spatial–Musical Association of Response Codes) to identify our tendency to associate high pitch and the upward direction together. This was investigated in a cross-comparison between English and Catalan speakers; in the experiment, the words for spatial elevation and pitch description were congruent and incongruent (Fernandez-Prieto et al., 2017). Results showed a longer response time for speakers when the word and spatial elevation direction were incongruent. Even though infants prefer congruent stimuli—that is, a rising pitch and a rising motion, and a falling pitch and a falling motion are more engaging—it takes longer developmentally to be able to discriminate between ascending and descending pitch stimuli. In work by Stalinski et al. (2008), people in five age groups (children aged 5, 6, 8, 11, and adults) were given three-note sequences to identify a difference where only the middle tone was higher or lower. Performance improved from the five- to the eight-year-olds, reaching the adult level at eight years. For all age groups, noting whether two pitch profiles were the same or different was easier than direction identification. The study also reports that, “Twenty-two percent of the children gave verbal or gestural responses to the examples that were considered correct. However, only 5 percent of the children used the traditional music terms “up and down” or “high and low” correctly.” (Hair, 1977).

These studies seem to show that although the association of pitch and vertical height might precede language acquisition, the embedding of pitch descriptors as vertical motion establishes this further. However, we must note that most of these studies were carried out using synthetic ‘pitch’, usually sine tones or MIDI notes, played on a piano. As such, matters can get complicated once we enter the domain of melodic phrases, especially those sung.

Phrases themselves are categorized in many languages and cultures as having inherent contour properties, which means that their internal pitch relationships are interpreted as shapes. As such, we may ignore some notes in favor of preserving structure. For example, a phrase with a long appoggiatura may simply be interpreted as bow shaped due to the weight of the last tonic landing, even if it is small in duration. On the other hand, a vibrato with a constantly changing pitch parameter of even up to a tone might be interpreted simply as a single note—a stable horizontal line-shaped contour. This means that in addition to pitch properties, some kind of a weighting of pitch values based on context is essential to understand the shapes of melodies.

Notation

In the earliest notation systems in western classical music, tiny squiggly lines above texts indicated the melodic contours that were to be sung to the text, in order to remember the melodies of the songs. Gradually, this is said to have evolved into an indicator line, which set the stage for modern notation as shown in Figure 2.5. This indicator line later developed to five several lines on the staff—each position on or between the lines is related to fixed pitches. Ornamental symbols are still notated using shapes; for example, the trill, mordant, glissando, all have shapes that look like representations of the contour shape.

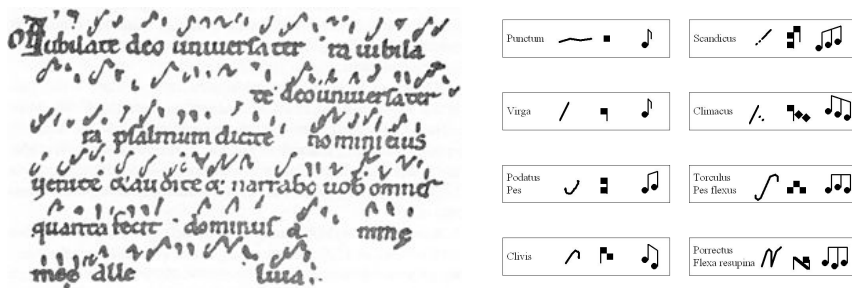


Figure 2.5: On the left, we see Palestrina's "Iubilare deo universa terra" psalm verses in neumes first published in 1593 in *Offertoria totius anni*, no. 14. It is argued that neumatic notation is derived from cheironomic hand gestures indicating changes in pitch. On the right, is an illustration of how neumes evolved into mensuration notation, and finally as modern notation.

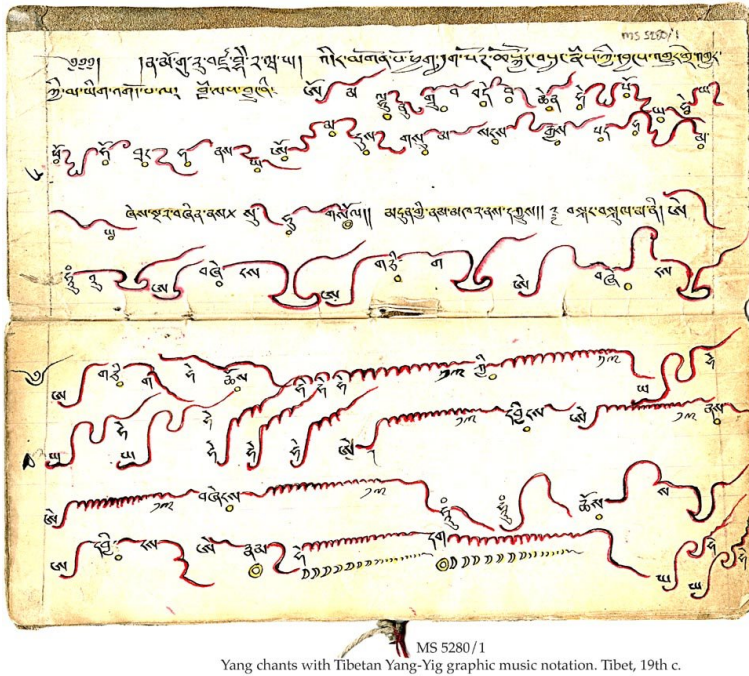
Modern notation with staff lines, and by extension MIDI-based notation, forms the majority of data sets for musical content retrieval and analysis. Contour analysis typologies are also based on notation, with discrete pitch and duration values. However, not all music can be transcribed into staff notation. Moreover, modern composers within the classical tradition have relied heavily on graphical notation to communicate musical ideas. My favorite example of this is John Cage's vocal piece, *Aria*, where the notation is just a collection of flowing colorful squiggly lines that are to be interpreted by a singer. We also find the use of contour-based notations in Tibetan chants, as shown in Figure 2.6 (Collection, 2019). Material in these chants is often rhythmically stretched.

2.4 Other Aspects of Melody

So far, I have explored the idea of melody by its defining aspects of pitch, organization, recognizability, succession, and shape. There are other aspects that influence melodic perception that I would also like to highlight. Primarily, this includes the special place of vocal melodies, and vocal learning in the cognition of melodies.

2.4.1 Voice and Melody

The definition of melody centers on our ability to sing—in essence, the ability to collapse a large quantity of acoustic and musical information into one hummable line. When the melodic line is clearly differentiated from other instruments, this task is trivial. But there are several musical cases where the melodic line is shared among many instruments, two or more melodic lines exist, or the timbre of the melody unclear and so on.



Yang chants with Tibetan Yang-Yig graphic music notation. Tibet, 19th c.

Figure 2.6: Tibetan Yang notation from silkscreen prints, ca. nineteenth century. This form of notation goes back to the sixth century (Collection, 2019)

An early study observing spontaneous singing by children was conducted by (Moorhead and Pond, 1941, p.2). Some of their main findings are interesting to begin this discussion with:

“(a) Experimentation with vocalizations, and with songs was the norm and that children do not organize their music in conventional tonalities as adults do, (b) children’s songs were plaintive compared to the songs adults would have them sing, (c) chant appeared under conditions of freedom-alone or in a group, (d) physical activity was directly related to rhythmic chant, (e) solo chants were like heightened speech in which the music conformed to the words, group chants were in duple meter and the words conformed to the rhythmic structure, (f) children explored instruments first before melodic and rhythmic patterns emerged, (g) instrumental improvisation was characterized by asymmetrical meter, followed by duple and triple meters, then steady beat.”

Comparison studies of spontaneous singing in children are hard to carry out because of various differences between children’s musical and verbal abilities, regardless of age. Spontaneous songs are also harder to keep track of and record.

2. Melody

However, human beings need exposure, play, and experimentation to develop their vocal apparatus. This is explained by the vocal learning hypothesis.

Vocal Learning Hypothesis

The main claim of the vocal learning hypothesis is that human beings (as well as songbirds) learn from a process of vocal imitation, relying immensely on auditory feedback. This is in contrast with contextual learning in vocal communication, where we learn to identify or react to sounds as a result of our experiences with them. Early vocalizations have little to do with mature song or speech, but form the building blocks of language cognition and musicality (Hansen, 1979).

Research has also demonstrated that we are better at remembering vocal melodies than instrumental melodies (Trehub et al., 1984; Weiss et al., 2012, 2015). This hypothesis has been tested time and again with speakers of different languages; it has been replicated across age groups, levels of expertise, and so on. This implies that the voice is functionally and biologically significant. Studies also suggest that there is greater pupil dilation in response to vocal stimuli, showing increased attention or arousal (Weiss et al., 2016).

Yet another interesting aspect of the voice is the margin of error we tolerate in the precision of vocal intonation. Dubbed as the ‘vocal generosity’ effect, it describes the tolerance of intonation differences in comparison with the tolerance for other instruments (Hutchins et al., 2012). Perhaps the note positions in exaggerated vibratos are a by-product of such an effect.

Subvocal Rehearsal

Subvocal rehearsal, or the vocal imitation of speech and sound, is an important aspect of the auditory working memory. Studies have shown that suppressing subvocal rehearsal directly affects our capacity to remember sequences (Rizzolatti and Craighero, 2004; Jacquemot and Scott, 2006; Kohler et al., 2002). Recent developments in electromyography have also enabled us to build systems to recognize subvocal speech (Jorgensen and Binsted, 2005). I offer more on this in the section on auditory working memory in Chapter 4.

Birdsong

The study of birdsong changed in the 1950s, after World War II, when the sound spectrograph became available (Marler and Slabbekoorn, 2004, p. 1). Since then, it has become possible to represent the fundamental frequencies in birdsong using ‘sonograms’. One reason that birdsong is relevant to music and melodic cognition is because both birds and humans are ‘vocal learners’. This means that songbirds require auditory input to learn ‘normal’ birdsong.

There is, however, a critical difference. In a 2016 paper, (Bregman et al., 2016) measured the ability of starlings to recognize pitch sequences that remove spectral structure. Humans are generally good at recognizing transpositions of melodies as instances of the same melodies. However, this paper finds that spectral shapes,

rather than melodic contours, predicted whether starlings recognized melodies. Shannon (2016) argues that this indicates that birdsong is more analogous to speech perception rather than music, although it is clear that humans and birds may have evolved to develop an acuity for melodic perception, albeit in different ways.

2.5 Discussion

Several studies have remarked on the nature of signification in melodic contour in a number of song styles. Mangaoang et al. (2018) write, in their analysis of Philippine disco-opera, how the contours of the heroine's song are constructed to remind one of the Disney princess songs. Tolbert (1990) also writes about Karelian lament songs, and talks about descending, terraced contours as being a contour-motivic device to communicate grief in these songs. Researchers have proposed and demonstrated the links between music perception and the body time and again (Leman, 2008; Godøy, 2003; Gritten and King, 2006, 2011; Godøy, 2010; Jensenius et al., 2006), as I will discuss in the following chapters.

In the sections above, I have established that contours are a feature unique to melodies in the following ways:

- it is important for us to remember and index melodies
- it is a low-level feature acquired early in childhood
- it is closely related to prosody and speech–affect comprehension, and
- contour stability is closely related to the perception of melodic stability.

Perception of pitch and melody are highly interrelated, and I want to understand melodic contour and its perception through the body in this thesis. I think a systematic study of this can shed light on cross-modal mechanisms, and help us understand pitch perception in the context of melody. The objective is, thus, to focus on the perception of dynamic pitch. In this chapter, I hope to have established that melodic perception is the ecological site of studying dynamic pitch perception. As such, I have demonstrated some peculiarities of melodic cognition to suggest that melodies are understood as holistic entities.

2.6 Summary

In this chapter, I have presented a review of melodies as speech, and then musical melodies. I have described their essential properties, and how melodic contour plays an important role in melodic recognition and identification in both speech and music; although studying speech contours and musical contours typically has different approaches. Identifiable melodies in speech and music laid out on a spectrum demonstrate the various levels between the binary of speech and musical melodies. Melodic entities, and their belongingness to various melodic 'families' is culturally determined, as a result of which, their similarity and

2. Melody

categorization also have various flavors. Vocal melodies are special, and melodic character is contingent on the ‘humability’ of melodic lines. Finally, I have highlighted how acknowledging the cross-model imagery of contour as shape can contribute to our understanding of melodic perception. In the next chapter, I will discuss auditory and motor imagery, going further into this cross-modal connection.

Chapter 3

Auditory and Motor Imagery

*Shape without form, shade without colour,
Paralysed force, gesture without motion;*

...
*Between the motion
And the act
Falls the Shadow*

- *The Hollow Men*, T. S. Elliot, 1925

3.1 Introduction

In *The Hollow Men*, one of the most famous poems of the twentieth century, T. S. Elliot emphasizes the gaps between the ideas of shape and form, gesture and motion, idea and reality, and motion and action, saying that between these ideas lies uncertainty. Having shape and gesture without form and motion, lacking intentionality, becomes a characteristic of the 'hollow' men. According to the poem, it is the act and quality of movement that defines one's "human-ness", or how much of a person one is.

The central aim of this thesis is to explore the multimodal relationship between melody and motor imagery, by discussing the cognition of melodic contour as a trace. In order to discuss the various aspects of this relationship, I would like to first discuss how humans perceives shapes within the visual and motor modalities, and why the notion of shape is essential to cognition.

Imagery, or the reproduction of multimodal sensations in the mind in the absence of stimuli, forms the basis for the theoretical framework used in this chapter. Psychologists have been studying mental imagery from the time of the Ancient Greeks to the Modernists (Kosslyn et al., 2006, p.4), although much of the research in this area has dealt with visual imagery rather than other modalities. In this chapter, I will focus my attention on auditory and motor imagery, and establish the context for auditory and motor cognition within the phenomenology of imagery.

3.1.1 Terminology

The following terms occur frequently while discussing visual and motor imagery in relation to contours:

1. *Line*: Contrary to the notion of a line in Cartesian geometry, where 'line' refers only to straight lines, line here simply refers to a continuous path of points. In this sense, a melodic line refers to the continuous arrangement of notes in succession.

3. Auditory and Motor Imagery

2. *Shape*: Shape generally refers to a geometric arrangement—an autonomous entity accompanied by a mathematical definition. However, when referring to melodic shape, I allude more to the shape-like properties of individual melodies, as invoked in a geometrical sense.
3. *Trajectory*: In the context of motion capture studies, trajectory in this thesis refers to the path that a particular marker traces in space. As such, it is different from a line because it refers to the actual motion history of a moving agent, while a line refers more to the intended trajectory.
4. *Contour*: The contour properties of melodies include the visual and gestural images invoked by directional features.

3.2 Auditory Imagery

It is a very common experience to be able to hear a tune in one's head, or to hear somebody's voice in one's head. Just as the 'mind's eye' is said to represent visual imagery, what we see when we are imagining something visually, the 'mind's ear' is a term used for auditory imagery, and when applicable, *musical imagery* (Godøy and Leman, 2010). Many studies have demonstrated the activation of the auditory cortex when the mind is engaged in auditory imagery (Zatorre and Halpern, 2005). Recalling melodic excerpts engages the auditory cortex in a similar way, as has been demonstrated earliest through PET scans of the brain (Halpern and Zatorre, 1999).

Contour representations and auditory imagery have been explored in experimental studies on the recall of melody (Reisberg, 2014, p.2). The debates on mental imagery revolve around the nature of internal representations of melodies, and how these representations are accessible to us.

3.2.1 Auditory Scene Analysis

It is often said that unlike the eyes, the ears have no lids. As such, we are constantly hearing what is happening in our environment; we can locate ourselves in space; and we understand the density of moving objects around us, the sounds they create, and their material properties. We can focus auditory attention on a specific stream within the spectrum of sounds in our hearing. In his book, *Auditory Scene Analysis*, Bregman (1994) attempts to explain how we segregate an 'auditory scene' (unfiltered auditory information that we hear at any time) into separate 'auditory streams'. For example, we are able to segregate the sound of someone speaking from the noise of a fan in the room. Further, we can distinguish a known voice from a cacophony of many simultaneous speakers. At the same instant, we can also identify the speaker, and perhaps even their temporary physiological state; for example, we know from their voice if they are joyful, or are tired. Bregman's analysis is about the auditory stream, and not explicitly about musicality, but it illuminates analysis of musical hearing as an auditory stream.

3.2.2 Ecological Pitch Perception

To address pitch perception from an ecological perspective, we have to question as to what end contour memory might have been selected for. What advantages does it give us? The way we perceive a fundamental pitch in a complex sound appears related to how the partials in a sound are distributed in terms of loudness. When we hear strong or loud harmonics, we tend to recognize the lowest loudest partial as the fundamental. A musical illusion called ‘the missing fundamental’ shows how we perceive a fundamental when the harmonics are strong, even if the fundamental partial is removed from the spectrum—our cognition compensates for this missing element.

Isolated tone perception and melodic perception are to be understood as separate because of many temporal and perceptual factors. First, upon rapid presentation, individual tones seem to fuse together to form iconic melodic motifs that we remember much better than series of tones whose succession is artificially broken or disturbed. As de Cheveigne writes, absolute pitch in humans is rare; although it can be acquired with a lot of training. The computationally harder and more abstract task of analyzing relative pitch, however, is actually a natural ability. We are naturally good at relative pitch perception, and interval perception. However, computational models of pitch detection based on relative pitch need to be developed (de Cheveigne, 2005).

Melody relies on pitch as the main variational material, or changes in pitch over time are primarily what makes a melody. Even though this is true, we perceive a plethora of features from melodic stimuli that influence pitch perception. For example, our perception of pitch is different in sounds with simultaneously increasing pitch and loudness, and where this relationship is inverse. However, it is noteworthy that in debrief interviews of experiments asking participants to trace melodies, participants claim to be tracing pitch, and do not mention timbre, dynamic envelopes and other sound features. However this is a circular argument. If we accept pitch as the aspect in melodic music that moves, then we have to rethink how pitch extraction algorithms tend to model the fundamental frequency as pitch.

3.2.3 Conceptual Shapes and Music Theory

Godøy (1999) suggests that thinking of musical sounds as shapes is a valuable way to develop a holistic approach to the analysis of musical sounds. He stresses the importance of holistic approaches to musical sound and musical cognition which is more than traditional music theory offers. Godøy defines conceptual shapes as the spectral distribution and temporal evolution of various aspects of a musical object.

In ‘Knowledge in music theory by shapes of musical objects and sound-producing actions’, Godøy (1997, p. 90) discusses the notion of shapes in musical objects, and how this idea of shape could be used to understand the holistic perception of musical objects. Here, shape is invoked as in common parlance; the drawing of dynamic envelopes is one example, as are wave forms used to represent

3. Auditory and Motor Imagery

different sounds. Godøy suggests that thinking in shapes is already the everyday working method of accessing useful categories in various domains; within sound and music, for example, there are sound envelopes, topological shapes such as helices, and graphical analyses of compositional structures. Melodic shapes, as has been explained in the previous chapter, are ubiquitous too, and could be explored analysis-by-synthesis (Godøy and Leman, 2010, p.243).

This thesis examines melody as the site of execution of our cognition of contour. Speech intonation is another area where our cognition of contour is useful to us—it aids our understanding of languages, dialects, affect, and so on. As such, melodic music and melodies are not primarily used in research on the cognition of linguistic contours.

3.3 Gestalt and Shape

Mathematician and phenomenologist René Thom writes, “the first objective is to characterize a phenomenon as a form, as a ‘spatial’ form. To understand means then, first of all, to geometrize” (Thom, 1983, p.6). Thus, we first present most models and architectures geometrically, to give them spatial meaning. To arrange them in relation to one another is to clarify the conceptualization of the model itself, whether it relates to abstract concept visualization or more concrete and data-driven methods. This is perhaps why we draw figures and conceptual maps to elaborate on concepts in conversation, and most analytical models have a visual or a geometric component.

3.3.1 Melodic Gestalt

Christian von Ehrenfels, one of the founders of Gestalt theory, uses melodies as a starting point to illustrate how Gestalt perception might function, as (Leman and Schneider, 1997, p. 46) has described. Ehrenfels writes:

Let us suppose, on the one hand, that the series of tones $t_1, t_2, t_3, \dots, t_n$ on being sounded, is apprehended by a conscious subject S as a tonal Gestalt (so that the memory-images of all the tones are simultaneously present to him); and let us suppose also that the sum of these n tones, each with its particular temporal determination, is brought to presentation by n unities of consciousness in such a way that each of these n individuals has in his consciousness only one single tone-presentation. Then the question arises whether the consciousness S , in apprehending the melody, brings more to his presentation than the n distinct individuals taken together. (Von Ehrenfels, 1988, p.85)

Gibson (1960, p.698), while elaborating upon ecological psychology, also refers to the use of melodic stimuli in psychology experiments :

“The Gestalt psychologists pointed out that a melody is perceived, but they never suggested that a melody was a stimulus. The notes

of the melody were taken to be the stimuli. But what about the transitions between notes, or the ‘transients’ of acoustical engineering? Are they stimuli? The investigators of speech sounds seem to think so, but the auditory literature of sensation is vague on this question. And if a short transition is a stimulus, why not a long transition or temporal pattern?”

In the Gestalt framework, continuity plays a big role in perception. This means that we are likely to perceive as salient, sensory events or objects that have the semblance of continuous motion, rather than those that jump haphazardly from state to state.

Bod (2002) argues with empirical evidence that Gestalt principles might not be sufficient to propose a comprehensive theory of memory, and that several of these principles are flouted when tested for segmentation, or grouping boundaries. For example, they find that segmentation boundaries can exist even when there is no note change. Further, boundaries can often also appear before or after large pitch intervals than right at those intervals, suggesting that large pitch intervals do not necessarily indicate phrasal shifts.

3.3.2 Contour Continuity and Gestalt

When our visual system encounters gaps along the edges of incomplete shapes, it is able to fill in the gaps, perceiving subjective contours of objects that are absent in the details of the stimulus.

Mathematical models of ‘mental gap-filling’, or contour perception of incomplete shape patterns, are modeled on the following operative principles:

1. Isotropy: The contours that are filled in are insensitive to rotation, translation, and scaling of the figures.
2. Smoothness: Except for corners, the contour apprehension is smooth, or differentiable at least once.
3. Minimum curvature: The filled in contours have a minimum curvature.
4. Locality: The filled-in contours are operative only to edge shapes, and not entire edges in cases of slight deviations.

An illustration of some of these concepts is made in Figure 3.1. Models like this can help us understand how contour filling in the mind’s eye validates Gestalt principles. The models explained here clearly refer to actual and not metaphorical shapes. However, similar principles should hold true for melodic contour as well.

I wish to make a distinction between the geometric and motion aspects of melodic memory and imagery on the basis of the extent and types of distortions that we are able to tolerate when it comes to melodic objects. For example, so long as the inter-event durations within a melody remain the same, we are able to retain the stability of a melodic object even when the tempo is dramatically

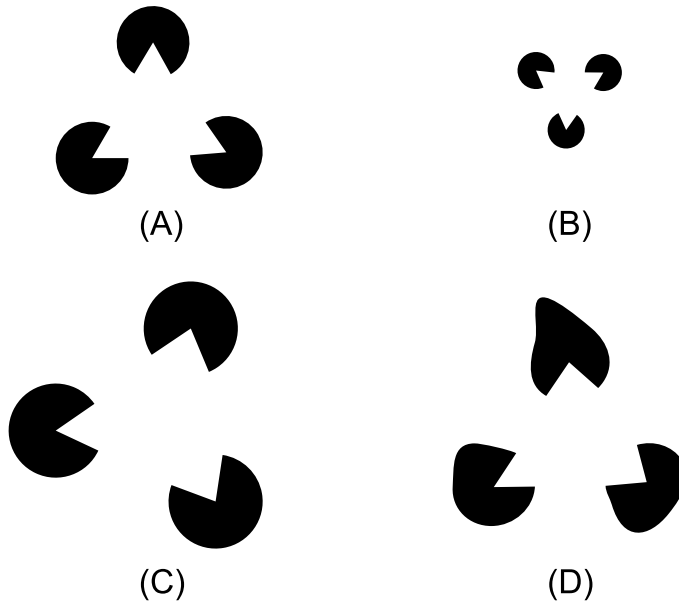


Figure 3.1: A and B represent the property of isotropy in this illusion; the triangle is perceived despite scaling and rotation. In C and D, with the change of an angle, we perceive the third edge of the triangle to be slightly curved. This alludes to the properties of smoothness and locality. The lines at the corners are still straight lines, but the effect is of a curved edge instead.

different. Further, so long as the relative contour remains constant, we are able to recognize a melody even if it is transposed heavily. Contour properties are thus resilient to temporal compression, expansion, and spectral change in our minds.

In his early writings on shape and *morphe*, Aristotle discusses grouping the same ‘types’ of objects by their apparent shapes. In Aristotle’s phenomenology of shape, these groupings are used as a measure of salience, and as a way to discriminate between different categories by appearance (Long, 2007). Aristotle distinguishes between the shapes of objects and their formative material, or essence; as two distinct ideas. ‘Dogness’, ‘catness’, and other properties of living beings are also invoked primarily by their external shapes.

3.4 Motor Imagery

Mental imagery is largely described as recreations of mental or sensory experiences in the mind, without external stimuli; for example, imagining someone’s face, or being able to ‘hear’ a piece of music you like in your ‘mind’s ear’. There

have been debates in the cognitive sciences about whether mental imagery is representational or purely experiential for many years. However, neural imaging studies have successfully demonstrated that we engage the same parts of the brain involved in the perception of a stimulus, and imagining the same stimulus. In other words, when we hear a song in the mind's ear, we engage the auditory cortex. Enactive theories of cognition explain perception as something an organism is actively 'doing', instead of being merely passive (Clarke, 2001). Motor imagery, referring to the mental representations of movement, could be considered necessary to planning and execution of actions (Jeannerod, 1995).

Visual and motor imagery are thus related. On registering visual information across time, we get a sense of the process of creation of the objects, and their affordances. Our ability to respond to visual stimuli, in part, uses this ability of ours according to this view.

3.4.1 Phenomenology of Shape

In common parlance, when referring to shapes, we mean circles, triangles, rectangles, and other common geometric figures that we learn about as children, as the building blocks of our first experiments with geometric abstraction. In geometry, shapes have definite forms that are describable by mathematical rules. However, when we say something is circular, for instance, we do not usually mean a perfect circle. Thus, our own perceptions of mathematically strict shapes can be adaptable to varying contexts. For example, in everyday parlance, every instance of drawing an imperfect circle, is understood as being the same shape.

The stability of shapes in our minds is explained through the study of *morphogenesis*. In (Bourguin and Lesne, 2010, p.2), the authors emphasize that to understand a shape requires an understanding of its formation or morphogenesis. In this view, such things as fractals, sand dunes, and waves are also shapes, in that they have well understood geometric forms and morphogeneses. In other words, if we understand how something is formed, we can tolerate variations in that shape. Our perception of human and nonhuman shapes might also be different. Natural formations, such as sand dunes or waves in the ocean, might be perceived more as 'living shapes', than man-made formations, such as cubes. Gestalt psychologists use melody as an example to explain what we think of as shape and contour. I explain this more in Chapter 4.

An accessible example to illustrate this point is the use of diagrams in mathematics and computer science. The virtual or unseen is often represented, especially in mathematical education, through movement metaphors and abstract representations of mathematical objects. In studies involving mathematics and gesture, authors have pointed to aspects of virtual shape apprehension that would be impossible without understanding the geometric relationships between objects, and movements that describe the morphogenesis of abstract visual geometry.

The connections between the spatial representations, illustrated by gesturing while teaching mathematics, have been written about in detail by Châtelet who remarks, 'The virtual requires the gesturing hand' (Châtelet, 1993, p.15).

3. Auditory and Motor Imagery

Châtelet insists that the gesture and the diagram are incomplete without each other, and that they each participate in the provisional ontology of the other (De Freitas and Sinclair, 2012). Châtelet suggests that diagrams are ‘physico-mathematical’ beings.

In this discussion, this notional, phenomenological shape is of interest to us in the context of the experimental work that has been conducted in this thesis. The idea of musical shape is multimodal, but quite accessible, as I have demonstrated in the chapter on melody. However, understanding how this shape is executed as an action merits a discussion on the phenomenology of shape abstraction and the perception of movement within static, abstract visuo-gestural objects.

3.4.2 Neural Correlates of Phenomenology of Shape

Human beings live in dynamic environments and are in constant motion. To receive and process visual information, our eyes make tiny saccadic movements several times in a second, and our heads turn. Despite the jerky eye movements, we can successfully integrate images of objects into stable entities, and apprehend the motion properties of objects. Another property of shape perception and imagery is our ability to scale, rotate, and transform objects. According to Biederman (2013), object invariance, also referred to as *shape constancy*, is the most striking feature of shape perception.

Shape constancy is modeled using shape percepts, or *geons*, that are proto-shapes or shape-images; we are able to break down objects that we visually perceive into a finite number of simpler shapes Biederman (1987). In his original theory, Biederman tries to understand object categorization through the process of geometrical deconstruction in our visuo-motor system. He proposes a set of volumetric geons (*geometric ions*) that serve as non-accidental primitives for shape perception, and by extension, object perception.

Shape-based object representation, and its relation to visual perception, has been the focus of several benchmark studies (Kobatake and Tanaka, 1994). These findings have been modeled using neural networks in other studies (Hummel and Biederman, 1992). In a review paper, Biederman (2013) describes theories of shape-based object representation in the *inferior temporal cortex* and the *lateral occipital complex*. Lesion and functional Magnetic Resonance Imaging (fMRI) studies of the brain have determined that these two areas exhibit a high degree of invariance to geon representation.

3.4.3 Lines and Movement

How do perceived shapes in musical objects relate to the phenomenology of shape perception itself? Lines that we observe in static images, also seem to have a direction of movement. They seem to emerge from somewhere, and go in another direction. Even when not explicitly specified, through, for example, arrowheads, we tend to think of lines as traversing different visual spaces. Wallach (1935) noted, that without context, the motion direction of a line itself is ambiguous. It’s also important to note that lines, in this context doesn’t refer to the notion of

lines in Euclidean geometry, but lines as we recognize them in visual perception, which is lines that we perceive to pass between objects. But in the presence of context, the motion direction of a line is determined by the shape of the aperture that it traverses. A simple example of this is the *barberpole illusion*, where diagonal lines appear to move perpendicularly due to the rotation of the barberpole, either upwards or downwards. This has been later experimentally analyzed by Wuerger et al. (1996).

If we were perceiving melodies as lines, the aperture, or canvases upon which we imagine them have a bearing on our perception on the direction of motion, and therefore would change the way in which we trace the perceived melodic lines themselves. This is why understanding shape perception and shape phenomenology is critical in trying to understand sound tracings. There has not been a lot of literature that deals with this explicitly in the context of sound-tracings.

The experiments conducted as a part of this thesis rely on cross-modal conversions between melodic lines and movement. This movement is different from rhythmic musical entrainment in that the synchrony of gestural imagery is less obvious. In the papers that comprise this thesis, we see that people have different associations and strategies to respond to sound with movement.

3.5 Summary

In this chapter, I have elaborated on the principles of auditory and motor imagery, which is our ability of being able to see and hear ‘in our minds’. Imagery forms one of the foundational concepts behind picturing melodic contours as lines, and being able to execute the line as movement. Our experience in the world, with overlapping stimuli, and intertwined visual and auditory scenes is discussed in Ecological Psychology, which is an important theoretical model for this thesis. I outlined how melody as shape is a foundational concept in gestalt psychology. Recent work on motor-mimesis and gestural sensations in audition are also discussed. Several theories for shape perception are outlined, and the cognitive dimensions of shape-perception from a phenomenological perspective are explained. The experiments in the thesis lean on the methodology of ‘sound-tracing’, which starts the next chapter, proceeding to a discussion of body movement, and how embodiment informs music cognition.

Chapter 4

Body Movement

*... that mind-body split, you know?
The head is good, body bad. Head is ego, body id.
When we say "I," - as when Rene Descartes said,
"I think therefore I am," - we mean the head.
And as David Lee Roth sang ..."I ain't got no body."
- Levine (2002)*

4.1 Introduction

The mind-body split is a problem that has long engaged philosophers and researchers interested in consciousness. In the traditionalist view of cognitive science cognition is seen as something that happens in the mind, while action happens in the body. In this we take embodied cognition as the point of departure, which takes the body with its active and perceptual capacities as the starting point. In such a perspective, music listening is also grounded in *behavior*.

Drawing upon this perspective, Clarke's early work on embodied cognition in music demonstrates the motivations to consider motion as an important aspect of music perception (Clarke, 2001). In recent years, several important works on embodied music cognition, both empirical, and theoretical have been published as we will see in this chapter.

This thesis combines embodied cognition and music perception, with analysis techniques borrowed from music information retrieval, biomechanical control, and interactive music interfaces. In this chapter, I talk about the key frameworks in my study of human body motion and its fundamental role in cognition in general, and speech and music cognition in particular. These include key frameworks in ecological psychology, auditory scene analysis, embodied music cognition, and gestural imagery related to musical movement. The core methodology used in the experiments for my work are rooted in sound-tracing, and so I will begin this chapter discussing it first.

4.2 Sound-Tracing

Sound-tracing is an experimental paradigm that is used to investigate spatial representations of sound. In a way, our perceptual system acts as a transducer to 'translate' between two modalities. Sound-tracing studies analyze participants' spontaneous renderings of melodies to movement, capturing their instantaneous multimodal associations. Typically, participants are asked to "draw" (or "trace") a short musical excerpt in air as they listen to it. Several studies have been carried out using digital tablets as the transducer or medium of recording the

4. Body Movement

data (Godøy et al., 2006; Glette et al., 2010; Godøy et al., 2005; Jensenius, 2007; Küssner, 2013; Roy et al., 2014; Kelkar, 2015). One restriction associated with using tablets is that the size of the rendering space is limited. Furthermore, as the dimensionality does not evolve over time, it represents a narrow bandwidth of possible movements.

Melodic sound-tracings, or manual renderings of sound-as-shape is not a task that is taught or rehearsed. A theoretically robust model of sound-tracing must consider the cognition of a melodic entity as a whole. By applying metaphors of motion, on a purely linguistic level, we often ascribe physical object properties to sonic features. Roughness or grain, for example, is the textural property of a rigid body, which is available to us from our experiences of touching rough objects. As an existing metaphor, it is readily transferable to our perception of texture in a sound object.

An alternative to tablet-based sound-tracing is full-body motion capture. This may be seen as a variation of ‘air performance’ studies, in which participants try to imitate sound-producing actions of the music they hear (Godøy et al., 2005). Nymoen et al. (2011) carried out a series of sound-tracing studies focusing on hand movements across many studies (Nymoen et al., 2013), elaborating several feature extraction methods to be used within the larger sound-tracing methodology.

A substantial amount of work in music perception and embodiment concentrates on rhythmic entrainment, tapping, and other time-synchronous responses. Approaches to melodic shape cognition that take an explicitly embodied perspective are fewer. Küssner (2014), investigates many properties of sound-tracings alongside their musical attributes. A significant takeaway from this work is the differences in representation between musicians and non-musicians’s tracings with modeling hyperparameters. Non-linearity of mapping was also reported in the tracings. Participants mapped an attribute to a physical axis, but they also end up representing feelings or other characteristics that the music invokes.

The experiments conducted for this thesis deal with motion capture data from participants, who trace the shapes of short melodies while listening to them for the second time. This experimental paradigm of sound-tracing is a development based on several theoretical constructs regarding body movement, and the relationships between music and action.

In my view, the interactions here are between audition, movement, and geometry. That pitched melodies can be thought of as lines is the interaction of audition and geometry; however, their execution through movements of the hands and body represents an interaction of geometry and movement with time—it involves manual control and action planning. In essence, sound-tracing involves translating the perception of auditory stimuli into an action that has its own geometry, and sometimes even a recognizable shape.

An action or a series of actions performed as spontaneous rendering of sounds can be described as tracings. We suppose that these movements reflect music-related gestural sonic imagery, and try to understand how the dynamics of control can influence motor action in sound-tracings. By definition, sound-tracing is

a many-to-one mapping methodology; mapping parameters can vary and are regulated or defined by several occurrences in the music.

4.3 Embodied Music Cognition

Embodied music cognition is a theoretical paradigm grounded in the understanding that the body is critically important to our perception of music. The theoretical framework draws from embodied cognition, in which perception is explained as a feature of embodiment, and where bodily interaction is seen as central to perception. Embodiment in musical contexts refers not only to active motor responses such as tapping to the beat, but also to motor constraints and actions that actively frame how we perceive musical stimuli (Lesaffre et al., 2017). In this model, action and perception form a feedback–feedforward loop where action and perception simultaneously inform and frame each other.

The fundamental claim of the embodied music cognition paradigm is that bodily involvement is crucial in human interactions with music, and therefore, also in our understanding of that interaction (Leman et al., 2018). Leman points out that there is no explicit theory that is in direct opposition to *embodied* music cognition which claims that music and the body are *not* related in music cognition. However, this embodied music cognition framework allows us to make explicit the role of the body in perceiving music. Music perception, in this paradigm, is considered a reconstruction of our bodies' interactions with the world. The notion that sound is closely related to physical materials and actions is central to understanding this concept. We learn early on that sounds have sources. These sources are present in our physical environment, and this physical world informs our ideas about acoustic expectations to a great degree.

4.3.1 Early Research on Music Embodiment

Historical work on embodied music can be traced back to music in the context of music accompanying exercise (Truslit, 1938), but empirical investigations of embodiment are much newer. One of the first methods to systematically study kinesthetic representation in relation to music and literary works was called Schallanalyse; it was developed by Eduard Sievers. He distinguished two classes of curves—general or 'Becking' curves and specific or filler curves (Taletfiillcurven), which described types of musical movement based on dynamic expression (related to loudness), and voice type (Shove and Repp, 1995, p.67). Becking and Truslit distinguish between three basic types of movement curves: 'open', 'closed', and 'winding' (Shove and Repp, 1995, p.71). These curves are not conducting movements—they are supposed to be executed with outstretched arms and are a means of portraying dynamics in space; the speed of the movement and the musical tension affects the curvature of the motion path.

In 1977, Clynes developed a notion of essentic forms, which are dynamic shapes that characterize basic emotions (Clynes, 1977). He developed an apparatus called a sentograph, which captures finger pressure in two directions

4. Body Movement

while listening to music. The study showed that subjects produced different ‘sentographs’ while listening to music that portrayed different feelings. He developed a computer program that enabled him to play music with different agogic (related to note-stretching) and dynamic patterns. Later, in 1992, Todd began studying the motion of the whole body rather than just the limbs or fingers. He found evidence that motion and music were interrelated based on how the ventromedial and lateral systems that control posture and motion, respond to music. His study of motoric expression in the form of ‘expressive body sway’ is significant.

4.3.2 Recent Research on Music Embodiment

Movement that accompanies music is understood now as an important phenomenon in music perception and embodied cognition (Leman, 2008). Research on the close relationship between sound and movement has shed light on the understanding of action as sound (Godøy, 2003) and sound as action (Jensenius, 2007; Jensenius et al., 2006). Interaction with music is a bodily activity, influencing perception and even contributing to it. Cross-modal correspondence is a phenomenon with a tight interactive loop—the body is a mediator that engages in both perceptual and performative roles (Gritten and King, 2006, 2011). Some of these interactions cause motor cortex activation even while simply listening to music (Molnar-Szakacs and Overy, 2006). This has led to empirical studies on how music and movements of the body share a common structure that affords universal emotional expression (Sievers et al., 2013). Mazzola and Andreatta (2007) have also worked on a topological understanding of musical space and the topological dynamics of musical gesture. Buteau and Mazzola (2000) analyzed contour similarity models as motivic topologies for specific excerpts, proposing a motivic evolution tree (MET), modeling changes in contour dynamics. These models are implemented in the Rubato software (Mazzola and Zahorka, 1994).

Studies on Hindustani music show that singers use a wide variety of movements and gestures during spontaneous improvisation (Clayton and Leante, 2013; Clayton et al., 2005; Rahaim, 2012). These movements are culturally codified; they are used in the performance space to aid improvisation and musical thought, and they convey musical information to listeners. Performers also use a variety of imaginary ‘objects’ with differing physical properties to illustrate their musical thought. Some other examples of research on body movements and melody include Huron’s studies on how height to which eyebrows are raised while singing are a cued response to melodic height (Huron and Shanahan, 2013). There are also studies suggesting that arch-shaped melodies especially have biological origins that are related to motor constraints (Tierney et al., 2011). According to the motor constraint hypothesis, melodic contours that have an arch shape is the most energetically efficient to produce by the vocal apparatus due to the increase and decrease in subglottal pressure during vocalization (Savage et al., 2017, p.332).

Auditory memory plays a key role in the discussion of melodic perception. The interplay between time scales of auditory ‘units’ and how we remember them is discussed below.

4.3.3 Auditory Working Memory

Working memory is the system responsible for the temporary storage and simultaneous manipulation of information in the brain (Schulze et al., 2018). People distinguish between working memory (WM) and short-term memory (STM); WM is the system responsible for temporary storage and simultaneous manipulation of information in the brain (Schulze et al., 2018), while short-term memory (STM) is used primarily for temporary storage. Demonstrably, it has been shown in experiments that information being processed in the WM is important in higher functional planning, and for long-term memory (LTM). WM functioning with respect to auditory cognition is studied in a number of ways, using verbal, tonal, and speech stimuli.

Baddeley and Hitch’s model of working memory, first proposed in 1974 (Baddeley and Hitch, 1974), is typically used in connection with music and language. The following concepts are useful in this discussion of Baddeley’s model of working memory:

1. Visuo-spatial sketchpad: A sort of whiteboard for the mind that plays a role in physical action planning and control.
2. Episodic buffer: This was added as a final explanatory component, as late as 2000 (Baddeley, 2000), to explain the transfer of temporal objects from the WM to the LTM.
3. Phonological loop: The phonological loop is the STM store and articulatory rehearsal location in the WM. This enables phonological as opposed to visual storage; the loop also engaged during subvocal rehearsal, which involves motor activation, and even in micro-movements of the vocal apparatus while engaging in memory storage or recall. When subvocal rehearsal is not possible in, for example, articulatory suppression conditions, verbal material is not remembered as well.

(Mora, 2000, p.150) explains the importance of singing to memorize something, where melody acts as a path towards remembering. This is also related to the phenomena of having a song stuck in your head. She quotes, “Murphey (1990) defines the ‘song-stuck-in-my-head’ phenomenon as a melodic Din, as an (in)voluntary musical and verbal rehearsal. Murphey also hypothesizes that the Din could be initiated by subvocal rehearsal. So, for example, we are able to rehear mentally the voice and words of a person with whom we have had an argument. Similarly, while reading the notes taken in a lecture, we will probably rehearse the lecturer’s voice, while at the same time we can mentally visualize the place from which s/he was talking and even her/his gestures or body movements.”

4.3.4 Multimodality

That we perceive the world using more than one sensory modality simultaneously seems to be the norm rather than the exception. This idea is described as ‘multimodality’—we perceive the world through several stimulus modes at the same time. When we see an action being performed, we expect to receive information about the action through many different sensory modalities—we see something, we expect a sound to accompany it, we can infer its touch, and so on. Particularly with sound, a range of experiments have shown that significant sensory illusions can occur as a result of our expectations. An example of this is the McGurk effect is an illusion involving the auditory and visual modalities, in which different, aligned visual and auditory speech stimuli, give rise to an ‘overriding’ effect of either the visual or auditory stimulus. For example, if we hear just the audio stream of a voice saying ‘Ga’, we apprehend the syllable accurately; however, when it is played back with a video of someone speaking ‘Da’, our visual system dominates (Rosenblum et al., 1997). Another example of this, dubbed the ‘cocktail party effect’, is an imaginary cocktail party situation, with several guests speaking simultaneously. We are able to understand someone’s speech better if we are looking at them directly (Arons, 1992). This seemingly simple effect is highly complex from the auditory cognition perspective.

For this thesis, audition and motor action are the areas of neural processing in special focus. For a long time, the motor and sensory cortices were thought to function separately from each other. However, a number of studies have demonstrated the role of the sensory and motor cortices in perception tasks (Cheung et al., 2016). Robust neural activity in the motor cortex is observed while listening to speech sounds.

The *motor theory of speech perception*, proposed in the 1960s, suggests that “objects of speech perception are the intended phonetic gestures of the speaker” (Liberman and Mattingly, 1985, p.2). Studies have since confirmed these findings using various methods. A review article published in 2010 compiles various studies on speech perception (Pulvermüller and Fadiga, 2010), and shows that in action studies, patients with lesions affecting the inferior frontal regions of the brain have difficulty comprehending phonemes and semantic categories.

4.3.5 Ecological Psychology

James Gibson’s work on visual perception in the 1970s laid the foundations for the ecological approach in psychology. This approach explores the connections between the ecological context, body movements, and the action-relevant information available to the perceptual framework. In this model, *affordances*—the possibilities for action, intervention, or use—are key to guiding perception. A drum is a well defined musical instrument, but in principle, tables, chairs, and other rigid surfaces around us become ‘drums’, or can be used as drums, and as such, having a rigid surface ‘affords’ an object to serve the function of a drum, because of the possibilities of action.

This model has action embedded in the core of our interaction with the world. Clark (1999) argues for how this model explains our interaction in the world parsimoniously. For example, in the visual domain, a good model for catching a ball would include understanding the velocity, trajectory, and direction of the incoming ball, and looking at the action with which the ball was thrown. If our perception was based on complicated mathematics, our response times to motor stimuli would be much longer than they are in reality.

4.3.6 Conceptual Metaphors

In their work, Lakoff and Johnson (1980) propose a theory for conceptual metaphors, suggesting that the words we use to describe phenomena drawn from a thematic vocabulary, rather than being employed at random (Lakoff and Johnson, 1980). For example, our word choices in certain phrases, such as ‘win’ an argument and look ‘ahead’ into the future, represent how we conceptualize arguments as war, and time as space. This theory is the foundation for the notion of conceptual metaphors, not only in language, but also in paralinguistic phenomena. These metaphors of space become apparent in gestures that accompany speech; for instance, someone might indicate ‘yesterday’ behind their back, and ‘tomorrow’ in front of the body.

Conceptual metaphor theory (Lakoff and Johnson, 1980; ?) first suggested that such linguistic metaphors show that many abstract ideas are expressed using metaphors for spatial concepts. The relationship between linguistic metaphor and spatial orientation might also extend to the link between language and gesture. David MacNeill (1992), in his book, *Hand and Mind: What Gestures Reveal About Thought*, he explores this relationship. In this book, he talks about the nature of co-speech gestures and idiosyncratic movements of the body that are associated with speech. He suggests that the microevolution of an utterance may, in turn, epitomize the macroevolution of the linguistic system from primitive gestures.

Metaphorics in gesture studies deals with gestures that serve as metaphors for abstractions in language. A good example of this is the ‘cup-of-meaning’ hand shape, where the speaker refers to an abstract idea with a cup-shaped gesture, which changes shape as the idea is refuted or changed. Data from several studies related to metaphorics in gestures suggest that speakers in demanding communicative situations—especially ones in which they had to express abstract ideas—conceptualize those ideas in space (Enfield, 2005; Nunez, 2004; Sweetser, 1998). Such metaphorical gestures may help in the conceptualization of abstract concepts by grounding them in space. These co-speech gestures are observable ‘referents’, even if the ideas they communicate are metaphorical or abstract. Music, however, is semantically void and cannot by itself make concrete references. The abstract nature of music makes it challenging to analyze gestures within any of these possible typographies. It also makes it hard to analyze the meanings of gestural associations. Do they represent cognitive schemas and back-end multimodal structures, or are they an epiphenomenon, reflecting habits formed during training?

4. Body Movement

In his book, *Understanding Music: Philosophy and Interpretation*, Scruton (2016, p.46) says:

“Of course the melody doesn’t *literally* move, it isn’t *literally* there. But we hear it all the same, by virtue of our capacity to hear metaphorically—in other words to organize our experience in terms of concepts that we do not literally apply. ”

This quote rings true in the qualitative observations of the participants’ movement in the experiments conducted for this thesis. Something in the melody is perceived to be moving metaphorically, and can be represented through hand-movements, using a variety of action metaphors.

4.4 Music Related Movement

Action produces sound, and action and sound are related to each other in a loop (Jensenius, 2007). Sound also invokes images of action. When we hear a loud drum stroke, we can imagine the force with which it might have been produced. In the following sections, I discuss gestural imagery in the service of musical imagery, and how musical gestures evoke sonic imagery.

4.4.1 Gestural Imagery

Godøy argues that the typology of sonorous objects can be extended to what he calls gestural-sonorous objects (Godøy, 2010). Theoretical approaches for this might include analyzing our conceptual apparatuses, our invocation of gestural metaphors, and so on in descriptions of sonorous objects; observation studies of music producing and music imitating actions; and sound-tracing studies.

In Schaeffer’s work, sound types are presented with the following categories (Schaeffer et al., 1967b):

1. Impulsive: The overall energy envelope is based on a sharp attack with a decaying resonance.
2. Sustained: The energy envelope is based on a continuous energy transfer that results in a continuously changing sound.
3. Iterative: The excitation pattern is built on a series of rapid and discontinuous energy transfers, or a series of attacks that tend to fuse into one another.

Godøy (2018) proposed a three-part model consisting of gesture sensations that are influenced by, and also influence, continuous sound and multimodal gesture sensations. Godøy argues that Schaeffer’s categories apply to music related movement and not just for analyzing sound, leading to phrase transitions when playing between sounds in any category in the aforementioned typology. For example, an impulsive drum stroke turning into an iterative drum trill has

these features present both in the sound and the action producing the sound. For music related movement, he proposed the following categories:

1. Sound producing: These movements are necessary in producing sound; for instance, directly striking a drum.
2. Sound accompanying: Movements such as dance, walking, and sound-tracings, which are representational or imitating movements.
3. Ancillary: Movements facilitating sound production.
4. Communicative: These movements communicate specific musical moments to the audience.

In Godøy's model of gestural-sonorous objects, our mental images of musical sound are based on an incessant process of mentally following, or tracing, the features and qualities of a sound. In this model, gestural sensations have visual and motor components based on the biomechanical affordances of the sound objects, and the physical constraints of our bodies.

It might seem to some that this model represents a listening scenario that is inorganic, or not representative of our real-world musical listening behaviors. Removed from their electro-acoustic contexts, these images of movements and their tracings would not be accurate representations of the music. However, while researching improvised gestures of North Indian classical music, Godøy's model applied even though that context has not been a part of the original propositions of motor-mimesis.

Ornaments as Multimodal Melodic–Gestural Objects

We are never completely still until we die. Body motion is studied from many angles and within different disciplines, including medicine, robotics, sports sciences, phenomenology, and psychology. Embodied and enactive cognition claims that much of human cognition is based on our ability to move and act in our environments.

The production of sound is intrinsically related to the performance of an action. We have to do something in order to produce a sound—hitting something produces a sound that is intrinsically related to the properties of the materials that contact each other, and the action that was involved in producing that sound influences it. For example, hitting a drum with a mallet with great force, from a distance, or gently, will produce different sonic possibilities. The coupling of action and sound is obvious in this simplest example of sound production—a collision of objects—and holds true for more complex actions that require the coordination of several muscle groups and body parts; for example, singing a tune or playing the piano.

Direct sound-producing actions are not, however, all that we do to produce music. There are many subtler, postural, and communicative dimensions of music related motion; for example, a pianist's body swaying while they play a particularly delicate passage. When we listen to an energetic drum sequence,

4. Body Movement

for example, we mentally simulate the movements of the percussionist. In this chapter, I unpack the different ways in which our perception of sound is multimodal, and how we are also able to demonstrate the sound–shape associations using our bodies.

While we listen to music, the act of tapping our feet to the rhythm—so-called rhythmic entrainment—has been researched in great detail. The timing precision and limbic control in our responses to different sound envelopes and different tapping effectors have been studied more over time than the the embodiment of melody.

In order to become proficient in playing an instrument, it is typical to practise ornaments that constitute the building blocks of the style over and over again. To play these ornaments intertwined in longer phrases, our execution of them has to be ‘without thinking’. Ornamental symbols, notated using shapes; for example, the trill, mordant, glissando, resemble the contour shapes of the sound produced.

4.4.2 Gesture and Musical Gestures

A *gesture* can be defined as a movement of a part of the body, especially a hand or the head, to express an idea or meaning. All body movements are not gestures; there is an implicit signification associated with gestures that is not generally associated with the terms actions or movements. Signification may include specific hand shapes and their associated movement qualities. Figure 4.1 illustrates hand shapes that are used as mnemonic devices while teaching children, which represent the seven notes of the diatonic scale.

Shrugging is a great example of a gesture that has a clear, universal meaning. It is a short movement—the shoulders are raised quickly, and the palms face upwards, as if to show that you are empty-handed. If our hands are otherwise occupied, this gesture can be transposed into a shrug that incorporates only the shoulders and a tilt of the head. A shrug has cultural connotations and its associated notional description is to indicate that the person does not have anything (in their hands). We can also execute this gesture in several time scales, using various effectors. Thus, communicating and understanding the gesture encompasses its culturally understood meaning, and its smooth execution.

We talk about ‘muscle memory’ in reference to tasks that we perform as automatisms, without explicit conscious awareness. Of course, our muscles do not actually hold memories, and procedural memory in the brain is what enables the performance of these tasks. While gesturing, and especially while making gestures that have communicative intent, we are able to perform these seemingly difficult-to-describe actions—such as ‘wiggling your fingers’—easily, which is perhaps a result of much practice. Rosenbaum points out that an innocuous gesture such as ‘wiggling your fingers’ might seem simple to do, but once we start to describe how to do it, the description seems complicated (Rosenbaum, 2017, p. 65).

The study of gestures involves multiple disciplines: their semiotic meanings and associations are studied as language, while their physiology and neurology

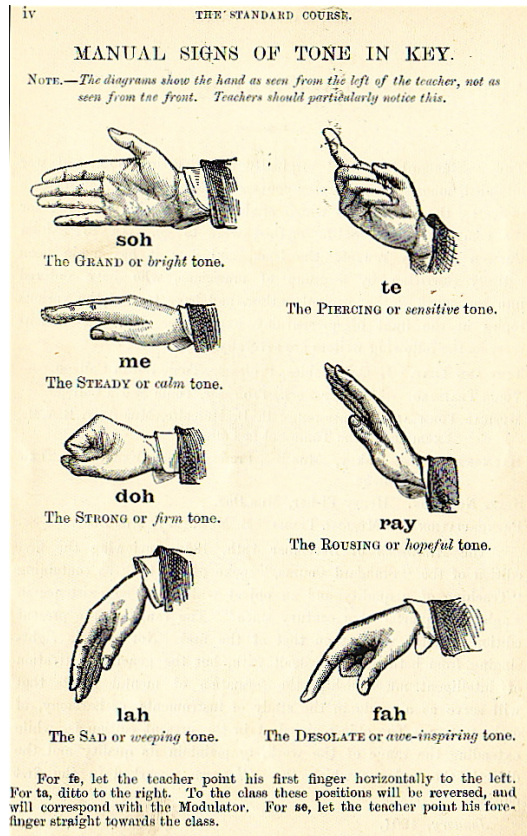


Figure 4.1: Solfa gestures are used to help memorize intervals in a major scale. This method has been used quite often to teach students of singing.

are studied to understand manual control. There may or may not be anything biological about the ‘okay’ gesture, in which the thumb and index finger form a ring and the other three fingers are splayed. But this type of gesture is understood by, for example, the speakers of a language, or those belonging to a particular culture. Tasks such as grasping objects can be scientifically studied in the context of biologically important fine motor skills.

Co-Speech Gesture

Speech is often accompanied by co-speech gestures such as movements of the body, particularly of the hands. Co-speech gestures are considered paralinguistic, and have only recently been studied in the realm of formal linguistics. Gesturing usually begins before speech, and plays a central role in the development of language. The simplest example is understanding references from pointing. Gesture studies have proposed several sub-types—iconic gestures indicate

4. Body Movement

the shapes and sizes of the referenced objects; metaphoric gestures indicate metaphorical qualities; and beat gestures emphasize particular words in speech. Deictic or pointing gestures can indicate an actual object in space, an abstract reference, or an idea. Many people think of emblems as co-speech gestures; however, these are simply gestural icons in the shared vocabulary of speakers of a language or members of a cultural group—for example, the ‘thumbs up’ gesture.

Kendon (2004), in *Gesture: Action as Visible Utterance*, poses the following question at its outset:

“How can a person, in creating an utterance, at one and the same time, use both a language system, and depictive pantomimic actions? As a close examination of the coordination with gesture and speech suggests, these two forms of expressions are integrated, produced together under the guidance of a single aim.” (Kendon, 2004, p.2)

Kendon defines gesture as ‘visible action as utterance’ (Kendon, 2004), and describes the long neglected body in linguistics as a central component in the affective processing of speech. Why do we make co-speech gestures, what is their typology and function, and how can we study them better? The ‘noise’ is the data, as it were. Researchers in the early twentieth century were preoccupied with studying the origins of language through gesture. In the middle of the century, however, this interest was deemed ‘unscientific’ (Kendon, 2004, p. 64). The more recent evolution of interest in co-speech gestures and embodiment has enabled us to further our understanding of music and acoustic perception.

What types of gestures do we use to accompany speech? In the study of co-speech gestures, the following typology is proposed by McNeill (1992):

1. Beats: Rhythmic gestures that mark words or phrases as significant to the discourse/pragmatic content.
2. Deictics: These gestures point at concrete entities or particular spaces.
3. Iconics: They depict the form or movement of physical entities, or the physical relationship between them.
4. Metaphorics: They represent an abstract idea as if it could be held or was located around the speaker—for example, a small object you hold in your hands to mean ‘this idea’.

While studying co-musical gestures, in improvised, especially vocal music, we find that such a typology of co-speech gestures can prove useful to understand and describe gestures accompanying music (Pearson, 2016).

4.5 Action–Sound Perception

Jensenius proposes a model of action–sound couplings, and the chain of production of action–sound for sounds that we perform intentionally. In this

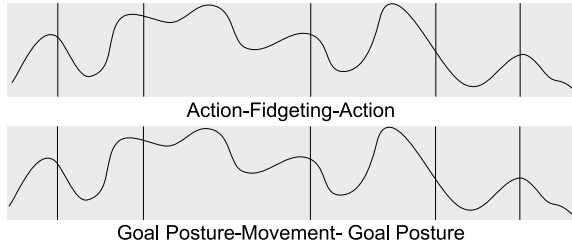


Figure 4.2: Similarities between the action–fidgeting model and the posture-based motion planning theory.

model, a sound producing action has an excitation phase that is preceded by a prefix and followed by a suffix. Godøy (2001) provides a broader background of sound producing action.

In Jensenius (2007)’s model of music-related movement, in controlling a musical instrument, movement phases, actions, and fidgeting might come together to form a chain of several movement units, action units, and fidgeting units. An action is distinguished from fidgeting in that the former is specific and goal-oriented. In this model, the term *gesture* is totally avoided, in order to not conflate it with its multiple meanings. Instead, a four-part model is provided:

1. Movement: The act of changing the physical position of a body part or object
2. Action: It denotes a movement unit or chunk. This is a goal-directed movement with a particular outcome.
3. Fidgeting: Non goal-directed movements, perhaps unintentional or subconscious movements, are classified as fidgeting, or ‘movement noise’.
4. Interaction: The reciprocal influence of the moving parts in an action stream.

In this type of model, the goal-directed nature of certain but not all movements is key. Jensenius represents this as a stream of action–fidgeting–action. This feature of the model is similar to a different theory of bimanual control, called posture-based motion planning theory, which I will explain in the following subsections.

4.5.1 Ideomotor Theory

Ideomotor theory proposes that whenever an action is executed, the mental representation of the movement itself is linked to the representation of the effects in the mind (Herbort and Butz, 2012). This means that once this relationship is

4. Body Movement

established, merely anticipating the effect allows us to create the appropriate movement. A review of the current state of work on ideomotor theory can be found in (Shin et al., 2010).

Laboratory experiments have shown that holding an object, such as a gun, in one's hand greatly increases the chances of spotting the same object in videos. Another experiment showed that visual targets are perceived as being nearer when they can be touched with a handheld tool (Rosenbaum, 2017, p. 97).

An experimental analysis of sound-tracings performed while simultaneously listening must include a discussion of ideomotor theory, and action-first approaches to perception. The intersecting ideas in this theory are that perception is greatly influenced by our intention to interact with objects in the real world.

4.5.2 Motor Control

David Rosenbaum's theory of posture-based motion planning (PBT) describes the motion planning phases in the execution of manual control (Rosenbaum, 2017). The main claim of this theory is that voluntary movements, instead of being produced directly in 'one fell swoop', are generated first by our nervous system using a series of 'goal postures'. Then, movements translate the body from one goal posture to the next. In this discourse, a posture is a musculo-skeletal state.

Rosenbaum raises the question of whether there is a point below which we cannot control our motor responses. For example, to what extent do we control our hands when we clap? We might have conscious control over executing a clap, and may even choose how to position our hands, or adjust the intended loudness. Once we initiate the clap, however, it seems to perform itself. This image of control can differ depending upon the tasks at hand, the extent of practice, and a number of other factors. Rosenbaum calls mental representations of action states 'images of achievement', likening them to reference conditions of feedback control theory. Without images of achievement, motor imagery does not have ideal states to which to aspire.

For a sequence of motor actions to precede purely perceptual states 'in the mind', a lot of well-rehearsed action sequences must take place. In order to perform a task such as grasping, the following mental apprehensions are needed: distance calculation, a mental image of achievement, or a goal state, and a rehearsed aim with hand-eye coordination. Chunking or stacking of multiple goal states in skilled tasks such as, say arpeggio playing, will influence not just the performance of the action, but the perception of the objects in relation to our bodies.

In the execution of an action such as a sound-tracing, goal motor-states could be pre-planned and contour trajectories 'filled in', particularly if sound stimuli are heard multiple times.

4.6 Summary

In this chapter, I have looked at the different ways in which music related motion is executed, and how action-aware models of musical perception can help us understand music as an embodied phenomenon. I believe that sound-tracing as an experimental paradigm brings together the multimodal mappings of pitched sound, gestural imagery evoked by these sounds, and defining geometries of these contours. In this chapter, I discuss gesturing during speaking, and how this field can also contribute to thinking about musical motion. I have also outlined some important mechanisms for action–sound associations, as well as some theories of motor planning and control. In the next chapter, I will explain the experiments performed for this thesis, with the details of the stimuli, and the data sets created as a result.

Chapter 5

Data Sets and Experiments

5.1 Introduction

For the experiments in this thesis, I used a single stimulus set consisting of 16 melodies. Using these stimuli, I conducted three experiments as detailed in this chapter. The data sets are described below, and they were released as mentioned in the references.

One stimulus set was used for two experiments, resulting in three data sets, as shown in Figure 5.1.

5.2 Stimulus Set

The stimulus set consists of 16 melodies from four music traditions that involve singing without words. For this set, we picked the genres classical vocalise, jazz scat, North Indian classical music, and Sámi Joik. These musical traditions also represent the author’s interests and experiences. We picked a short melodic fragment from a performance in each of the four musical traditions. As explained in the previous chapter, vocal melodies hold a closer relationship to the body, and they are often described in terms of our ability to hum them.

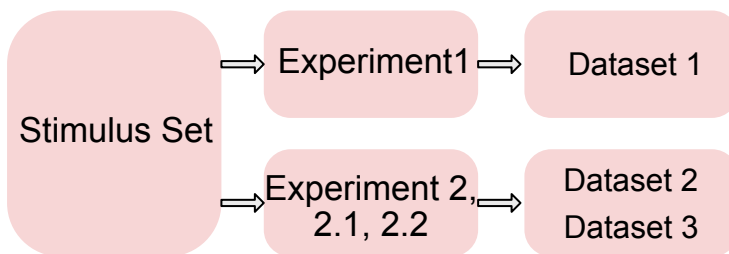


Figure 5.1: A stimulus set containing 16 melodies was used for two motion capture experiments, resulting in three data sets: one for melody–motion pairs, the second for repetitions in sound-tracings, and the third for singing back melodies after one hearing.

5. Data Sets and Experiments

We specifically chose vocals without lyrics to avoid participants' perception of contours being influenced by words and word meanings. In most studies dealing with melodic contours, the role of text and lyrics has not been explicitly explored. However, the inter-dependency of prosody and text has been studied in compositional contexts.

Here, we also selected stimuli with little to no accompaniment, in order to maintain the focus of the experiment on melodic contour and the monodic line only.

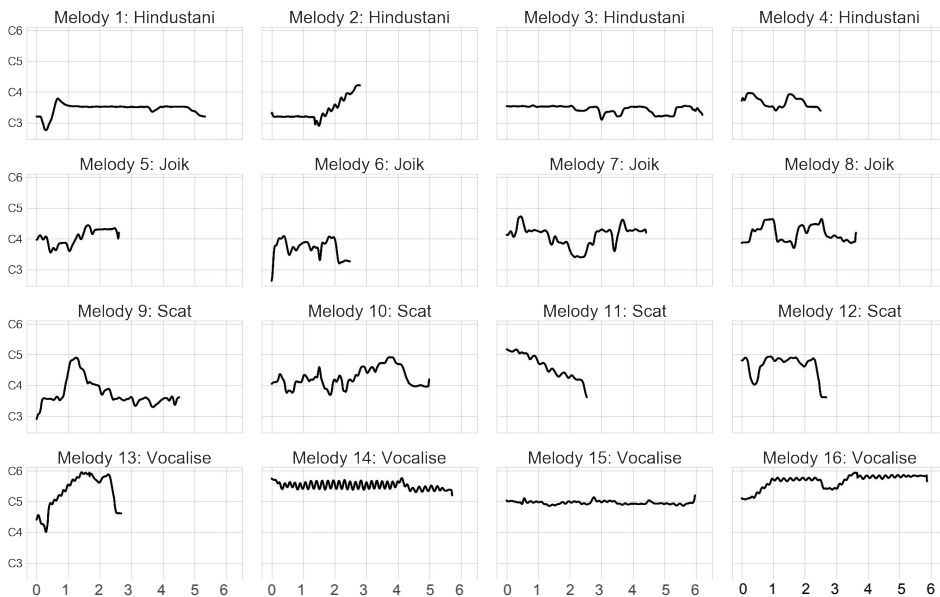


Figure 5.2: The 16 melodies used as the stimulus set for all the experiments in the thesis come from four different music cultures and contain no words. The X-axis represents time, and Y-axis represents pitch height in MIDI notation.

5.2.1 Descriptions of the Musical Styles

Here I will explain the cultural contexts of the musical styles chosen. The four musical systems were some where there is a tradition of singing without words. This was to make sure that melody occupies the central position in the composition.

North-Indian Classical Music At the foundation of North-Indian classical music, lies concept of a 'raga', which includes a scale and a grammar to guide and aid improvisation. In the modern system of 'khyal' singing, musicians improvise with the raga by establishing the grammatical form, and then singing over it on

a vowel. Nonsense syllables such as ‘re-ne-na-tom’ are also used in some schools of music in the tradition, and in forms related to khyal singing, such as *drupad*, which relies on long improvisations.

The phrases chosen for the stimulus set come from lecture demonstrations by Pt K. G. Ginde, available in the University of Washington Ethnomusicology Archives (of Washington Ethnomusicology Archives, 1991). The raga that of the melodic stimuli is *Bhoopali*.

Joik The joik is a song style of the Sámi people from northern Norway, Sweden, Finland, and parts of Russia. The joik is a sung, melodic form that is meant to evoke an element in the landscape, a person, an animal, or any other kind of entity. A joik might be composed based on any number of characteristics of the entity being evoked. Usually, a joik is sung without words, using syllables to aid and propel the melody.

The joiks chosen for this data set come from the Smithsonian Folkways collection, *Lappish Joik Songs from Northern Norway*, produced and recorded by Wolfgang Laade and Dieter Christensen (1956). The three pieces chosen are by Per Henderek Haetta (Quarja), Inga Susanne Haetta (Markel Joavna Piera), and Nils N. Eira (Track 49) (Laade and Christensen, 1956).

Jazz Scat Scatting is a technique in jazz singing in which meaningless syllables are sung to improvised melodies. Some trace the origins of scatting to people attempting to recreate percussion patterns using the voice. The syllables used and their composition within improvisations depend on individual performance techniques, time periods in scat singing, and other factors.

The scat stimuli used for this thesis come from Ella Fitzgerald’s recording of “One Note Samba”, composed by Antonio Carlos Jobim, from a 1969 recording at the *Montreaux Jazz Festival*. The scat sections in this performance have little to no accompaniment.

Vocalize A vocalize is a piece or set of exercises in the western classical singing tradition that is explicitly sung without words. Western classical music is, more often than not, composed with words. However, some improvised sections, such as the *coda*—a passage that functions as an extended cadence—use this technique of singing a vowel without text.

For this stimulus set, we used excerpts from the piece, “Si, Ferite Il Chieggo”, from Rossini’s 1820 opera, *Maometto II*. The chosen recording is sung by soprano, June Anderson, in a performance with the Ambrosian Philharmonic Orchestra in 1983.

5. Data Sets and Experiments

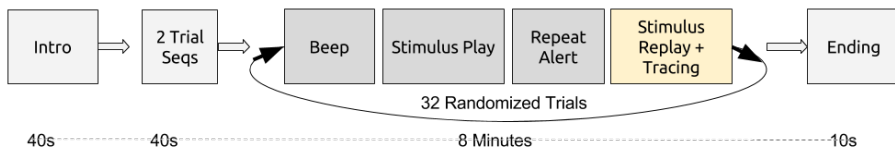


Figure 5.3: Experimental flow for experiment 1.

5.3 Experiments

5.3.1 Experiment 1: Melodic Tracings

A total of 32 subjects (17 female and 15 male) were recruited, with a mean age of 31 years ($SD = 9$ years). The participants were mainly university students and employees, with and without musical training. Their musical experience was quantified using the OMSI (Ollen Musical Sophistication Index) questionnaire (Ollen, 2006); they were asked about their familiarity with the musical genres, and their dancing experience. The mean OMSI score was 694 ($SD = 292$), indicating that the general musical proficiency of the subjects in this data set was quite high. The average familiarity with western classical music was 4.03 out of a possible 5 points, 3.25 for jazz music, 1.87 for joik, and 1.71 for Indian classical music. Thus, two genres (vocalize and scat) were more familiar than the two others (North indian and joik). All participants provided written consent before they participated in the study, and they were free to withdraw at any point during the experiment. The study obtained ethical approval (on 22 August 2016; project code 49258) from the Norwegian Centre for Research Data (NSD).

Procedure Each subject participated in the experiment alone; the total duration was around 10 minutes. Participants were instructed to move their hands as if to create the melody with their movements. The use of the term 'creating', instead of 'representing', was purposeful, as in earlier studies (Nymoen et al., 2012, 2013), to avoid the terms playing or singing. Subjects could stand freely, anywhere in the room, and face whichever direction they liked; nearly all of them faced the speakers and chose to stand in the center of the lab. The room lighting was dimmed to help the subjects feel comfortable and to encourage them to move as they pleased. The stimuli were played at a comfortable listening level on two Genelec 8020 speakers, placed 3 m in front of the subjects at a height of approximately 1.5 m. Each session consisted of an introduction, two example sequences, 32 trials, and a conclusion, as shown in Figure 2. Each melody was played twice, with a two-second pause between instances. During the first presentation, participants were asked to listen to the stimuli; during the second presentation, they were asked to trace the melody. A long beep preceding the melody indicated the first presentation of the stimulus; a short beep indicated the repetition of the stimulus. All the instructions and required guidelines were recorded and played back through the speaker so as not to interrupt the flow of

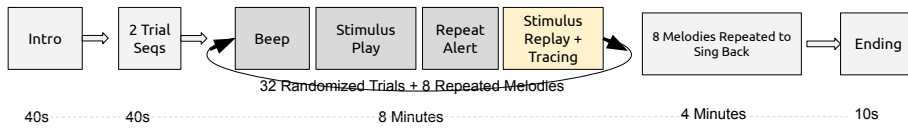


Figure 5.4: Flow of Experiment 2. The repetitions and sung sections are included in this experiment.

the experiment as illustrated in Figure 5.3

5.3.2 Experiment 2: Melodic Tracings, Tracing Repetitions, Singing Melodies

For this data set, 20 subjects participated twice in the same experiment described above, with eight stimuli, which were chosen randomly each time. As such, there were two responses to each tracing by each participant in this data set. There were 9 male and 11 female participants in this study; the mean age was 28 years ($SD = 7.4$ years). They were mainly university students and employees, with and without musical training. Their musical experience was quantized using the OMSI questionnaire (Ollen, 2006), and they were asked about their familiarity with the four musical genres, and dancing experience. The mean OMSI score was 540.89 ($SD = 342.92$). The average familiarity with the four genres was as follows: 3.62, 2, 1,87, 2.93. All participants provided written consent before they participated in the study, and they were free to withdraw at any point during the experiment. The study obtained ethical approval (on 26 June 2017; project code 54653) from the Norwegian Centre for Research Data (NSD).

The flow of the experiment is as shown in Figure 5.3. The details of the laboratory environment and the playback are the same as explained for Experiment 1 in Paragraph 5.3.1. This second experiment included two sub-parts that were not included in the first experiment:

1. All 20 participants traced eight of the melodies in their stimulus set twice during the experiment. Although there have been several studies on sound-tracings in general, very few have looked at whether people repeat their own tracings.

2. Ten participants were chosen at random to perform another task, in which they repeated eight melodies in the experiment, by singing them back after one hearing. The melodies are complex, and hard for even trained singers to remember in their entirety after hearing them just once. The singing was recorded using a microphone in the lab. Motion capture was used to determine whether participants naturally tended to use body movements to remember and reproduce the melodies. One participant had to be excluded due to technical problems with recording quality.

5.4 Data Sets Used

Data Set 1: Melody–Motion Correspondences through Sound-Tracings

This data set was collected in two stages, during which participants traced 16 melodies freely with their hands in two conditions; participants had 21 markers on their bodies, and were recorded using eight infrared motion capture cameras. The data set contains a total of 32 subjects (17 female and 15 male), with a mean age of 31 years ($SD = 9$ years). They were mainly university students and employees, with and without musical training. Their musical experience was quantized using the OMSI questionnaire (Ollen, 2006), and they were asked about their familiarity with the musical genres, and their dancing experience. The mean OMSI score was 694 ($SD = 292$), indicating that the general musical proficiency in this data set was on the higher side. The average familiarity with western classical music was 4.03 out of a possible 5 points, 3.25 for jazz music, 1.87 for joik, and 1.71 for north Indian music. All participants provided written consent before they participated in the study; they were free to withdraw at any point during the experiment.

After post-processing, this data set contained a total of 794 tracings; it has been released as supplementary material in the article, ‘Analyzing Free-Hand Tracings of Melodic Phrases’ (Kelkar and Jensenius, 2018).

Data Set 2: Repetitions of Sound-Tracings

This dataset consists of 20 participants tracing eight melodies in two iterations. The eight melodies were randomly selected from the stimulus set. The repetitions occur at the end after tracing all 32 stimuli first.

Data Set 3: Singing Complex Melodies Back After One Hearing

For this data set, nine participants sang into a microphone eight randomly selected melodies from the 32 melodic fragments in the main stimulus set. Only eight melodies were selected, in order to limit the duration of the experiment. A total of 72 melodic fragments were obtained as a result. There were 9 participants in this study; 6 male and 3 female. The mean age was 27 years ($SD = 5.8$ years). They were mainly university students and employees, with and without musical training. Their musical experience was quantized using the OMSI questionnaire (Ollen, 2006), and they were asked about their familiarity with the four musical genres, and dancing experience. The mean OMSI score was 540.89 ($SD = 342.92$). The average familiarity with the four genres was as follows: 3.78, 2.33, 1.89, 3. All participants provided written consent before participation, and they were free to withdraw at any time during the experiment.

The post-processed data set contained the pitch extracted from each of the recorded phrases.

5.5 Summary

In this section, I described the stimulus set, experiments, and data collected. The same stimulus set was used to perform two experiments, which resulted in three data collections. The experiments involve participants listening to melodies twice, and in the second iteration, moving to the melody. Examining these data involves analyzing melodic material, motion capture data, and the comparisons and correlations between melody–motion pairs. This is described in detail in Chapter 6.

Chapter 6

Methods

“...if an algorithm performs well on a certain class of problems then it necessarily pays for that with degraded performance on the set of all remaining problems.”

– (Wolpert et al., 1997, p.1)

6.1 Introduction

The data types and analysis methods used in the papers included in the thesis are elaborated on within this chapter. Since the work handles two types of data—music data as sound signals and movement data gathered from an infrared motion capture system—as explained in Section 6.2 and Section 6.3, the analysis methods for each of the two data types are separate, but there are several ways to analyze them together, as explained in Section 6.4. Figure 6.1 explains a summary of the signals, representations, analysis methods, and outputs that are explained in this chapter.

6.2 Sound Analysis

The stimulus material and the analysis of melodies in this thesis depend upon the analysis of sound and music in various formats of representation—primarily audio files and symbolic data. The stimuli are presented as audio and recorded material. There are several audio file formats.

For a deeper analysis of melody and melodic contour, we extract the fundamental melody from each of these audio files using the YIN algorithm (De Cheveigné and Kawahara, 2002). It has been shown that melodic contours are an abstraction of pitch and pitch patterns in music. There is little reason to suggest that sheet music should be the ideal source of comparison to rely on contour models, as the perception of contour does may or may not correspond to the algorithmic representation of pitch height, and melody may or may not be understood only as discrete pitches.

6.2.1 Computational Representation and Features of Melody

Computational work in music hinges on the representation of music as data, and the way in which this is done determines and limits how research can be conducted. That which is often considered ‘noise’ in one data set may have a lot of meaning, relevance, and applicability in another analysis or interpretation. For example, the insights we gain from score-based representations of musical traditions that do not use scores to encode or perform music may be limited.

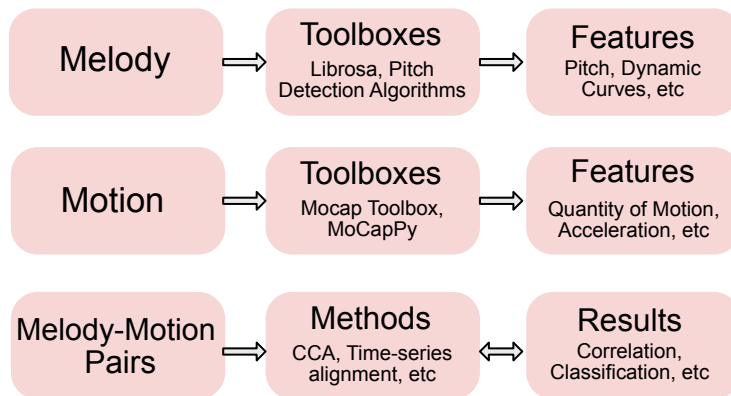


Figure 6.1: An illustration of different levels of data handling and analyses in the experiments. The three types of data each have their own signal

That said, the kind of representation is usually chosen to address and answer specific questions.

Aucouturier and Bigand (2012) have put forth a dilemma of human and machine listening, and their pitfalls and merits in their paper. This question is very important to try to answer as the fields of music information retrieval and music psychology progress.

To analyze melodies from musical material, one of two approaches are typically used to represent musical material: symbolic or signal-based methods. Symbolic methods rely on transcribing melodies into their constituent pitch values using one of many notation types. This may include transcribing pitch into western notation, describing pitch classes, transcribing relative interval distances in semitones, and so on. Signal-based analysis of melodic material involves analyzing music or melodic features from recorded material using signal processing. Both these approaches have strengths and weaknesses. While a symbolic model allows for generalizations by pointing out exactly what we want to see in the data, the signal-based approach allows us to see the musical content as more than just the notes that we have deemed important. These two approaches were developed with completely different applications in mind. Symbolic data-based methods are popular in the analysis of western classical music because the practice of creating symbolic notation is common and accessible for that type of music. However, signal-based approaches allow for us to truly expand our data sets to include music from anywhere in the world, and for non-musical material to be analyzed as music. The development of many symbolic notation methods confines the region of interest considerably, as most music in the world do not use notation as the main pedagogical or representational method. In this section, I will describe both of these approaches in detail, focusing on how they facilitate

melodic contour analysis.

A melody can be represented in many ways: a musical excerpt sung, played, recorded, or recorded and played back; its structural features (such as notes); extracted contours; approximations or deviations of other typological shapes; a structural operation like transposition; or as belonging to a family of melodic groups like modes, ragas, and so on. Each of these melodic families excludes something that could be important for another melodic family. For example, a mode and a raga both have scale properties, but do not share the same concept of melodic grammar. What forms a melodic family is both culturally and musically determined. All of these features together describe the concept of a melody, but how we choose to represent melody in research is a deliberate choice.

Deciding on the level of representation is the first and most crucial choice in any empirical work related to music analysis. In *Languages of Art*, Goodman (1968) emphasizes that representing artistic or musical material is an act of curation, which includes breaking down the features and picking ones that we desire in our system, and leaving behind undesirable ones. In western classical music, for example, it is commonplace to think of a score as a faithful representation of that which is most important to music—in a way, the skeletal or formative representation of that which is music. This idea translates poorly to another musical cultures; for example, in North Indian music, to confine the pitch continuum into the chosen discrete representations on a musical staff would be to take away from the essence of that music. Therefore, the act of choosing a method of representation also determines what is important or worthy of analysis, and by extension, might disregard the cultural values of another musical system.

How does the choice of representation of music relate to our perceptual apparatus? Research in pitch and pitch organization has focused on pitch with and without musical context. In his chapter, ‘Pitch and Pitch Structures’, Schmuckler argues for ecologically valid perceptual contexts in pitch perception studies (Schmuckler, 2004), arguing that studying pitch outside its musical context presents only half the picture. Arguing in favor of the same contextual setup, for the studies in this thesis, using melodies in all kinds of contexts are important for us to formulate an ecologically valid picture. Therefore, I focused on analyzing material directly from musical recordings, as opposed to generating isotonic melodic material specifically for testing purposes. One consequence of this approach is that without reliable tools to describe this musical material, its analysis and comparison are impossible. Music information retrieval and signal processing have come a long way in understanding musical content; we use that body of knowledge as the basis for understanding all of our musical stimuli.

6.2.2 Symbolic Approaches

The history of musicology in the West relies on the study and analysis of scores that are part of a symbolic system. Notes are transcribed into discrete time, timbre, and pitch units. Rhythmic notation separates into discrete time units related to beat length; pitch units are separated into diatonic note positions,

defined by a one-to-one relation with an absolute frequency; and timbre is typically written for certain instrument or instrument families. The evolution of this notation led first to the representation of the ornaments over certain notes in the music, at which point only specific ornaments could be represented, and later more were added in response to the need to notate more complex intonation, and freer rhythm and dynamic qualities. The expansion of western notation in this way does not mean that it stops being symbolic, but that the types of symbols and the relations between them expand. While the opposite of a symbolic approach would be accurate signal-based representation, the latter cannot really be classified as notation. A notational system lies in the representational space somewhere between actualization and abstraction, allowing for enough abstraction to represent the intentions of the composer, but not to the extent that different versions of a piece are impossible.

This history is important because western notation informs, to a large extent, conceptions we have about musical abstraction—for instance, how information is stored and manipulated for experimentation in musicology, psychology, and computer science. The reason a discussion of symbolic musical systems is important to this thesis is because a large amount of research done on melodic contour analysis and melodic grammars relies on symbolic music.

Using symbolic data for melodic analysis involves feeding notes into a computer. This process divides musical time into discrete units using one of many available options. Using MIDI scores is a very common way of inputting symbolic musical data. MIDI music in its early days made it possible for electronic instruments to be able to communicate with each other. In a review, Wiggins et al. (1993) presented a framework for the analysis of symbolic notation systems popular at the time (many of them still survive) on two axes: structured generality, and expressive completeness. In any notation system, there is a trade-off between expressive completeness and structured generality.

Annotation and Transcription

Symbolic methods mainly rely on annotation and transcription. When corpuses contain data in the form of music that was written first, such as classical music, it is easier than dealing with, for example, improvisation and other musical forms. In trying to analyze folk music—even European folk music—(Van Kranenburg et al., 2007), symbolic approaches rely on annotations of folk melodies. However, in the living tradition, these melodies are not necessarily played exactly as they are notated, and taught more often by ear than by reading score.

Symbolic Music and Contour Analysis

Several studies focusing on melodic contour analysis begin by analyzing symbolic data in the form of western notation, pitch class, or interval numbers. Contour analysis methods developed with symbolic music in mind often use up and down, or + and – signs, to represent the direction each note takes in relation to the previous. Although this method is primarily used for western music analysis, it

is often applied to other types of music as well (Eerola et al., 2006; Eerola and Bregman, 2007).

6.2.3 Music Information Retrieval

Music information retrieval (MIR) is an interdisciplinary domain that draws from signal processing, electrical engineering, machine learning, and music cognition, among other fields, to build a computational understanding of musical content. As a domain, MIR represents a scope of problems that can together help us solve a magnitude of cross-comparison based issues with finding, searching for, and indexing music and music metadata. The MIR domain of problems is most widely used by music libraries and digital music servicing platforms that connect millions of listeners to millions of songs. The most common use case for several applications is building recommendation systems, but the impact of these systems on music psychology studies cannot be understated.

Some examples of the use of MIR techniques include studies on emotional affect using mood detection algorithms (Juslin et al., 2014), where the MIR toolbox (Lartillot and Toiviainen, 2007) is used to enumerate the acoustic characteristics of the stimulus set, which are evaluated against listeners' ratings of affect. Knox et al. (2011) analyzed pain relieving music using MIR techniques to study acoustic characteristics. Malandrakis et al. (2011) used MIR techniques for emotion tracking in film music. MIR-based methods and tools are used on both symbolic and signal-based data sets.

Downie (2003) explains the facets of MIR systems and their challenges. The simplest and the first challenge involves *transcription of pitch* from audio recordings. This includes melodic extraction from polyphonic recordings, to their generating accurate transcriptions, and matching or comparison of melodic contours from one another. The *temporal* challenges include detection and identification of pitch durations, and beat durations. A level above, there are challenges with tempo detection, detection of non-standard tempo behavior such as rubato or accelerandos, and so on. The *harmonic facet* includes problems of detection of harmony from sound recordings, or from score. The difficulties in this domain mainly lie in the diversity of harmonic content, and also the window of interpretation of harmonic analysis. The *timbral facet* deals with the ability to identify between several instruments, but can also be thought of as important for detecting instrumental expressivity. The *editorial challenges* include retrieving information from details of the sound signal. For example markings on a score that could be possible to detail from hearing. This may include dynamic changes, annotations of song sections, identification of structure and its tagging, and so on. The *textual facet* includes detection of lyrics, synchronising lyrics with sound recordings or scores, matching melodies to text translations, and so on are typical problems within this domain. Finally the *bibliographic challenges* include curation of metadata and information about the ontologies and tagging of sound files.

These challenges are further complicated by a range of problems associated with using MIR algorithms in several different scenarios. Downie underlines

some of them 2003:

1. Multi-representational challenge: Music data has many formats including sound recordings, score, metadata, and motion data. Within each data type, there are several encoding standards, several physical formats such as tapes, CDs; and digital formats. To have these data types be accessible in different types of analyses presents a challenge of representation.
2. Multicultural challenge: Although the majority of early MIR algorithms and interests serve the western classical canon, it is obvious that different musical cultures with their own epistemologies and data pose a range of challenges in the umbrella of MIR.
3. Multi-experiential challenge: Creative users of music, both musicians and users of a range of applications developed for musical interaction, can use MIR-like methods in a range of generative or creative applications. This invites comparisons of music data with a large quantity of other kinds of data, such as physiological, meta annotations, and so on.
4. Multi-disciplinarity challenge: Inherently, MIR represents an area of interest in music information, rather than a range of techniques and methods. This means that multi-disciplinarity is an ever-present challenge, even though a range of MIR tasks and procedures have been standardized and accepted by the MIR community.

It is most important to remember that no single algorithm or group of algorithms can ‘solve’ these tasks in the way that we are able to, with our ears and brain, enumerate a range of information from an auditory stream. However, it is promising how far music information tasks have come, and how much they have diversified by integrating with techniques such as AI.

Motion Capture and MIR

In their 2009 paper, Godoy and Jensenius make a case for a body movement based model for MIR tasks. The authors suggest that bodily sensations are a key feature of understanding musical style. In order for us to be able to harness this model of music listening for music-retrieval, movement-inducing cues, and taxonomies of multimodal responses to music. This thesis takes a step towards this goal by modeling body movement to melody.

6.2.4 Extraction of Pitch Contours and Contour Analysis

A pitch detection algorithm is used to estimate the fundamental pitch or frequency of a sound signal or musical recording. Generally speaking, pitch detection algorithms are in the time or frequency domains, or in both. Pitch detection algorithms vary considerably, depending on the task at hand. Extracting pitch from a monophonic signal is a relatively simpler task than extracting melodic contours from recordings of many instruments playing together, for which a variety of other strategies are used (Bittner et al., 2017; Salamon et al., 2012).

6.2.5 Pitch Detection Algorithms

Generally, two approaches can be taken towards pitch detection: time-domain and frequency-domain analysis. The following is an overview of some commonly used algorithms and their descriptions:

1. Zero crossing: The simplest idea for pitch detection for a quasi-periodic signal in the time domain is created by low-pass filtering the signal and then detecting peaks, or zero crossings. Linear predictive coding is often used to calculate f_0 using this method.
2. Auto-correlation: This method implements the auto-correlation method of Boersma (1993) and is available in Praat. Essentially, this algorithm involves computing correlation of a signal with a delayed copy of itself. Here, r is the autocorrelation function, w

$$r(\tau) = \int_{-\infty}^{\infty} \chi(t) \chi(t + \tau) dt$$

3. Cepstrum: is obtained by computing the inverse Fourier transform of the logarithm of the estimated spectrum of a signal, as defined below.

$$C(\tau) = |F(\log |F(x(t))^2|)|$$

4. Average Magnitude Difference Function: The AMDF pitch detector forms a function which is the compliment of the autocorrelation function, in that it measures the difference between the waveform and a lagged version of itself.

$$(\tau) = \int_{-\infty}^{\infty} |\chi(\tau) - \chi(t + \tau)|^b dt, b = 1$$

5. Spectral subharmonic summation: This method, also available in Praat, is described as a spectral subharmonic summation according to an algorithm by Hermes (1998). The idea of this algorithm is to arrive at a summation of each of the subharmonic components in the spectrum.
6. eSRPD: Enhanced super resolution pitch detector algorithm by (Bagshaw et al., 1993).
7. YIN: YIN pitch estimator is an improvement on the autocorrelation function with an additional step of a cumulative mean normalized difference function.

For the stimuli in the experiments, we use the YIN algorithm. The methods described above are to arrive at a representation of the pitch in a melody. Methods to model the contour, however, have to have a more general understanding of pitch trajectories.

6.2.6 Contour Analysis Methods

Although melodic contour has been established as an interesting and important feature of musical processing, only a few quantitative models or formal descriptions of contour have been proposed. I will outline some of them here:

1. Parson's Code for Contour Description Parson's directory for melodic indexing is based on the idea that every subsequent note in a melody can be represented using just three letters to denote direction: up (U), down (D), and repeat (R), to generate a unique contour description for each melody (Parsons, 1975). The book, which represents several thousand melodies in this way, claims to help people easily identify a piece of music from its melodic identity. For example, Parson's code for the theme of Beethoven's Fifth Symphony would be:

*RRD URRD

This seemingly parsimonious model is surprisingly good at identifying a large number of melodies, given all the possible permutations and combinations of a long enough phrase. This model is also used in Musipedia, an online repository with many in-built melodic search algorithms (Irwin, 2008).

2. Adam's Contour Typology Adam's contour typology, proposed in 1976, provides a way to think about melodic contours using the following features: the relative positions of the initial (I), final (F), highest (H), and lowest (L) pitches in a melodic line. These are described as the minimal boundaries of a melodic segment. Using these minimal boundaries, it is possible to identify a typology of 12 different ways in which these four features relate to each other, where three kinds of relationships are possible between each pair: $<$, $>$, or $=$. Using these 12 relationships, he defines a set of primary features of melodies: slope, deviation, and reciprocal. Here, *slope* is the relationship between the initial and final notes of a melody; deviation is a change of direction in the slope of the contour; and reciprocal is defined in terms of the first and only deviation. This gives us a total of 15 different primary features for melodic shape. Secondary melodic features or contour shapes contain finer details about a melodic fragment; the author explains this as being the difference between describing any rectangle, versus listing the properties of a particular rectangle. Repetition and the descriptions of minimal boundaries form secondary melodic features. The author also proposes the use of graphs to further elaborate melodic contours.

3. Morris' Model of Contour Relations This model presents an algorithm for reducing a complex melody to its salient contour, similar to the Upline of Schenkerian analysis (Katz, 1935), but without tonality as the central defining concept (Morris, 1993). This model presents 25 typologies of frequently occurring melodic contour patterns and another combination model.

4. Quinn's Combinatorial Model for Pitch Contour Quinn's combinatorial model 1999 creates a matrix representation of all note transitions to one another. These matrices can then be compared with each other using matrix similarity related methods, giving us a model of melodic similarity that does not rely upon accurate pitches being described.

5. Friedmann’s Contour Adjacency Series (CAS) and Contour Class model (CC) Both these models rely on creating a matrix for all the constituent notes in a melodic phrase, looking at their transitional probabilities from one to another.

6. Time Pitch Beat Model Time-pitch-beat, or a TPB number triplet is coded in this model for each melodic fragment. Subsequently, a resolution vector Q , records the intervallic changes between each consecutive note. Q is therefore an n dimensional vector, for $n+1$ notes in the melody (Kim et al., 2000). The TPB vector is then used to find melodic similarities.

7. Fourier Analysis Techniques Schmuckler (1999) proposes conducting a Fourier analysis of the sound signal as a method to compare the contours of different melodies. Since it shifts contour perception from the time to the frequency domain, and in the experiments, it has worked well. In Paper I, I analyze the melodies in the stimulus set using some of these models, to try to compare if they would work simultaneously for tracings and melodies.

6.3 Motion Analysis

Motion capture, often shortened as *MoCap* is a general term referring to a group of methods used to measure motion. There are a number of techniques that can be thought of as motion capture, and a recent overview can be found at (Jensenius, 2018). Verbal descriptions and annotations of movement are the simplest ways to capture motion; codes for describing gestures are often used in annotating gestures accompanying speech (Poggi, 2002). Photography, video analysis, and light-sensing have also been used to capture motion (Jensenius, 2007). The following methods for motion tracking are most important, and have been detailed by (Nymoen, 2013):

1. Acoustical tracking: The use of phase differences to measure reflections of sound, and hence, the changing distance of a moving object (Bishop et al., 2001).
2. Mechanical tracking: Tracking or measuring the angles between two mechanical parts that are changes the length of a resistor to infer the distance or angle of the tracked object, for example as used by De Laubier (1998). A joystick is a good example of this.
3. Magnetic tracking: Tracking the magnetic field of a moving electromagnet, measuring its distance from a stationary electromagnet (Bishop et al., 2001).
4. Inertial tracking: Using accelerometers and gyroscopes to track the position of an object based on its interaction with gravity (Bishop et al., 2001). Inertial Motion Units (IMU) are an example of this type of tracking (Tanenhaus and Lipeles, 2009), as are, for example Xsens suits.



Figure 6.2: Pictures from the FourMs MoCap Lab where the experiments are conducted. On the left is the lab with the cameras and speakers mounted as shown, on the right is a participant wearing reflective markers.

5. Optical tracking: Using infrared (IR), regular video, stereo video, thermal, or depth cameras to estimate an object's position in space. Using optical tracking requires post-processing using computer vision algorithms.

The main methods for motion capture in this thesis revolve around the infrared capture of motion data. Optical tracking systems with infrared cameras transmit and receive infrared light, measuring the location coordinates of light reflective markers. Motion capture systems work calibrating a known configuration of distances between markers and the shape of the layout, to lay down a grid. Thereafter, new markers can be introduced and computed by calculating distances as determined by each of the different cameras. This produces a time-sensitive trace of the infrared markers as seen by all of the cameras to measure movement accurately in space and time. Motion capture is used extensively to create geometric models of three-dimensional beings for animation and computer-generated graphics (CGI). Owing to its precision and sensitivity, it has been a great tool to analyze performing bodies.

6.3.1 Tracking Data through Infrared Motion Capture Systems

In infrared motion capture systems, tracking data are usually recorded as three-dimensional (3D) coordinates for every point that is tracked by reflective markers. The coordinate axes should be situated in space using calibration kits. The calibration kits are pre-programmed for the system to understand their placement and calculate the tracking of all other points using a camera system based on this calibration. This initial placement of the calibration kit results in the creation of a local three-dimensional coordinate system. Once this framework is in place,

the subsequent tracking of markers in the motion capture system is recorded along this three-dimensional local coordinate system (LCS), sometimes called the laboratory coordinate system.

On entering the lab, a participant is required to wear a MoCap suit, which is dark; reflective markers, representing the points that need to be tracked, are placed on the suit.

6.3.2 Details for the Experiments in the thesis

The experiments conducted in this thesis were carried out in the fourMs motion capture lab at the University of Oslo using a Qualisys Infrared Motion Capture System. The system consists of eight Oqus 300 cameras surrounding the space, and one regular video camera (Canon XF 105). In Qualisys, the LCS is usually a Cartesian coordinate system—the XY plane is usually the floor, and the Z axis represents vertical motion.

For the experiments conducted in this thesis, each participant wore a MoCap suit with 21 reflective markers on joints (Figure 6.3.3). The labeling scheme is explained in detail in Appendix A. Most of the studies in this thesis involve analyzing hand marker positions; however, the full body was tracked in order to retain additional information, in case of further analyses, and for qualitative analysis and visualizations of the entire body.

The system captures data up to 500 Hz. For these experiments, we used a capture rate of 200 Hz, because it is sufficient to record movement in this context. We also made a video recording of all the participants, so that we could hear the sounds played in the lab during post-processing.

6.3.3 Post Processing

Once the data were recorded, they were post-processed and cleaned before analysis. The post-processing phase consisted of marker labeling, removal of ghost markers, gap-filling, and smoothing. Thereafter, the data were exported to a file format that I discuss below.

Marker Labeling Marker labeling is essential to understanding which joint or part of body each marker represents. This was done using a list of codes—a shorthand system to easily understand which body part each marker refers to, and on which side of the body, as shown in Figure 6.3.3. The complete list of labels is explained in Appendix B.

Ghost Markers A MoCap system might often track ghost markers, mistaking small glitches or reflections as markers; these must be removed manually. Many ghost markers are present for very short amounts of time, making it easy to remove them.

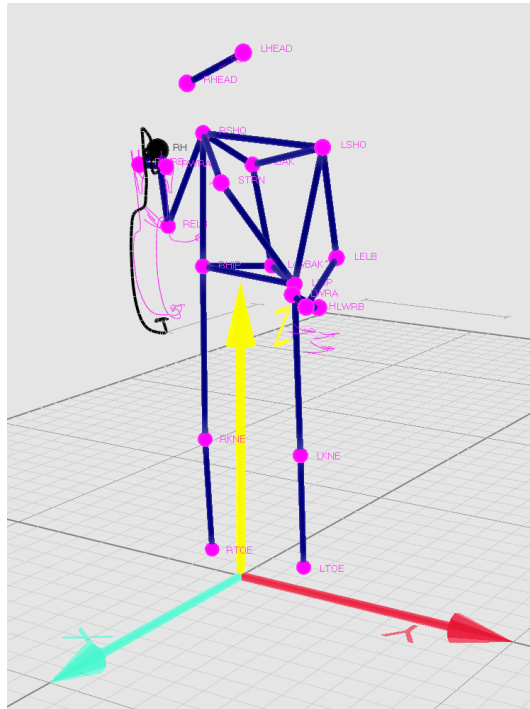


Figure 6.3: An example of a post-processed motion capture stick figure. A detailed list of marker labels can be found in Appendix B.

Gap Filling Lastly, MoCap recordings may contain missing frames in the data because of tracking errors, drops in the data packets sent over a network, occlusions, a moving body part overlapping a marker at a certain time point. Usually, these gaps are only a few frames long, and can be filled using interpolation and smoothing.

Common techniques for gap-filling include nearest neighbor interpolation, linear interpolation, or polynomial interpolation. Nearest neighbor interpolation looks for the nearest points around the gap, and does a step-interpolation between them. Linear interpolation aims to connect the available points before and after a gap using an interpolated line that passes through the points on either side of the gap.

In the post-processing of this data set, polynomial interpolation was used for gap-filling to ensure smooth trajectories.

6.3.4 File Formats for Motion Capture Files

Once the data are labeled, cleaned, and gap-filled, they can be exported into one of many formats for analysis. The Qualisys Track Manager (QTM) system is a tool offers visualizations of the data for inspection, but detailed analysis

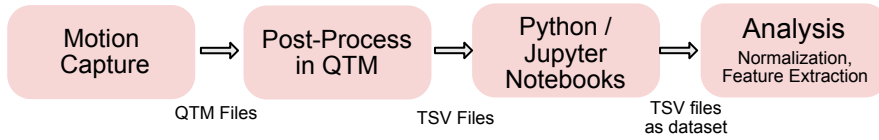


Figure 6.4: An illustration of the flow of mocap data. The mocap files are exported from QTM, imported into python. Normalized files are re-exported, and analyzed in python for obtaining features.

requires exporting the files into a scripting or programming environment. Below, I list some common file formats used for handling MoCap data.

1. TSV (tab-separated values): The output of this file format includes a list of captured attributes, such as capture rate, number of markers, length of capture, and marker list. Below these details, this format outputs a column per axis per marker of data. Each row represents a frame of capture.
2. C3D (coordinate 3D): C3D is commonly used as a standard file format in biomechanical research. The C3D specification expects physical measurements to be one of two types—positional information (3D coordinates) or numeric data (serial information). Each 3D coordinate is stored as raw (X, Y, Z) data samples with information about the sample—accuracy (the average error or residual) and camera contribution (which specific cameras were used to produce the data).
3. MAT (MATLAB files): The .MAT file format can be used to export data so that it is easily readable into MATLAB. The data are output in a STRUCT file, with matrices for the 3D coordinates of each point.
4. AVI (audio video interleave): Video files can be output from QTM into AVI. These are useful in video analyses.
5. BVH (bioVision hierarchy): Although QTM does not currently output to it, BVH is a common file format for 3D animation. BVH is compatible with a variety of software that use motion capture for artistic or animation purposes.

For our analysis, we chose to export data as TSV files. this made it easy to import them into Python, or Matlab, and analyze easily. Since the released dataset contains hand movement data, TSV also makes for a compact format. Figure 6.4 illustrates the flow of data for mocap through this whole process.

6.3.5 Feature Extraction

The motion capture data were imported into Python (v2.7.12) and MATLAB (R2013b, MathWorks, Natick, MA., USA). All of the 10-minute recordings were segmented using automatic windowing, and each of the segments were manually annotated for further analysis .

It is useful to understand the features from motion capture recordings to grasp how motion data can be processed and analyzed. In my own work, I like to differentiate between physical and perceptual features of motion. This is similar to the physical and psychoacoustic properties of sound features. While physical attributes refer to mathematical formulae directly applied to motion capture data, the perceptual features of motion identify something about the quality with which a movement is performed; for example, the perceived smoothness of a movement. For this analysis, we focus on physical features and not perceptual ones.

There may also be features that relate particularly to the task at hand. For example, in the experiments conducted for this thesis, most people used the hands as the ‘primary effectors’ for movement representations of melodies. This meant that the distances between hands, the symmetry between hands, and so on were important features that needed to be measured. For example, we compute the following features:

1. Velocity: The first derivative of position data describes the velocity of a marker. This can be described along any of the three dimensions.
2. Acceleration: The second derivative from position data describes the acceleration of a marker.
3. Jerk: The third derivative of position data describes the ‘jerk’. Jerk data typically represents sudden changes in acceleration.
4. Jounce: The fourth derivative from position data describes the jounce, describing sudden changes in jerk.
5. Quantity of motion (QoM): This describes the level of movement. It is calculated as the average of the vector magnitude for each sample.
6. Range: The range of a marker describes the minimum and maximum values of a marker for the duration of calculation across each axis.
7. Cumulative distance: The cumulative distance traveled is calculated as the summation of the Euclidean distance that each marker travels.

Features for Hand Markers

Table 6.1 includes a description of features specifically developed to analyze the motion data of hand markers. In Table 6.2, features are made specifically considering the nature of strategies used in data exploration, as is explained in Paper II.

	Motion Features	Description
1	VerticalMotion	z -axis coordinates at each instant of each hand
2	Range	(min, max) tuple for each hand
3	Hand Distance	The Euclidean distance between the 2D coordinates of each hand
4	Quantity of Motion	The sum of absolute velocities of all the markers
5	Distance Traveled	Cumulative Euclidean distance traveled by each hand per sample
6	Absolute Velocity	Uniform linear velocity of all dimensions
7	Absolute Acceleration	The derivative of the absolute velocity
8	Smoothness	The number of knots of a quadratic spline interpolation fitted to each tracing
9	VerticalVelocity	The first derivative of the z -axis in each participant's tracing
10	CubicSpline10Knots	10 knots fitted to a quadratic spline for each tracing

Table 6.1: The features extracted from the motion capture data to describe hand movements.

6.3.6 Toolboxes for MoCap Data

The BTK (Biomechanical ToolKit) has a stand-alone application—Mokka, or Motion Kinematic and Kinetic Analyzer (Barre, 2013)—which facilitates easy analysis of MoCap data on a graphical user interface. This application has tools for 2D and 3D visualizations. Mokka is built with Java, and provides a timeline view. C3D files can be imported, and Electromyography or EMG analysis can be integrated into Mokka.

MoCap Toolbox for MATLAB

The MoCap Toolbox, developed at the University of Jyväskylä, includes a variety of possible algorithms for feature extraction from MoCap data. This toolbox was developed especially for working with music or dance related motion; it can be used to calculate several motion features.

Python Toolboxes Overview

Several toolboxes in Python have been written specifically to analyze MoCap data. A small overview is listed below:

1. BTK/Biomechanical ToolKit: It was developed as a Python wrapper along with the stand-alone application Mokka.

6. Methods

#	Strategy	Distinguishing Features	Description
1	Dominant hand as needle	Right hand QoM much greater than left QoM	$QoM(LHY) \gg QoM(RHY)$
2	Changing inter-palm distance	Root mean squared difference of left and right hands in x	$RMS(LHX) - RMS(RHX)$
3	Lateral symmetry between hands	Nearly constant difference between left and right hands	$RHX - LHX = C$
4	Manipulating a small object	Right and left hands follow similar trajectories in x	$RH(x, y, z) = LH(x, y, z) + C$
5	Drawing arcs along circles	Fit of (x, y, z) for left and right hands to a sphere	$x^2 + y^2 + z^2$
6	Percussive asymmetry	Dynamic time warp of (x, y, z) of left, right hands	$dtw(RH(x, y, z), LH(x, y, z))$

Table 6.2: Quantitative motion capture features that match the qualitatively observed strategies. QoM refers to *quantity of motion*.

2. PyMo: This toolbox was created by Omid Alemi. It is a Python library for machine learning research on motion capture data. The analysis tools are written to process Blender BVH files.
3. C3D wrapper: This toolbox reads and writes C3D data into Python.

Additional Contributions

As a part of this thesis, I built a motion capture toolbox in Python. It is not fully developed yet, but I have included its data structures and functions in Appendix B.

6.4 Motion-Sound Analysis

Up until this point, I have described the analysis methods for (1) motion capture data and (2) audio files and music related data. In the next section, I will describe the methods used to interpret the data as melody–motion pairs, as suggested in (Godøy and Jensenius, 2009).

6.4.1 Visual Inspection and Visualization

Visualization and visual inspection are powerful tools for understanding multimodal data, especially while analyzing them together. While developing techniques to analyze new data, visualization is often the first method used to consider the possible patterns in the data. Good visualization techniques also help communicate the quantitative findings more comprehensively, both for other researchers and a general interested audience. Figure 6.4.1 demonstrates visualization of Quantities of motion for different melodies by the same participant.

In sound-tracing and other data handled in this thesis, the development of new models to analyze and explain the data depends upon how we can first explore and understand what lies in the data. In order to understand the critical elements present in data, visualization is an important first step. Motion data visualization can be tricky but important. The first reason for this is that the data are inherently multidimensional, and interpreting them requires us to resist seeing the motion tracings as human movements alone, but also to look at the nuances of the tracings. The second reason is that we often consider combining meta-level analytical features from these data to understand what is going on. For example, a simple 3D visualization may not bring out what an acceleration plot can.

Toiviainen and Eerola (2006) present an overview of visualization methods for various music information related tasks. Visualization techniques can also include methods such as principal component analysis (PCA) visualization, which offers a better interpretation of trends in the data. The authors then use self-organizing maps (SOMs), an unsupervised learning method to organize melodies in a hyperspace. An SOM calculates and arranges data points with respect to each other in the learning space.

6.4.2 Statistical Testing

Statistical hypothesis testing is a method of statistical inference drawn from a series of mathematical measures to accept or reject a hypothesis. There are many tests, and which test should be chosen according to the type of data, and the distribution of data. Since I have used t-testing in evaluating a portion of the data in Paper II, I describe this method shortly below.

t-testing

A *t*-test is one of the simplest tools to calculate whether the difference between random samples from two populations is statistically significant. The process involves defining a null hypothesis and stating that the means of the populations are equal; then, the null hypothesis is supported or rejected by the *t*-test. The power of a *t*-test lies in its ability to test competing hypotheses on a smaller sample set, by approximating the normal distribution function of the data set. A *t*-test that shows statistical significance will be able to establish a difference

6. Methods

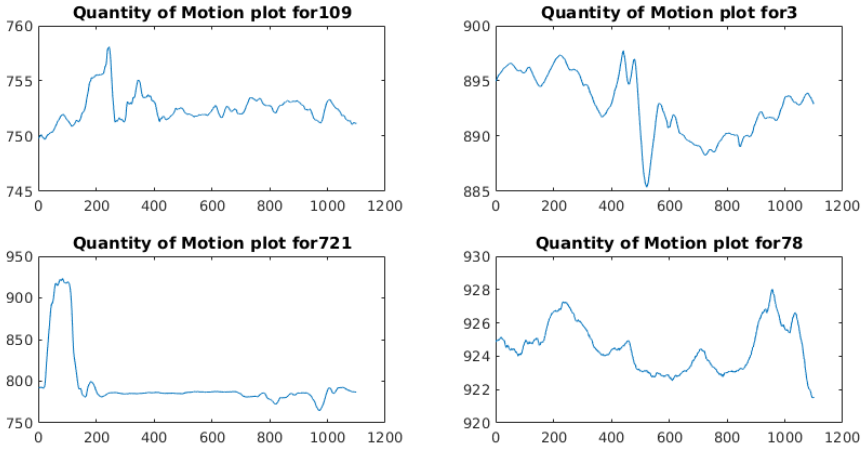


Figure 6.5: Visualizations of quantities of motion. Visual inspection

between two sample sets. The t -statistic is calculated using the following formula:

$$t = (\bar{X} - \mu) / \left(\frac{\sigma}{\sqrt{n}} \right)$$

Here, t is the test statistic, μ stands for the population mean, σ , the standard deviation, where \bar{X} is the sample mean from X_1, X_2, \dots, X_n are n samples in the data. From the t -statistic, the standard deviation and mean values are calculated. In statistical significance testing, the p -value is used to determine statistical significance by comparing a known probability distribution to the statistics obtained from the data, at a significance level that is most meaningful for the analysis. Typical levels for statistical significance are between 0.001 and 0.05, which gives us a confidence rating of 99 to 95 percent that the null hypothesis can be rejected.

6.4.3 Machine Learning and Artificial Intelligence

Broadly, machine learning describes a set of mathematical paradigms for performing discrimination and pattern recognition tasks. In the simplest of these tasks, a machine can tell apart two different categories of data, as could a basic perceptron machine, which was invented in the late 1950s. The evolution of algorithms and techniques within machine learning, an exponential growth in computational power, and an ever improving range of data sets have enabled this field to grow exponentially.

A widely quoted definition of machine learning is by Samuel (1959), [Machine Learning is a] “Field of study that gives computers the ability to learn without

being explicitly programmed”. The notion of explicit programming is to find conditionals in the data such that the computer can execute code without a model of the categories it is able to discriminate among.

Machine learning problems deal with data sets, which are representations of real world phenomena in terms of data. These data can come in many possible formats, and a learning task may be in one of several different classes of problems. The objective for learning a particular quality of the data also might vary—we might need to predict a variable in comparison to another, detect the presence of a certain characteristic, generate a string of words based on a prompt, and so on. The only reason all of these various classes of problems are included under the umbrella of learning is because, instead of explicit instructions, some variables, features, weights, or other paradigms are approximated by the computer. Feature sets are thus commonly encountered in this paradigm, referring to the properties of each data element. Machine learning pipelines often include ‘training’ and ‘testing’ phases—training is carried out on a smaller portion of the same data set, and the performance of the algorithm is tested on the rest of it.

Broadly speaking, here are some categories of machine learning methods (Bishop, 2006, p.3):

1. Supervised learning: These methods include pipelines where the training set includes category labels. This means that the learning happens through examples.
2. Unsupervised learning: When category labels are absent in the training, learning methods are described as unsupervised.
3. Reinforcement learning: This set of methods is based on the idea that an agent in the world learns from positive and negative feedback on its decisions.

The following can be described as typical learning problems:

1. Classification: An example of a classification task is to distinguish between different categories, given some features. The simplest task is a binary classification; for example, to differentiate between two instruments based on their sound streams. A classification task with multiple classes, for example to classify handwritten digits.
2. Regression: An example of a regression task is to predict the change of a variable with respect to a different variable. An example of a regression task is to analyze the correlation between melodic data and motion trajectories, to measure their covariance.
3. Clustering: A clustering task is an unsupervised method to represent the data in a hyperspace so that it separates into k clusters. An ideal clustering task would involve maximizing the distances between clusters and minimizing their distances from their respective centroids.

4. Retrieval: An example of this type of task is to find a melody, given a contour profile; for example, as seen on Musipedia, where a user can input a contour profile based on Parson's code (Parsons, 1975) of contour directions to retrieve melodies that fit the contour profile (Irwin, 2008).

Neural Networks and Deep Neural Networks A neural network is an architecture of artificial 'neurons' that simulates the inter-connectedness of neurons in the brain. Neural network architectures are a particular class of learning algorithms. Neural networks have also been applied for generating new material; for example, interpolated faces, or melodic material.

A neural network must be trained for a classification or generation task using a sufficient number of data examples. Usually, differences among architectures is how the error is propagated through the different neuronal 'layers' (the error can be passed forwards, backwards, in both directions, and so on), and other sub-networks can be added to specific layers. A deep neural network represents the addition of many more layers to the network architecture. The presence of several additional layers means that a network is better equipped to deal with more non-linearity.

Multimodal retrieval For the data sets and experiments described in the previous chapter, the analysis compares two types of data. In computation, this is referred to as multimodal analysis; for example, the joint analysis of image and text pairs. In the multimodal retrieval paradigm, different types of data are handled together. The objective is to learn a set of mapping functions that project the different modalities onto a common metric space, to be able to retrieve relevant information in one modality through a query in another. This paradigm is often used in the retrieval of image from text and text from image. The multimodal retrieval method relevant to this thesis is primarily canonical correlation analysis or CCA, explained in detail in Section 6.4.5

6.4.4 Time Series Distance Metrics

A simple way to analyze sound-tracing related movement data would be to compare different time series. Several tasks in the appended papers involve measuring the distance between two time series. In essence, time series are matrices. Several methods have been proposed for time series distance matching, which can be imagined as the extent of the variance between two different time series. Broadly, the methods can be classified as follows:

1. Reduced dimensionality measures: These are useful when time series are large or tightly sampled and therefore, time series distance cannot be measured by sample-to-sample distances. Reduced dimensionality measures can include, for example, discrete Fourier transform, and then a comparison between the transforms. Discrete cosine transform (DCT), Chebyshev polynomials are some ways to measure time series distance using reduced dynamics.

2. Distance measures: The easiest way to think about distance measures is to calculate the Euclidean or edit distance, which would be the distance per sample between each series. Edit distance measures can also be improved by introducing edit distance penalties, or using a weighted alignment before distance is calculated.
3. Longest common subsequence (LCS): LCS is a set of problems to detect the longest common subsequences in any two sequences of data. This problem can be adapted to string commonalities and continuous time series.
4. Data adaptive measures: Data adaptive measures try to fit first to the series data before computing distance; for example, polynomial interpolation or regression using piecewise polynomials or splines. In these methods, the approximated functions are then compared to find distances. For symbolic data, such methods can include substrings, case modification, or other kinds of approximation.
5. Non data adaptive measures: Wavelet transforms, spectral measures such as discrete Fourier transform (DFT), DCT, and piecewise aggregation approximation methods can be used.

Time Series Similarity Measures

We can also take another approach; instead of trying to find the distance between two time series in a hyperspace, we can measure their similarities. Broad classifications of similarity measures are as follows:

1. Lock step measures: Distance measures such as L1 L2 norms.
2. Elastic measures: DTW, LCS, edit distance, and sequence weighted alignment can be thought of as elastic distance measures because they aim to match subsequences or subsections of data.
3. Threshold based measures: TQuEST or the threshold query execution method.
4. Pattern based measures: Spatial assembling distribution, also called SpADe.

6.4.5 Canonical Correlation Analysis (CCA)

Canonical correlation analysis (CCA) is a common tool used to investigate linear relationships among two sets of variables. In a review paper, Wang et al. analyze several models for cross-modal retrieval (Wang et al., 2016). Firstly, CCA allows us to approach the data through real-value based methods. The method also allows us to approach the data in a pairwise retrieval paradigm. Just like unsupervised methods, the CCA method uses similar pairs to learn meaningful distance metrics between different modalities. CCA has also been

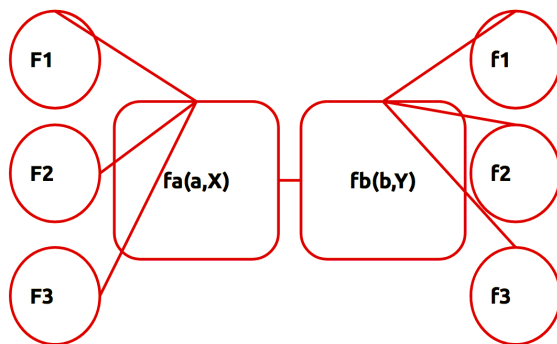


Figure 6.6: A representation of the CCA algorithm used; fa and fb represent the two different data sets—melodic and motion features, respectively.

used previously to show music and brain imaging inter-relationships (Berger and Dmochowski, 2017).

A previous study analyzing tracings to pitched and non pitched sounds also used CCA to understand sound-motion relationships. In the paper, the authors describe the inherent non-linearity in the mappings, despite finding intrinsic sound-action relationships (Nymoen et al., 2011). This work was extended in a new paper, where CCA is used to interpret how different features correlate (Nymoen et al., 2013). Pitch and vertical motion have linear relationships in this analysis, although it is important to note that the sound samples used were short and synthetic.

CCA has also been used in research related to the analysis and retrieval of music-motion (Nymoen et al., 2011, 2013; Caramiaux et al., 2009; Caramiaux and Tanaka, 2013). The biggest reservations in analyzing music-motion data using CCA is that non-linearity cannot be represented, and the method is highly dependent on time synchronization. The temporal evolution of motion and sound remains linear over time (Caramiaux and Tanaka, 2013). To get around this, kernel-based methods can be used to introduce non-linearity. Ohkushi et al. (2011) present a study that used kernel-based CCA methods to analyze motion and music features together using video sequences from classical ballet, and optical flow based clustering. Bozkurt et al. present a CCA based system to analyze and generate speech and arm motion for a prosody-driven synthesis of the ‘beat gesture’, which is used to emphasize prosodically salient points in speech (Bozkurt et al., 2016). CCA is used to explore the data sets in this thesis due to the previous successes associated with using this family of methods. The same data are also analyzed using deep CCA, a neural network approximation of CCA, to better understand the non-linear mappings.

CCA is a statistical method used to find linear combinations of two variables, $X = (x_1, x_2, \dots, x_n)$ and $Y = (y_1, y_2, \dots, y_m)$, with n and m independent variables as vectors, a and b , such that their correlation, $\rho = \text{corr}(aX, bY)$, of the transformed variables is maximized. Furthermore, more linear vectors, a' and b' ,

can be found—they maximize the correlation and are not correlated with the previous transformed variables. This process can be repeated till $d = \min(m, n)$ dimensions.

6.4.6 Deep CCA

Deep CCA is a neural network based approximation of CCA. By introducing neuronal layers, we are able to approximate non-linearity in the problem. The CCA can be extended to include non-linearity by using a neural network to transform the X and Y variables, as in the case of deep CCA (Andrew et al., 2013). Given the network parameters, θ_1 and θ_2 , the objective is to maximize the correlation, $\text{corr}(f(X, \theta_1), f(Y, \theta_2))$. The network is trained by following the gradient of the correlation objective as estimated from the training data.

6.4.7 Template Matching

Template matching represents a series of methods from image processing that are used for tasks where a stimulus is chosen as a template to match to other query images. This ensures that the algorithm retains some robustness as a result of searching only for relevant template-related features; it helps to avoid such problems as noisy backgrounds. In Müller’s book on information retrieval for motion data, a method is described for using motion data to create motion templates, which act as compact and explicit matrix representations of motion data for a search and retrieval algorithm (Müller, 2007).

6.5 Summary

In this chapter, I present a summary of the methods used first in sound analysis, then motion analysis, and in a joint analysis of sound and motion. Sound analysis deals with several levels, starting with storing the data as sound files, signal processing algorithms for pitch detection, and methods for contour representation analysis. I summarize motion capture technologies, focusing on infrared motion capture, and how data is obtained and analyzed. To cross-correlate the data, I use a range of methods from statistical hypothesis testing to canonical correlation analysis. Other methods for comparing multimodal time-series data have also been explained for context. I have presented an overview of the computational tools and possibilities, and their contributions to this thesis. Using these technologies, the datasets are analyzed, and a summary of the papers, and a discussion of the results is laid out in the next chapter.

Chapter 7

Conclusions

*Embodied cognition is all fine, but tell me
whose body are we talking about?
- A well-meaning friend*

7.1 Research Summary

Using the methods as described above, four papers form the core of this thesis, and are appended towards the end. Each of the four papers explores a dimension of melodic contour: verticality, motion metaphors, body use, and multi-feature correlational analysis. Some emergent findings from the data are also discussed in detail, relating to: verticality, imagery, voice, body use, and cultural considerations. I hope the analysis can inform some areas for research in melodic perception, as well as building systems for indexing, and generating music.

7.1.1 Paper I

Reference: Kelkar, T., & Jensenius, A. R. (2017). Exploring melody and motion features in “sound-tracings”. In Proceedings of the 14th *Sound and Music Computing Conference* (pp. 98-103). Aalto University.

Abstract

Pitch and spatial height are often associated when describing music. In this paper, we present results from a sound-tracing study in which we investigated such sound–motion relationships. Subjects were asked to move as if they were creating the melodies they heard, and their motion was captured with an infrared, marker-based camera system. The analysis focused on calculating feature vectors typically used in melodic contour analyses. We used these features to compare melodic contour typologies with motion contour typologies. This is based on using proposed feature sets that were made for melodic contour similarity measurements. We applied these features to the melodies and motion contours to establish whether there is a correspondence between the two, and to find the features that match the most. We found a relationship between vertical motion and pitch contour when evaluating according to features, rather than simply comparing contours.

Discussion

The first question we explored was whether models of melodic contour that have been explored in previous research Schmuckler (2010, 1999) can be applied to movement data. The reasoning behind this is to test the assumption of pitch verticality in a full-body free hand sound-tracing study. We test three contour features that describe many contour models: 1. Signed interval distances, describing the directions of subsequent notes; 2. Initial, final, highest, lowest vector containing these four notes from the melody, and 3. Signed relative distances, including the intervals in semitones and the directions. We try to perform a simple retrieval experiment for the 3rd feature with k-nearest neighbors, and find that the recognition accuracy is not significant for any contour profile. Although these features appear in many contour models, vertical motion alone does not sufficiently explain tracings drawn by the participants. In qualitative observations we notice that people use several types of metaphorical representations instead. This is investigated in Paper II.

7.1.2 Paper II

Reference: Kelkar, T., & Jensenius, A. R. (2017, June). Representation strategies in two-handed melodic sound-tracing. In Proceedings of the 4th International Conference on Movement Computing (p. 11). ACM.

Abstract

This paper describes an experiment in which subjects participated in a sound-tracing task to vocal melodies. They could move their hands freely in the air; their motion was captured using an infrared, marker-based system. We present a typology of the distinct strategies that the participants used to represent their perception of the melodies. These strategies appear to be ways to represent time and space through the finite motion possibilities of two hands moving freely in space. We observed these strategies and present their typologies through qualitative analysis. Then, we numerically verified the consistency of these strategies by conducting tests of significance between labeled and random samples.

Discussion

In this paper, we describe the main strategies used by most participants to express melodic contour in motion metaphors. As discussed in the previous paper, vertical motion is related but not sufficient to describe sound-tracings of participants. This is because people use various metaphors to represent the changes they hear in melodic motion. By metaphors, I am referring to the use of bimanual movement as if the two hands were carrying or manipulating objects having different properties. We isolate 6 metaphorical strategies used, and demonstrate how these can be quantitatively differentiated from one another. We perform automated windowing of tracings in the motion data of each participant,

and using statistical hypothesis testing, calculate how different metaphors can be identified.

7.1.3 Paper III

Reference: Kelkar, T., & Jensenius, A. (2018). Analyzing free-hand sound-tracings of melodic phrases. *Applied Sciences*, 8(1), 135.

Abstract

In this paper, we report on a free-hand motion capture study in which 32 participants ‘traced’ 16 melodic vocal phrases in the air with their hands, in two experimental conditions. Melodic contours are often thought of as correlating with vertical movement (up and down) in time—this was our initial expectation. We found an arch shape in most of the tracings, although this did not directly correspond to the melodic contour. Furthermore, the representation of pitch in the vertical dimension was but one of a diverse range of movement strategies used to trace melodies. Six different mapping strategies were observed; these strategies were quantified and statistically tested. The conclusion is that metaphorical representation is much more common than a ‘graph-like’ rendering in such a melodic sound-tracing task. Other findings include a clear gender difference in some of the tracing strategies, and the unexpected representation of melodies in terms of a small object for some of the north Indian music examples. The data also show a tendency of participants to move within a shared ‘social box’.

Discussion

In this article, we analyze Dataset 1 further in order to understand the distribution of the strategies of motion metaphors identified in the previous article. We find that there is a clear arch shape when looking at the averages of the motion capture data, regardless of the general shape of the melody itself. This may support the idea of a motor constraint hypothesis that has been used to explain the similar arch-like shape of sung melodies. I will explain this in more detail in section 7.3. We find a gender difference for some of the strategies. This was most evident for representing small objects using both hands, which women performed more than men. The use of this strategy was also found to be more common for melodies from north Indian music even when participants who had no or little exposure to this musical genre. This type of metaphor has previously been reported to have been used by practitioners of this genre (Clayton, 2001). The data show a tendency of moving within a shared ‘social box’. This may be thought of as an invisible space that people constrain their movements to, even without any exposure to the other participants’ tracings. We find an inverse relationship between the participants’ heights and the use of the extreme periphery of the body to represent the highest notes of the melodic stimuli.

7.1.4 Paper IV

Reference: Kelkar, T., Roy, U., & Jensenius, A. R. (2018). Evaluating a collection of Sound-Tracing Data of Melodic Phrases. In *Proceedings of the 19th International Society for Music Information Retrieval Conference*, Paris, France. (pp. 74-81)

Abstract

Melodic contour—the ‘shape’ of a melody—is a common way to visualize and remember a musical piece. Studies on sound-tracings investigate the relationship between people’s ability to draw such melodic contours on paper or in the air. Understanding such relationships between music and motion is interesting from a cognitive perspective, and it can also be useful in the development of retrieval systems. The purpose of this paper was to explore the building blocks of a future ‘gesture-based’ melody retrieval system. We present a data set containing 16 melodic phrases from four musical styles, with a large range of contour variability. This is accompanied by full-body motion capture data of 26 participants performing sound-tracing to the melodies, which have multiple label sets. The data set was examined using canonical correlation analysis (CCA), and its neural network variant (deep CCA), to understand how melodic contours and sound-tracings correlate. The analyses reveal non-linear relationships between sound and motion. The link between pitch and verticality does not appear strong enough for complex melodies. We also found that descending melodic contours have the lowest correlations with tracings.

Discussion

In this paper, we analyze Dataset 1 from a multimodal retrieval approach. This method has been used in previous studies to analyze sound tracings (Nymoen et al., 2013). We introduce a new feature for calculating the longest run lengths of individual tracings to improve the correlations of the system. We test different category labels for melodies. Genres show the least amount of agreeability and improvement. With all melody and all motion features, we find an overall correlation of 0.44 with Deep CCA, for both the longest ascend and longest descend features. This supports the view that non-linearity is inherent to tracings. This will be explained further in 7.3.

7.2 Discussion

In Chapter 1, I established that the key research question of the thesis was to understand the role of embodiment in melodic contour perception using sound-tracing as the experimental methodology. I elaborate on the findings below.

RQ1. How do listeners represent melodic motion through body movement, and how can we analyze motion representations of melodic contours?

From the experiments conducted, we find that both expert musicians and non-experts are able to trace melodies that they hear. The relative ease of the task points to the fact that representing melodic stimuli through visual and gestural imagery is not a task that requires specialized training or access to musical knowledge. Most participants reported that the task became rather intuitive to them as the experiment progressed, even if they anticipated it being difficult. This confirms the findings of previous studies, as well as the hypothesis that the connection between visual or motor imagery and melodic stimuli is quite obvious. All participants were able to perform every melody, outliers in the data sets are purely results of marker dropouts or other technical difficulties.

In Paper III, while doing qualitative analysis of the data, we found differences in body use that had little to do with the experiments themselves. That body height influences how participants trace melodies is a new finding, where height negatively correlates with the perceived pitch–height of the highest musical note. Previous sound-tracing studies have often looked at tracings on a tablet, thus missing the context associated with representing melodic movement using the whole body.

RQ2. What are characteristics, and applications of motion representations of melodic contour?

An important characteristic of sound tracings in this context is the representation of melodies as metaphors, where the movement resembles an interaction with imaginary object that act as stand-ins for the perceived movement in the music.

In Paper I and II, I analyze and quantify metaphors found in melodic sound-tracings. Specifically, in the Paper I, I compare symbolic contour typologies and motion features in sound-tracings to see if vertical motion justifiably explains the shape of contour tracings. Evidently, when the paradigm of melodic representation moves from vertical placement to motion representation, contour representations change.

Although much of the literature points to correlations between melodic pitch and vertical movement, our findings reveal a much more complex picture. For example, relative pitch height appears to be more important than absolute pitch height. People seem to think about vocal melodies as actions, rather than interpreting pitch purely in one (vertical) dimension over time. The literature on contour features emphasizes that while tracing melodies through an allocentric representation of the listening body, the notion of pitch height representation matter much less than previously thought. Therefore, contour features cannot be extracted merely from cross-modal comparisons of two data sets. We propose that other strategies can be used for the analysis of contour representations, but this will have to be developed in future research.

7. Conclusions

One interesting finding is that there are gender-based differences in participants' sound-tracing strategies. Women seem to show a greater diversity of strategies in general, and they use object-like representations more often than do men.

RQ 3: How can we test if motion related to melodic contours is consistent: a) within participants, and b) across participants?

In visual inspection of the data set, we find that body use within participants is significantly different, depending upon a variety of factors. To study this, we perform a correlational analysis of many features with each other: melody labels, contour profile, genre, presentation (recorded vs resynthesized melodies), and to detect the presence of motivic repetition and vibrato.

In Paper IV, we study the correlations among all participants for the above mentioned categories of stimuli. Through these patterns we find that the highest agreement was for representation of a vibrato, motivic repetition, and for melody labels. This means that within participants, these are the features that are most similarly traced.

In Paper III, we also present a frequency diagram for representation strategies used by participants for tracing melodies, and find that there are some general trends with some melodies represented more often by some strategies.

The control of various melodic features seems to follow a hierarchy in most participants. Vowel changes are most often represented by palm shapes, whereas melodic leaps are more often represented at the shoulder joint level. The scale at which melodies are represented relative to the body is also consistent across participants.

In B, I explain the python functions created to analyze the features of the melody-motion pairs.

RQ 4: Can we build a system to retrieve specific melodies based on sound-tracings?

In Paper IV, I present the beginnings of a CCA-based system to retrieve melodies from sound-tracings. I have tried to integrate the findings from the empirical experiments with an experiment on melodic indexing and contour retrieval through body movement. Two versions of canonical correlation analysis are used to demonstrate the highly correlating factors between melodies and their respective motions.

Previous studies in this area include shorter sound stimuli or synthetically generated isochronous music samples. The strength of this particular study is in its use of shorter melodies, and that the performed tracings are not iconic or symbolic, but spontaneous. As such the data set might bring us closer to understanding contour perception in melodies. Hopefully the data set will prove useful for pattern mining—it presents the community with novel multi-modal possibilities, and could be employed in user-centric retrieval interfaces.

7.3 General Discussion

This thesis has tied experimental methods in embodied music cognition to computational analysis of melody. Within computational musicology, methods that use sheet music or symbolic data exclusively limit us to certain musical styles, kinds of analysis pertaining to certain inferences drawn from tonality, and so on. There is also a significant bias toward working with music data sets in generative and interactive music as these data sets are more available. Despite the fact that most extensive data sets are available in MIDI and symbolic notation, I have specifically addressed the limitations of studying contour perception using symbolic methods in Paper I of this thesis.

I will now discuss four main aspects of the findings: 1. Verticality, 2. Imagery, 3. Voice, and 4. The Body, and discuss their details.

7.3.1 Verticality

Arch-Shape In Paper III, we plot the average vertical trajectory of each melody, to discover that the the average vertical trajectories all resemble an arch shape. It has been speculated before (Savage et al., 2017) that an arch shape represents the shape that would support the motor constraint hypothesis, that the effort required to produce songs would follow an arch shaped trajectory. It is interesting to note that this shape is found even in purely ascending or purely descending melodies.

Having said this, there are differences within the contour and acceleration profiles of what is drawn between the ascent and descent. A significant thing to note here is that this shape is observed only for the vertical profile of contour tracing. Although vertical profile is where we suspect the most amount of motion corresponding to contour, this is not what we actually see in the data.

Extreme Periphery How close to the body do traced melodies lie? We find that the use of the space around the body is closely related to pitch height. However, we find an inverse relationship between the participants' heights and the use of the extreme periphery of the body to represent the highest notes of the melodic stimuli. It is interesting to note that this region around our bodies is reserved for the very highest notes, and some participants also reach beyond this region by extending their toes, for example. Similarly, melodies seldom extend below the center of mass, there are very few examples in the data of participants stopping down to represent contour.

Relative vs Absolute Representation For sound tracings relative pitch height appears to be more important than absolute pitch height. This means that the same height in space might be used to represent two very different pitches depending upon the ambit and the context of the melodic fragment that is traced. Our ability to zoom-in on context is highlighted by this result. This could tie back into the arch-shaped average tracings, because in essence every sung phrase occupies then, the same procession of events.

7. Conclusions

Research points towards gesture and body movement encoding an ‘effort’ dimension. As we ascend up in our own pitch range, we need more effort to be able to reach higher, and this effort can be perceived both in the sound as well as the movement. Representing relative as opposed to absolute pitch might be related to the perception of effort in a singular phrase.

Non-Linearity In Paper IV, we present a CCA of various features of the melodies. Nymoen et al. (2013) previously highlighted non-linearity in sound tracing features. This points towards two things: 1. Features take precedence over one another, given that sound tracing is already an act of dimensionality reduction, and 2. Context-sensitive features make absolute correlational analysis not yield good results.

When trying to model tracings, a non-linear network gave better results than a vanilla version, as described in Paper IV.

Contour Properties The above results indicate that verticality in contour tracings is not the most important aspect, at least in sound tracings. Previous studies have shown that verticality is just one dimension of pitch-mapping for us, and variations over languages and cultures exist (Eitan and Timmers, 2010). In observations of tracings, we find that the general arch shape also exists as phrase-termination annotation—the hands are placed at a location that represents the beginning of the phrase, and brought down once the phrase is completed. However, contour features are represented in the middle of a sharp initial ascent and a sharp final descent in the movement. In order to focus on contour features, acceleration and jerk plots are more useful to analyze.

7.3.2 Imagery

Experience with dance, sign language, or musical traditions with simultaneous musical gestures (such as conducting) all contribute to the interpretation of music as motion. Had we chosen participants from any one pool of experience with music-motion, it may have been easier to model sound-tracings, but despite the diversity of participants’s experiences, there is a considerable amount of consistency between subjects. Research has also shown that the representation of time itself depends upon linguistic factors—speakers of different languages conceive the physical representation of a ‘timeline’ very differently from each other (Boroditsky, 2000; Fuhrman et al., 2011). The agreement of which directions around the speaker time seems to travel could influence how melodies traveling through time are represented and shaped.

Canvas We see that positing a ‘timeline’ relative to the body is implicitly imagined while performing this task. Some might adopt the strategy of a timeline ahead of them, much like a sketchbook or a graph, while others might imagine a cylindrical canvas around their bodies. This allocentric placement of melody on

a canvas in front of the body is consistent for each participant's tracing data, but varies across participants.

Metaphor The idea of movement as metaphor, in the context of this discussion borrows a lot from the conceptual metaphors in speech-gestures. The idea is that we refer to conceptual objects as if they had spatial properties and locations. We might say 'this idea', indicating to something as if it were present in our hands. In the qualitative observations of the study, it became clear that melodic movement could easily be represented in this form.

In Papers I and II, I have delved deeper into the gestural metaphors used to represent the sense of movement in melody. The use of metaphorical gestural representations, as shown, is not unusual in the sound-producing movements of some vocal styles (Paschalidou et al., 2016; Pearson, 2016).

According to the gestural affordances of musical sound theory (Godøy, 2010), several gestural representations can exist for the same sound, but there is a limit to how much they can differ. In other words, gestural representations might vary a lot, but that does not mean that any gestural representation response can perfectly fit a single stimulus. Our data support this idea of a number of possible and overlapping action strategies. Several spatial and visual metaphors are used by a wide range of people, such as the use of objects with various properties such as rigidity, elasticity, and so on.

Looking at the features of Melody 4, the intervals steadily descend, then they ascend, and finally come down again in the same step-lengths. This arguably resembles an object that slips smoothly along a slope, and could be a probable reason for the overwhelming representation of this particular melody as an object. In future studies, it would be interesting to see whether we can recreate this mapping in other melodies, or understand how perceived melodic motion can resemble the sounds generated by physical motion of objects around us.

7.3.3 Voice

We expected musical genre to have some impact on the results. For example, given that western vocalizes are sung in a higher pitch range than the rest of the genres in this data set, as expected, on average, people represent western vocalize tracings spatially higher than tracings in the other genres. In acceleration plots for each melody, this difference is stark.

Motif As explained in the appendix, some melodies had a repeating motif. In Paper IV, CCA analysis reveals that one of the best performing indices of the network is for distinction between phrases containing motifs and those not containing motifs. Motivic repetition reflected in tracings is an indication of consistency in the imagery of contour.

Vowels Participants were tasked with imagining as if their movements are sound-producing, which can have some connotations of 'control' through

7. Conclusions

movement embedded in it. We find that typically, palm shapes are used to identify changes in vowels of the melody, whereas pitch leaps are represented in joints closer to the body, such as the shoulders. When vowels became absent, as in the case of resynthesized melodies, the formant shape is interpreted as being small and narrow.

Vibrato We found that melodies with the maximum amount of vibrato (Melodies 14 and 16 in Figure 5.2) are represented with the largest changes in acceleration in the motion capture data. This implies that although the pitch deviation in this case is not so large, the perception of a moving melody is a much stronger influence on contour shape compared to melodies that have larger pitch changes. It could be argued that both Melodies 4 and 16 contain motivic repetitions that cause this pattern. However, repeating motifs are as much a part of Melodies 6 and 8 (joik). The effect of the vowels used in these melodies can thus be negated, due to the similarities between the tracings for original and resynthesized stimuli. There are some melodies that stand out for use of certain hand strategies. Melody 4 (north Indian) is, curiously, primarily represented as a small object.

Phrasing What constitutes a melodic phrase, and how to define the boundary conditions of melodic phrases is a problem with many possible solutions. If we were to use sound-tracing as an annotation system for not only non-western, but also non-standard or non-notatable musical pieces, we would be able to create a model of melodic phrasing that is more related to how we listen rather than to just understanding rules about tonality and pitch relationships.

7.3.4 Body

It is worth noting the several limitations of the current experimental methodology and analysis. Any laboratory study of this kind should present subjects with an unfamiliar and non-ecological environment. The results are presumably also, to a large extent, influenced by the habitus of body use in general—the direction of written scripts in the native languages of the participants, their familiarity with scientific graphs, and so on.

The extreme periphery of the body, reached in this case by extending the shoulders, seems to be infrequently used simply for participants that are taller. In Paper III, I call this effect as a ‘social box’ wherein people try to occupy the same space regardless of the differences in their own body dimensions.

Gender We found some differences between the tracings belonging to men and women. Women had a greater diversity of hand strategies in general, as demonstrated in Paper III. Moreover, some strategies are used more frequently by women than men, most interestingly that of representing the movement in melodies as an imaginary small object. It is unclear why this might be, and beyond the scope of this project.

7.3.5 Cultural Considerations

Despite using melodic stimuli from different musical cultures, the objective was not to find cultural differences or musical universals. The research presents embodied cognition as an approach that can help situate melodic perception in the body. I have tried to address the *why* of studying contour typologies earlier by examining what makes them interesting for study.

I would like to borrow the *why* of studying contour typologies to further discuss the intentions of data collection practices in computational musicology that might have underpinnings that we might forget to consider while analyzing the data. The fields of computer science and systematic musicology are currently diversifying and there is a growing interest in understanding ‘other’ cultures. In an ever growing, more frequently migrating world, our need to apply theory to every kind of practice is urgent. This means that the links between methods and models, and their origins and history, weaken over time as technology advances. The original motivations for building systems for musical indexing and melodic modelling, are very different from the contexts in which the data are currently used—primarily to find cultural differences or musical universals.

In his 1983 book, *The Study of Ethnomusicology*, Nettl reminds us that the study of musical universals was popular among nineteenth century musicologists, such as Wilhelm Wundt; this line of inquiry made a comeback around the 1970–80s. The idea of looking for universals persists strongly in computational musicology, which offers several tools to understand the ‘feature dispersion’ of the musics of the world.

The methodical study of melodic contour extends further back in history than modern computational methods, which model pitch contours in different ways and apply the results to generalize and generate music. Early researchers have noted that melodic contour is the most stable element in differentiating melodic identities, styles, and canons (Dowling, 1972). Melodic contour typologies have also been proposed in several of these early works, along with generalized descriptions of contour behaviors. Much of the early literature on melodic contour analyzed them using cultural tropes such as ‘white’ or ‘primitive’ music, as is seen in many works in ethnomusicology (Nettl, 1956; Sachs, 2012; Roberts, 1922). Curt Sachs’s 2012 work on melodic contours and ideas about melodic cascades, in the *Wellsprings of Music*, have since fallen out of fashion. The book did not age well because of the presence of many misguided cultural meta-theories in the writing. Susanne Youngerman critiqued his work in a detailed article that I paraphrase in this section 1974. Youngerman’s article still serves as a concise warning, against using culture as a broad term to mean different methods of aesthetic appreciation. A longer description of various challenges can be found in Appendix A.

In studies of cultural differences in early ethnomusicology, most analyses are conducted with the help of notated music, whether or not notated music applies to the musical form or musical culture in question. Contour typologies of all kinds are based on fixed note positions, even if the instrument does not produce notes in that manner, for example, the voice. This aspect of the study of

7. Conclusions

melodic contours stays with us even today. Through the methods in this thesis, I have demonstrated that looking at melodic contours from the perspectives of sound-signal and body motion can provide insight on various perceptual patterns associated with melodic listening.

While trying to understand the history of contour classification, we cannot adopt a perspective that elides the goal of contour typologies—to essentialize cultural-melodic phenotypes through qualitative methods, one of which is contour typologies. I would like to state that developing contour typologies is not a goal of the research conducted for this thesis. It is instead to see how contours are understood and represented as shapes.

7.4 Impact and Future Work

The studies conducted for this thesis could inform several areas: 1. research in melody and prosody perception, 2. search and retrieval systems for music, 3. interactive music generation

7.4.1 Research in Melody and Prosody Perception

Treating sound-tracings as annotation material provides a new way of thinking about music analysis and studies on segmentation and phrasing. Through the experiments conducted, I have shown that people demonstrate a remarkable amount of consistency in their spatial and motor representations of music and sonic objects.

Experiments akin to the ones conducted for this thesis could inform our understanding of melodic phrasing, and the differences in people’s phrasing boundaries in various musical contexts. This could help us build a more sophisticated understanding of pitch perception in musical or melodic contexts.

7.4.2 Search and Retrieval Systems

Several projects on melodic content retrieval using intuitive and multimodal representations of musical data have been developed for use on different data sets. The oldest example of this is the 1975 project, ‘Directory of Tunes and Musical Themes’, in which the author used a simplified contour notation method involving letters to denote contour directions to create a dictionary of musical themes, from which one may look up a tune they remember (Parsons, 1975). The descriptions of this book suggested that people would always be able to look up a tune stuck in their head so long as they could describe the contour direction of the melody.

Parsons’ model has been adopted for melodic contour retrieval by Musipedia (Irwin, 2008). Another system has been proposed in the recent project called SoundTracer, in which the motion of a user’s mobile phone is used to retrieve tunes from a music archive (Lartillot, 2018). A critical difference between these approaches is how they handle contour and musical mappings, especially variations in time scales and time representations. Most of these methods do

not have ground truth models of contours, and instead use one of several ways of mappings, each with its own assumptions.

7.4.3 Interactive Music Generation

Interactive systems to generate music—that fill in the gaps, as it were, from skeletal notions of melodic contour—are popular. Typically, the interface involves a way for a user to generate a contour, on to which generative models are mapped, to give a possible musical solution. Several studies on embodied music cognition investigate the idea of drawing or tracing as a method of understanding the mappings of shapes and lines to musical material. Common modes of interaction for explicitly melodic interfaces include tracing, drawing, sketching, and hand-waving for contour depiction. Mappings from these contour depictions take various forms.

Generated melodies are usually harder to evaluate for goodness and fit in music generation, than, for example harmonic adherence to a particular style. Depending on the task at hand, melodic generation tasks may or may not require continuity and development, and some idea of self-similarity—a musical piece generally has a sense of identity maintained through variations that within a range of thematic . Although melodic generation spans a large range of mathematical, parametric, and learning methods, generally melodic generation is constrained to the symbolic domain. Majority of the systems for generation are thus in the symbolic domain, and inapplicable to musical cultures and styles where either datasets for symbolic music do not exist; or symbolic generation is insufficient to capture the musical style.

Typically, gesture–melody interfaces start with some assumption of melodic grammars, goals, strategies, styles; gestural interfaces start with some baseline measurements of user actions. There is a mid-level representation of the input and output, at which point mapping might be carried out. The goals of some of these instruments might either be the real-time generation of melodies or complete songs based on one or a few factors mapped from the gesture. There are also differences in generation, based on whether MIDI output is preferred to continuous-pitch output. Continuous and particularly on-the-fly interfaces often do not even use a melodic model, preferring instead to map a range of inputs to pitch and timbre characteristics, which places the interface control fully in the hands of the user. There is, therefore, a spectrum from fully generative systems with little mapping control to fully mapped systems with no generation model, between which gesture–melody interfaces lie.

The arrival of Magenta pre-trained melodic generation models has made the mid-level of melodic generation more accessible to several post-Magenta papers. A proof-of-concept for exploring this was recently published as part of Google’s Magenta project offshoots (Donahue et al., 2018); contour enumeration happens through a series of press buttons, reducing the 81 piano keys to four, and mapping improvisation using generation on top of these contours. A system of sketching melodies on paper is also described in Kitahara et al. (2017), where people use a tablet to draw melodic contours, on a grid that resembles the

7. Conclusions

piano roll format. TrAP uses a tablet-based interface to map four types of contour profiles by segmentation (Roy et al., 2014). The MicroJam application by Martin et al. includes a sketch-based interface to input melodies and rhythm; the application can also be used for collaborative performances and modeling (Martin and Tørresen, 2017).

He is an interface for generating melodic material based on brush calligraphy (Kang and Chien, 2010). The system uses computer vision techniques to analyze brush stroke features and ink thickness and map them on to a pentatonic scale. The Melodic Brush interface by Huang et al. is another example of the use of calligraphy inputs, with the specific intention of generating Chinese music, using a kinect depth camera (Huang et al., 2012).

Interactive interfaces for generating music use multimodal and movement-based methods very often. Hopefully the material in this thesis provides new ideas in this area. To extend some of the work in this thesis is also my personal goal for future work.

7.4.4 Future Work

The analysis of Dataset 2 and Dataset 3 have not yet been published yet in the time constraint available. In the future, I plan to analyze these and publish them. I also plan to release the results of the full body data collected. The tools and notebooks released as a part of the thesis also will be improved to accommodate various use cases of motion capture data for musical motion. Similarly, I am in the process of exploring the movement of other body parts except the hands, and what we can infer from, for example mirroring of hand movements with head movements, and so on.

Although canonical correlation is a way to model and understand correspondences between melody and movement, there is no gold standard, or right answer for what movement is to be performed, and how much does it fit. The only way we know when we see a cartoon film where music accompanies motion, and so on, is that some features in the sound and movement become naturally corresponded to us. In order to bring this out in the sound tracing paradigm, I wish to create a system to synthesize melody—motion pairs by training a network to analyze this data set, and to conduct a study in which users evaluate system generated music—motion pairs in a forced-choice paradigm.

Speech prosody and its relationship with musical melody is also an area that I have grazed upon several times, but not in the experiments. In the future, I would like to incorporate some of the findings related to metaphorical movement back into analysis of speech gesture.

Bibliography

- C. R. Adams. Melodic contour typology. *Ethnomusicology*, pages 179–215, 1976.
- G. Andrew, R. Arora, J. Bilmes, and K. Livescu. Deep canonical correlation analysis. In *International Conference on Machine Learning*, pages 1247–1255, 2013.
- M. W. Andrews, W. J. Dowling, J. C. Bartlett, and A. R. Halpern. Identification of speeded and slowed familiar melodies by younger, middle-aged, and older musicians and nonmusicians. *Psychology and aging*, 13(3):462, 1998.
- B. Arons. A review of the cocktail party effect. *Journal of the American Voice I/O Society*, 12(7):35–50, 1992.
- J.-J. Aucouturier and E. Bigand. Mel cepstrum & ann ova: The difficult dialog between mir and music cognition. In *Proceedings of the 13th Conference of the International Society for Music Information Retrieval*, pages 397–402. Citeseer, 2012.
- A. Baddeley. The episodic buffer: a new component of working memory? *Trends in cognitive sciences*, 4(11):417–423, 2000.
- A. D. Baddeley and G. Hitch. Working memory. In *Psychology of learning and motivation*, volume 8, pages 47–89. Elsevier, 1974.
- P. C. Bagshaw, S. Hiller, and M. A. Jack. Enhanced pitch tracking and the processing of f0 contours for computer aided intonation teaching. In *Third European Conference on Speech Communication and Technology*, 1993.
- M. Baroni and C. Jacoboni. Proposal for a grammar of melody: The bach chorales. 1978.
- A. Barre. Mokka: Motion kinetic and kinematic analyzer, 2013. URL <http://biomechanical-toolkit.github.io/mokka/>.
- W. L. Bell, D. L. Davis, A. Morgan-Fisher, and E. D. Ross. Acquired aprosodia in children. *Journal of Child Neurology*, 5(1):19–26, 1990.
- N. G. B. K. J. Berger and J. P. Dmochowski. Decoding neurally relevant musical features using canonical correlation analysis. In *Proceedings of the 18th International Society for Music Information Retrieval Conference, Souzhou, China*, 2017.
- I. Biederman. Recognition-by-components: a theory of human image understanding. *Psychological review*, 94(2):115, 1987.

- I. Biederman. Psychophysical and neural correlates of the phenomenology of shape. *Handbook of Experimental Phenomenology: Visual Perception of Shape, Space and Appearance*, pages 415–436, 2013.
- C. M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg, 2006. ISBN 0387310738.
- G. Bishop, G. Welch, and B. D. Allen. Tracking: Beyond 15 minutes of thought. *SIGGRAPH Course Pack*, 11, 2001.
- R. M. Bittner, J. Salamon, J. J. Bosch, and J. P. Bello. Pitch contours as a mid-level representation for music informatics. In *Audio Engineering Society Conference: 2017 AES International Conference on Semantic Audio*, Jun 2017. URL <http://www.aes.org/e-lib/browse.cfm?elib=18756>.
- R. Bod. Memory-based models of melodic analysis: Challenging the gestalt principles. *Journal of New Music Research*, 31(1):27–36, 2002.
- D. Boer and R. Fischer. Towards a holistic model of functions of music listening across cultures: A culturally decentred qualitative approach. *Psychology of Music*, page 0305735610381885, 2011.
- P. Boersma. Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. In *Proceedings of the institute of phonetic sciences*, volume 17, pages 97–110. Amsterdam, 1993.
- C. Bohak and M. Marolt. Calculating similarity of folk song variants with melody-based features. In *Proceedings of the 10th Conference of the International Society for Music Information Retrieval*, pages 597–602, 2009.
- D. Bolinger and D. L. M. Bolinger. *Intonation and its parts: Melody in spoken English*. Stanford University Press, 1986.
- L. Boroditsky. Metaphoric structuring: Understanding time through spatial metaphors. *Cognition*, 75(1):1–28, 2000.
- P. Bourgin and A. Lesne. *Morphogenesis: origins of patterns and shapes*. Springer Science & Business Media, 2010.
- E. Bozkurt, Y. Yemez, and E. Erzin. Multimodal analysis of speech and arm motion for prosody-driven synthesis of beat gestures. *Speech Communication*, 85:29–42, 2016.
- A. S. Bregman. *Auditory scene analysis: The perceptual organization of sound*. MIT press, 1994.
- M. R. Bregman, A. D. Patel, and T. Q. Gentner. Songbirds use spectral shape, not pitch, for sound pattern recognition. *Proceedings of the National Academy of Sciences*, 113(6):1666–1671, 2016.

- B. H. Bronson. Melodic stability in oral transmission. *Journal of the International Folk Music Council*, 3:50–55, 1951.
- C. Buteau and G. Mazzola. From contour similarity to motivic topologies. *Musicae Scientiae*, 4(2):125–149, 2000.
- E. Cambouropoulos. Melodic cue abstraction, similarity, and category formation: A formal model. *Music Perception: An Interdisciplinary Journal*, 18(3):347–370, 2001. ISSN 0730-7829. DOI: 10.1525/mp.2001.18.3.347. URL <http://mp.ucpress.edu/content/18/3/347>.
- E. Cambouropoulos. Musical parallelism and melodic segmentation. *Music Perception: An Interdisciplinary Journal*, 23(3):249–268, 2006.
- B. Caramiaux and A. Tanaka. Machine learning of musical gestures. In *Proceedings of the 13th International Conference on New Interfaces for Musical Expression*, pages 513–518, 2013.
- B. Caramiaux, F. Bevilacqua, and N. Schnell. Towards a gesture-sound cross-modal analysis. In *International Gesture Workshop*, pages 158–170. Springer, 2009.
- G. Châtelet. Les jeux du mobile mathématique, physique, philosophie. 1993.
- T. Cheng, S. Fukayama, and M. Goto. Comparing RNN parameters for melodic similarity. *Proceedings of the 19th International Society for Music Information Retrieval Conference*, pages 1–8, 2018. URL <http://staff.aist.go.jp/m.goto/>.
- C. Cheung, L. S. Hamilton, K. Johnson, and E. F. Chang. The auditory representation of speech sounds in human motor cortex. *Elife*, 5:e12577, 2016.
- A. Clark. An embodied cognitive science? *Trends in cognitive sciences*, 3(9):345–351, 1999.
- E. Clarke. Meaning and the specification of motion in music. *Musicae Scientiae*, 5(2):213–234, 2001.
- M. Clayton. *Time in Indian Music: Rhythm, Metre, and Form in North Indian Rag Performance: Rhythm, Metre, and Form in North Indian Rag Performance*. Oxford University Press, 2001.
- M. Clayton and L. Leante. Embodiment in music performance. 2013.
- M. Clayton, R. Sager, and U. Will. In time with the music: the concept of entrainment and its significance for ethnomusicology. In *European meetings in ethnomusicology.*, volume 11, pages 1–82. Romanian Society for Ethnomusicology, 2005.
- M. Clynes. *Sentics: The touch of emotions*. Anchor Press, 1977.

Bibliography

- T. S. Collection. Tibetan yang-yig notation: Yang chants with tibetan graphic music notation, 2019. URL <https://www.schoyencollection.com/music-notation/graphic-notation/tibetan-yang-yig-ms-5280-1>.
- J. R. Cowdery. *The melodic tradition of Ireland*. Kent State University Press, 1990.
- L. L. Cuddy. On hearing pattern in melody. *Psychology of Music*, 10(1):3–10, 1982.
- M. E. Curtis and J. J. Bharucha. Memory and musical expectation for tones in cultural context. *Music Perception: An Interdisciplinary Journal*, 26(4): 365–375, 2009.
- S. Dalla Bella, I. Peretz, and N. Aronoff. Time course of melody recognition: A gating paradigm study. *Perception & Psychophysics*, 65(7):1019–1028, 2003.
- A. de Cheveigne. Pitch perception models. In *Pitch*, pages 169–233. Springer, 2005.
- A. De Cheveigné and H. Kawahara. Yin, a fundamental frequency estimator for speech and music. *The Journal of the Acoustical Society of America*, 111(4): 1917–1930, 2002.
- E. De Freitas and N. Sinclair. Diagram, gesture, agency: Theorizing embodiment in the mathematics classroom. *Educational Studies in Mathematics*, 80(1-2): 133–152, 2012.
- S. De Laubier. The meta-instrument. *Computer Music Journal*, 22(1):25–29, 1998.
- D. Deutsch, R. Lapidis, and T. Henthorn. The speech-to-song illusion. *The Journal of the Acoustical Society of America*, 124(4):2471–2471, October 2008. ISSN 0001-4966.
- D. Deutsch, T. Henthorn, and R. Lapidis. Illusory transformation from speech to song. *The Journal of the Acoustical Society of America*, 129(4):2245–2252, 2011.
- C. Donahue, I. Simon, and S. Dieleman. Piano genie. *arXiv preprint arXiv:1810.05246*, 2018.
- W. J. Dowling. Recognition of melodic transformations: Inversion, retrograde, and retrograde inversion. *Perception & Psychophysics*, 12(5):417–421, 1972.
- W. J. Dowling. Scale and contour: Two components of a theory of memory for melodies. *Psychological review*, 85(4):341, 1978.
- J. S. Downie. Music information retrieval. *Annual review of information science and technology*, 37(1):295–340, 2003.

- T. Eerola and M. Bregman. Melodic and contextual similarity of folk song phrases. *Musicae Scientiae*, 11(1 suppl):211–233, 2007.
- T. Eerola, P. Toiviainen, and C. Krumhansl. Real-time prediction of melodies: continuous predictability judgements and dynamic models. In *Proceedings of the 7th international conference on music perception and cognition*, pages 473–476. SA: Causal Productions Adelaide, 2002.
- T. Eerola, T. Himberg, P. Toiviainen, and J. Louhivuori. Perceived complexity of western and african folk melodies by western and african listeners. *Psychology of Music*, 34(3):337–371, 2006.
- Z. Eitan and R. Timmers. Beethoven’s last piano sonata and those who follow crocodiles: Cross-domain mappings of auditory pitch in a musical context. *Cognition*, 114(3):405–422, 2010.
- I. Fernandez-Prieto, C. Spence, F. Pons, and J. Navarra. Does language influence the vertical representation of auditory pitch and loudness? *i-Perception*, 8(3): 2041669517716183, 2017.
- M. Ferrand, P. Nelson, and G. Wiggins. Memory and melodic density: a model for melody segmentation. In *Proceedings of the XIV Colloquium on Musical Informatics (XIV CIM 2003)*, pages 95–98, 2003.
- B. W. Frankland, S. McAdams, and A. J. Cohen. Parsing of melody: Quantification and testing of the local grouping rules of lerdahl and jackendoff’s a generative theory of tonal music. *Music Perception: An Interdisciplinary Journal*, 21(4):499–543, 2004.
- P. Fuchs. *Oxford Handbook of Auditory Science: The Ear*. OUP Oxford, 2010.
- O. Fuhrman, K. McCormick, E. Chen, H. Jiang, D. Shu, S. Mao, and L. Boroditsky. How linguistic and cultural forces shape conceptions of time: English and mandarin time in 3d. *Cognitive science*, 35(7):1305–1328, 2011.
- J. J. Gibson. The concept of the stimulus in psychology. *American psychologist*, 15(11):694, 1960.
- K. H. Glette, A. R. Jensenius, and R. I. Godøy. Extracting action-sound features from a sound-tracing study. 2010.
- R. I. Godøy. Imagined action, excitation, and resonance. *Musical imagery*, 239: 52, 2001.
- R. I. Godøy. Motor-mimetic music cognition. *Leonardo*, 36(4):317–319, 2003.
- R. I. Godøy. Gestural affordances of musical sound. *Musical gestures: Sound, movement, and meaning*, pages 103–125, 2010.
- R. I. Godøy. Sonic object cognition. In *Springer Handbook of Systematic Musicology*, pages 761–777. Springer, 2018.

Bibliography

- R. I. Godøy and A. R. Jensenius. Body movement in music information retrieval. In *10th International Society for Music Information Retrieval Conference*, 2009.
- R. I. Godøy and M. Leman. *Musical gestures: Sound, movement, and meaning*. Routledge, 2010.
- R. I. Godøy, E. Haga, and A. R. Jensenius. Playing “air instruments”: mimicry of sound-producing gestures by novices and experts. In *International Gesture Workshop*, pages 256–267. Springer, 2005.
- R. I. Godøy, E. Haga, and A. R. Jensenius. Exploring music-related gestures by sound-tracing: A preliminary study. 2006.
- R. I. Godøy. Knowledge in music theory by shapes of musical objects and sound-producing actions. In *Music, gestalt, and computing*, pages 89–102. Springer, 1997.
- R. I. Godøy. Cross-modality and conceptual shapes and spaces in music theory. *Music and signs*, pages 85–98, 1999.
- N. Goodman. *Languages of art: An approach to a theory of symbols*. Hackett publishing, 1968.
- A. Gritten and E. King. *Music and gesture*. Ashgate Publishing, Ltd., 2006.
- A. Gritten and E. King. *New perspectives on music and gesture*. Ashgate Publishing, Ltd., 2011.
- H. I. Hair. Discrimination of tonal direction on verbal and nonverbal tasks by first grade children. *Journal of Research in Music education*, 25(3):197–210, 1977.
- A. R. Halpern and R. J. Zatorre. When that tune runs through your head: a pet investigation of auditory imagery for familiar melodies. *Cerebral cortex*, 9(7):697–704, 1999.
- P. Hansen. Vocal learning: its role in adapting sound structures to long-distance propagation, and a hypothesis on its evolution. *Animal Behaviour*, 1979.
- O. Herbot and M. V. Butz. Too good to be true? ideomotor theory from a computational perspective. *Frontiers in psychology*, 3:494, 2012.
- D. J. Hermes. Measuring the perceptual similarity of pitch contours. *Journal of Speech, Language, and Hearing Research*, 41(1):73–82, 1998.
- M. Hood. *The ethnomusicologist*, volume 140. Kent State Univ Pr, 1982.
- M. X. Huang, W. W. W. Tang, K. W. K. Lo, C. K. Lau, G. Ngai, and S. Chan. MelodicBrush. *Proceedings of the Designing Interactive Systems Conference on - DIS '12*, (July):418, 2012. DOI: 10.1145/2317956.2318018. URL <http://dl.acm.org/citation.cfm?doid=2317956.2318018>.

- J. E. Hummel and I. Biederman. Dynamic binding in a neural network for shape recognition. *Psychological review*, 99(3):480, 1992.
- D. Huron and D. Shanahan. Eyebrow movements and vocal pitch height: evidence consistent with an ethological signal. *The Journal of the Acoustical Society of America*, 133(5):2947–2952, 2013.
- S. Hutchins, C. Roquet, and I. Peretz. The vocal generosity effect: How bad can your singing be? *Music Perception: An Interdisciplinary Journal*, 30(2): 147–159, 2012.
- K. Irwin. Musipedia: The open music encyclopedia. *Reference Reviews*, 22(4): 45–46, 2008.
- C. Jacquemot and S. K. Scott. What is the relationship between phonological short-term memory and speech processing? *Trends in cognitive sciences*, 10 (11):480–486, 2006.
- P. Janata, J. L. Birk, J. D. Van Horn, M. Leman, B. Tillmann, and J. J. Bharucha. The cortical topography of tonal structures underlying western music. *science*, 298(5601):2167–2170, 2002.
- M. Jeannerod. Mental imagery in the motor context. *Neuropsychologia*, 33(11): 1419–1432, 1995.
- A. R. Jensenius. Action-sound: Developing methods and tools to study music-related body movement. 2007.
- A. R. Jensenius. *Methods for Studying Music-Related Body Motion*, pages 805–818. Springer Berlin Heidelberg, Berlin, Heidelberg, 2018. DOI: 10.1007/978-3-662-55004-5_38.
- A. R. Jensenius, T. Kivifte, and R. I. Godøy. Towards a gesture description interchange format. In *Proceedings of the 2006 conference on New interfaces for musical expression*, pages 176–179. IRCAM—Centre Pompidou, 2006.
- C. Jorgensen and K. Binsted. Web browser control using emg based sub vocal speech recognition. In *Proceedings of the 38th Annual Hawaii International Conference on System Sciences*, pages 294c–294c. IEEE, 2005.
- P. N. Juslin, L. Harmat, and T. Eerola. What makes music emotionally significant? exploring the underlying mechanisms. *Psychology of Music*, 42(4):599–623, 2014.
- L. Kang and H.-Y. Chien. Hé: Calligraphy as a musical interface. In *NIME*, pages 352–355, 2010.
- A. T. Katz. Heinrich schenker’s method of analysis. *The Musical Quarterly*, 21 (3):311–329, 1935.

Bibliography

- T. Kelkar. *Applications of Gesture and Spatial Cognition in Hindustani Vocal Music*. PhD thesis, International Institute of Information Technology, 2015.
- T. Kelkar and A. R. Jensenius. Analyzing free-hand sound-tracings of melodic phrases. *Applied Sciences*, 8(1):135, 2018.
- A. Kendon. *Gesture: Visible action as utterance*. Cambridge University Press, 2004.
- M. Kennedy, T. Rutherford-Johnson, and J. Kennedy. *The Oxford dictionary of music*. Oxford University Press, 2013.
- R. Kent. Coarticulation in recent speech production. *Journal of Phonetics*, 5(1): 15–133, 1977.
- Y. E. Kim, W. Chai, R. Garcia, and B. Vercoe. Analysis of a contour-based representation for melody. In *Proceedings of the 2nd Conference of the International Society for Music Information Retrieval*, 2000.
- M. Kimmel. Properties of cultural embodiment: Lessons from the anthropology of the body. *Body, language and mind*, 2:77–108, 2008.
- T. Kitahara, S. I. Giraldo, and R. Ramírez. Jamsketch: a drawing-based real-time evolutionary improvisation support system. In *Proceedings of the 17th International Conference on New Interfaces for Musical Expression*, pages 505–506, 2017.
- D. Knox, S. Beveridge, L. A. Mitchell, and R. A. MacDonald. Acoustic analysis and mood classification of pain-relieving music. *The Journal of the Acoustical Society of America*, 130(3):1673–1682, 2011.
- E. Kobatake and K. Tanaka. Neuronal selectivities to complex object features in the ventral visual pathway of the macaque cerebral cortex. *Journal of neurophysiology*, 71(3):856–867, 1994.
- E. Kohler, C. Keysers, M. A. Umiltà, L. Fogassi, V. Gallese, and G. Rizzolatti. Hearing sounds, understanding actions: action representation in mirror neurons. *Science*, 297(5582):846–848, 2002.
- S. M. Kosslyn, W. L. Thompson, and G. Ganis. *The case for mental imagery*. Oxford University Press, 2006.
- C. L. Krumhansl, P. Toivanen, T. Eerola, P. Toiviainen, T. Järvinen, and J. Louhivuori. Cross-cultural music cognition: Cognitive methodology applied to north sami yoiks. *Cognition*, 76(1):13–58, 2000.
- K. Kuiper and D. Haggio. Livestock auctions, oral poetry, and ordinary language. *Language in society*, 13(2):205–234, 1984.
- M. Kussner. *Shape, drawing and gesture: Cross-modal mappings of sound and music*. PhD thesis, King’s College London, 2014.

- M. B. Küssner. Music and shape. *Literary and Linguistic Computing*, 28(3): 472–479, 2013.
- W. Laade and D. Christensen. Lappish joik songs from northern norway, 1956.
- D. R. Ladd. *Intonational phonology*. Cambridge University Press, 2008.
- G. Lakoff and M. Johnson. Conceptual metaphor in everyday language. *The journal of Philosophy*, 77(8):453–486, 1980.
- O. Lartillot. Soundtracer, 2018. URL <https://itunes.apple.com/us/app/soundtracer/id1320398109?mt=8>.
- O. Lartillot and P. Toiviainen. A matlab toolbox for musical feature extraction from audio. In *International conference on digital audio effects*, pages 237–244. Bordeaux, 2007.
- M. Leman. *Embodied music cognition and mediation technology*. Mit Press, 2008.
- M. Leman and A. Schneider. Origin and nature of cognitive and systematic musicology: An introduction. In *Music, Gestalt, and Computing*, pages 11–29. Springer, 1997.
- M. Leman, P.-J. Maes, L. Nijs, and E. Van Dyck. *What Is Embodied Music Cognition?*, pages 747–760. Springer Berlin Heidelberg, Berlin, Heidelberg, 2018. ISBN 978-3-662-55004-5. DOI: 10.1007/978-3-662-55004-5_34.
- F. Lerdahl and R. Jackendoff. A generative theory of tonal music. 1987.
- M. Lesaffre, P.-J. Maes, and M. Leman. *The Routledge companion to embodied music interaction*. Taylor & Francis, 2017.
- D. Levertov. *Poems of Denise Levertov, 1960-1967*. New Directions Publishing, 1983.
- E. Levine. A theory of everything, February 2002. URL https://www.ted.com/talks/emily_levine_s_theory_of_everything.
- A. M. Liberman and I. G. Mattingly. The motor theory of speech perception revised. *Cognition*, 21(1):1–36, 1985.
- C. P. Long. Aristotle’s phenomenology of form: The shape of beings that become. *Epoché: A Journal for the History of Philosophy*, 11(2):435–448, 2007.
- N. Malandrakis, A. Potamianos, G. Evangelopoulos, and A. Zlatintsi. A supervised approach to movie emotion tracking. In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pages 2376–2379. IEEE, 2011.
- S. C. Mangaoang, Aine, J. Brackett, and K. M. Smith. Here lies love and the politics of disco-opera. In *The Routledge Companion to Popular Music Analysis*, pages 365–381. Routledge, 2018.

Bibliography

- P. R. Marler and H. Slabbekoorn. *Nature's music: the science of birdsong*. Elsevier, 2004.
- C. P. Martin and J. Tørresen. Microjam: An app for sharing tiny touch-screen performances. In *Proceedings of the International Conference on New Interfaces for Musical Expression*, pages 495–496. Aalborg University Copenhagen, 2017.
- G. Mazzola and M. Andreatta. Diagrams, gestures and formulae in music. *Journal of Mathematics and Music*, 1(1):23–46, 2007.
- G. Mazzola and O. Zahorka. The rubato workstation for musical analysis and performance. In *Proc. of the 3rd International Conference on Music Perception and Cognition (ICMPC)*. Liege, 1994.
- D. McNeill. *Hand and mind: What gestures reveal about thought*. University of Chicago press, 1992.
- J. Mehler and E. Dupoux. *Nacer sabiendo: Introducción al desarrollo cognitivo del hombre*, volume 5. Anaya-Spain, 1992.
- B. Mintzer. *Contemporary jazz etudes*. Miami, FL: Warner Bros, 2004.
- I. Molnar-Szakacs and K. Overy. Music and mirror neurons: from motion to e'motion. *Social cognitive and affective neuroscience*, 1(3):235–241, 2006.
- G. E. Moorhead and D. Pond. *Music of young children*. Pillsbury Foundation for Advancement of Music Education, 1941.
- C. F. Mora. Foreign language acquisition and melody singing. *ELT journal*, 54(2):146–152, 2000.
- R. D. Morris. New directions in the theory and analysis of musical contour. *Music Theory Spectrum*, 15(2):205–228, 1993.
- D. Müllensiefen, K. Frieler, et al. Cognitive adequacy in the measurement of melodic similarity: Algorithmic vs. human judgments. *Computing in Musicology*, 13(2003):147–176, 2004.
- M. Müller. *Information retrieval for music and motion*, volume 2. Springer, 2007.
- T. Murphey. The song stuck in my head phenomenon: a melodic din in the lad? *System*, 18(1):53–64, 1990.
- E. Narmour. *The analysis and cognition of melodic complexity: The implication-realization model*. University of Chicago Press, 1992.
- B. Nettl. *Music in primitive culture*. Harvard University Press Cambridge, MA, 1956.
- S. Nootboom. The prosody of speech: melody and rhythm. *The handbook of phonetic sciences*, 5:640–673, 1997.

- K. Nymoen. *Methods and Technologies for Analysing Links Between Musical Sound and Body Motion*. PhD thesis, University of Oslo, 2013.
- K. Nymoen, B. Caramiaux, M. Kozak, and J. Torresen. Analyzing sound tracings: A multimodal approach to music information retrieval. In *Proceedings of the 1st International Association for Computing Machinery Workshop on Music Information Retrieval with User-centered and Multimodal Strategies*, MIRUM '11, pages 39–44, New York, NY, USA, 2011. Association for Computing Machinery. ISBN 978-1-4503-0986-8. DOI: 10.1145/2072529.2072541. URL <http://doi.acm.org/10.1145/2072529.2072541>.
- K. Nymoen, J. Torresen, R. Godøy, and A. Jensenius. A statistical approach to analyzing sound tracings. *Speech, sound and music processing: Embracing research in India*, pages 120–145, 2012.
- K. Nymoen, R. I. Godøy, A. R. Jensenius, and J. Torresen. Analyzing correspondence between sound objects and body motion. *Association for Computing Machinery Trans. Appl. Percept.*, 10(2):9:1–9:22, June 2013. ISSN 1544-3558. DOI: 10.1145/2465780.2465783. URL <http://doi.acm.org/10.1145/2465780.2465783>.
- U. of Washington Ethnomusicology Archives. Ragamala lecture-demonstrations recordings: Pandit k.g. ginde, 1991. URL <http://archiveswest.orbiscascade.org/ark:/80444/xv80252>.
- H. Ohkushi, T. Ogawa, and M. Haseyama. Music recommendation according to human motion based on kernel cca-based relationship. *EURASIP Journal on Advances in Signal Processing*, 2011(1):121, 2011.
- J. E. Ollen. *A criterion-related validity test of selected indicators of musical sophistication using expert ratings*. PhD thesis, The Ohio State University, 2006.
- A. Pannekamp, U. Toepel, K. Alter, A. Hahne, and A. D. Friederici. Prosody-driven sentence processing: An event-related brain potential study. *Journal of cognitive neuroscience*, 17(3):407–421, 2005.
- D. Parsons. *The directory of tunes and musical themes*. Cambridge, Eng.: S. Brown, 1975.
- S. Paschalidou, T. Eerola, and M. Clayton. Voice and movement as predictors of gesture types and physical effort in virtual object interactions of classical indian singing. In *Proceedings of the 3rd International Symposium on Movement and Computing*, page 45. Association for Computing Machinery, 2016.
- A. D. Patel. *Music, language, and the brain*. Oxford university press, 2010.
- M. Pearce, D. Müllensiefen, and G. A. Wiggins. A comparison of statistical and rule-based models of melodic segmentation. In *Proceedings of the 9th Conference of the International Society for Music Information Retrieval*, pages 89–94, 2008.

Bibliography

- M. T. Pearce and G. A. Wiggins. Expectation in melody: The influence of context and learning. *Music Perception: An Interdisciplinary Journal*, 23(5): 377–405, 2006.
- M. T. Pearce, D. Müllensiefen, and G. A. Wiggins. Melodic grouping in music information retrieval: New methods and applications. In *Advances in music information retrieval*, pages 364–388. Springer, 2010.
- L. Pearson. *Gesture in Karnatak Music: Pedagogy and Musical Structure in South India*. PhD thesis, Durham University, 2016.
- G. Perle. *The Operas of Alban Berg, Volume II: Lulu*, volume 2. Univ of California Press, 1989.
- I. Poggi. Towards the alphabet and the lexicon of gesture, gaze and touch. In *Virtual Symposium on Multimodality of Human Communication*. <http://www.semioticon.com/virtuals/index.html>, 2002.
- F. Pulvermüller and L. Fadiga. Active perception: sensorimotor circuits as a cortical basis for language. *Nature reviews neuroscience*, 11(5):351, 2010.
- W. V. Quine. Identity, ostension, and hypostasis. *The Journal of Philosophy*, 47(22):621–633, 1950.
- I. Quinn. The combinatorial model of pitch contour. *Music Perception: An Interdisciplinary Journal*, 16(4):439–456, 1999.
- M. Rahaim. *Musicking Bodies: Gesture and Voice in Hindustani Music*. Wesleyan University Press, 2012.
- D. Reisberg. *Auditory imagery*. Psychology Press, 2014.
- G. Rizzolatti and L. Craighero. The mirror-neuron system. *Annu. Rev. Neurosci.*, 27:169–192, 2004.
- H. H. Roberts. New phases in the study of primitive music. *American Anthropologist*, 24(2):144–160, 1922.
- D. A. Rosenbaum. *Knowing hands: The cognitive psychology of manual control*. Cambridge University Press, 2017.
- L. D. Rosenblum, M. A. Schmuckler, and J. A. Johnson. The mcgurk effect in infants. *Perception & Psychophysics*, 59(3):347–357, 1997.
- E. D. Ross. Nonverbal aspects of language. *Neurologic Clinics*, 11(1):9–23, 1993.
- U. Roy, T. Kelkar, and B. Indurkha. Trap: An interactive system to generate valid raga phrases from sound-tracings. In *Proceedings of the 14th International Conference on New Interfaces of Musical Expression Conference*, pages 243–246, 2014.

- E. Rusconi, B. Kwan, B. L. Giordano, C. Umiltà, and B. Butterworth. Spatial representation of pitch height: the smarc effect. *Cognition*, 99(2):113–129, 2006.
- C. Sachs. *The wellsprings of music*. Springer Science & Business Media, 2012.
- J. Salamon, G. Peeters, and A. Röbel. Statistical characterisation of melodic pitch contours and its application for melody extraction. In *ISMIR*, pages 187–192, 2012.
- A. L. Samuel. Some studies in machine learning using the game of checkers. *IBM J. Res. Dev.*, 3(3):210–229, July 1959. ISSN 0018-8646. DOI: 10.1147/rd.33.0210. URL <http://dx.doi.org/10.1147/rd.33.0210>.
- M. Śarmā. *Tradition of Hindustani Music*. APH Publishing, 2006.
- P. E. Savage and Q. D. Atkinson. Automatic tune family identification by musical sequence alignment. In *Proceedings of the 16th Conference of the International Society for Music Information Retrieval*, volume 163, pages 162–168, 2015.
- P. E. Savage, A. T. Tierney, and A. D. Patel. Global music recordings support the motor constraint hypothesis for human and avian song contour. *Music Perception: An Interdisciplinary Journal*, 34(3):327–334, 2017.
- P. Schaeffer, F. B. Mâche, M. Philippot, F. Bayle, L. Ferrari, I. Malec, and B. Parmegiani. *La musique concrète*. Presses universitaires de France, 1967a.
- P. Schaeffer, G. Reibel, B. Ferreyra, H. Chiarucci, F. Bayle, A. Tanguy, J.-L. Ducarme, J.-F. Pontefract, and J. Schwarz. *Solfège de l’objet sonore*. INA/GRM, 1967b.
- E. G. Schellenberg, A. M. Krysciak, and R. J. Campbell. Perceiving emotion in melody: Interactive effects of pitch and rhythm. *Music Perception: An Interdisciplinary Journal*, 18(2):155–171, 2000.
- M. A. Schmuckler. Testing models of melodic contour similarity. *Music Perception: An Interdisciplinary Journal*, 16(3):295–326, 1999.
- M. A. Schmuckler. Pitch and pitch structures. *Ecological psychoacoustics*, pages 271–315, 2004.
- M. A. Schmuckler. Melodic contour similarity using folk melodies. *Music Perception: An Interdisciplinary Journal*, 28(2):169–194, 2010.
- K. Schulze, S. Koelsch, and V. Williamson. *Auditory Working Memory*, pages 461–472. Springer Berlin Heidelberg, Berlin, Heidelberg, 2018. ISBN 978-3-662-55004-5. DOI: 10.1007/978-3-662-55004-5_24. URL https://doi.org/10.1007/978-3-662-55004-5_24.
- R. Scruton. *Understanding music: Philosophy and interpretation*. Bloomsbury Publishing, 2016.

Bibliography

- C. Seeger. On the moods of a music-logic. *Journal of the American Musicological Society*, 13(1/3):224–261, 1960.
- R. V. Shannon. Is birdsong more like speech or music? *Trends in cognitive sciences*, 20(4):245–247, 2016.
- S. Shayan, O. Ozturk, and M. A. Sicoli. The thickness of pitch: Crossmodal metaphors in farsi, turkish, and zapotec. *The Senses and Society*, 6(1):96–105, 2011.
- Y. K. Shin, R. W. Proctor, and E. J. Capaldi. A review of contemporary ideomotor theory. *Psychological bulletin*, 136(6):943, 2010.
- P. Shove and B. H. Repp. Musical motion and performance: Theoretical and empirical perspectives. *The practice of performance: Studies in musical interpretation*, pages 55–83, 1995.
- B. Sievers, L. Polansky, M. Casey, and T. Wheatley. Music and movement share a dynamic structure that supports universal expressions of emotion. *Proceedings of the National Academy of Sciences*, 110(1):70–75, 2013.
- S. M. Stalinski, E. G. Schellenberg, and S. E. Trehub. Developmental changes in the perception of pitch contour: Distinguishing up from down. *J. Acoust. Soc. Am.*, 124(3):1759–1763, 2008.
- M. E. Tanenhaus and J. L. Lipeles. Miniaturized inertial measurement unit and associated methods, Apr. 28 2009. US Patent 7,526,402.
- D. Temperley. What’s key for key? the krumhansl-schmuckler key-finding algorithm reconsidered. *Music Perception: An Interdisciplinary Journal*, 17(1):65–100, 1999.
- D. Temperley. Composition, perception, and schenkerian theory. *Music Theory Spectrum*, 33(2):146–168, 2011.
- J. Tenney and L. Polansky. Temporal gestalt perception in music. *Journal of Music Theory*, 24(2):205–241, 1980.
- R. Thom. *Paraboles et catastrophes*. Flammarion, 1983.
- A. T. Tierney, F. A. Russo, and A. D. Patel. The motor origins of human and avian song structure. *Proceedings of the National Academy of Sciences*, 108(37):15510–15515, 2011.
- B. Tillmann and E. Bigand. Musical structure processing after repeated listening: schematic expectations resist veridical expectations. *Musicae Scientiae*, 14(2 suppl):33–47, 2010.
- B. Tillmann, J. J. Bharucha, and E. Bigand. Implicit learning of tonality: a self-organizing approach. *Psychological review*, 107(4):885, 2000.

- P. Toiviainen and T. Eerola. Visualization in comparative music research. In *COMPSTAT*, pages 209–221. Springer, 2006.
- E. Tolbert. Women cry with words: Symbolization of affect in the karelian lament. *Yearbook for Traditional Music*, 22:80–105, 1990.
- L. J. Trainor, C. M. Austin, and R. N. Desjardins. Is infant-directed speech prosody a result of the vocal expression of emotion? *Psychological science*, 11(3):188–195, 2000.
- S. E. Trehub, D. Bull, and L. A. Thorpe. Infants’ perception of melodies: The role of melodic contour. *Child development*, pages 821–830, 1984.
- A. Truslit. Creation and movement in music: A tale of musical performance and its moving figures and music [shape and movement in music]. *Berlin-Lichtenfelde: Vieweg*, 1938.
- J. Urbano, J. Lloréns, J. Morato, and S. Sánchez-Cuadrado. Mirex 2012 symbolic melodic similarity: Hybrid sequence alignment with geometric representations. *Music Information Retrieval Evaluation eXchange*, pages 3–6, 2012.
- P. Van Kranenburg, J. Garbers, A. Volk, F. Wiering, L. Grijp, and R. Veltkamp. Towards integration of music information retrieval and folk song research. In *Proceedings of the 8th International Conference on Music Information Retrieval*, pages 505–8, 2007.
- V. Vasant. Sangeet visharad, 1998.
- A. Volk, J. Garbers, P. Van Kranenburg, F. Wiering, R. C. Veltkamp, and L. P. Grijp. Applying rhythmic similarity based on inner metric analysis to folksong research. In *Proceedings of the 8th Conference of the International Society for Music Information Retrieval*, pages 293–296, 2007.
- C. Von Ehrenfels. On “gestalt qualities.”. *B. Smith (Ed. & Trans.), Foundations of Gestalt theory*, pages 82–117, 1988.
- D. T. Vuvan and M. A. Schmuckler. Tonal hierarchy representations in auditory imagery. *Memory & cognition*, 39(3):477–490, 2011.
- P. Walker, J. G. Bremner, U. Mason, J. Spring, K. Mattock, A. Slater, and S. P. Johnson. Preverbal infants are sensitive to cross-sensory correspondences much ado about the null results of lewkowicz and minar (2014). *Psychological science*, 25(3):835–836, 2014.
- H. Wallach. Über visuell wahrgenommene bewegungsrichtung. *Psychologische Forschung*, 20(1):325–380, 1935.
- K. Wang, Q. Yin, W. Wang, S. Wu, and L. Wang. A comprehensive survey on cross-modal retrieval. *arXiv preprint arXiv:1607.06215*, 2016.

- M. W. Weiss, S. E. Trehub, and E. G. Schellenberg. Something in the way she sings: Enhanced memory for vocal melodies. *Psychological Science*, 23(10):1074–1078, 2012. DOI: 10.1177/0956797612442552. URL <https://doi.org/10.1177/0956797612442552>. PMID: 22894936.
- M. W. Weiss, P. Vanzella, E. G. Schellenberg, and S. E. Trehub. Pianists exhibit enhanced memory for vocal melodies but not piano melodies. *The Quarterly Journal of Experimental Psychology*, 68(5):866–877, 2015.
- M. W. Weiss, S. E. Trehub, E. G. Schellenberg, and P. Habashi. Pupils dilate for vocal or familiar music. *Journal of Experimental Psychology: Human Perception and Performance*, 42(8):1061, 2016.
- R. J. West and R. Fryer. Ratings of suitability of probe tones as tonics after random orderings of notes of the diatonic scale. *Music Perception: An Interdisciplinary Journal*, 7(3):253–258, 1990.
- B. W. White. Recognition of distorted melodies. *The American journal of psychology*, 73(1):100–107, 1960.
- G. Wiggins, E. Miranda, A. Smaill, and M. Harris. A framework for the evaluation of music representation systems. *Computer Music Journal*, 17(3):31–42, 1993.
- H. Wittmann. Intonation in glottogenesis. *The melody of language: Festschrift Dwight L. Bolinger*, pages 315–29, 1980.
- D. H. Wolpert, W. G. Macready, et al. No free lunch theorems for optimization. *IEEE transactions on evolutionary computation*, 1(1):67–82, 1997.
- S. Wuerger, R. Shapley, and N. Rubin. “on the visually perceived direction of motion” by hans wallach: 60 years later. *Perception*, 25(11):1317–1367, 1996.
- S. Youngerman. Curt sachs and his heritage: A critical review of world history of the dance with a survey of recent studies that perpetuate his ideas. *Dance Research Journal*, 6(2):6–19, 1974.
- B. Yung. *Cantonese opera: performance as creative process*. Cambridge University Press, 1989.
- B. Yung. The relationship of text and tune in chinese opera. In *Music, language, speech and brain*, pages 408–418. Springer, 1991.
- R. J. Zatorre and S. R. Baum. Musical melody and speech intonation: Singing a different tune. *PLoS biology*, 10(7):e1001372, 2012.
- R. J. Zatorre and A. R. Halpern. Mental concerts: Musical imagery and auditory cortex. *Neuron*, 47(1):9 – 12, 2005. ISSN 0896-6273. DOI: <https://doi.org/10.1016/j.neuron.2005.06.013>. URL <http://www.sciencedirect.com/science/article/pii/S0896627305005180>.

Papers

Paper I

Exploring melody and motion features in “sound-tracings”

Tejaswinee Kelkar, Alexander Refsum Jensenius

Published in *In Proceedings of the 14th Sound and Music Computing Conference*, July 2017, pp. 98–103.

EXPLORING MELODY AND MOTION FEATURES IN “SOUND-TRACINGS”

Tejaswinee Kelkar

University of Oslo, Department of Musicology
tejaswinee.kelkar@imv.uio.no

Alexander Refsum Jensenius

University of Oslo, Department of Musicology
a.r.jensenius@imv.uio.no

ABSTRACT

Pitch and spatial height are often associated when describing music. In this paper we present results from a sound-tracing study in which we investigate such sound–motion relationships. The subjects were asked to move as if they were creating the melodies they heard, and their motion was captured with an infra-red, marker-based camera system. The analysis is focused on calculating feature vectors typically used for melodic contour analysis. We use these features to compare melodic contour typologies with motion contour typologies. This is based on using proposed feature sets that were made for melodic contour similarity measurement. We apply these features to both the melodies and the motion contours to establish whether there is a correspondence between the two, and find the features that match the most. We find a relationship between vertical motion and pitch contour when evaluated through features rather than simply comparing contours.

1. INTRODUCTION

How can we characterize melodic contours? This question has been addressed through parametric, mathematical, grammatical, and symbolic methods. The applications of characterizing melodic contour can be for finding similarity in different melodic fragments, indexing musical pieces, and more recently, for finding motifs in large corpora of music. In this paper, we compare pitch contours with motion contours derived from people’s expressions of melodic pitch as movement. We conduct an experiment using motion capture to measure body movements through infra-red cameras, and analyse the vertical motion to compare it with pitch contours.

1.1 Melodic Similarity

Marsden disentangles some of our simplification of concepts while dealing with melodic contour similarity, explaining that the conception of similarity itself means different things at different times with regards to melodies [1]. Not only are these differences culturally contingent, but also dependent upon the way in which music is represented as data. Our conception of melodic similarity can

be compared to the distances of melodic objects in a hyperspace of all possible melodies. Computational analyses of melodic similarity have also been essential for dealing with issues regarding copyright infringement [2], “query by humming” systems used for music retrieval [3, 4], and for use in psychological prediction [5].

1.2 Melodic Contour Typologies

Melodic contours serve as one of the features that can describe melodic similarity. Contour typologies, and building feature sets for melodic contour have been experimented with in many ways. Two important variations stand out — the way in which melodies are represented and features are extracted, and the way in which typologies are derived from this set of features, using mathematical methods to establish similarity. Historically, melodic contour has been analysed in two principal ways, using (a) symbolic notation, or (b) recorded audio. These two methods differ vastly in their interpretation of contour and features.

1.3 Extraction of melodic features

The extraction of melodic contours from symbolic features has been used to create indexes and dictionaries of melodic material [6]. This method simply uses signs such as +/-/=, to indicate the relative movement of each note. Adams proposes a method through which the key points of a melodic contour — the high, low, initial, and final points of a melody — are used to create a feature vector that he then uses to create typologies of melody [7]. It is impossible to know with how much success we can constrain melodic contours in finite typologies, although this has been attempted through these methods and others. Other methods, such as that of Morris, constrain themselves to tonal melodies [8], and yet others, such as Friedmann’s, rely on relative pitch intervals [9]. Aloupis et. al. use geometrical representations for melodic similarity search. Although many of these methods have found robust applications, melodic contour analysis from notation is harder to apply to diverse musical systems. This is particularly so for musics that are not based on western music notation. Ornaments, for example, are easier to represent as sound signals than symbolic notation.

Extraction of contour profiles from audio-based pitch extraction algorithms has been demonstrated in several recent studies [10, 11], including specific genres such as flamenco voice [12, 13]. While such audio-based contour extraction may give us a lot of insight about the musical data at hand,

Copyright: © 2017 Author1 et al. This is an open-access article distributed under the terms of the [Creative Commons Attribution 3.0 Unported License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

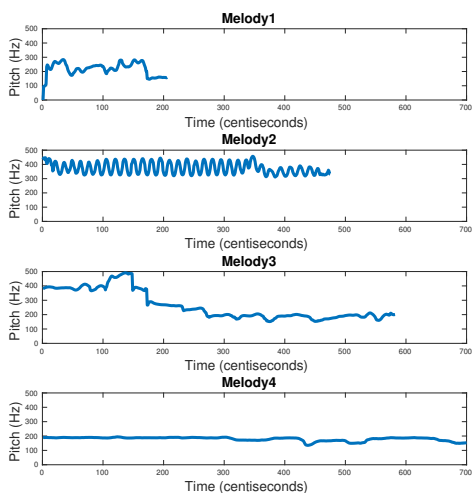


Figure 1. Examples of Pitch features of selected melodies, extracted through autocorrelation.

the generalisability of such a method is harder to evaluate than those of the symbolic methods.

1.4 Method for similarity finding

While some of these methods use matrix similarity computation [14], others use edit distance-based metrics [15], and string matching methods [16]. Extraction of sound signals to symbolic data that can then be processed in any of these ways is yet another method to analyse melodic contour. This paper focuses on evaluating melodic contour features through comparison with motion contours, as opposed to being compared to other melodic phrases. This would shed light on whether the perception of contour as a feature is even consistent, measurable, or whether we need other types of features to capture contour perception.

Yet another question is how to evaluate contours and their behaviours when dealing with data such as motion responses to musical material. Motion data could be transposed to fit the parameters required for score-based analysis, which could possibly yield interesting results. Contour extraction from melody, motion, and their derivatives could also demonstrate interesting similarities between musical motion and melodic motion. This is what this paper tries to address: looking at the benefits and disadvantages of using feature vectors to describe melodic features in a multimodal context. The following research questions were the most important for the scope of this paper:

1. Are the melodic contours described in previous studies relevant for our purpose?
2. Which features of melodic contours correspond to features extracted from vertical motion in melodic tracings?

In this paper we compare melodic movement, in terms of pitch, with vertical contours derived from motion capture recordings. The focus is on three features of melodic contour, using a small dataset containing motion responses of 3 people to 4 different melodies. This dataset is from a larger experiment containing 32 participants and 16 melodies.

2. BACKGROUND

2.1 Pitch Height and Melodic Contour

This paper is concerned with *melody*, that is, sequences of pitches, and how people trace melodies with their hands. Pitch appears to be a musical feature that people easily relate to when tracing sounds, even when the timbre of the sound changes independently of the pitch [17–19]. Melodic contour has been studied in terms of symbolic pitch [20, 21]. Eitan explores the multimodal associations with pitch height and verticality in his papers [22, 23]. Our subjective experience of melodic contours in cross cultural contexts is also investigated in Eerola’s paper [24].

The ups and downs in melody have often been compared to other multimodal features that also seem to have up-down contours, such as words that signify verticality. This attribute of pitch to verticality has also been used as a feature in many visualization algorithms. In this paper, we focus particularly on the vertical movement in the tracings of participants, to investigate if there is, indeed, a relationship with the vertical contours of the melodies. We also want to see if this relationship can be extracted through features that have been explored to represent melodic contour. If the features proposed for melodic contours are not enough, we wish to investigate other methods that can be used to represent a common feature vector between melody and motion in the vertical axis. All 4 melodies in the small dataset that we create for the purposes of this experiment are represented as pitch in Figure 1.

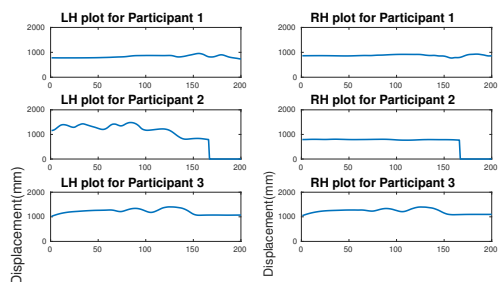


Figure 2. Example plots of some sound-tracing responses to Melody 1. Time (in frames) runs along the x-axes, while the y-axes represent the vertical position extracted from the motion capture recordings (in millimetres). LH=left hand, RH=right hand.



Figure 3. A symbolic transcription of Melody 1, a sustained vibrato of a high soprano. The notated version differs significantly from the pitch profile as seen in Figure 2. The appearance of the trill and vibrato are dimensions that people respond through in motion tracings, that don't clearly appear in the notated version.

	Feature 1	Feature 3
Melody1	[+, -, +, -, +, -, -]	[0, 4, -4, 2, -2, 4, 0, -9]
Melody2	[+, -, -]	[0, 2, -2, -2, 0, 0]
Melody3	[+, -, -, -, -, -, -]	[0, -2, -4, -1, -1, -1, -4, -2, -3, 0, 0, 0]
Melody4	[+, -, +, -, -, +, -, -]	[0, -2, 2, -4, 2, -2, 4, -2, -2]

Table 1. Examples of Features 1 and 3 for all 3 melodies from score.

2.2 Categories of contour feature descriptors

In the following paragraphs, we will describe how the feature sets selected for comparison in this study are computer. The feature sets that come from symbolic notation analysis are revised to compute the same features from the pitch extracted profiles of the melodic contours.

2.2.1 Feature 1: Sets of signed pitch movement direction

These features are described in [6], and involve a description of the points in the melody where the pitch ascends or descends. This method is applied by calculating the first derivatives of the pitch contours, and assigning a change of sign whenever the spike in the velocity is greater than or less than the standard deviation of the velocity. This helps us come up with the transitions that are more important to the melody, as opposed to movement that stems from vibratos, for example.

2.2.2 Feature 2: Initial, Final, High, Low features

Adams, and Morris [7, 8] propose models of melodic contour typologies and melodic contour description models that rely on encoding melodic features using these descriptors, creating a feature vector of those descriptors. For this study, we use the feature set containing initial, final, high and low points of the melodic and motion contours computed directly from normalized contours.

2.2.3 Feature 3: Relative interval encoding

In these sets of features, for example as proposed in Friedman, Quinn, Parsons, [6, 9, 14], the relative pitch distances are encoded either as a series of ups and downs, combined with features such as operators ($i, =, i_c$) or distances of relative pitches in terms of numbers. Each of these methods employs a different strategy to label the high and low



Figure 4. Lab set-up for the Experiment with 21 markers positioned on the body. 8 Motion capture cameras are hanging on the walls.

points of melodies. Some rely on tonal pitch class distribution, such as Morris's method, which is also analogous to Schenkerian analysis in terms of ornament reduction; while others such as Friedmann's only encode changes that are relative to the ambit of the current melodic line. For the purposes of this study, we pick the latter method given as all the melodies in this context are not tonal in the way that would be relevant to Morris.

3. EXPERIMENT DESCRIPTION

The experiment was designed so that subjects were instructed to perform hand movements as if they were creating the melodic fragments that they heard. The idea was that they would "shape" the sound with their hands in physical space. As such, this type of free-hand sound-tracing task is quite different from some sound-tracing experiments using pen on paper or on a digital tablet. Participants in a free-hand tracing situation would be less fixated upon the precise locations of all of their previous movements, thus giving us an insight of the perceptually salient properties of the melodies that they choose to represent.

3.1 Stimuli

We selected 16 melodic fragments from four genres of music that use vocalisations without words:

1. Scat singing
2. Western classical vocalise
3. Sami joik
4. North Indian music

The melodic fragments were taken from real recordings, containing complete phrases. This retained the melodies in the form that they were sung and heard in, thus preserving their ecological quality. The choice of vocal melodies was both to eliminate the effect of words on the perception of music, but also to eliminate the possibility of imitating the sound-producing actions on instruments ("air-instrument" performance) [25].

There was a pause before and after each phrase. The phrases were an average of 4.5 seconds in duration (s.d. 1.5s). These samples were presented in two conditions: (1) the real recording, and (2) a re-synthesis through a saw-tooth wave from an autocorrelation analysis of the pitch profile. There was thus a total of 32 stimuli per participant.

The sounds were played at comfortable listening level through a Genelec 8020 speaker, placed 3 metres ahead of the participants at a height of 1 meter.

3.2 Participants

A total of 32 participants (17 female, 15 male) were recruited to move to the melodic stimuli in our motion capture lab. The mean age of the participants was 31 years ($SD=9$). The participants were recruited from the University of Oslo, and included students, and employees, who were not necessarily from a musical background.

The study was reported to and obtained ethical approval from the Norwegian Centre for Research Data. The participants signed consent forms and were free to withdraw during the experiment, if they wished.

3.3 Lab set-up

The experiment was run in the fourMs motion capture lab, using a Qualisys motion capture system with eight wall-mounted Oqus 300 cameras (Figure 3.1, capturing at 200 Hz). The experiment was conducted in dim light, with no observers, to make sure that participants felt free to move as they liked. A total of 21 markers were placed on the body of the participants: the head, shoulders, elbows, wrists, knees, ankles, the torso, and the back of the body. The recordings were post-processed in Qualisys Track Manager (QTM), and analysed further in Matlab.

3.4 Procedure

The participants were asked to trace all 32 melody phrases (in random order) as if their hand motion was ‘producing’ the melody. The experiment lasted for a total duration of 10 minutes. After post processing the data from this experiment, we get a dataset for motion of 21 markers while the participants performed sound-tracing. We take a subset of this data for further analysis of contour features. In this step, we extract the motion data for the left and the right hands from a small subset of 4 melodies performed by 3 participants. We focus on the vertical movement of both the hands given as this analysis pertains to verticality of pitch movement. We process these motion contours along with the pitch contours for the 4 selected melodies, through 3 melodic features as described in section 2.2.

4. MELODIC CONTOUR FEATURES

For the analysis, we record the following feature vectors through some of the methods mentioned in section 1.2. The feature vectors are calculated as mentioned below:

Feature 1 Signed interval distances: The obtained motion and pitch contours are binned iteratively to calculate average values in each section. Mean vertical motion for all participants is calculated. This mean motion is then binned in the way that melodic contours are binned. The difference between the values of the successive bins is calculated. The sign of this difference is concatenated to form a feature vector composed of signed distances.

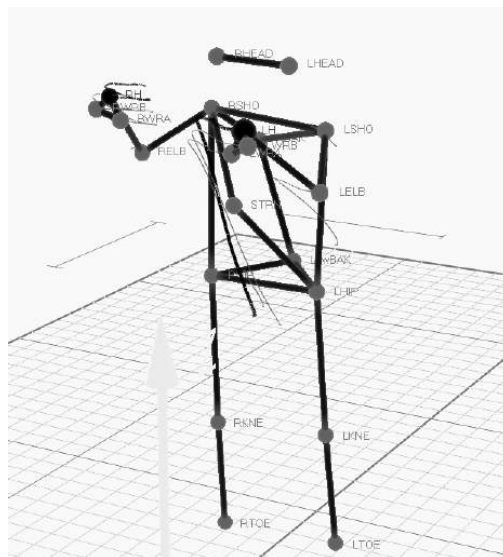


Figure 5. Example of post-processed Motion Capture Recording. The markers are labelled and their relative positions on the co-ordinate system is measured.

Feature 2 Initial, Final, Highest, Lowest vector: These features were obtained by calculating the four features mentioned above as indicators of the melodic contour. This method has been used to form a typology of melodic contours.

Feature 3 Signed relative distances: The obtained signs from Feature 1 are combined with relative distances of each successive bin from the next. The signs and the values are combined to give a more complete picture. Here we considered the pitch values at the bins. These did not represent pitch class sets, and therefore made the computation “genre-agnostic.”

Signed relative distances of melodies are then compared to signed relative distances of average vertical motion to obtain a feature vector.

5. RESULTS

5.1 Correlation between pitch and vertical motion

Feature 3, which considered an analysis of signed relative distances had the correlation coefficient of 0.292 for all 4 melodies, with a p value of 0.836 which does not show a confident trend. Feature 2, containing a feature vector for melodic contour typology, performs with a correlation coefficient of 0.346, indicating a weak positive relationship, with a p value of 0.07, which indicates a significant positive correlation. This feature performs well, but is not robust in terms of its representation of the contour itself, and fails when individual tracings are compared to melodies, yielding an overall coefficient of 0.293.

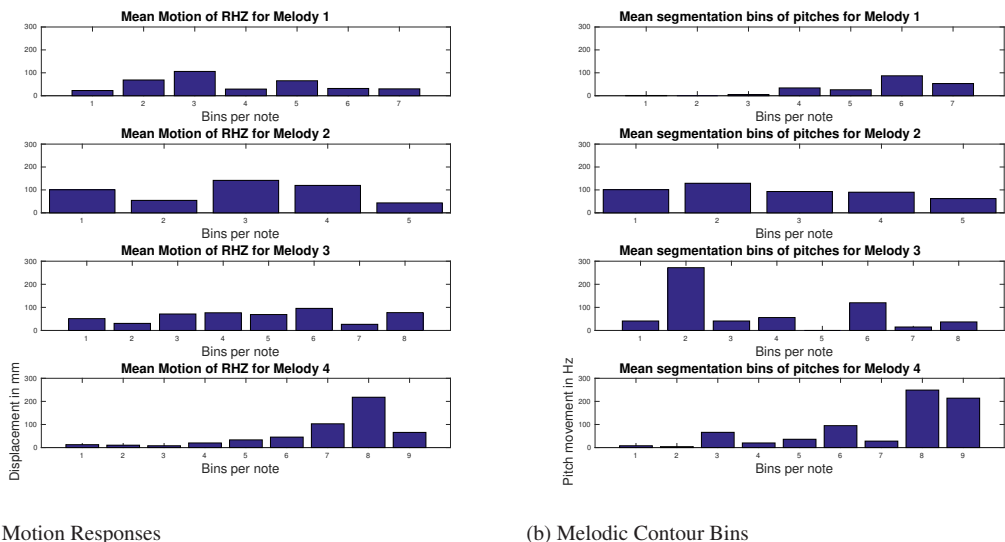


Figure 6. Plots of the representation of features 1 and 3. These features are compared to analyse similarity of the contours.

5.2 Confusion between tracing and target melody

As seen in the confusion matrix in Figure 7, the tracings are not clearly classified as target melodies by direct comparison of contour values itself. This indicates that although the feature vectors might show a strong trend in vertical motion mapping to pitch contours, this is not enough for significant classification. This demonstrates the need for having feature vectors that adequately describe what is going on in music and motion.

6. DISCUSSION

A significant problem when analysing melodies through symbolic data is that a lot of the representation of texture, as explained regarding Melody 2, gets lost. Vibratos, ornaments, and other elements that might be significant for the perception of musical motion can not be captured efficiently through these methods. However, these ornaments certainly seem salient for people’s bodily responses. Further work needs to be carried out to explain the relationship of ornaments and motion, and this relationship might have little or nothing to do with vertical motion.

We also found that the performance of a tracing is fairly intuitive to the eye. The decisions for choosing particular methods of expressing the music through motion do not appear odd when seen from a human perspective, and yet characterizing what are significant features for this cross-modal comparison is a much harder question.

Our results show that vertical motion seems to correlate with pitch contours in a variety of ways, but most significantly when calculated in terms of signed relative values. Signed relative values, as in Feature 3, also maintain the context of the melodic phrase itself, and this is seen to be significant for sound-tracings. Interval distances matter

less than the current ambit of melody that is being traced.

Other contours apart from pitch and melody are also significant for this discussion, especially timbral and dynamic changes. However, the relationships between those and motion were beyond the scope of this paper. The interpretation of motion other than just vertical motion is also not handled within this paper.

The features that were shown to be significant can be applied for the whole dataset to see relationships between vertical motion and melody. Contours of dynamic and timbral change can also be interesting to compare with the same methods against melodic tracings.

7. REFERENCES

- [1] A. Marsden, “Interrogating melodic similarity: a definitive phenomenon or the product of interpretation?” *Journal of New Music Research*, vol. 41, no. 4, pp. 323–335, 2012.
- [2] C. Cronin, “Concepts of melodic similarity in music copyright infringement suits,” *Computing in musicology: a directory of research*, no. 11, pp. 187–209, 1998.
- [3] A. Ghias, J. Logan, D. Chamberlin, and B. C. Smith, “Query by humming: musical information retrieval in an audio database,” in *Proceedings of the third ACM international conference on Multimedia*. ACM, 1995, pp. 231–236.
- [4] L. Lu, H. You, H. Zhang *et al.*, “A new approach to query by humming in music retrieval.” in *ICME*, 2001, pp. 22–25.

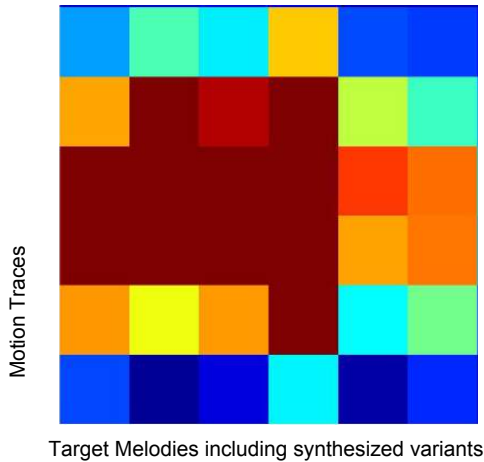


Figure 7. Confusion matrix for Feature 3, to analyse the classification of raw motion contours with pitch contours for 4 melodies.

- [5] N. N. Vempala and F. A. Russo, "Predicting emotion from music audio features using neural networks," in *Proceedings of the 9th International Symposium on Computer Music Modeling and Retrieval (CMMR)*. Lecture Notes in Computer Science London, UK, 2012, pp. 336–343.
- [6] D. Parsons, *The directory of tunes and musical themes*. Cambridge, Eng.: S. Brown, 1975.
- [7] C. R. Adams, "Melodic contour typology," *Ethnomusicology*, pp. 179–215, 1976.
- [8] R. D. Morris, "New directions in the theory and analysis of musical contour," *Music Theory Spectrum*, vol. 15, no. 2, pp. 205–228, 1993.
- [9] M. L. Friedmann, "A methodology for the discussion of contour: Its application to schoenberg's music," *Journal of Music Theory*, vol. 29, no. 2, pp. 223–248, 1985.
- [10] J. Salamon and E. Gómez, "Melody extraction from polyphonic music signals using pitch contour characteristics," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 6, pp. 1759–1770, 2012.
- [11] R. M. Bittner, J. Salamon, S. Essid, and J. P. Bello, "Melody extraction by contour classification," in *Proc. ISMIR*, pp. 500–506.
- [12] E. Gómez and J. Bonada, "Towards computer-assisted flamenco transcription: An experimental comparison of automatic transcription algorithms as applied to a cappella singing," *Computer Music Journal*, vol. 37, no. 2, pp. 73–90, 2013.
- [13] J. C. Ross, T. Vinutha, and P. Rao, "Detecting melodic motifs from audio for hindustani classical music." in *ISMIR*, 2012, pp. 193–198.
- [14] I. Quinn, "The combinatorial model of pitch contour," *Music Perception: An Interdisciplinary Journal*, vol. 16, no. 4, pp. 439–456, 1999.
- [15] G. T. Toussaint, "A comparison of rhythmic similarity measures." in *ISMIR*, 2004.
- [16] D. Bainbridge, C. G. Nevill-Manning, I. H. Witten, L. A. Smith, and R. J. McNab, "Towards a digital library of popular music," in *Proceedings of the fourth ACM conference on Digital libraries*. ACM, 1999, pp. 161–169.
- [17] K. Nymoen, "Analyzing sound tracings: a multimodal approach to music information retrieval," in *Proceedings of the 1st international ACM workshop on Music information retrieval with user-centered and multimodal strategies*, 2011.
- [18] M. B. Küssner and D. Leech-Wilkinson, "Investigating the influence of musical training on cross-modal correspondences and sensorimotor skills in a real-time drawing paradigm," *Psychology of Music*, vol. 42, no. 3, pp. 448–469, 2014.
- [19] G. Athanasopoulos and N. Moran, "Cross-cultural representations of musical shape," *Empirical Musicology Review*, vol. 8, no. 3-4, pp. 185–199, 2013.
- [20] M. A. Schmuckler, "Testing models of melodic contour similarity," *Music Perception: An Interdisciplinary Journal*, vol. 16, no. 3, pp. 295–326, 1999.
- [21] J. B. Prince, M. A. Schmuckler, and W. F. Thompson, "Cross-modal melodic contour similarity," *Canadian Acoustics*, vol. 37, no. 1, pp. 35–49, 2009.
- [22] Z. Eitan and R. Timmers, "Beethovens last piano sonata and those who follow crocodiles: Cross-domain mappings of auditory pitch in a musical context," *Cognition*, vol. 114, no. 3, pp. 405–422, 2010.
- [23] Z. Eitan and R. Y. Granot, "How music moves," *Music Perception: An Interdisciplinary Journal*, vol. 23, no. 3, pp. 221–248, 2006.
- [24] T. Eerola and M. Bregman, "Melodic and contextual similarity of folk song phrases," *Musicae Scientiae*, vol. 11, no. 1 suppl, pp. 211–233, 2007.
- [25] R. I. Godøy, E. Haga, and A. R. Jensenius, "Playing air instruments: mimicry of sound-producing gestures by novices and experts," in *International Gesture Workshop*. Springer, 2005, pp. 256–267.

Paper III

Analyzing free-hand sound-tracings of melodic phrases

Tejaswinee Kelkar, Alexander Refsum Jensenius

Published in *Applied Sciences*, January 2018, pp. 137-157.



Article

Analyzing Free-Hand Sound-Tracings of Melodic Phrases

Tejaswinee Kelkar *  and Alexander Refsum Jensenius 

University of Oslo, Department of Musicology, RITMO Centre for Interdisciplinary Studies in Rhythm, Time and Motion, 0371 Oslo, Norway; a.r.jensenius@imv.uio.no

* Correspondence: tejaswinee.kelkar@imv.uio.no; Tel.: +47-4544-8254

Academic Editor: Vesa Valimäki

Received: 31 October 2017; Accepted: 15 January 2018; Published: 18 January 2018

Abstract: In this paper, we report on a free-hand motion capture study in which 32 participants ‘traced’ 16 melodic vocal phrases with their hands in the air in two experimental conditions. Melodic contours are often thought of as correlated with vertical movement (up and down) in time, and this was also our initial expectation. We did find an arch shape for most of the tracings, although this did not correspond directly to the melodic contours. Furthermore, representation of pitch in the vertical dimension was but one of a diverse range of movement strategies used to trace the melodies. Six different mapping strategies were observed, and these strategies have been quantified and statistically tested. The conclusion is that metaphorical representation is much more common than a ‘graph-like’ rendering for such a melodic sound-tracing task. Other findings include a clear gender difference for some of the tracing strategies and an unexpected representation of melodies in terms of a small object for some of the Hindustani music examples. The data also show a tendency of participants moving within a shared ‘social box’.

Keywords: motion; melody; shape; sound-tracing; multi-modality

1. Introduction

How do people think about melodies as shapes? This question comes out of the authors’ general interest in understanding more about how spatiotemporal elements influence the cognition of music. When it comes to the topic of melody and shape, these terms often seem to be interwoven. In fact, the Concise Oxford Dictionary of Music defines melody as: “A succession of notes, varying in pitch, which has an organized and recognizable shape.” [1]. Here, shape is embedded as a component in the very definition of melody. However, what is meant by the term ‘melodic shape’, and how can we study such melodic shapes and their typologies?

Some researchers have argued for thinking of free-hand movements to music (or ‘air instrument performance’ [2]) as visible utterances similar to co-speech gestures [3–5]. From the first author’s experience as an improvisational singer, a critical part of learning a new singing culture was the physical representation of melodic content. This physical representation includes bodily posture, gestural vocabulary and the use of the body to communicate sung phrases. In improvised music, this also includes the way in which one uses the hands to guide the music and the expectation of a familiar audience from the performing body. These body movements may refer to spatiotemporal metaphors, quite like the ones used in co-speech gestures.

In their theory of cognitive metaphors, Lakoff and Johnson point out how the metaphors in everyday life represent the structure through which we conceptualize one domain with the representation of another [6]. Zbikowski uses this theory to elaborate how words used to describe pitches in different languages are mapped onto the metaphorical space of the ‘vertical dimension’ [7]. Descriptions of melodies often use words related to height, for example: a ‘high’- or ‘low’-pitched

voice, melodies going ‘up’ and ‘down’. Shayan et al. suggest that this mapping might be more strongly present in Western cultures, while the use of other metaphors in other languages, such as thick and thin pitch, might explain pitch using other non-vertical one-dimensional mapping schemata [8]. The vertical metaphor, when tested with longer melodic lines, shows that we respond non-linearly to the vertical metaphors of static and dynamic pitch stimuli [9]. Research in music psychology has investigated both the richness of this vocabulary and its perceptual and metaphorical allusions [10]. However, the idea that the vertical dimension is the most important schema of melodic motion is very persistent [11]. Experimentally, pitch-height correspondences are often elicited by comparing two or three notes at a time. However, when stimuli become more complex, resembling real melodies, the persistence of pitch verticality is less clear. For ‘real’ melodies, shape descriptions are often used, such as arches, curves and slides [12].

In this paper, we investigate shape descriptions through a sound-tracing approach. This was done by asking people to listen to melodic excerpts and then move their body as if their movement was creating the sound. The aim is to answer the following research questions:

1. How do people present melodic contour as shape?
2. How important is vertical movement in the representation of melodic contour in sound-tracing?
3. Are there any differences in sound-tracings between genders, age groups and levels of musical experience?
4. How can we understand the metaphors in sound-tracings and quantify them from the data obtained?

The paper starts with an overview of related research, before the experiment and various analyses are presented and discussed.

2. Background

Drawing melodies as lines has seemed intuitive across different geographies and time periods, from the Medieval neumes to contemporary graphical scores. Even the description of melodies as lines enumerates some of their key properties: continuity, connectedness and appearance as a shape. Most musical cultures in the world are predominantly melodic, which means that the central melodic line is important for memorability. Melodies display several integral patterns of organization and time-scales, including melodic ornaments, motifs, repeating patterns, themes and variations. These are all devices for organizing melodic patterns and can be found in most musical cultures.

2.1. Melody, Prosody and Memory

Musical melodies may be thought of as closely related to language. For example, prosody, which can also be described as ‘speech melody,’ is essential for understanding affect in spoken language. Musical and linguistic experiences with melody can often influence one another [13]. Speech melodies and musical melodies are differentiated on the basis of variance of intervals, delineation and discrete pitches as scales [14]. While speech melodies show more diversity in intonation, there is lesser diversity in prosodic contours internally within a language. Analysis of these contours is used for recognition of languages, speakers and dialects in computation [15].

It has been argued that tonality makes musical melodies more memorable than speech melodies [16], while contour makes them more recognizable, especially in unfamiliar musical styles [17]. Dowling et al. suggest that adults use contour to recognize unfamiliar melodies, even when they have been transposed or when intervals are changed [16]. There is also neurological evidence supporting the idea that contour memory is independent of absolute pitch location [18]. Early research in contour memory and recognition demonstrated that acquisition of memory for melodic contour in infants and children precedes memory for intonation [17,19–22]. Melodic contour is also described as a ‘coarse-grained’ schema that lacks the detail from musical intervals [14].

2.2. Melodic Contour

Contour is often used to refer to sequences of up-down movement in melodies, but there are also several other terms that in different ways touch upon the same idea. Shape, for example, is more generally used for referring to overall melodic phrases. Adams uses the terms contour and shape interchangeably [23] and also adds melodic configuration and outline to the mix of descriptors. Tierney et al. discuss the biological origins and commonalities of melodic shape [24]. They also note the predominance of arch-shaped melodies in music, the long duration of the final notes of phrases, and the biases towards smooth pitch contours. The idea of shapes has also been used to analyze melodies of infants crying [25]. Motif, on the other hand, is often used to refer to a unit of melody that can be operated upon to create melodic variation, such as transposition, inversion, etc. Yet another term is that of melodic chunk, which is sometimes used to refer to the mnemonic properties of melodic units, while *museme* is used to indicate instantaneous perception. Of all these terms, we will stick with contour for the remainder of this paper.

2.3. Analyzing Melodic Contour

There are numerous analysis methods that can be used to study melodic contour, and they may be briefly divided into two main categories: signal-based or symbolic representations. When the contour analysis uses a signal-based representation, a recording of the audio is analyzed with computational methods to extract the melodic line, as well as other temporal and spectral features of the sound. The symbolic representations may start from notated or transcribed music scores and use the symbolic note representations for analysis. Similar to how we might whistle a short melodic excerpt to refer to a larger musical piece, melodic dictionaries have been created to index musical pieces [26]. Such indexes merit a thorough analysis of contour typologies, and several contour typologies were created to this end [23,27,28]. Contour typology methods are often developed from symbolic representations and notated as discrete pitch items. Adam's method for contour typology analysis, for example, codes the initial and final pitches of the melodies as numbers [23]. Parson's typology, on the other hand, uses note directions and their intervals as the units of analysis [26]. There are also examples of matrix comparison methods that code pitch patterns [27]. A comparison of these methods to perception and memory is carried out in [29,30], suggesting that the information-rich models do better than more simplistic ones. Perceptual responses to melodic contour changes have also been studied systematically [30–32], revealing differences between typologies and which ones come closest to resembling models of human contour perception.

The use of symbolic representations makes it easier to perform systematic analysis and modification of melodic music. While such systematic analysis works well for some types of pre-notated music, it is more challenging for non-notated or non-Western music. For such non-notated musics, the signal-based representations may be a better solution, particularly when it comes to providing representations that more accurately describe continuous contour features. Such continuous representations (as opposed to more discrete, note-based representations) allow the extraction of information from the actual sound signal, giving a generally richer dataset from which to work. The downside, of course, is that signal-based representations tend to be much more complex, and hence more difficult to generalize from.

In the field of music perception and cognition, the use of symbolic music representations, and computer-synthesized melodic stimuli, has been most common. This is the case even though the ecological validity of such stimuli may be questioned. Much of the previous research into the perception of melodic contour also suffers from a lack of representation of non-Western and also non-classical musical styles, with some notable exceptions such as [33–35].

Much of the recent research into melodic representations is found within the music information retrieval community. Here, the extraction of melodic contour and contour patterns directly from the signal is an active research topic, and efficient algorithms for extraction of the primary melody have been tested and compared in the MIREX (Music Information Retrieval Evaluation Exchange) competitions

for several years. Melodic contour is also used to describe the instrumentation of music from audio signals, for example in [36,37]. It is also interesting to note that melodic contour is used as the first step to identify musical structure in styles such as in Indian Classical music [38], and Flamenco [39].

2.4. Pitch and the Vertical Dimension

As described in Section 1, the vertical dimension (up-down) is a common way to describe pitch contours. This cross-modal correspondence has been demonstrated in infants [40], showing preferences for concurrence of auditory pitch ‘rising,’ visuospatial height, as well as visual ‘sharpness’. The association with visuospatial height is elaborated further with the SMARC (Spatial-Musical Association of Response Codes) effect [11]. Here, participants show a shorter response time for lower pitches co-occurring with left or bottom response codes, while higher pitches strongly correlated with response codes for right or top. A large body of work tries to tease apart the nuances of the suggested effect. Some of the suggestions include the general setting of the instruments and the bias of reading and writing being from left to right in most of the participants [41], as contributing factors to the manifestation of this effect.

The concepts of contour rely upon pitch height being a key feature of our melodic multimodal experience. Even the enumeration of pitch in graphical formats plays on this persistent metaphor. Eitan brings out the variety of metaphors for pitch quality descriptions, suggesting that up and down might only be one of the ways in which cross-modal correspondence with pitch appears in different languages [9,10]. Many of the tendencies suggested in the SMARC effect are less pronounced when more, and more complex, stimuli appear. These have been tested in memory or matching tasks, rather than asking people to elicit the perceived contours. The SMARC effect may here be seen in combination with the STEARC (Spatial-Temporal Association of Response Codes) effect, stating that timelines are more often perceived as horizontally-moving objects. In combination, these two effects may explain the overwhelming representation of vertical pitch on timelines. The end result is that we now tend to think of melodic representation mostly in line-graph-based terms, along the lines of what Adams ([23], p. 183) suggested:

There is a problem of the musical referents of the terms. As metaphoric depictions, most of these terms are more closely related to the visual and graphic representations of music than to its acoustical and auditory characteristics. Indeed, word-list typologies of melodic contour are frequently accompanied by ‘explanatory’ graphics.

This ‘problem’ of visual metaphors, however, may actually be seen as an opportunity to explore multimodal perception that was not possible to understand at the time.

2.5. Embodiment and Music

The accompaniment of movement to music is understood now as an important phenomenon in music perception and cognition [42]. Research studying the close relationship between sound and movement has shed light on the mechanism to understand action as sound [43] and sound as action [44,45]. Cross-modal correspondence is a phenomenon with a tight interactive loop with the body as a mediator for perceptual, as well as performative roles [46,47]. Some of these interactions show themselves in the form of motor cortex activation when only listening to music [48]. This has further led to empirical studies of how music and body movement share a common structure that affords equivalent and universal emotional expression [49]. Mazzola et al. have also worked on a topological understanding of musical space and the topological dynamics of musical gesture [50].

Studies of Hindustani music show that singers use a wide variety of movements and gestures that accompany spontaneous improvisation [4,51,52]. These movements are culturally codified; they appear in the performance space to aid improvisation and musical thought, and they also convey this information to the listener. The performers also demonstrate a variety of imaginary ‘objects’ with various physical properties to illustrate their musical thought.

Some other examples of research on body movement and melody include Huron's studies of how eyebrow height accompany singing as a cue response to melodic height [53], and studies suggesting that especially arch-shaped melodies have common biological origins that are related to motor constraints [24,54].

2.6. Sound-Tracings

Sound-tracing studies aim at analyzing spontaneous rendering of melodies to movement, capturing instantaneous multimodal associations of the participants. Typically, subjects are asked to draw (or trace) a sound example or short musical excerpt as they are listening. Several of these studies have been carried out with digital tablets as the transducer or the medium [2,44,55–59]. One restriction of using tablets is that the canvas of the rendering space is very restrictive. Furthermore, the dimensionality does not evolve over time and represents a narrow bandwidth of possible movements.

An alternative to tablet-based sound-tracing setups is that of using full-body motion capture. This may be seen as a variation of 'air performance' studies, in which participants try to imitate the sound-producing actions of the music to which they listen [2]. Nymoen et al. carried out a series of sound-tracing studies focusing on movements of the hands [60,61], elaborating on several feature extraction methods to be used for sound-tracing as a methodology. The current paper is inspired by these studies, but extending the methodology to full-body motion capture.

3. Sound-Tracing of Melodic Phrases

3.1. Stimuli

Based on the above considerations and motivations, we designed a sound-tracing study of melodic phrases. We decided to use melodic phrases from vocal genres that have a tradition of singing without words. Vocal phrases without words were chosen so as to not introduce lexical meaning as a confounding variable. Leaving out instruments also avoids the problem of subjects having to choose between different musical layers in their sound-tracing.

The final stimulus set consists of four different musical genres and four stimuli for each genre. The musical genres selected are: (1) Hindustani (North Indian) music, (2) Sami joik, (3) scat singing, (4) Western classical vocalize. The melodic fragments are phrases taken from real recordings, to retain melodies within their original musical context. As can be seen in the pitch plots in Figure 1, the melodies are of varying durations with an average of 4.5 s (SD = 1.5 s). The Hindustani and joik phrases are sung by male vocalists, whereas the scat and vocalize phrases are sung by female vocalists. This is represented in the pitch range of each phrase as seen in Figure 1. The Hindustani and joik melodies are mainly sung in a strong chest voice in this stimulus set. Scat vocals are sung with a transition voice from chest to head. The Vocalizes in this set are sung by a soprano, predominantly in the head register. Hindustani and vocalize samples have one dominant vowel that is used throughout the phrase. The Scat singing examples use traditional 'shoobi-doo-wop' syllables, and joik examples in this set predominantly contain syllables such as 'la-la-lo'.

To investigate the effects of timbre, we decided to create a 'clean' melody representation of each fragment. This was done by running the sound files through an autocorrelation algorithm to create phrases that accurately resemble the pitch content, but without the vocal, timbral and vowel content of the melodic stimulus. These 16 re-synthesized sounds were added to the stimulus set, thus obtaining a total of 32 sound stimuli (summarized in Table 1).

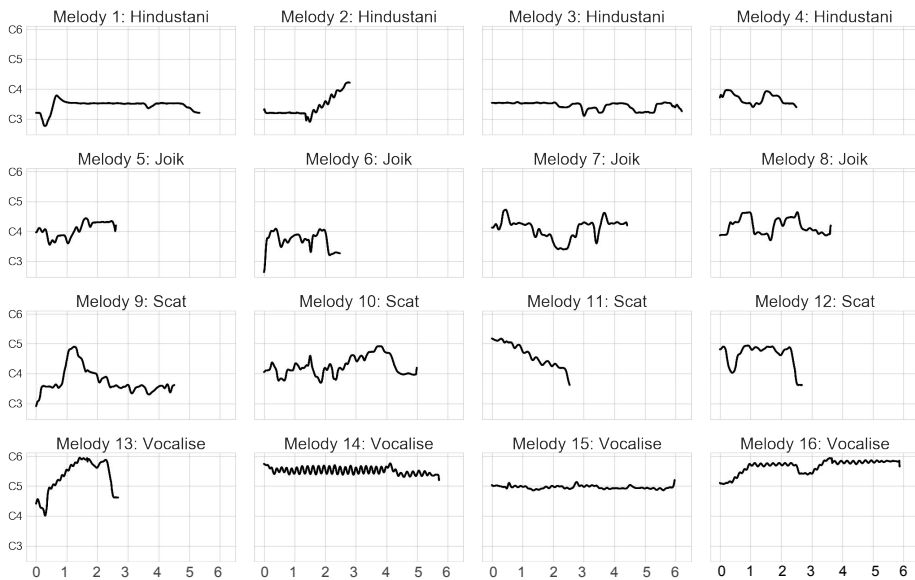


Figure 1. Pitch plots of all the 16 melodic phrases used as experiment stimuli, from each genre. The x axis represents time in seconds, and the y axis represents notes. The extracted pitches were re-synthesized to create a total of 32 melodic phrases used in the experiment.

Table 1. An overview of the 32 different stimuli: four phrases from each musical genre, all of which were presented in both normal and re-synthesized versions.

Type	Hindustani	Joik	Scat	Vocalize
Normal	4	4	4	4
Re-synthesized	4	4	4	4

3.2. Subjects

A total of 32 subjects (17 female, 15 male) was recruited, with a mean age of 31 years (SD = 9 years). The participants were mainly university students and employees, both with and without musical training. Their musical experience was quantized using the OMSI (Ollen Musical Sophistication Index) questionnaire [62], and they were also asked about the familiarity with the musical genres, and their experience with dancing. The mean of the OMSI score was 694 (SD = 292), indicating that the general musical proficiency in this dataset was on the higher side. The average familiarity with Western classical music was 4.03 out of a possible 5 points, 3.25 for jazz music, 1.87 with joik, and 1.71 with Indian classical music. Thus, two genres (vocalize and scat) were more familiar than the two others (Hindustani and joik). All participants provided their written consent for inclusion before they participated in the study, and they were free to withdraw during the experiment. The study obtained ethical approval from the Norwegian Centre for Research Data (NSD), with the project code 49258 (approved on 22 August 2016).

3.3. Procedure

Each subject performed the experiment alone, and the total duration was around 10 min. They were instructed to move their hands as if their movement was creating the melody. The use of the term creating, instead of representing, is purposeful, as in earlier studies [60,63], to avoid the act of

playing or singing. The subjects could freely stand anywhere in the room and face whichever direction they liked, although nearly all of them faced the speakers and chose to stand in the center of the lab. The room lighting was dimmed to help the subjects feel more comfortable to move as they pleased.

The sounds were played at a comfortable listening level through a Genelec 8020 speaker, placed 3 m in front of the subjects. Each session consisted of an introduction, two example sequences, 32 trials and a conclusion, as sketched in Figure 2. Each melody was played twice with a 2-s pause in between. During the first presentation, the participants were asked to listen to the stimuli, while during the second presentation, they were asked to trace the melody. A long beep indicated the first presentation of a stimulus, while a short beep indicated the repetition of a stimuli. All the instructions and required guidelines were recorded and played back through the speaker to not interrupt the flow of the experiment.

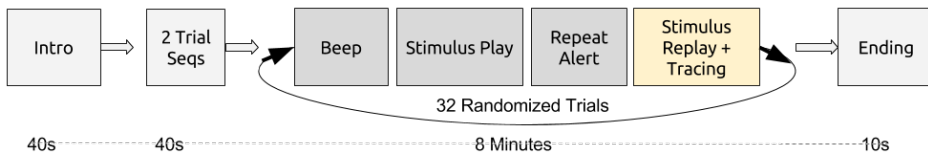


Figure 2. The experiment flow, with an approximate total duration of 10 min

The subjects' motion was recorded using an infrared marker-based motion capture system from Qualisys AB (Gothenburg, Sweden), with 8 Oqus 300 cameras surrounding the space (Figure 3a) and one regular video camera (Canon XF105 (manufactured in Tokyo, Japan)), for reference. Each subject wore a motion capture suit with 21 reflective markers on each joint (Figure 3b). The system captured at a rate of 200 Hz. The data were post-processed in the Qualisys Track Manager software (QTM, v2.16, Qualisys AB, Gothenburg, Sweden), which included labeling of markers and removal of ghost-markers (Figure 3c). We used polynomial interpolation to gap-fill the marker data, where needed. The post-processed data was exported to Python (v2.7.12 and MATLAB (R2013b, MathWorks, Natick, MA, USA) for further analysis. Here, all of the 10-min recordings were segmented using automatic windowing, and each of the segments were manually annotated for further analysis in Section 6.

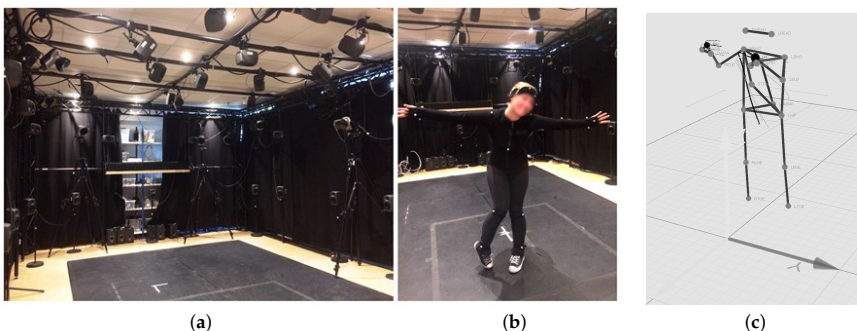


Figure 3. (a) The motion capture lab used for the experiments. (b) The subjects wore a motion capture suit with 21 reflective markers. (c) Screenshot of a stick-figure after post-processing in Qualisys Track Manager software (QTM).

4. Analysis

Even though full-body motion capture was performed, we will in the following analysis only consider data from the right and left hand markers. Marker occlusions from six of the subjects were difficult to account for in the manual annotation process, so only data from 26 participants were used in the analysis that will be presented in Section 5. This analysis is done using comparisons of means and distribution patterns. The occlusion problems were easier to tackle with the automatic analysis, so the analysis that will be presented in Section 6 was performed on data from all 32 participants.

4.1. Feature Selection from Motion Capture Data

Table 2 shows a summary of the features extracted from the motion capture data. Vertical velocity is calculated as the first derivative of the z-axis (vertical motion) for each tracing over time. ‘Quantity of motion’ is a dimensionless quantity representing the overall amount of motion in any direction from frame to frame. Hand distance is calculated as the euclidean distance between the x,y,z coordinates for each marker for each hand. We also calculate the sample-wise distance traveled for each hand marker.

Table 2. The features extracted from the motion capture data.

	Motion Features	Description
1	VerticalMotion	z-axis coordinates at each instant of each hand
2	Range	(Min, Max) tuple for each hand
3	HandDistance	The euclidean distance between the 2d coordinates of each hand
4	QuantityofMotion	The sum of absolute velocities of all the markers
5	DistanceTraveled	Cumulative euclidean distance traveled by each hand per sample
6	AbsoluteVelocity	Uniform linear velocity of all dimensions
7	AbsoluteAcceleration	The derivative of the absolute velocity
8	Smoothness	The number of knots of a quadratic spline interpolation fitted to each tracing
9	VerticalVelocity	The first derivative of the z-axis in each participant’s tracing
10	CubicSpline10Knots	10 knots fitted to a quadratic spline for each tracing

4.2. Feature Selection from Melodic Phrases.

Pitch curves from the melodic phrases were calculated using the autocorrelation algorithm in Praat (v6.0.30, University of Amsterdam, The Netherlands), eliminating octave jumps. These pitch curves were then exported for further analysis together with the motion capture features in Python. Based on contour analysis from the literature [23,30,64], we extracted three different melodic features, as summarized in Table 3.

Table 3. The features extracted from the melodic phrases.

	Melodic Features	Description
1	SignedIntervalDirection	Interval directions (up/down) calculated for each note
2	InitialFinalHighestLowest	Four essential notes of a melody: initial, final, highest, lowest
3	SignedRelativeDistances	Feature 1 combined with relative distances of each successive note from the next, only considering the number of semitones for each successive change.

5. Analysis of Overall Trends

5.1. General Motion Contours

One global finding from the sound-tracing data is that of a clear arch shape when looking at the vertical motion component over time. Figure 4 shows the average contours calculated from the z-values of the motion capture data of all subjects for each of the melodic phrases. It is interesting to note a clear arch-like shape for all of the graphs. This fits with suggestions of a motor constraint

hypothesis suggesting that melodic phrases in general have arch-like shapes [24,54]. In our study, however, the melodies have several different types of shapes (Figure 1), so this may not be the best explanation for the arch-like motion shapes. A better explanation may be that the subjects would start and end their tracing from a resting position in which the hands would be lower than during the tracing, thus leading to the arch shapes seen in Figure 4.

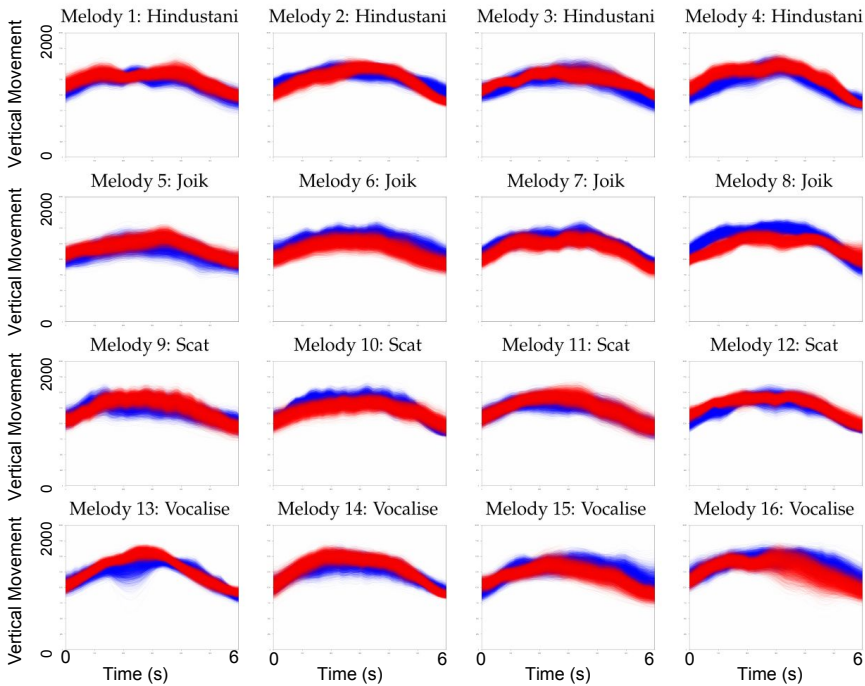


Figure 4. Average contours plotted from the vertical motion capture data in mm (z-axis) for each of the melodies (red for the original and blue for the re-synthesized versions of the melodies). The x-axis represents normalized time, and the y-axis represents aggregated tracing height for all participants.

5.2. Relationship between Vertical Movement and Melodic Pitch Distribution

To investigate more closely the relationship between vertical movement and the distribution of the pitches in the melodic fragments, we may start by considering the genre differences. Figure 5 presents the distribution of pitches in each genre in the stimulus set. These are plotted on a logarithmic frequency scale to represent the perceptual relationships between them. In the plot, each of the four genres are represented by their individual distributions. The color distinction is on the basis of whether the melodic phrase has one direction or many. Phrases closer to being ascending, descending, or stationary are coded as not changing direction. We see that in all of these conditions, the vocalize phrases in the dataset have the highest pitch profiles and the Hindustani phrases have the lowest.

If we then turn to look at the vertical dimension of the tracing data, we would expect to see a similar distribution between genres as that for the pitch heights of the melodies. Figure 6 shows the distribution of motion capture data for all tracings, sorted in the four genres. Here, the distribution of maximum z-values of the motion capture data shows a quite even distribution between genres.

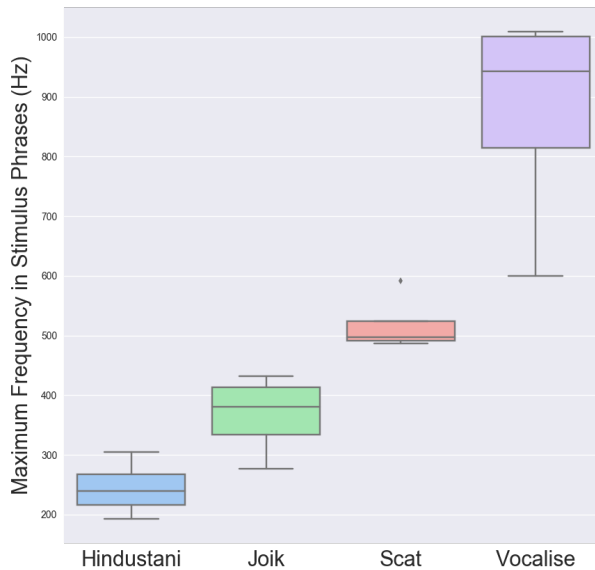


Figure 5. Pitch distribution for each genre based on mean pitches in each phrase. If movement tracings were an accurate representation of absolute pitch height, movement plots should resemble this plot.

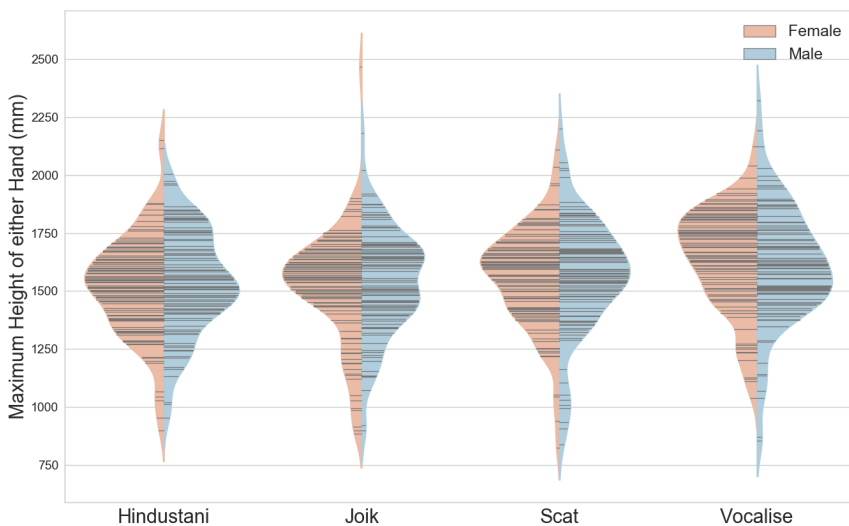


Figure 6. Plots of the maximum height of either hand for each genre. The left/pink distributions are from female subjects, while the right/blue distributions are from the male subjects. Each half of each section of the violin plot represents the probability distribution of the samples. The black lines represent each individual data point.

5.3. Direction Differences

The direction differences in the tracings can be studied by calculating the coefficients of variation of movement in all three axes for both the left (LH) and right (RH) hands. These coefficients are

found to be LHvar $(x,y,z) = (63.7,45.7,26.6)$; RHvar $(x,y,z) = (56.6,87.8,23.1)$, suggesting that the amount of dispersion on the z-axis (the vertical) is the most consistent. This suggests that a wide array of representations in the x and y-axes are used.

The average standard deviations for the different dimensions were found to be LHstd = (99 mm, 89 mm, 185 mm); RHstd = (110 mm, 102 mm, 257 mm). This means that most variation is found in the vertical movement of the right hand, indicating an effect of right-handedness among the subjects.

5.4. Individual Subject Differences

Plots of the distributions of the quantity of motion (QoM) for each subject for all stimuli show a large degree of variation (Figure 7). Subjects 4 and 12, for example, have very small diversity in the average QoM for all of their tracings, whereas Subjects 2 and 5 show a large spread. There are also other participants, such as 22, who move very little on average for all their tracings. Out of the two types of representations (original and re-synthesized stimuli), we see that there is in general a larger diversity of movement for the regular melodies as opposed to the synthesized ones. However, the statistical difference between synthesized and original melodies was not significant.

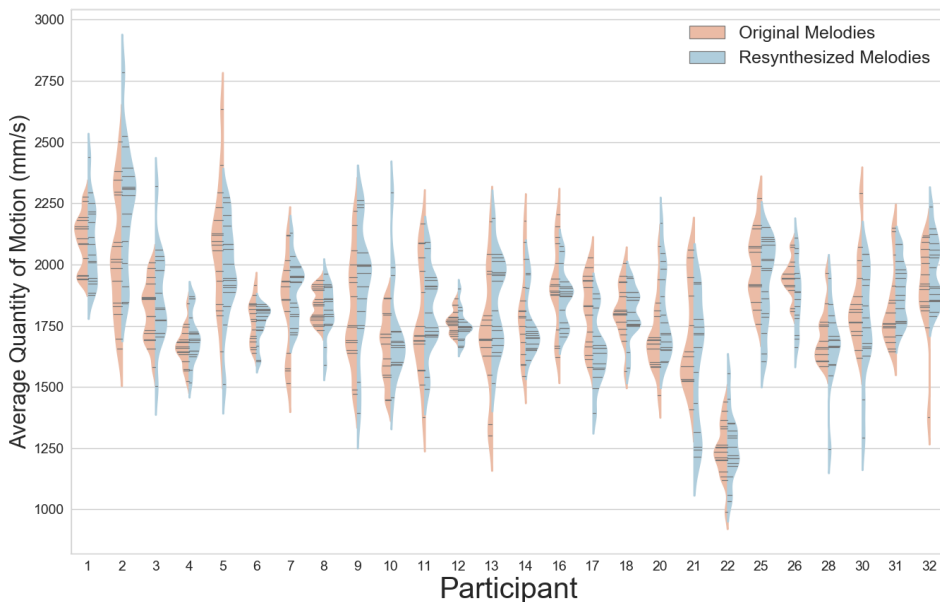


Figure 7. Distribution of the average quantity of motion for each participant. Left/red distributions are of the synthesized stimuli, while the right/green distributions are of the normal recordings.

5.5. Social Box

Another general finding from the data that is not directly related to the question at hand, but that is still relevant for understanding the distribution of data, is what we call a shared ‘social box’ among the subjects. Figure 6 shows that the maximum tracing height of the male subjects were higher than those of the female subjects. This is as expected, but a plot of the ‘tracing volume’ (the spatial distribution in all dimensions) reveals that a comparably small volume was used to represent most of the melodies (Figure 8). Qualitative observation of the recordings reveal that shorter subjects were more comfortable about stretching their hands out, while taller participants tended to use a more restrictive space relative to their height. This happened without there being any physical constraints of

their movements, and no instructions that had pointed in the direction of the volume to be covered by the tracings.

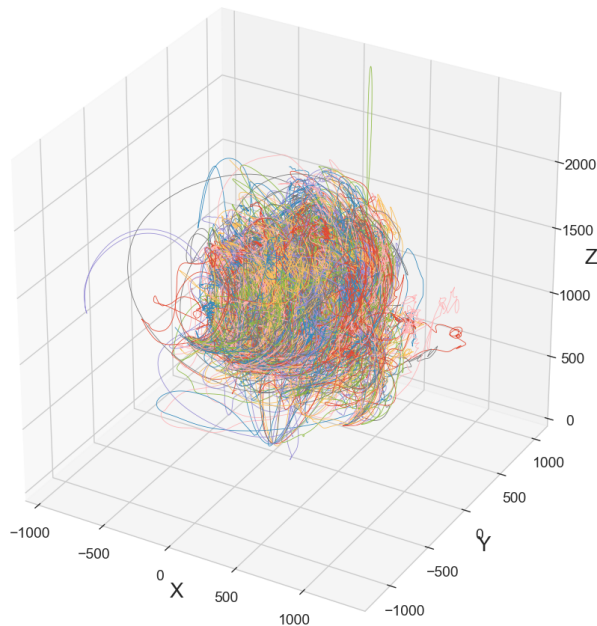


Figure 8. A three-dimensional plot of all sound-tracings for all participants reveal a fairly constrained tracing volume, a kind of ‘social box’ defined by the subjects. Each color represents the tracings of a single participant, and numbers along each axis are millimetres.

It is almost as if the participants wanted to fill up an invisible ‘social box,’ serving as the collective canvas on which everything can be represented. This may be explained by the fact that we share a world together that has the same dimensions: doors, cars, ceilings, and walls are common to us all, making us understand our body as a part of the world in a particular way. In the data from this experiment, we explore this by analyzing the range of motion relative to the heights of the participants through linear regression. The scaled movement range in the horizontal plane is represented in Figure 9, and shows that the scaled range reduces steadily over time as the height of the participants increases. Shorter participants occupy a larger area in the horizontal plane, while taller participants occupy a relatively smaller area. The R^2 coefficient of regression is found to be 0.993, meaning that this effect is significant.

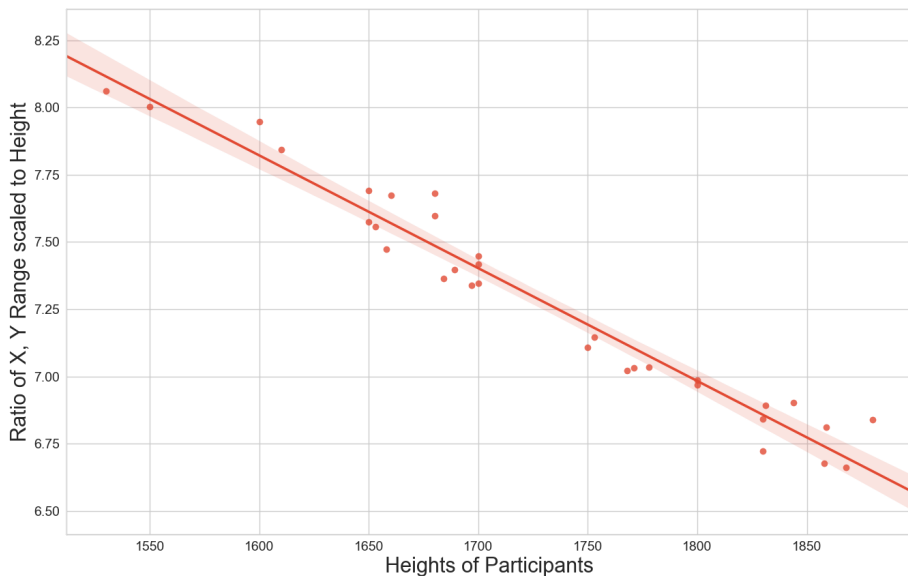


Figure 9. Regression plot of the heights of the participants against scaled (x, y) ranges. There is a clearly decreasing trend for the scaled range of movements in the horizontal plane. The taller a participant, the lower is their scaled range.

5.6. An Imagined Canvas

In a two-dimensional tracing task, such as with pen on paper, the ‘canvas’ of the tracing is both finite and visible all the time. Such a canvas is experienced also for tasks performed with a graphical tablet, even if the line of the tracing is not visible. In the current experiment, however, the canvas is three-dimensional and invisible, and it has to be imagined by the participant. Participants who trace by just moving one hand at a time seem to be using the metaphor of a needle sketching on a moving paper, much like an analogue ECG (Electro CardioGram) machine. Depending on the size of the tracing, the participants would have to rotate or translate their bodies to move within this imagined canvas. We observe different strategies when it comes to how they reach beyond the constraints of their kinesphere, the maximum volume you can reach without moving to a new location. They may step sideways, representing a flat canvas placed before them, or may rotate, representing a cylindrical canvas around their bodies.

6. Mapping Strategies

Through visual inspection of the recordings, we identify a limited number of strategies used in the sound-tracings. We therefore propose six schemes of representation that encompass most of the variation seen in the hands’ movement, as illustrated in Figure 10 and summarized as:

1. One outstretched hand, changing the height of the palm
2. Two hands stretching or compressing an “object”
3. Two hands symmetrically moving away from the center of the body in the horizontal plane
4. Two hands moving together to represent holding and manipulating an object
5. Two hands drawing arcs along an imaginary circle
6. Two hands following each other in a percussive pattern

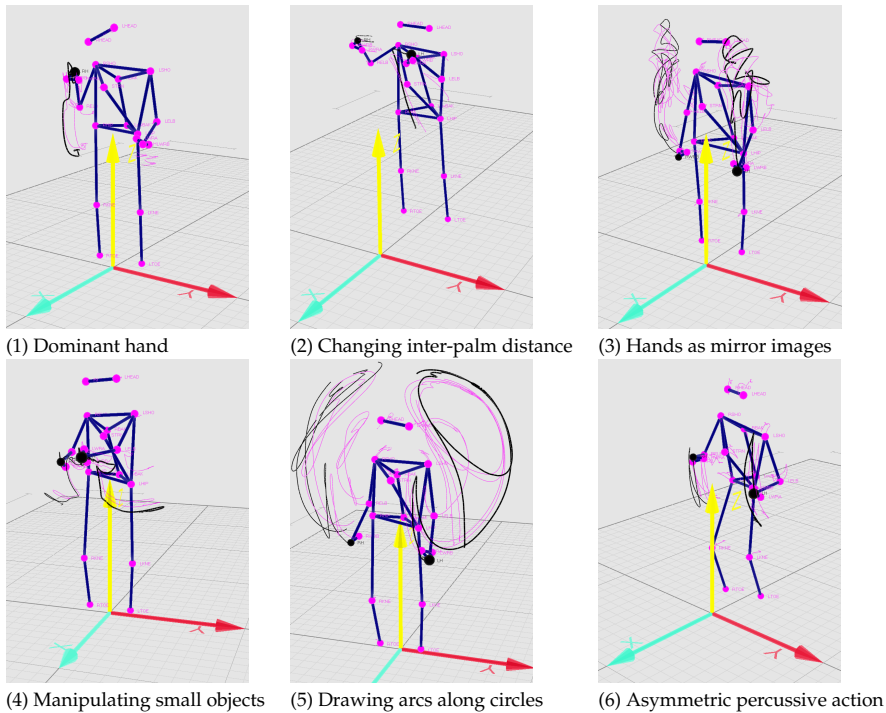


Figure 10. Motion history images exemplifying the six dominant sound-tracing strategies. The black lines from the hands of the stick figures indicate the motion traces of each tracing.

These qualitatively derived strategies were the starting point for setting up an automatic extraction of features from the motion capture data. The pipeline for this quantitative analysis consists of the following steps:

1. Feature selection: Segment the motion capture data into a six-column feature vector containing the (x,y,z) coordinates of the right palm and the left palm, respectively.
2. Calculate quantity of motion (QoM): Calculate the average of the vector magnitude for each sample.
3. Segmentation: Trim data using a sliding window of 1100 samples in size. This corresponds to 5.5 s, to accommodate the average duration of 4.5 s of the melodic phrases. The hop size for the windows is 10 samples, to obtain a large set of windowed segments. The segments that have the maximum mean values are then separated out to get a set of sound-tracings.
4. Feature analysis: Calculate features from Table 4 for each segment.
5. Thresholding: Minimize the six numerical criteria by thresholding the segments based on two-times the standard deviation for each of the computed features.
6. Labeling and separation: Obtain tracings that can be classified as dominantly belonging to one of the six strategy types.

Table 4. Quantitative motion capture features that match the qualitatively selected strategies. QoM, quantities of motion.

#	Strategy	Distinguishing Features	Description	Mean	SD
1	Dominant hand as needle	Right hand QoM much greater than left QoM	$QoM(LHY) \gg QoM(RHY)$	0.50	0.06
2	Changing inter-palm distance	Root mean squared difference of left, right hands in x	$RMS(LHX) - RMS(RHX)$	0.64	0.12
3	Lateral symmetry between hands	Nearly constant difference between left and right hands	$RHX - LHX = C$	0.34	0.11
4	Manipulating a small object	Right and left hands follow similar trajectories in x	$RH(x,y,z) = LH(x,y,z) + C$	0.72	0.07
5	Drawing arcs along circles	Fit of (x,y,z) for left and right hands to a sphere	$x^2 + y^2 + z^2$	0.17	0.04
6	Percussive asymmetry	Dynamic time warp of (x,y,z) of Left, Right Hands	$dtw(RH(xyz), LH(xyz))$	0.56	0.07

After running the segmentation and labeling on the complete data set, we performed a *t*-test to determine whether there is a significant difference between the labeled samples and the other samples. The results, summarized in Table 5 show that the selected features demonstrate the dominance of one particular strategy for many tracings. All except Feature 4 (manipulating a small ‘object’) show significant results compared to all other tracing samples for automatic annotation of hand strategies. While this feature cannot be extracted from the aforementioned heuristic, the simple feature for euclidean distance between two hands proves effective to be able to explain this strategy.

Table 5. Significance testing for each feature against the rest of the features.

Strategy #	<i>p</i> -Value
Strategy 1 vs. rest	0.003
Strategy 2 vs. rest	0.011
Strategy 3 vs. rest	0.005
Strategy 4 vs. rest	0.487
Strategy 5 vs. rest	0.003
Strategy 6 vs. rest	0.006

In Figure 11, we see that hand distance might be an effective way to compare different hand strategies. Strategy 2 performs the best on testing for separability. The hand distance for Strategy 4, for example is significantly lower than the rest. This is because this tracing style represents people who use the metaphor of an imaginary object to represent music. This imaginary object seldom changes its physical properties—its length and breadth and general size is usually maintained.

Taking demographics into account, we see that the distribution of the female subjects’ tracings for vocalizes have a much wider peak than the rest of the genres. In the use of hand strategies, we observe that women use a wider range of hand strategies as compared to men (Figure 11). Furthermore, Strategy 5 (drawing arcs) is done entirely by women. The representation of music as objects is also seen to be more prominent in women, as is the use of asymmetrical percussive motion. Comparing the same distribution of genders for genres, we do not find a difference in overall movement or general body use between the genders. If anything, the ‘social box effect’ makes the height differences of genres smaller than they are.

In Figure 12, we visualize the use of these hand strategies for every melody by all the participants. Strategy 2 is used in 206 tracings, whereas Strategy 5 is used for only 8 tracings. Strategies 1, 3, 4 and 5 are used 182, 180, 161 and 57 times, respectively. Through this heat map in 12, we also can find some outliers for the strategies that are more infrequently used. For example, we see that Melodies 4, 13 and 16 show specially dominant use of some hand strategies.



Figure 11. Hand distance as a feature to discriminate between tracing strategies.



Figure 12. Heat map of representation of hand strategy per melody.

7. Discussion

In this study, we have analyzed people’s tracings to melodies from four musical genres. Although much of the literature points to correlations between melodic pitch and vertical movement, our findings show a much more complex picture. For example, relative pitch height appears to be much more important than absolute pitch height. People seem to think about vocal melodies as actions, rather than interpreting the pitches purely in one dimension (vertical) over time. The analysis of contour features from the literature shows that while tracing melodies through an allocentric representation of the

listening body, the notions of pitch height representations matter much less than previously thought. Therefore contour features cannot be extracted merely by cross-modal comparisons of two data sets. We propose that other strategies can be used for contour representations, but this is something that will have to be developed more in future research.

According to the gestural affordances of musical sound theory [65], several gestural representations can exist for the same sound, but there is a limit to how much they can be manipulated. Our data support this idea of a number of possible and overlapping action strategies. Several spatial and visual metaphors are used by a wide range of people. One interesting finding is that there are gender differences between the representations of the different sound-tracing strategies. Women seem to show a greater diversity of strategies in general, and they also use object-like representations more often than men.

We expected musical genre to have some impact on the results. For example, given that Western vocalizes are sung with a pitch range higher than the rest of the genres in this dataset (Figure 5), it is interesting to note that, on average, people do represent vocalize tracings spatially higher than the rest of the genres. We also found that the melodies with the maximum amount of vibrato (melodies 14 and 16 in Figure 5) are represented with the largest changes of acceleration in the motion capture data. This implies that although the pitch deviation in this case is not so significant, the perception of a moving melody is much stronger by comparison to other melodies that have larger changes in pitch. It could be argued that both melody 4 and 16 contain motivic repetition that cause this pattern. However, repeating motifs are as much parts of melodies 6 and 8 (joik). The values represented by these melodies are applicable to their tracings as original as well as synthesized phrases. The effect of the vowels used in these melodies can also thus be negated. As seen in Figure 12, there are some melodies that stand out for some hand strategies. Melody 4 (Hindustani) is curiously highly represented as a small object. Melody 12 is overwhelmingly represented by symmetrical movements of both hands, while Melodies 8 and 9 are overwhelmingly represented by using 1 hand as the tracing needle.

We find it particularly interesting that subjects picked up on the idea of using small objects as a representation technique in their tracings. The use of small objects to represent melodies is well documented in Hindustani music [52,66–68]. However, the subjects' familiarity score with Hindustani music was quite low, so familiarity can not explain this interesting choice of representation in our study. Looking at the melodic features of melody 4, for example, it is steadily descending in intervals until it ascends again and comes down the same intervals. This may be argued to resemble an object that smoothly slips on a slope, and could be a probable reason for the overwhelming object representation of this particular melody. In future studies, it would be interesting to see whether we can recreate this mapping in other melodies, or model melodies in terms of naturally occurring melodic shapes born out of physical forces interacting with each other.

It is worth noting that there are several limitations with the current experimental methodology and analysis. Any laboratory study of this kind would present subjects with an unfamiliar and non-ecological environment. The results would also to a large extent be influenced by the habitus of body use in general. Experience in dance, sign language, or musical traditions with simultaneous musical gestures (such as conducting), all play a part in the interpretation of music as motion. Despite these limitations, we do see a considerable amount of consistency between subjects.

8. Conclusions

The present study shows that there are consistencies in people's sound-tracing to the melodic phrases used in the experiment. Our main findings can be summarized as:

- There is a clear arch shape when looking at the averages of the motion capture data, regardless of the general shape of the melody itself. This may support the idea of a motor constraint hypothesis that has been used to explain the similar arch-like shape of sung melodies.
- The subjects chose between different strategies in their sound-tracings. We have qualitatively identified six such strategies and have created a set of heuristics to quantify and test their reliability.

- There is a clear gender difference for some of the strategies. This was most evident for Strategy 4 (representing small objects), which women performed more than men.
- The ‘obscure’ strategy of representing melodies in terms of a small object, as is typical in Hindustani music, was also found in participants who had no or little exposure to this musical genre.
- The data show a tendency of moving within a shared ‘social box’. This may be thought of as an invisible space that people constrain their movements to, even without any exposure to the other participants’ tracings. In future studies, it would be interesting to explore how constant such a space is, for example by comparing multiple recordings of the same participants over a period of time.

In future studies, we want to investigate all of these findings in greater detail. We are particularly interested in taking the rest of the body’s motion into account. It would also be relevant to use the results from such studies in the creation of interactive systems, ‘reversing’ the process, that is, using tracing in the air as a method to retrieve melodies from a database. This could open up some exciting end-user applications and also be used as a tool for music performance.

Supplementary Materials: The following are available online at Available online: <http://www.mdpi.com/2076-3417/8/1/135/s1>, supplementary archive consists of data files of the following nature: segmented motion tracings of 26 participants annotated with participant number, melody traced, and hand strategy used. The melodies are from 1 to 16, and the pitch data and sound stimuli are separately provided as well. More information about the same and code can be provided upon request.

Acknowledgments: This work was partially supported by the Research Council of Norway through its Centres of Excellence scheme, Project Number 262762.

Author Contributions: T.K. and A.R.J. both conceived and designed the experiments; T.K. performed the experiments and analyzed the data; both authors contributed analysis tools, and wrote the paper.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Kennedy, M. Rutherford-Johnson, T.; Kennedy, J. *The Oxford Dictionary of Music*, 6th ed.; Oxford University Press: Oxford, UK, 2013. Available online: <http://www.oxfordreference.com/view/10.1093/acref/9780199578108.001.0001/acref-9780199578108> (accessed on 16 January 2018).
2. Godøy, R.I.; Haga, E.; Jensenius, A.R. *Playing “Air Instruments”: Mimicry of Sound-Producing Gestures by Novices and Experts*; International Gesture Workshop; Springer: Berlin/Heidelberg, Germany, 2005; pp. 256–267.
3. Kendon, A. *Gesture: Visible Action as Utterance*; Cambridge University Press: Cambridge, UK, 2004.
4. Clayton, M.; Sager, R.; Will, U. In time with the music: The concept of entrainment and its significance for ethnomusicology. In *European Meetings in Ethnomusicology*; ESEM Counterpoint 1; Romanian Society for Ethnomusicology: Bucharest, Romania, 2005; Volume 11, pp. 1–82.
5. Zbikowski, L.M. Musical gesture and musical grammar: A cognitive approach. In *New Perspectives on Music and Gesture*; Ashgate Publishing Ltd.: Farnham, UK, 2011; pp. 83–98.
6. Lakoff, G.; Johnson, M. Conceptual metaphor in everyday language. *J. Philos.* **1980**, *77*, 453–486.
7. Zbikowski, L.M. Metaphor and music theory: Reflections from cognitive science. *Music Theory Online* **1998**, *4*, 1–8.
8. Shayan, S.; Ozturk, O.; Sicoli, M.A. The thickness of pitch: Crossmodal metaphors in Farsi, Turkish, and Zapotec. *Sens. Soc.* **2011**, *6*, 96–105.
9. Eitan, Z.; Schupak, A.; Gotler, A.; Marks, L.E. Lower pitch is larger, yet falling pitches shrink. *Exp. Psychol.* **2014**, *61*, 273–284.
10. Eitan, Z.; Timmers, R. Beethoven’s last piano sonata and those who follow crocodiles: Cross-domain mappings of auditory pitch in a musical context. *Cognition* **2010**, *114*, 405–422.
11. Rusconi, E.; Kwan, B.; Giordano, B.L.; Umiltà, C.; Butterworth, B. Spatial representation of pitch height: The SMARC effect. *Cognition* **2006**, *99*, 113–129.
12. Huron, D. The melodic arch in Western folksongs. *Comput. Musicol.* **1996**, *10*, 3–23.

13. Fedorenko, E.; Patel, A.; Casasanto, D.; Winawer, J.; Gibson, E. Structural integration in language and music: Evidence for a shared system. *Mem. Cognit.* **2009**, *37*, 1–9.
14. Patel, A.D. *Music, Language, and the Brain*; Oxford University Press: New York, NY, USA, 2010.
15. Adami, A.G.; Mihaescu, R.; Reynolds, D.A.; Godfrey, J.J. Modeling prosodic dynamics for speaker recognition. In Proceedings of the 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'03), Hong Kong, China, 6–10 April 2003; Volume 4.
16. Dowling, W.J. Scale and contour: Two components of a theory of memory for melodies. *Psychol. Rev.* **1978**, *85*, 341–354.
17. Dowling, W.J. Recognition of melodic transformations: Inversion, retrograde, and retrograde inversion. *Percept. Psychophys.* **1972**, *12*, 417–421.
18. Schindler, A.; Herdener, M.; Bartels, A. Coding of melodic gestalt in human auditory cortex. *Cereb. Cortex* **2012**, *23*, 2987–2993.
19. Trehub, S.E.; Becker, J.; Morley, I. Cross-cultural perspectives on music and musicality. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **2015**, *370*, 20140096.
20. Trehub, S.E.; Bull, D.; Thorpe, L.A. Infants' perception of melodies: The role of melodic contour. *Child Dev.* **1984**, *55*, 821–830.
21. Trehub, S.E.; Thorpe, L.A.; Morrongiello, B.A. Infants' perception of melodies: Changes in a single tone. *Infant Behav. Dev.* **1985**, *8*, 213–223.
22. Morrongiello, B.A.; Trehub, S.E.; Thorpe, L.A.; Capodilupo, S. Children's perception of melodies: The role of contour, frequency, and rate of presentation. *J. Exp. Child Psychol.* **1985**, *40*, 279–292.
23. Adams, C.R. Melodic contour typology. *Ethnomusicology* **1976**, *20*, 179–215.
24. Tierney, A.T.; Russo, F.A.; Patel, A.D. The motor origins of human and avian song structure. *Proc. Natl. Acad. Sci. USA* **2011**, *108*, 15510–15515.
25. Díaz, M.A.R.; García, C.A.R.; Robles, L.C.A.; Altamirano, J.E.X.; Mendoza, A.V. Automatic infant cry analysis for the identification of qualitative features to help opportune diagnosis. *Biomed. Signal Process. Control* **2012**, *7*, 43–49.
26. Parsons, D. *The Directory of Tunes and Musical Themes*; S. Brown: Cambridge, UK, 1975.
27. Quinn, I. The combinatorial model of pitch contour. *Music Percept. Interdiscip. J.* **1999**, *16*, 439–456.
28. Narmour, E. *The Analysis and Cognition of Melodic Complexity: The Implication-Realization Model*; University of Chicago Press: Chicago, IL, USA, 1992.
29. Marvin, E.W. A Generalized Theory of Musical Contour: Its Application to Melodic and Rhythmic Analysis of Non-Tonal Music and Its Perceptual and Pedagogical Implications. Ph.D. Thesis, University of Rochester, Rocheste, NY, USA, 1988.
30. Schmuckler, M.A. Testing models of melodic contour similarity. *Music Percept. Interdiscip. J.* **1999**, *16*, 295–326.
31. Schmuckler, M.A. Melodic contour similarity using folk melodies. *Music Percept. Interdiscip. J.* **2010**, *28*, 169–194.
32. Schmuckler, M.A. Expectation in music: Investigation of melodic and harmonic processes. *Music Percept. Interdiscip. J.* **1989**, *7*, 109–149.
33. Eerola, T.; Himberg, T.; Toiviainen, P.; Louhivuori, J. Perceived complexity of western and African folk melodies by western and African listeners. *Psychol. Music* **2006**, *34*, 337–371.
34. Eerola, T.; Bregman, M. Melodic and contextual similarity of folk song phrases. *Musicae Sci.* **2007**, *11*, 211–233.
35. Eerola, T. Are the emotions expressed in music genre-specific? An audio-based evaluation of datasets spanning classical, film, pop and mixed genres. *J. New Music Res.* **2011**, *40*, 349–366.
36. Salamon, J.; Peeters, G.; Röbel, A. Statistical Characterisation of Melodic Pitch Contours and its Application for Melody Extraction. In Proceedings of the ISMIR, Porto, Portugal, 8–12 October 2012; pp. 187–192.
37. Bittner, R.M.; Salamon, J.; Bosch, J.J.; Bello, J.P. Pitch Contours as a Mid-Level Representation for Music Informatics. In Proceedings of the Audio Engineering Society Conference: 2017 AES International Conference on Semantic Audio, Erlangen, Germany, 22–24 June 2017.
38. Rao, P.; Ross, J.C.; Ganguli, K.K.; Pandit, V.; Ishwar, V.; Bellur, A.; Murthy, H.A. Classification of melodic motifs in raga music with time-series matching. *J. New Music Res.* **2014**, *43*, 115–131.
39. Gómez, E.; Bonada, J. Towards computer-assisted flamenco transcription: An experimental comparison of automatic transcription algorithms as applied to a cappella singing. *Comput. Music J.* **2013**, *37*, 73–90.

40. Walker, P.; Bremner, J.G.; Mason, U.; Spring, J.; Mattock, K.; Slater, A.; Johnson, S.P. Preverbal infants are sensitive to cross-sensory correspondences much ado about the null results of Lewkowicz and Minar (2014). *Psychol. Sci.* **2014**, *25*, 835–836.
41. Timmers, R.; Li, S. Representation of pitch in horizontal space and its dependence on musical and instrumental experience. *Psychomusicol. Music Mind Brain* **2016**, *26*, 139–148.
42. Leman, M. *Embodied Music Cognition and Mediation Technology*; MIT Press: Cambridge, MA, USA, 2008.
43. Godøy, R.I. Motor-mimetic music cognition. *Leonardo* **2003**, *36*, 317–319.
44. Jensenius, A.R. Action-Sound: Developing Methods and Tools to Study Music-Related Body Movement. Ph.D. Thesis, University of Oslo, Oslo, Norway, 2007.
45. Jensenius, A.R.; Kvifte, T.; Godøy, R.I. Towards a gesture description interchange format. In Proceedings of the 2006 Conference on New Interfaces for Musical Expression, IRCAM—Centre Pompidou, Paris, France, 4–8 June 2006; pp. 176–179.
46. Gritten, A.; King, E. *Music and Gesture*; Ashgate Publishing, Ltd.: Hampshire, UK, 2006.
47. Gritten, A.; King, E. *New Perspectives on Music and Gesture*; Ashgate Publishing, Ltd.: Surrey, UK, 2011.
48. Molnar-Szakacs, I.; Overy, K. Music and mirror neurons: From motion to 'e'motion. *Soc. Cognit. Affect. Neurosci.* **2006**, *1*, 235–241.
49. Sievers, B.; Polansky, L.; Casey, M.; Wheatley, T. Music and movement share a dynamic structure that supports universal expressions of emotion. *Proc. Natl. Acad. Sci. USA* **2013**, *110*, 70–75.
50. Buteau, C.; Mazzola, G. From contour similarity to motivic topologies. *Musicae Sci.* **2000**, *4*, 125–149.
51. Clayton, M.; Leante, L. Embodiment in Music Performance. In *Experience and Meaning in Music Performance*; Oxford University Press: New York, NY, USA, 2013; pp. 188–207.
52. Rahaim, M. *Musicking Bodies: Gesture and Voice in Hindustani Music*; Wesleyan University Press: Middletown, CT, USA, 2012.
53. Huron, D.; Shanahan, D. Eyebrow movements and vocal pitch height: Evidence consistent with an ethological signal. *J. Acoust. Soc. Am.* **2013**, *133*, 2947–2952.
54. Savage, P.E.; Tierney, A.T.; Patel, A.D. Global music recordings support the motor constraint hypothesis for human and avian song contour. *Music Percept. Interdiscip. J.* **2017**, *34*, 327–334.
55. Godøy, R.I.; Haga, E.; Jensenius, A.R. Exploring Music-Related Gestures by Sound-Tracing: A Preliminary Study. 2006. Available online: https://www.duo.uio.no/bitstream/handle/10852/26899/Godxy_2006b.pdf?sequence=1&isAllowed=y (accessed on 16 January 2018).
56. Glette, K.H.; Jensenius, A.R.; Godøy, R.I. Extracting Action-Sound Features From a Sound-Tracing Study. 2010. Available online: https://www.duo.uio.no/bitstream/handle/10852/8848/Glette_2010.pdf?sequence=1&isAllowed=y (accessed on 16 January 2018).
57. Küssner, M.B. Music and shape. *Lit. Linguist. Comput.* **2013**, *28*, 472–479.
58. Roy, U.; Kelkar, T.; Indurkha, B. TrAP: An Interactive System to Generate Valid Raga Phrases from Sound-Tracings. In Proceedings of the NIME, London, UK, 30 June–3 July 2014; pp. 243–246.
59. Kelkar, T. Applications of Gesture and Spatial Cognition in Hindustani Vocal Music. Ph.D. Thesis, International Institute of Information Technology, Hyderabad, India, 2015.
60. Nymoen, K.; Godøy, R.I.; Jensenius, A.R.; Torresen, J. Analyzing Correspondence Between Sound Objects and Body Motion. *ACM Trans. Appl. Percept.* **2013**, *10*, 9.
61. Nymoen, K.; Caramiaux, B.; Kozak, M.; Torresen, J. Analyzing Sound-Tracings: A Multimodal Approach to Music Information Retrieval. In Proceedings of the 1st International ACM Workshop on Music Information Retrieval with User-centered and Multimodal Strategies (MIRUM '11), Scottsdale, AZ, USA, 30 November 2011; ACM: New York, NY, USA, 2011; pp. 39–44.
62. Ollen, J.E. A Criterion-Related Validity Test of Selected Indicators of Musical Sophistication Using Expert Ratings. Ph.D. Thesis, The Ohio State University, Columbus, OH, USA, 2006.
63. Nymoen, K.; Torresen, J.; Godøy, R.; Jensenius, A. A statistical approach to analyzing sound-tracings. In *Speech, Sound and Music Processing: Embracing Research in India*; Springer: Berlin/Heidelberg, Germany, 2012; pp. 120–145.
64. Schmuckler, M.A. Components of melodic processing. In *Oxford Handbook of Music Psychology*; Oxford University Press: Oxford, UK, 2009; p. 93.
65. Godøy, R.I. Gestural affordances of musical sound. In *Musical Gestures: Sound, Movement, and Meaning*; Routledge: New York, NY, USA, 2010; pp. 103–125.

66. Paschalidou, S.; Clayton, M.; Eerola, T. Effort in Interactions with Imaginary Objects in Hindustani Vocal Music—Towards Enhancing Gesture-Controlled Virtual Instruments. In Proceedings of the 2017 International Symposium on Musical Acoust, Montreal, QC, Canada, 18–22 June 2017.
67. Paschalidou, S.; Clayton, M. Towards a sound-gesture analysis in Hindustani Dhrupad vocal music: Effort and raga space. In Proceedings of the ICMEM, University of Sheffield, Sheffield, UK, 23–25 March 2015; Volume 23, p. 25.
68. Pearson, L.; others. Gesture in Karnatak Music: Pedagogy and Musical Structure in South India. Ph.D. Thesis, Durham University, Durham, UK, 2016.



© 2018 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).

Paper IV

Evaluating a collection of Sound-Tracing Data of Melodic Phrases

Tejaswinee Kelkar, Udit Roy, Alexander Refsum Jensenius

Published in *In Proceedings of the 19th International Society for Music Information Retrieval Conference, Paris, France.*, September 2018, pp. 74-81.

EVALUATING A COLLECTION OF SOUND-TRACING DATA OF MELODIC PHRASES

Tejaswinee Kelkar

RITMO, Dept. of Musicology

University of Oslo

tejaswinee.kelkar@imv.uio.no

Udit Roy

Independent Researcher

udit.roy@alumni.iiit.ac.in

Alexander Refsum Jensenius

RITMO, Dept. of Musicology

University of Oslo

a.r.jensenius@imv.uio.no

ABSTRACT

Melodic contour, the ‘shape’ of a melody, is a common way to visualize and remember a musical piece. The purpose of this paper is to explore the building blocks of a future ‘gesture-based’ melody retrieval system. We present a dataset containing 16 melodic phrases from four musical styles and with a large range of contour variability. This is accompanied by full-body motion capture data of 26 participants performing sound-tracing to the melodies. The dataset is analyzed using canonical correlation analysis (CCA), and its neural network variant (Deep CCA), to understand how melodic contours and sound tracings relate to each other. The analyses reveal non-linear relationships between sound and motion. The link between pitch and verticality does not appear strong enough for complex melodies. We also find that descending melodic contours have the least correlation with tracings.

1. INTRODUCTION

Can hand movement be used to retrieve melodies? In this paper we use data from a ‘sound-tracing’ experiment (Figure 1) containing motion capture data to describe music–motion cross-relationships, with the aim of developing a retrieval system. Details about the experiment and how motion metaphors come to play a role in the representations are presented in [19]. While our earlier analysis was focused on the use of the body and imagining metaphors for tracings [17, 18], in this paper, we will focus on musical characteristics and study music–motion correlations. The tracings present a unique opportunity for cross-modal retrieval, because a direct correspondence between tracing and melodic contour presents an inherent ‘ground-truth.’

Recent research in neuroscience and psychology has shown that *action* plays an important role in perception. In phonology and linguistics, the co-articulation of action and sound is also well understood. Theories from embodied music cognition [22] have been critical to this exploration of multimodal correspondences.

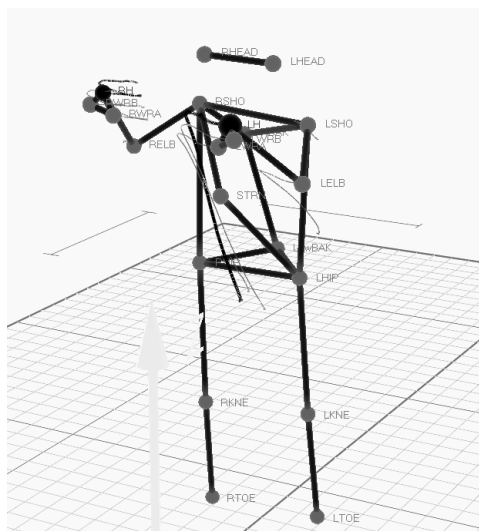


Figure 1. An example of post-processed motion capture data from a sound-tracing study of melodic phrases.

Contour perception is a coarse-level musical ability that we acquire early during childhood [30, 33, 34]. Research suggests that our memory for contour is enhanced when melodies are tonal, and when tonal accent points of melodies co-occur with strong beats [16], making melodic memory a salient feature in musical perception. More generally, it is easier for people to remember the general shape of melody rather than precise intervals [14], especially if they are not musical experts. Coarse representations of melodic contour, such as with drawing or moving hands in the air may be intuitive to capturing musical moments of short time scales [9, 25].

1.1 Research Questions

The inspiration for our work mainly comes from several projects on melodic content retrieval using intuitive and multi-modal representations of musical data. The oldest example of this is the 1975 project titled ‘Directory of Tunes and Musical Themes,’ where the author uses a simplified contour notation method, involving letters for de-



© Tejaswinee Kelkar, Udit Roy, Alexander Refsum Jensenius. Licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0). **Attribution:** Tejaswinee Kelkar, Udit Roy, Alexander Refsum Jensenius. “Evaluating a collection of Sound-Tracing Data of Melodic Phrases”, 19th International Society for Music Information Retrieval Conference, Paris, France, 2018.

noting contour directions, to create a dictionary of musical themes where one may look up a tune they remember [29]. This model is adopted for melodic contour retrieval in Musipedia.com [15]. Another system is proposed in the recent project SoundTracer, in which a user’s motion of their mobile phone is used to retrieve tunes from a music archive [21]. A critical difference between these approaches is how they handle mappings between contour information and musical information, especially differences between time-scales and time-representations. Most of these methods do not have ground-truth models of contours, and instead use one of several ways of mappings, each with its own assumptions.

Godøy et al. has argued for using motion-based, graphical, verbal, and other representations of motion data in music retrieval systems [10]. Liem et al. make a case for using multimodal user-centered strategies as a way to navigate the discrepancy between audio similarity and music similarity [23], with the former referring to more mathematical features, and the latter to more perceptual features. We proceed with this as the point of departure for describing our dataset and its characteristics, to approach the goal of making a system for classifying sound-tracings of melodic phrases with the following specific questions:

1. Are the mappings between melodic contour and motion linearly related?
2. Can we confirm previous findings regarding correlation between pitch and the vertical dimension?
3. What categories of melodic contour are most correlated for sound-tracing queries?

2. RELATED WORK

Understanding the close relationship between music and motion is vital to understanding subjective experiences of performers and listeners, [7, 11, 12]. Many empirical experiments aimed at investigating music–motion correspondences deal with stimulus data that is made to explicitly observe certain mappings, for example pitched and non-pitched sound, vertical dimension and pitch, or player expertise [5, 20, 27]. This means that the music examples themselves are sorted into types of sound (or types of motion). We are more interested in observing how a variety of these mapping relationships change in the content of melodic phrases. For this we use multiple labeling strategies as explained in section 3.4. Another contribution of this work is the use of musical styles from various parts of the world, including those that contain microtonal inflections.

2.1 Multi-modal retrieval

Multi-modal retrieval is the paradigm of information retrieval used to handle different types of data together. The objective is to learn a set of mapping functions that project the different modalities into a common metric space, to be able to retrieve relevant information in one modality

through a query in another. We see that this paradigm is used often in the retrieval of image from text and text from image. Canonical Correlation Analysis (CCA) is a common tool for investigating linear relationships of two sets of variables. In the review paper by Wang et al. for cross modal retrieval [35], several implementations and models are analyzed. CCA is also previously used to show music and brain imaging cross relationships [3].

A previous paper analyzing tracings to pitched and non pitched sounds also used CCA to understand music–motion relationships [25], where the authors describe inherent non-linearity in the mappings, despite finding intrinsic sound-action relationships. This work was extended in [26], in which CCA was used to interpret how different features correlate with each other. Pitch and vertical motion have linear relationships in this analysis, although it is important to note that the sound samples used for this study were short and synthetic.

The biggest reservations in analyzing music–motion data through CCA is that non-linearity cannot be represented, and the dependence of the method on time synchronization is high. The temporal evolution of motion and sound remains linear over time [6]. To get around this, kernel-based methods can be used to introduce non-linearity. Ohkushi et al., present a paper that uses Kernel-based CCA methods to analyze motion and music features together using video sequences from classical ballet, and optical flow based clustering. Bozkurt et al. present a CCA based system to analyze and generate speech and arm motion for prosody-driven synthesis of the ‘beat-gesture’ [4], which is used for emphasizing prosodically salient points in speech. We explore our dataset through CCA due to the previous successes of using this family of methods. We will analyze the same data using Deep CCA, a neural-network approximation of CCA, to understand better the non-linear mappings.

2.2 Canonical Correlation Analysis

CCA is a statistical method to find a linear combination of two variables $X = (x_1, x_2, \dots, x_n)$ and $Y = (y_1, y_2, \dots, y_m)$ with n and m independent variables as vectors a and b such that their correlation $\rho = corr(aX, bY)$ of the transformed variables is maximized. Linear vectors a' and b' can be found such that $a', b' = \operatorname{argmax}_{a,b} corr(a^T X, b^T Y)$. We can then find the second set of coefficients which maximize the correlation of the variables $X' = aX$ and $Y' = bY$ with the additional constraint to keep (X, X') and (Y, Y') uncorrelated. This process can be repeated till $d = \min(m, n)$ dimensions.

The CCA can be extended to include non-linearity by using a neural network to transform the X and Y variables as in the case of Deep CCA [2]. Given the network parameters θ_1 and θ_2 , the objective is to maximize the correlation $corr(f(X, \theta_1), f(Y, \theta_2))$. The network is trained by following the gradient of the correlation objective as estimated from the training data.

3. EXPERIMENT DESCRIPTION

3.1 Procedure

The participants were instructed to move their hands as if their movement was creating the melody. The use of the term ‘creating,’ instead of ‘representing,’ is purposeful, as shown in earlier studies [26,27], to be able to access sound-production as the tracing intent. The experiment duration was about 10 minutes. All melodies were played at a comfortable listening level through a Genelec 8020 speaker, placed 3m in front of the subjects. Each session consisted of an introduction, two example sequences, 32 trials and a conclusion. Each melody was played twice with a 2s pause in between. During the first presentation, the participants were asked to listen to the stimuli, while during the second presentation, they were asked to trace the melody. All the instructions and required guidelines were recorded and played back through the speaker. Their motions are tracked using 8 infra-red cameras from Qualisys (7 Oqus 300 and 1 Oqus 410). We then post-process the data in Qualisys Track Manager (QTM) first by identifying and labeling each marker for each participant. Thereafter, we create a dataset containing Left and Right hand coordinates for all participants.

Six participants in the study had to be excluded due to too many marker dropouts, giving us a final dataset containing 26 participants tracing 32 melodies: 794 tracings for 16 melodic categories.

3.2 Subjects

The 32 subjects (17 females, 15 males) had a mean age of 31 years ($SD = 9$ years). They were mainly university students and employees, both with and without musical training. Their musical experience was quantized using the OMSI (Ollen Musical Sophistication Index) questionnaire [28], and they were also asked about the familiarity with the musical genres, and their experience with dancing. The mean of the OMSI score was 694 ($SD = 292$), indicating that the general musical proficiency in this dataset was on the higher side. The average familiarity with Western classical music was 4.03 out of a possible 5 points, 3.25 for jazz music, 1.87 with Sami joik, and 1.71 with Hindustani music. None of the participants reported having heard any of the melodies played to them. All participants provided their written consent for inclusion before they participated in the study, and they were free to withdraw during the experiment. The study design was approved by the National ethics board (NSD).

3.3 Stimuli

In this study, we decided to use melodic phrases from vocal genres that have a tradition of singing without words. Vocal phrases without words were chosen so as to not introduce lexical meaning as a confounding variable. Leaving out instruments also avoids the problem of subjects having to choose between different musical layers in their sound-tracing. The final stimulus set consists of four different

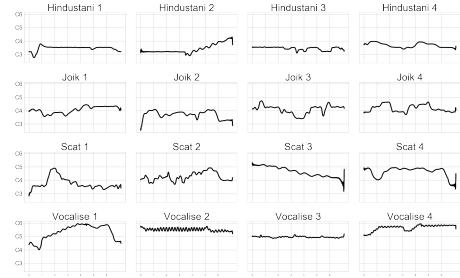


Figure 2. Pitch plots of all the 16 melodic phrases used as experiment stimuli, from each genre. The x axis represents time in seconds, and the y axis represents notes. The extracted pitches were re-synthesized to create a total of 32 melodic phrases used in the experiment.

musical genres and four stimuli for each genre. The musical genres selected are: (1) Hindustani music, (2) Sami joik, (3) jazz scat singing, (4) Western classical vocalise. The melodic fragments are phrases taken from real recordings, to retain melodies within their original musical context. As can be seen in the pitch plots in Figure 2, the melodies are of varying durations with an average of 4.5 s ($SD = 1.5$ s). The Hindustani and joik phrases are sung by male vocalists, whereas the scat and vocalise phrases are sung by female vocalists. This is represented in the pitch range of each phrase as seen in Figure 2.

Seeger	xx	xy	xyy	xyx
Schaeffer	Impulsive	Iterative	Sustained	
Varna	Ascending	Descending	Stationary	Varying
Hood	Arch	Bow	Tooth	Diagonal
Adams	Repetition	Recurrence		

Figure 3. Contour Typologies discussed previously in melodic contour analysis. This figure is representative, made by the authors.

Melodic contours are overwhelmingly written about in terms of pitch, and so we decided to create a ‘clean’ pitch-only representation of each melody. This was done by running the sound files through an autocorrelation algorithm to create phrases that accurately resemble the pitch content, but without the vocal, timbral and vowel content of the melodic stimulus. These 16 re-synthesized sounds were added to the stimulus set, thus obtaining a total of 32 sound stimuli.

	ID	Description
1	All	16 Melodies
2	IJSV	4 Genres
3	ADSC	Ascending, Descending, Steady or Combined
4	OrigVSyn	Original vs Synthesized
5	VibNonVib	Vibrato vs No Vibrato
6	MotifNonMotif	Motif Repetition Present vs Not

Table 1. Multiple labellings for melodic categories: we represent the 16 melodies using 5 different label sets. This helps us analyze which features are best related to which contour classes, genres, or melodic properties.

3.4 Contour Typology Descriptions

We base the selection of melodic excerpts on the descriptions of melodic contour classes as seen in Figure 3. The reference typologies are based on the work of Seeger [32], Hood [13], Schaeffer [8], Adams [1], and the Hindustani classical Varna system. Through these typologies, we hope to cover commonly understood contour shapes and make sure that the dataset contains as many of them as possible.

3.4.1 Multiple labeling

To represent the different contour types and categories that these melodies represent, we create multiple labels that explain the differences. This enables us to understand how the sound tracings actually map to the different possible categories, and makes it easier to see patterns from the data. We describe these labels as seen in Table 3.4.1. Multiple labels allow us to see what categories does the data describe, and which features or combination of features can help retrieve which labels. Some of these labels are categories, while some are one-versus-rest. Category labels include individual melodies, genres, and contour categories, while one-versus-rest correlations are computed for finding whether vibrato, motivic repetitions exist in the melody, and whether the melodic sample is re-synthesized or original.

4. DATASET CREATION

4.1 Preprocessing of Motion Data

We segment each phrase that is traced by the participants, label participant and melody numbers, and extract the data for left and right hand markers for this analysis, since the instructions asked people to trace using their hands. To analyze this data, we are more interested in contour features and shape information than time-scales. We therefore time-normalize our datasets so that every melodic sample and every motion tracing is the same length. This makes it easier to find correlations between music and motion data using different features.

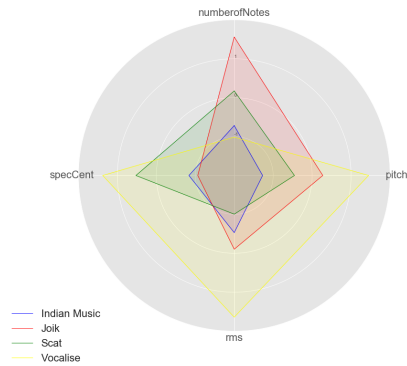


Figure 4. Feature distribution of melodies for each genre. We make sure that a wide range of variability in the features, as described in Table 2 is present in the dataset.

	Feature	Calculated by
1	Pitch	Autocorrelation function using PRAAT
2	Loudness	RMS value of the sound using Librosa
3	Brightness	Spectral Centroid using Librosa
4	Number of Notes	Number of notes per melody

Table 2. Melody features extracted for analysis, and details of how they are extracted.

5. ANALYSIS

5.1 Music

Since we are mainly interested in melodic correlations, the most important feature describing melodies is to extract pitch. For this, we use autocorrelation algorithm available in the PRAAT phonetic program. We use Librosa v0.5.1 [24] to compute the RMS energy (loudness), and the brightness using Spectral Centroid. We transcribe the melodies to get the number of notes per melody. The distribution of these features can be seen for each genre in the stimulus set in Figure 4. We have tried to be true to the musical styles used in this study, most of which do not have written notation as an inherent part of their pedagogy.

5.2 Motion

For tracings, we calculate 9 features that describe various characteristics of motion. We record only X and Z axes, as maximum motion is found along these directions. The derivatives of motion (velocity, acceleration, jerk) and quantity of motion (QoM) which is a cumulative velocity quantity are calculated. Distance between hands, cumulative distance, and symmetry features are calculated as indicators of contour-supporting features, as found in previous studies.

	Feature	Description
1	X-coordinate (X)	Axis corresponding to the direction straight ahead of the participant
2	Z-coordinate (Z)	Axis corresponding to the upwards direction
3	Velocity (V)	First derivative of vertical position
4	Acceleration (A)	Second derivative of vertical position
5	Quantity of Motion	Sum of absolute velocities for all markers
6	Distance between Hands	Sample-wise Euclidean distance between hand markers
7	Jerk	Third derivative of vertical position
8	Cumulative Distance Traveled	Euclidean distance traveled per sample per hand
9	Symmetry	Difference between the left and right hand in terms of vertical position and horizontal velocity

Table 3. Motion features used for analysis. 1-5 are for the dominant hand, while 6-9 are features for both hands.

5.3 Joint Analysis

In this section we present our analysis on our dataset with these two feature sets. We analyze the tracings for each melody as well as utilize the multiple label sets to discover interesting patterns in our dataset which are relevant for a retrieval application.

5.3.1 Dynamic Time Warping

Dynamic Time Warping (DTW) is a method to align sequences of different lengths using substitution, addition and subtraction costs. It is a non-metric method giving us the distance between two sequences after alignment.

In recent research, vertical motion has been shown to correlate with pitch in the past for simple sounds. Some form of non-alignment is also observed between the motion and pitch signals. We perform the same analysis on our data: compute the correlation between pitch and motion in the Z axis before and after alignment with DTW for the 16 melodies and plot their mean and variance in Figure 5.

5.3.2 Longest Run-lengths

While observing the dataset, we find that longest ascending and descending sequences in the melodies are most often reliably represented in the motions, although variances in stationary notes, and ornaments is likely to be much higher. To exploit this feature in tracings, we use ‘‘Longest Run-lengths’’ as a measure. We find multiple subsequences following a pattern which can possess discriminative qualities. For our analysis, we use the ascending and descending patterns, thus finding the subsequences

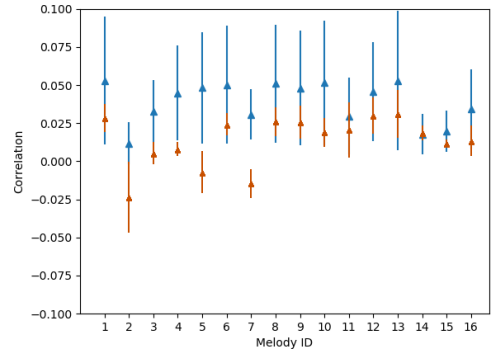


Figure 5. Correlations of pitch with raw data (red) vs after DTW-alignment (blue). Although a DTW alignment improves the correlation, we observe that correlation is still low suggesting that vertical motion and pitch height are not that strongly associated.

from the feature sequence which are purely ascending or descending. We then rank the subsequences and build a feature vector from the lengths of the top N results. This step is particularly advantageous when comparing features from motion and music sequences as it captures the overall presence of the pattern in the sequence remaining invariant to the mis-alignment or lag between the sequences from different modalities. As an example, if we select the Z-axis motion of the dominant hand and the melody pitch as our sequences and retrieve top 3 ascending subsequence lengths. To make the features robust, we do a low pass filtering of the sequence as a preprocessing step.

We analyze our dataset by computing the features for few combinations of motion and music features for ascending and descending patterns. Thereafter, we perform CCA and show the resulting correlation of first transformed dimension in Table 4. We utilize the various label categories generated for the melodies, and show the impact of the features on the labels from each category in Tables 4 and 5. We select the top four run lengths as our feature for each music–motion feature sequence. For Deep CCA analysis, we use a two layered network (same for both motion and music features) with 10 and 4 neurons. A final round of linear CCA is also performed on the network output.

6. RESULTS AND DISCUSSION

Figure 5 shows correlations with raw data and after DTW alignment between the vertical motion and pitch for each melody. Overall, the correlation improves after DTW alignment, suggesting phase lags and phase differences between the timing of melodic peaks and onsets, and those of motion. We see no significant differences between genres, although the improvement in correlations for the vocalize examples is the least pre and post DTW. This could be because of the continuous vibrato in these examples, causing people to use more ‘shaky’ representations which are most

Motion	Music	All		ADSC		IJSV	
Ascend Pattern		CCA	Deep CCA	CCA	Deep CCA	CCA	Deep CCA
Z	Pitch	0.19	0.23	0.25 0.16 0.09 0.05	0.24 0.17 0.12 0.13	0.16 -0.13 0.01 0.37	0.19 0.21 0.08 0.36
Z + V	Pitch	0.21	0.27	0.26 0.09 0.15 0.10	0.30 0.03 0.05 0.17	0.22 -0.13 -0.01 0.35	0.24 0.25 0.15 0.34
All	All	0.33	0.44	0.31 0.14 0.19 0.29	0.44 0.29 0.01 0.36	0.30 0.28 0.23 0.42	0.38 0.43 0.27 0.52
Descend Pattern							
Z	Pitch	0.18	0.21	0.16 -0.11 0.15 0.20	0.17 0.19 0.09 0.19	0.22 0.21 -0.04 0.23	0.22 0.18 0.08 0.28
Z + V	Pitch	0.21	0.31	0.23 0.03 0.14 0.22	0.28 0.28 0.30 0.32	0.26 0.23 0.10 0.24	0.42 0.18 0.34 0.17
All	All	0.35	0.44	0.39 0.12 0.20 0.25	0.38 0.02 0.37 0.37	0.35 0.25 0.12 0.36	0.40 0.22 0.14 0.52

Table 4. Correlations for all samples in the dataset and the two major categorizations of music labels, using ascend and descend patterns as explained in Section 5.3.2, and features from Tables 3 and 2

Motion	Music	MotifNonMotif		OrgSyn		VibNonVib	
Ascend Pattern		CCA	Deep CCA	CCA	Deep CCA	CCA	Deep CCA
Z	Pitch	0.05 0.23	0.13 0.26	0.19 0.19	0.22 0.25	0.33 0.07	0.33 0.13
Z + V	Pitch	0.10 0.24	0.17 0.31	0.19 0.22	0.24 0.31	0.33 0.09	0.32 0.20
All	All	0.29 0.34	0.36 0.47	0.30 0.35	0.42 0.45	0.38 0.29	0.49 0.40
Descend Pattern							
Z	Pitch	0.20 0.17	0.19 0.21	0.20 0.16	0.23 0.18	0.20 0.17	0.24 0.18
Z + V	Pitch	0.22 0.22	0.32 0.29	0.24 0.20	0.35 0.26	0.22 0.22	0.14 0.34
All	All	0.25 0.40	0.37 0.45	0.38 0.33	0.45 0.44	0.33 0.35	0.54 0.35

Table 5. Correlations for two-class categories, using ascend and descend patterns as explained in Section 5.3.2 with features from Tables 3 and 2

consistent between participants. The linear mappings of pitch and vertical motion are limited, making the dataset challenging. This also means that the associations between pitch and vertical motion, as described in previous studies, are not that clear for this stimulus set, especially as we use musical samples that are not controlled for being isochronous, nor equal tempered.

Thereafter, we conduct CCA and Deep CCA analysis as seen in Tables 4, 5. Overall, Deep CCA performs better than its linear counterpart. We find better correlation with all features from Table 3, as opposed to just using vertical motion and velocity. With ascending and descending longest run-lengths, we are able to achieve similar results for correlating all melodies with their respective tracings. However, descending contour classification does not have similar success. There is more general agreement on contour with some melodies than others, with purely descending melodies having particularly low correlation. There is some evidence that descending intervals are harder to identify than ascending intervals [31], and this could explain a low level of agreement in this study amongst people for descending melodies. Studying differences between ascending and descending contours requires further study.

While using genre-labels (IJSV) for correlation, we find that scat samples show the least correlation, and the least improvement. Speculatively, this could be related to the high number of spoken syllables in this style, even though the syllables are not words. Deep CCA also gives an overall correlation of 0.54 for recognizing melodies containing vibrato from the dataset. This is an indication that sonic

textures are well represented in such a dataset.

With all melody and all motion features, we find an overall correlation of 0.44 with Deep CCA, for both the longest ascend and longest descend features. This supports the view that non-linearity is inherent to tracings.

7. CONCLUSIONS AND FUTURE WORK

Interest in cross-modal systems is growing in the context of multi-modal analysis. Previous studies in this area include shorter time scales or synthetically generated isochronous music samples. The strength of this particular study is in using musical excerpts as are performed, and that the performed tracings are not iconic or symbolic, but spontaneous. This makes the dataset a step closer to understanding contour perception in melodies. We hope that the dataset will prove useful for pattern mining, as it presents novel multimodal possibilities for the community and could be used for user-centric retrieval interfaces.

In the future, we wish to create a system to synthesize melody–motion pairs based on training a network to this dataset, and conducting a user evaluation study, where users evaluate system generated music–motion pairs in a forced–choice paradigm.

8. ACKNOWLEDGMENTS

Partially supported by the Research Council of Norway through its Centres of Excellence scheme (262762 & 250698), and the Nordic Sound and Music Computing Network funded by the Nordic Research Council.

9. REFERENCES

- [1] Charles R Adams. Melodic contour typology. *Ethnomusicology*, pages 179–215, 1976.
- [2] Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu. Deep canonical correlation analysis. In *International Conference on Machine Learning*, pages 1247–1255, 2013.
- [3] Nick Gang Blair Kaneshiro Jonathan Berger and Jacek P Dmochowski. Decoding neurally relevant musical features using canonical correlation analysis. In *Proceedings of the 18th International Society for Music Information Retrieval Conference, Suzhou, China, 2017*.
- [4] Elif Bozkurt, Yücel Yemez, and Engin Erzin. Multimodal analysis of speech and arm motion for prosody-driven synthesis of beat gestures. *Speech Communication*, 85:29–42, 2016.
- [5] Baptiste Caramiaux, Frédéric Bevilacqua, and Norbert Schnell. Towards a gesture-sound cross-modal analysis. In *International Gesture Workshop*, pages 158–170. Springer, 2009.
- [6] Baptiste Caramiaux and Atau Tanaka. Machine learning of musical gestures. In *NIME*, pages 513–518, 2013.
- [7] Martin Clayton and Laura Leante. Embodiment in music performance. 2013.
- [8] Rolf Inge Godøy. Images of sonic objects. *Organised Sound*, 15(1):54–62, 2010.
- [9] Rolf Inge Godøy, Egil Haga, and Alexander Refsum Jensenius. Exploring music-related gestures by sound-tracing: A preliminary study. 2006.
- [10] Rolf Inge Godøy and Alexander Refsum Jensenius. Body movement in music information retrieval. In *10th International Society for Music Information Retrieval Conference*, 2009.
- [11] Anthony Gritten and Elaine King. *Music and gesture*. Ashgate Publishing, Ltd., 2006.
- [12] Anthony Gritten and Elaine King. *New perspectives on music and gesture*. Ashgate Publishing, Ltd., 2011.
- [13] Mantle Hood. *The ethnomusicologist*, volume 140. Kent State Univ Pr, 1982.
- [14] David Huron. The melodic arch in western folksongs. *Computing in Musicology*, 10:3–23, 1996.
- [15] K Irwin. Musipedia: The open music encyclopedia. *Reference Reviews*, 22(4):45–46, 2008.
- [16] Mari Riess Jones and Peter Q Pfordresher. Tracking musical patterns using joint accent structure. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, 51(4):271, 1997.
- [17] Tejaswinee Kelkar and Alexander Refsum Jensenius. Exploring melody and motion features in sound-tracings. In *Proceedings of the SMC Conferences*, pages 98–103. Aalto University, 2017.
- [18] Tejaswinee Kelkar and Alexander Refsum Jensenius. Representation strategies in two-handed melodic sound-tracing. In *Proceedings of the 4th International Conference on Movement Computing*, page 11. ACM, 2017.
- [19] Tejaswinee Kelkar and Alexander Refsum Jensenius. Analyzing free-hand sound-tracings of melodic phrases. *Applied Sciences*, 8(1):135, 2018.
- [20] M Kussner. Creating shapes: musicians and non-musicians visual representations of sound. In *Proceedings of 4th Int. Conf. of Students of Systematic Musicology, U. Seifert and J. Wewers, Eds. epOs-Music, Osnabrück (Forthcoming)*, 2012.
- [21] Olivier Lartillot. Soundtracer, 2018.
- [22] Marc Leman. *Embodied music cognition and mediation technology*. MIT Press, 2008.
- [23] Cynthia Liem, Meinard Müller, Douglas Eck, George Tzanetakis, and Alan Hanjalic. The need for music information retrieval with user-centered and multimodal strategies. In *Proceedings of the 1st international ACM workshop on Music information retrieval with user-centered and multimodal strategies*, pages 1–6. ACM, 2011.
- [24] Brian McFee, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. librosa: Audio and music signal analysis in python. 2015.
- [25] Kristian Nymoen, Baptiste Caramiaux, Mariusz Kozak, and Jim Torresen. Analyzing sound tracings: A multimodal approach to music information retrieval. In *Proceedings of the 1st International ACM Workshop on Music Information Retrieval with User-centered and Multimodal Strategies*, MIRUM '11, pages 39–44. New York, NY, USA, 2011. ACM.
- [26] Kristian Nymoen, Rolf Inge Godøy, Alexander Refsum Jensenius, and Jim Torresen. Analyzing correspondence between sound objects and body motion. *ACM Trans. Appl. Percept.*, 10(2):9:1–9:22, June 2013.
- [27] Kristian Nymoen, Jim Torresen, Rolf Godøy, and Alexander Refsum Jensenius. A statistical approach to analyzing sound tracings. *Speech, sound and music processing: Embracing research in India*, pages 120–145, 2012.
- [28] Joy E Ollen. *A criterion-related validity test of selected indicators of musical sophistication using expert ratings*. PhD thesis, The Ohio State University, 2006.
- [29] Denys Parsons. *The directory of tunes and musical themes*. Cambridge, Eng.: S. Brown, 1975.

- [30] Aniruddh D Patel. *Music, language, and the brain*. Oxford university press, 2010.
- [31] Art Samplaski. Interval and interval class similarity: Results of a confusion study. *Psychomusicology: A Journal of Research in Music Cognition*, 19(1):59, 2005.
- [32] Charles Seeger. On the moods of a music-logic. *Journal of the American Musicological Society*, 13(1/3):224–261, 1960.
- [33] Sandra E Trehub, Judith Becker, and Iain Morley. Cross-cultural perspectives on music and musicality. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 370(1664):20140096, 2015.
- [34] Sandra E Trehub, Dale Bull, and Leigh A Thorpe. Infants' perception of melodies: The role of melodic contour. *Child development*, pages 821–830, 1984.
- [35] Kaiye Wang, Qiyue Yin, Wei Wang, Shu Wu, and Liang Wang. A comprehensive survey on cross-modal retrieval. *arXiv preprint arXiv:1607.06215*, 2016.

Appendices

Appendix A

Details of the Experiments

A.1 Marker Labeling

For the 21 markers place on the body, the following are the codes:

1. RTOE: Right foot marker
2. LTOE: Left foot marker
3. RKNE: Right knee marker
4. LKNE: Left knee marker
5. RHEAD: Right head marker
6. LHEAD: Left head marker
7. STRN: Sternum marker
8. RSHO: Right shoulder marker
9. LSHO: Left shoulder marker
10. RBAK: Right back marker (asymmetric)
11. RWRA: Right wrist inside marker
12. RWRB: Right wrist outside marker
13. LWRA: Left wrist inside marker
14. LWRB: Left wrist outside marker
15. RH: Right palm marker
16. LH: Left palm marker
17. RELB: Right elbow marker
18. LELB: Left elbow marker
19. LowBAK: Tailbone marker
20. RHIP: Right hip marker
21. LHIP: Left hip marker

A. Details of the Experiments

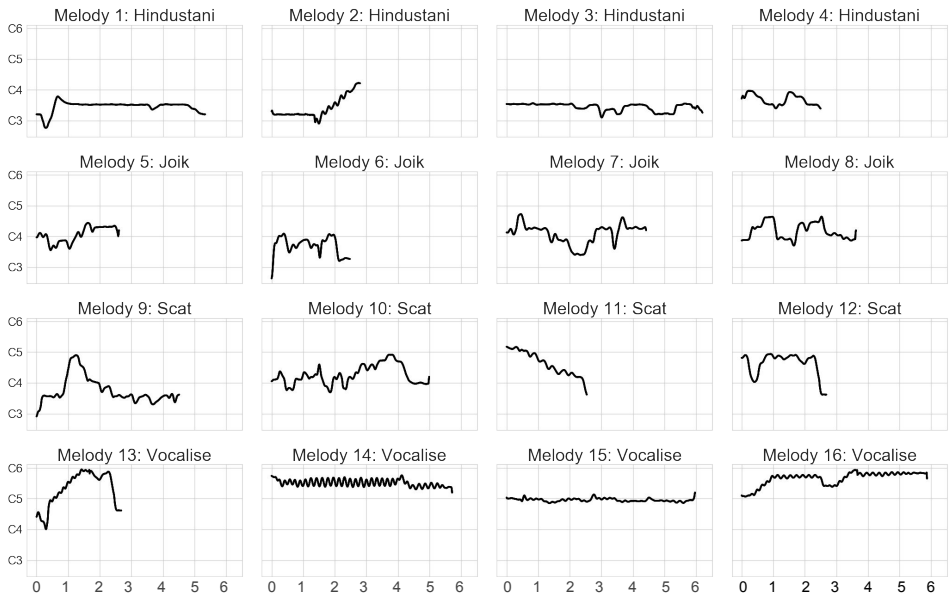


Figure A.1: The 16 melodies used as the stimulus set for all the experiments in the thesis come from four different music cultures and contain no words.

A.2 List of Melodic Stimuli

1. North Indian singing: All four melodies are sung by Ginde (of Washington Ethnomusicology Archives, 1991). The individual features are mentioned below.
 - Melody 1: stationary contour profile, lowest note of the stimulus set
 - Melody 2: ascending contour profile
 - Melody 3: stationary contour profile
 - Melody 4: motivic repetition
2. Joik: The individual features are mentioned below.
 - Melody 5: Per Henderek Haetta (Quarja), antecedent
 - Melody 6: Per Henderek Haetta (Quarja), consequent
 - Melody 7: Nils N. Eira (Track 49), audible vocal break
 - Melody 8: Inga Susanne Haetta (Markel Joavna Piera, motivic repetition)
3. Jazz Scat: All four melodies are sung by Fitzgerald as described in Chapter 5. The individual features for the choices are mentioned below.

- Melody 9: descending contour profile
 - Melody 10: no clear contour direction, rhythmic play
 - Melody 11: descending contour profile
 - Melody 12: ascending contour profile
4. Vocalise: All four melodies are sung by June Anderson, as described in Chapter 5. The individual features are mentioned below.
- Melody 13: ascending contour profile
 - Melody 14: extreme vibrato of a whole tone
 - Melody 15: stationary contour profile
 - Melody 16: motivic repetition, highest sustained note of the stimulus set

A.3 File Formats for Symbolic Music Notation

Some symbolic representation languages that inform melodic data sets are as follows:

1. MIDI Music exchange digital interface: This is a standard interchange format for music data, that carries event messages specifying, pitch, duration, velocity, vibrato, panning, and clock signals to set the tempo. MIDI messages have made it possible for digital musical instruments to communicate with each other.
2. MusicXML : This is an XML based representation of music notation for the web.
3. Humdrum Syntax: Humdrum is a protocol created by David Huron in 1980, that is now widely used as a set of command line tools for music analysis.
4. Pitch class: set of pitches described as numbers.
5. Music Encoding Initiative (MEI): MEI is a standard for encoding symbolic musical data
6. Notation Interchange File Format (NIFF): This is a file format to be able to use notation in various types of software for reading music notation seamlessly.

Appendix B

List of implemented functions and features

The datasets and codes can be found on

<http://tejaswineek.github.io>

As a part of the code developed for this thesis, the following functions have been implemented for mocap data:

B.1 Dependencies

The python toolbox in progress has dependencies in numpy, scipy, matplotlib, pandas.

B.2 Data Types

The datasets are a collection of tsv files. Each file encodes the following information: [Participant ID, Stimulus ID, Strategy ID].

B.3 Functions

B.3.1 Motion Features

1. Quantity of Motion
2. Range
3. Hand Distance
4. Upsampling
5. Downsampling
6. Velocity
7. Acceleration
8. Jerk
9. Normalize

We are able to plot the data in the following ways:

B. List of implemented functions and features

1. 2D plotting

2. 3D plotting

Additional functions for analysis:

1. Spline Interpolation

2. Movement Plane

3. Peak Detection