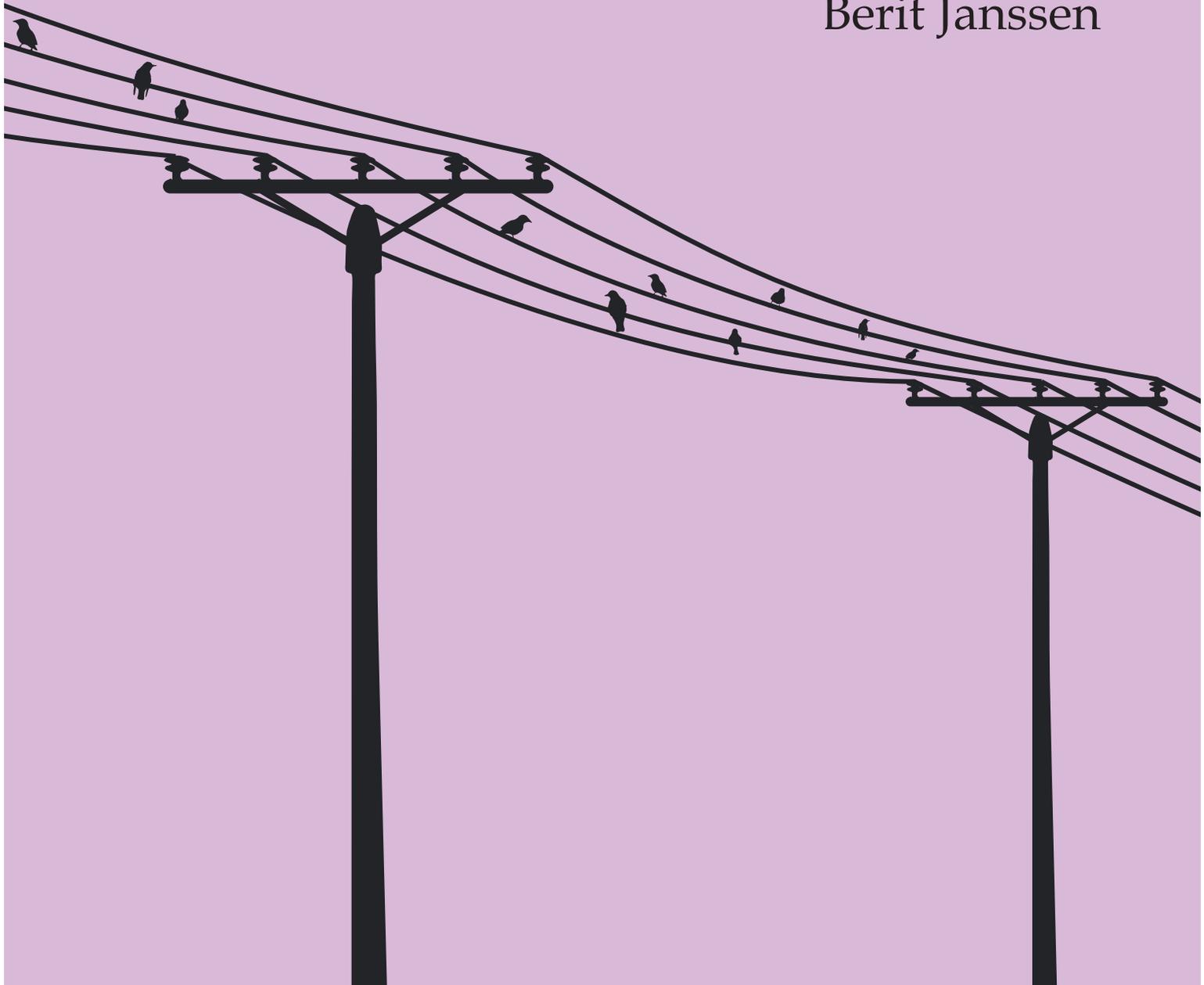


RETAINED OR LOST IN TRANSMISSION?

Analyzing and Predicting
Stability in Dutch Folk Songs

Berit Janssen



Deze uitgave is mede mogelijk
gemaakt door de
J.E. Jurriaanse Stichting.

RETAINED OR LOST IN TRANSMISSION?
Analyzing and Predicting Stability in Dutch Folk Songs

BERIT JANSSEN

ILLC Dissertation Series DS-2018-01



INSTITUTE FOR LOGIC, LANGUAGE AND COMPUTATION

For further information about ILLC-publications, please contact

Institute for Logic, Language and Computation
Universiteit van Amsterdam
Science Park 107
1098 XG Amsterdam
phone: +31-20-525 6051
e-mail: illc@uva.nl
homepage: <http://www.illc.uva.nl/>

Copyright © 2018 by Berit Janssen. Published under the Creative Commons
Attributions Licence, CC BY 4.0 (<https://creativecommons.org/licenses/by/4.0/>)



This document was typeset using the LaTeX template `classicthesis` developed by
André Miede (<http://code.google.com/p/classicthesis/>).

Cover design by Tessa Veldhorst (<http://www.deschaapjesfabriek.nl/>).

Printed and bound by OffPage.

ISBN: 978-94-6182-861-3

RETAINED OR LOST IN TRANSMISSION?
Analyzing and Predicting Stability in Dutch Folk Songs

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Universiteit van Amsterdam
op gezag van de Rector Magnificus
prof. dr. ir. K.I.J. Maex
ten overstaan van een door het College voor Promoties ingestelde
commissie, in het openbaar te verdedigen in de Aula der Universiteit
op 9 februari 2018, 13 uur

door

Berit Dorle Janssen

geboren te Brockel, Duitsland

Promotiecommissie

Promotor: Prof. dr. H.J. Honing Universiteit van Amsterdam

Copromotor: dr. ir. P. van Kranenburg Meertens Instituut

Overige leden: Prof. dr. J.J.E. Kursell Universiteit van Amsterdam

Prof. dr. L.W.M. Bod Universiteit van Amsterdam

Prof. dr. T. Meder Rijksuniversiteit Groningen

Dr. E. Gómez Universitat Pompeu Fabra

Dr. J.A. Burgoyne Universiteit van Amsterdam

Faculteit der Geesteswetenschappen

The research described in this dissertation was performed at the Meertens Institute, Amsterdam, and funded by the Computational Humanities programme of the Royal Netherlands Academy of Arts and Sciences (KNAW), under the auspices of the Tunes & Tales project.

I've been doing it for years, my goal is moving near;
It says "Look I'm over here", then it up and disappear.

Some say that knowledge is something sat in your lap,
Some say that knowledge is something you never have.

— Kate Bush, "Sat in Your Lap"

Dedicated to everyone who accompanied me on this longest journey.

CONTRIBUTIONS

CHAPTER 1, INTRODUCTION

Berit Janssen (BJ) wrote the Introduction and created the figures, with contributions by Henkjan Honing (HH) and Peter van Kranenburg (PvK). Theo Meder gave feedback on Section 1.1.1.

CHAPTER 2, ANALYZING STABILITY

BJ wrote this chapter and created the figures, with contributions by HH and PvK.

CHAPTER 3, MUSICAL PATTERN DISCOVERY

This literature overview of pattern discovery is based on two publications, (Janssen, de Haas, Volk, & van Kranenburg, 2013) and (Janssen, de Haas, Volk, & van Kranenburg, 2014). BJ performed the literature study and wrote the manuscripts, with contributions by Bas de Haas, PvK and Anja Volk (AV). BJ revised these publications by including the most recent literature, and modified Table 4.1, based on suggestions by Sally Wyatt and Andreas van Cranenburgh.

CHAPTER 4, FINDING OCCURRENCES OF MELODIC PHRASES IN FOLK SONGS

The first stage of this research has been presented at the 2015 International Conference of the Society for Music Information Retrieval (Janssen, van Kranenburg, & Volk, 2015). BJ designed the method, performed the analyses, created the figures and wrote the manuscript of this publication. PvK advised on the method and edited the manuscript. AV advised on the literature background and edited the manuscript. For a subsequent publication in the *Journal of New Music Research* (Janssen, van Kranenburg, & Volk, 2017), BJ modified the original method and evaluation, based on feedback by PvK and an anonymous reviewer. BJ created the figures and revised the manuscript, with contributions by PvK and AV. The current chapter is based on this later publication.

CHAPTER 5, PREDICTING STABILITY IN FOLK SONG TRANSMISSION

The analysis of stability has appeared in *Frontiers of Psychology* (Janssen, Burgoyne, & Honing, 2017). BJ performed the analyses of musical and statistical data, created the figures and wrote the manuscript. John Ashley Burgoyne advised on the statistical analysis, helped with Figure 5.5, and edited the manuscript. HH advised on the

analysis of musical data and edited the manuscript. BJ extended the chapter for the dissertation, based on feedback by PvK and HH.

CHAPTER 6, CONCLUSIONS AND FUTURE WORK

BJ wrote the Conclusion, with contributions by HH and PvK. Folgert Karsdorp gave feedback on an early draft on cultural transmission.

APPENDIX A, THE ROLE OF ABSOLUTE PITCH MEMORY IN THE ORAL TRANSMISSION OF FOLKSONGS

The appendix appeared in *Empirical Musicology Review* (Olthof, Janssen, & Honing, 2015). Based on a research idea by HH, Merwin Olthof performed the analyses and wrote the manuscript. BJ advised on the pitch analysis and statistical evaluation, created Figure A.3 and edited the manuscript. PvK advised on the automatic pitch analysis. HH advised on the statistical analysis, created Figures A.1 and A.2, and edited the manuscript.

CONTENTS

ACKNOWLEDGEMENTS	1
1 INTRODUCTION	3
1.1 Terminology	3
1.1.1 Musicological terminology	4
1.1.2 Computational terminology	5
1.2 Theories and studies on music transmission	5
1.2.1 Artificial transmission chains	7
1.2.2 Comparing variants in folk song collections	8
1.3 The studied material	9
1.4 Music representation	12
1.5 Outline of the dissertation	13
i QUANTIFYING STABILITY AND VARIATION	17
2 ANALYZING MUSICAL VARIATION	19
2.1 Musical aspects	19
2.2 Studies on musical variation with various musical aspects	21
2.2.1 Musical aspects in the comparison of musical traditions	21
2.2.2 Musical aspects for organizing European folk song collections	24
2.2.3 Musical aspects in diachronous studies	26
2.3 Research on quantifying variation of note sequences in folk songs	28
2.3.1 Units of transmission	28
2.3.2 Stability of note sequences	30
2.4 Conclusion	31
3 MUSICAL PATTERN DISCOVERY	33
3.1 Goals of musical pattern discovery	34
3.2 Pattern discovery methods	36
3.2.1 String-based, time-series and geometric methods	36
3.2.2 Exact or approximate matching	38
3.2.3 Recent developments and new challenges	39
3.3 Music representation	40
3.3.1 Recent developments and new challenges	41
3.4 Filtering	41
3.4.1 Filtering based on length	42
3.4.2 Filtering based on frequency	42
3.4.3 Filtering based on spacing	43
3.4.4 Filtering based on similarity	43
3.4.5 Recent developments and new challenges	44
3.5 Evaluation	45
3.5.1 Qualitative evaluation	45

3.5.2	Evaluation on speed	45
3.5.3	Evaluation on segmentation	45
3.5.4	Evaluation on classification	46
3.5.5	Evaluation on compression	46
3.5.6	Evaluation on annotated patterns	46
3.5.7	Recent developments and new challenges	47
3.6	Conclusion	48
4	FINDING OCCURRENCES OF MELODIC SEGMENTS IN FOLK SONGS	51
4.1	Material	53
4.2	Compared Similarity Measures	54
4.2.1	Similarity Measures Comparing Equal-Length Note Sequences	56
4.2.2	Similarity Measures Comparing Variable-Length Note Sequences	57
4.2.3	Similarity Measures Comparing Abstract Representations	59
4.3	Evaluation	60
4.3.1	Glass ceiling	62
4.3.2	Baselines	62
4.4	Comparison of similarity measures	63
4.4.1	Results	63
4.4.2	Discussion	65
4.5	Dealing with transposition and time dilation differences	66
4.5.1	Music representations	66
4.5.2	Results	67
4.5.3	Discussion	70
4.6	Combination of the best-performing measures	71
4.6.1	Method	71
4.6.2	Results	71
4.6.3	Discussion	72
4.7	Optimization and performance of similarity measures for data subsets	72
4.7.1	Method	73
4.7.2	Similarity thresholds	73
4.7.3	Agreement with ground truth	75
4.7.4	Discussion	76
4.8	Conclusion	76
ii	PREDICTING STABILITY	79
5	PREDICTING STABILITY IN FOLK SONG TRANSMISSION	81
5.1	Hypothesized predictors for stability	81
5.1.1	Phrase length	82
5.1.2	Phrase repetition	82
5.1.3	Phrase position	82
5.1.4	Melodic expectancy	83
5.1.5	Repeating motifs	85
5.2	Material	87

5.3	Formalizing hypotheses	87
5.3.1	Influence of phrase length	88
5.3.2	Influence of rehearsal	89
5.3.3	Influence of the primacy effect	89
5.3.4	Influence of expectancy	90
5.3.5	The influence of repeating motifs	94
5.4	Research method	97
5.4.1	Logistic regression	97
5.4.2	Generalized Linear Mixed Model	100
5.4.3	Model selection	101
5.5	Results	101
5.6	Discussion	103
6	CONCLUSIONS AND FUTURE WORK	107
6.1	Transmission	107
6.2	Quantifying stability and variation	109
6.3	Predicting stability	112
iii	APPENDIX	115
A	THE ROLE OF ABSOLUTE PITCH MEMORY IN THE ORAL TRANSMISSION OF FOLKSONGS	117
A.1	Background	118
A.1.1	Traditional Absolute Pitch Versus Absolute Pitch Memory	118
A.1.2	Related Work on Absolute Pitch Memory	118
A.1.3	Song Memory	119
A.1.4	Material	120
A.2	Dataset A: Between tune family analysis	121
A.2.1	Method	121
A.2.2	Quantitative analysis with circular statistics	122
A.2.3	Baseline	123
A.2.4	Results	123
A.3	Dataset B: Between and within tune family analysis	124
A.3.1	Method	124
A.3.2	Baseline	125
A.3.3	Results	125
A.4	Discussion	126
A.4.1	A Role for Absolute Pitch Memory in Oral Transmission of Folk Songs	126
A.4.2	Gender, Lyrics and Geographical Origins	129
A.5	Conclusions	130
B	SIMILARITY MEASURES AND MUSIC REPRESENTATIONS	133
	BIBLIOGRAPHY	135
	GLOSSARY	149

SAMENVATTING	157
SUMMARY	159
ZUSAMMENFASSUNG	161
BIOGRAPHY	163
TITLES IN THE ILLC DISSERTATION SERIES	165

ACKNOWLEDGEMENTS

I needed some time to consider when I got the offer to join the Tunes & Tales project in 2011. What a great chance and honour – but there was so much I did not know, so many new people to meet, and everything might just collapse if I didn't get along with my supervisors... I count myself lucky that I got the chance to learn so much; from colleagues who were generous with their time and wisdom; and from my supervisors who were my role models in terms of passion, attention to detail and good time management. Now it's finally time to drop the curtain on the project, and say my thanks.

Henkjan, dankjewel dat je op het juiste moment zei, “je mag koppig zijn!”, en mij vooral er in bekrachtigde om mijn eigen weg te vinden – zolang ik die maar kon verdedigen. Peter, dankjewel voor jouw grondige zorgvuldige manier van werken, jij hebt waarschijnlijk tijdens het project weken aan tijd besteed aan het lezen van mijn drafts, en mij erg geholpen om deze verdedigbaarheid van mijn methodes te bereiken. Louis, dankjewel dat je de stenen voor dit onderzoek aan het rollen bracht – jammer genoeg kan je het me niet meer zien afronden.

Ashley, thank you for thinking along on statistical methods, which helped me greatly with rounding off my research on predicting stability. Folgert, dank voor je tips tijdens mijn eerste stappen met Python, en voor jouw enthousiasme voor de Tunes kant van ons gezamenlijke verhaal.

Dear members of my reading committee – Julia, Rens, Theo, Emilia and Ashley – thank you for being willing to help me along in this final stage of my research by reading my manuscript, and for participating in my defence ceremony. I cannot wait to hear your questions. Fleur and Yvonne, mijn paranimfen, dank jullie wel voor de hulp bij de voorbereidingen en bij de verdediging.

Lieve Music Cognition Group medestrijders – Carlos, Makiko, Joey, Bastiaan, Paula, Aline, Ben – dank voor de vele discussies over muziekonderzoek, maar ook voor gezellige babbels over het leven, binnen en buiten de academische wereld. Lieve Meertens collega's, dank voor vele gezellige lunches in de Cola fabriek, en voor dat er altijd een bureau voor mij bleef staan, ook na de verhuizing in het nieuwe mooie gebouw in de Amsterdamse binnenstad. Lieve collega's van de eHumanities groep, dank voor het gezamenlijke DH'en, het was altijd inspirerend om elkaars onderzoek te volgen. Lieve collega's bij het Digital Humanities Lab in Utrecht: dank voor de gezellige sfeer op een geweldige nieuwe werkplek waar ik weer heel veel mag leren.

Veel dank vooral aan Martine, Ellen, Sanneke en Jorn, die veel werk hebben verzet om melodieën te annoteren, beschrijven en archiveren: zonder dit werk had mijn onderzoek niet kunnen plaatsvinden. Ook heel veel dank aan Merwin, die ik mocht begeleiden tijdens zijn onderzoeksstage aan het Meertens instituut: de uitkomst van de stage was een erg mooi paper dat in de appendix van dit proefschrift is opgenomen.

Liebe Eltern, liebe Isving, David, Enno und Tamme, liebe Ulfert und Tabea, liebe Sooke, Maren, Bjarne und Jacob, wie schön, in diesen Jahren die Familie wachsen zu sehen, und gemeinsam gemütliche Stunden in Brockel und Den Haag zu verbringen, die mir so viel Kraft und Gemütsruhe gegeben haben. Lieve Lowie en Liesbeth, liebe Janneke, Nick, Tygo en Loki, fijn ook om zo een geweldige schoonfamilie te hebben die met veel belangstelling en humor mijn leven zoveel rijker maakt.

Dank aan mijn geweldige vrienden in Nederland: Tessa (van wie het overmooie kaftontwerp is), Rutger, Mirjam en Tom, Evi en Joris, Lina en Floris, Karlijn en Martijn, Victor, Pieter, Sonja, Linda en Roeland, Anouk en Jelmer om namen te noemen, maar ook bedankt aan mijn medemuzikanten bij de Woodstreet big band en mijn mederoeiers bij RV de Laak. Door jullie gezelligheid heb ik enorm veel energie opgestoken. Liebe Chrissy, danke für Deine lange treue Freundschaft, und vielerlei Austausch. Ich habe durch Dich viel Tips bekommen, wie ich mehr schaffe und fokussierter arbeiten kann, gerade im letzten Stadium des Schreibens.

Last but not least, thank you Tijn. For coaching me through some difficult stages (because they happened, too), for teaching me to embrace the person that I am, and for being such a wonderful dad to our wonderful daughter. You and Leslie shine a light into every corner of my being.

INTRODUCTION

Music is often described as if it were alive: even when it cannot be physically heard, it seems to hang out in our minds. Sometimes it is the persistent background soundtrack caused by involuntary musical imagery, sometimes we will it back into being by singing or playing it. When music is performed, it changes. Sometimes, this change is subtle – some ornamentation here, failing to hit a note there – sometimes, change can also be extreme – such as a new interpretation of a pop ballad in punk style.

The current dissertation investigates this change introduced by remembering and performing music. I analyze a collection of folk songs to establish which parts of a melody change relatively little, or remain stable, and which parts of a melody show more change. My goal is to find underlying cognitive mechanisms which might account for the relative stability of some musical ideas, or the relative volatility of others. This leads me to formulate and test a number of hypotheses to predict which parts of a melody may remain stable.

My research is related to ethnomusicology, where various studies have addressed the phenomenon of stability. I quantify stability and variation in music transmission by drawing on the current possibilities of computational musicology. Computational musicology has been an active research field since the 1960s, seeking to approach musicological questions with computational methods, methods which are also a focus of interest for the Music Information Retrieval (MIR) community (c.f. Volk, Wiering, & van Kranenburg, 2011, for an overview of computational musicology and related research fields). Computational musicology fits into the broader context of the more recent research programme Digital Humanities, which approaches arts and social sciences with computational methods (Burdick, 2012). My hypotheses to predict stability are based on insights from cognition, and particularly, music cognition. Moreover, I employ statistical methods to test these hypotheses.

The following section introduces terminology from music theory, computational and statistical methods – terminology that I use repeatedly in this dissertation. The third section reviews ethnomusicological studies on transmission of folk song melodies; the fourth section introduces the folk song collection on which I tested my hypotheses on stability and variability; the final section of this chapter gives an overview of the structure of the dissertation.

1.1 TERMINOLOGY

This section introduces key terms from musicology, computing and statistics central to my research, which readers who are familiar with these domains are invited to skip. For reference in later chapters, I endeavoured to assemble all terminology in the glossary in the back of the book.

1.1.1 Musicological terminology

My research revolves around *monophonic* folk song melodies, i.e., melodies which are performed by one singer who is not accompanied by other musical instruments. The melodies consist of notes, which have a given *pitch*, or perceived height. In the time domain, notes are commonly described by their *onset*, or start, and their *duration*. In computational analysis, it is customary to describe timing of notes by their *inter-onset interval*, which is the distance between the onsets of consecutive notes. Between two consecutive onsets, also silence, or a *rest*, may occur. In the pitch domain, often the distance between adjacent pitches, or their *pitch interval*, is considered.

The studied folk songs can be subdivided into *phrases*: note sequences which are perceived as units, and are often demarcated by a rest, or a prolonged note, also known as *fermata*. Phrases may repeat within a melody, and their succession is known as the *form* of the melody. Melodies may also consist of *motifs*, very short collections of notes which are repeated over the course of a melody. In some computational studies, the term *motif* is also used more widely to refer to any collection of notes of a given length, which may or may not repeat. The end notes of a melody or phrase are referred to as a *cadence*, which often has a conclusive character.

Many times in this dissertation, melodies will be represented by their notation, in which duration of notes are indicated by the horizontal spacing and note type, and their pitch is indicated through vertical position. Figure 1.1 illustrates the relationship between pitch, duration, pitch interval and inter-onset interval, which are the most widely used concepts in the present dissertation.

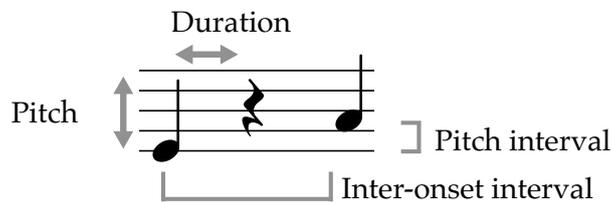


Figure 1.1: The relationship between pitch and duration of a note, and the pitch interval and inter-onset interval of two consecutive notes.

Folk song melodies can also be described in terms of their *scale*, or set of distinct pitches of a musical piece. In the time domain, melodies can be described by their *meter*, a description of how accented and unaccented notes follow each other in a melody. One cycle of unaccented and accented notes is delineated by *bars* in Western music notation. If the first note of a melody or phrase occurs before the first accent defined by the meter, it is known as an *anacrusis*.

Music is performed by singers or instrumentalists. I refer to humans performing music as singers or *musicians*, independent of whether they sing or play an instrument, and also independent of their expertise. This is in contrast to the use of terminology in some music studies, where only performance experts are called musicians. For mu-

sic notation in figures, often representing fragments from the studied folk songs as performed by singers, audio examples are available.¹

1.1.2 Computational terminology

To investigate how information on a melody can be quantified, I refer to pitches, durations, scales, meter, and other properties as a melody's *musical aspects*. Furthermore, I distinguish *local musical aspects*, such as single notes or chords, from *global musical aspects*, such as the underlying scale or meter of a piece.

The information relating to musical aspects can be either *binary* (i.e., an aspect is present or absent), *categorical* (i.e., an aspect can take various values), or *continuous* (i.e., it can take a range of values). An example for a binary musical aspect would be the occurrence of a phrase in a melody; an example for a categorical musical aspect would be different scales by which different melodies can be described; an example for a continuous musical aspect would be the pitch of a note. Musical pieces are often compared in terms of their musical aspects. This is commonly done through computing a *similarity measure*, which compares either how many musical aspects the pieces share (i.e., if they share many identical aspects, they are more similar than if they do not), or how similar the aspects themselves are (e.g., if two melodies consist of pitches which are close to each other, they might be considered more similar than if their pitches are far apart).

As sequences of musical aspects are compared, it may be beneficial to not only compare element by element (*fixed-length comparison*), but also allow that sequences may differ in length (*variable-length comparison*). For instance, one melody may be slightly longer than another and therefore contain more pitches. One well-known technique for variable-length comparison which has been used in various literature, as well as this dissertation, is *alignment*. It compares all items in two sequences, and finds the optimal correspondences between items, which may result in gaps in one sequence in relation to the other sequence.

For hypothesis testing, I make use of *regression*, which infers whether the development of a given *independent variable*, i.e. the property which is used to predict an observation, is statistically linked to the *dependent variable*, i.e. the observation being predicted. I also refer to the independent variables as *predictors*.

1.2 THEORIES AND STUDIES ON MUSIC TRANSMISSION

My research is inspired by folk song research, which has a long-standing tradition of studying music transmission between humans: Tappert (1890/1965) observed how melodies travel over borders, and from concert hall to public house, being subject to variation in the process. Bayard (1950) suggested that folk song variants originate from

¹ The digital version of my dissertation provides links to the audio versions, indicated by red text in the captions of figures. Alternatively, navigate to <http://beritjanssen.com/AudioExamples/> to find a list of the audio examples, sorted by the numbers of the figures.

transmission of an ancestor melody, which is why he proposed the term *tune family* for “a group of melodies showing basic interrelation by means of constant melodic correspondence, and presumably owing their mutual likeness to descent from a single air that has assumed multiple forms through processes of variation, imitation and assimilation” (Bayard, 1950, p.33).

Bayard’s tune family concept seems to suggest that if we were to identify melodies belonging to the same tune family from different time periods, we would be able to study how transmission shapes music over time, tracing change from one generation of melodies to the next. However, it is important to keep in mind that musical pieces may circulate in a musical tradition in various fashions: from one generation of musicians to another (vertical transmission), between different musicians within a generation (horizontal transmission), from one musician to another musician, from one musician to a group of learners, from a group of experts to one learner, or between groups of experts and learners. Therefore, the transmission path of a piece of music, represented by a number of variants, is not self-evident: variants may have copied from the same earlier piece of music, but may also have copied from different variants, or from each other. This means that the true relationship between a group of variants is often not known.

Still, the similarity between variants from different time periods may reveal trends of transmission. Nettl (2005) summarizes such types of music transmission: first, transmission in which no change occurs; second, transmission with a fixed tendency, such that every participant in a musical tradition changes a melody in the same way; third, transmission resulting in many variants, which can be visualized as a branching tree; fourth, transmission in which variants borrow from other, unrelated pieces, comparable to a “tree whose roots and limbs are attacked by shoots from elsewhere” (p. 298f.).

Nettl’s first type would point towards a musical tradition in which conformity is the ideal of transmission: participants in the tradition would strive to copy the most common form of a musical piece as closely as possible. The second type would reveal a bias towards a specific kind of variation: for instance, the most prestigious musicians might be copied, or the variants perceived as the most aesthetically pleasing by all participants in a musical tradition. Both of these types can be expected to lead to less variation over time, leading to only one version of a piece of music, according to models of cultural transmission (Henrich & Boyd, 2002).

The third type would represent unbiased transmission, in which all variants are equally accepted, and will be varied in turn. This might lead to quite severe changes over time, to such an extent that it is unclear whether two pieces are related. This would be even more true of Nettl’s fourth type, which entails influence across tune families, such that parts of one piece of music become adopted by the variant of another piece, leading, as it were, to musical chimeras.

Cowdery (1990) pointed out that for Irish music, which may be imagined as Nettl’s fourth type of transmission, the tune family concept in Bayard’s sense is not necessarily useful. He therefore contends that for some musical genres the tune family concept needs to be broadened to “a tune model as a field of possibilities within a basic con-

tour" (Cowdery, 1990, p.73), as melodic material is recombined in ways that would be difficult to capture in mutually exclusive categories.

Nettl's types illustrate that it is not easy to study how exactly songs and instrumental music spread within a musical tradition, considering that we know so little about the actual transmission paths of the songs. Below, I will show two approaches to studying how music transmission affects melodies: one approach controls the uncertainty of music transmission through constructing artificial transmission chains; another compares variants of existing folk song collections.

1.2.1 *Artificial transmission chains*

Artificial transmission chains were first introduced in folklore research: Bartlett (1920) simulated the transmission of folk tales by asking participants to read and then retell stories, based on the version by the previous participant. For the transmission of musical structure, the transmission chain paradigm was used in a recall study by Klusen, Moog, and Piel (1978). With the explicit goal of modelling the oral transmission of melodies, the researchers instructed participants to record their recall of a folk song. The participants were recruited from various social groups, such as students with or without musical training, civil servants, craftsmen and untrained workers, and balanced in gender. Klusen and colleagues compared a sequential design, where the chain was constructed along many singers, to a parallel design, where the variation between participants was considered after they had learned a folk song from the same source. In both the sequential and the parallel design, the participants heard the source melody three times before recording their own recall.

In the parallel design, participants heard four different versions of a folk song over the course of four weeks: each week, they would be presented with one variant and had to recall it. This setup was meant to study whether closely related melodies would become mixed in subsequent recall. The order in which the melodies were presented was varied: the group of 40 participants was divided into four groups of ten participants. Each group started the experiment with a different variant in the first week. In the sequential design, 40 participants were divided into four groups as well, forming a transmission chain of ten participants each. Each chain was initiated with the same melody, and the change from participant to participant was observed.

The results indicate a tendency to change some tones more than others, a phenomenon which Klusen and colleagues distinguished as "weak" and "strong" notes. Moreover, familiar melodies resembling the target melodies were reportedly contaminating the recall of the target. Figure 1.2 shows an example of a melody, redrawn after a figure by the authors, in which the number under the notes show how often the various notes in a melody have been changed by the participants, which is also reflected in the size of the note head: strong notes (i.e., few changes) have a bigger note head than weak notes. Overall, the perceived changes were greatest in the *melos*, i.e., variations in pitch, followed by rhythmic variations; variations in the song lyrics occurred only rarely. As for sociodemographic factors, those groups with musical training showed



Figure 1.2: The **result of the recall experiment** by Klusen et al. (1978). The size of the note shows its strength: stronger notes are bigger, weaker notes smaller. The numbers under the notes indicate how many of the 40 participants changed that particular note.

higher recall accuracy than those without, and those groups with higher education, e.g., civil servants, showed higher recall accuracy than, e.g., untrained workers. Gender did not influence recall accuracy.

1.2.2 Comparing variants in folk song collections

Bronson (1950) studied 100 folk song variants from a tune family of which some versions are known as *Edward*. To this end, he selected variants from British and Anglo-American folk song collections from the 16th to the 20th century. He determined which notes in the variants corresponded, and then identified stable notes, i.e. notes which most variants shared with each other: these stable notes were found in the cadence ending the first phrase, the first stressed note of the first and the second phrase, and the penultimate stressed notes of the first and the second phrase. He also found that the majority of songs exhibited a minor tonality, and noted a tendency to extend the duration of notes at phrase endings, which resulted in considerable variation of the notated meter around phrase endings.

Louhivuori (1990) applied a similar method as Bronson in his analysis of spiritual folk songs from Finnish Besecherism. He digitally encoded the collection of 1700 melodies, with 199 identified tune families, with an alphabet representing the notes, and hand-aligned the melodic variants of 25 tune families with each other, before computationally comparing bar by bar to find variations between them. He identifies sensitive areas for change in these tune families, and shows that variations are more frequent in the second bar of each phrase, and least frequent in the anacrusis and the last bar of a phrase.

Olthof et al. (2015) have used comparative analysis to show that the pitch chroma at which singers sing a melody may also be stable. Pitch chroma refers to the categories of distinct pitches, such as C, D, or E in Western music notation. Pitch chromas which are spaced an octave apart are considered highly similar by the human auditory system, such that men and women can sing together at different voice ranges but feel they hold the same melody. For more details on this research, refer to [Appendix A](#). To summarize the results, we observed that the tonic pitch chroma in two of the five analyzed tune families was centered around a mean pitch chroma, indicating a strong preference to recite the song centered on this pitch chroma. We also show that folk songs

exhibit increased tonic pitch chroma uniformity within specific geographic regions, or depending on the lyrics with which a melody is sung.

1.3 THE STUDIED MATERIAL

The Meertens Tune Collections provide a rich resource to study a well-documented musical tradition. The Meertens Tune Collections contain a corpus of instrumental music from the 18th century (INS), a corpus of 4125 folk songs from oral transmission (FS), and a subset of 360 folk songs from FS with annotations (ANN). In my research, I focus on the FS and ANN collections, which are relatively homogenous sets of folk songs representing a time interval of a few decades. The folk songs are monophonic, i.e., they only feature one melody line sung, in most cases, by one singer.

The FS corpus is the result of an extensive effort to collect Dutch folk songs in the 1950s to 1980s. This effort was started by Will Scheepers, and later continued by the researcher and radio presenter Ate Doornbosch. In his radio show *Onder de Groene Linde* he broadcast recordings from correspondents who sang songs they remembered from their childhood, and encouraged his listeners to contact him if they remembered different versions of such a song, or songs they had not yet heard in the programme. He would then visit the correspondents at home to record another item for the growing collection (Grijp, 2008).

As the largest part of the FS collection is the result of the recruitment of correspondents through the *Onder de Groene Linde* radio show, it is important to remark the bias towards a specific part of the Dutch population introduced through the traditionally strong link between geographic and social environment and the consumed media: the listeners of the broadcasting corporation VARA were predominantly left-wing. This means that some parts of the Netherlands, such as the Central and Eastern provinces in the South of the country, inhabitants of which traditionally listened to channels from other broadcasting corporations, are under-represented in the collection. Moreover, most of the correspondents had spent their working lives as farming and factory workers, and were at retirement age by the time of recording (Grijp, 2008). Furthermore, Doornbosch was mostly interested in ballads, which are songs with narrative lyrics. He tried to reconstruct how ballads spread throughout the Low Countries, e.g., along trade routes, or from migrant workers from Frisia or Germany (Grijp & Roodenburg, 2005, p. 47). This means that he scarcely recorded children's songs, songs related to seasons or holidays, or church songs. These biases mean that in terms of geography, age, social class and repertoire, the FS collection should not be assumed to represent the full range of Dutch folk song culture of the 20th century. Still, the collection provides an invaluable resource for my study of music transmission.

Of the recordings, more than 7000 in total, about half were transcribed in later years, both by Doornbosch and his co-workers, and by documentalists hired in projects in the 1990s and 2000s to make the songs available in a database, the Meertens Institute's *Liederenbank*, or Dutch song database.² The *Liederenbank* was established by Louis

² www.liederenbank.nl



Figure 1.3: Ate Doornbosch (right) recording a folk song in the home of an informant as part of the *Onder de Groene Linde* collection.

Grijp, an adept lutenist and researcher whose interest in the origins of Dutch song culture made him the spearhead of research on Dutch songs and instrumental music until his demise in 2016.

The Dutch song database started as a local collection of metadata about songs, but today is an internet resource in which information on more than 170,000 pieces from Dutch song and instrumental culture can be found. If available, music notation, either as scans from prints and manuscripts, or in a machine-readable format, and recordings for the songs from the above-mentioned fieldwork are also downloadable.

The Dutch folk song database is exceptionally well-documented. Most of the folk songs have been categorized into tune families by domain experts, or through extensive computational analysis (van Kranenburg, Volk, & Wiering, 2013). Machine-readable music notation has been produced by hand, and melodies have been subdivided into melodic *phrases*. Boundaries between such phrases are usually demarcated by rests in the music at which singers may breathe, and often by end rhyme of the lyrics' lines. Many books of songs and instrumental music have been acquired over the years, and its contents digitized, and linked to the songs from fieldwork. For many songs, infor-

mation on their musical origin is available, as melodies often originate from operas or well-known instrumental compositions.

The MTC-FS collection is a subcategory of the Dutch folk song database and contains exclusively songs whose music notation is available in machine-readable format, i.e., Humdrum ***kern* and MIDI. 2503 songs are transcribed fieldwork recordings, 1617 songs originate from song books known to contain variants of the fieldwork recordings. All songs have been subdivided into melodic phrases, which are mostly between six and twelve notes long, with an average length of nine notes.

Songs in the MTC-FS collection have been assigned an identifier, which is the string “NLB” for *Nederlandse Liederbank*, followed by six numbers, which indicate the record number, and followed by an underscore and another two numbers, which indicate the verse. Songs whose identifiers start with “NLBo7” and “NLBo8” are all based on transcriptions, songs whose identifiers start with “NLB1” are all based on song books. In cases where the verses in a folk song recording were musically very different, several verses may have been transcribed, and are indicated with ascending numbers, starting from “01” for the first verse, which is for most songs also the only transcribed verse.

The smaller MTC-ANN collection contains 360 songs from the MTC-FS collection, but some files have been renamed for the re-release of the MTC-ANN collection (version 2.0), used for this research, with the consequence that the datasets are not fully compatible (see van Kranenburg, Janssen, & Volk, 2016, for a full documentation). The MTC-ANN corpus was originally assembled for testing similarity relationships between folk songs, with the goal of facilitating categorization with computational analysis methods. To this end, domain specialists added annotations to the MTC-ANN corpus (version 1.0) in 2008, such as their form and pairwise similarity relationships between songs.

The three domain experts who provided information on music similarity stated that their judgement on the categorization of melodies into tune families was guided by the presence of characteristic motifs in the melodies (Volk & van Kranenburg, 2012). As a result, the experts were also asked to annotate such motifs signalling tune family membership, leading to 1229 annotated motifs of 94 motif classes in the 360 melodies. Some corrections and additions led to the current set of 1657 annotated characteristic motifs. These motifs vary considerably in length, and motifs belonging to the same motif class may be highly similar, but also not very similar at all. This is why these motif annotations are not used in the current dissertation.

I used the ANN corpus in this dissertation as a training set for the computational method to study evolution of musical structure, and for this purpose, Meertens documentalists annotated the similarity of phrases within the 26 tune families contained in the corpus, as described in [Chapter 4](#), and released with ANN2.0.³

³ www.liederenbank.nl/mtc

1.4 MUSIC REPRESENTATION

The Meertens Tune Collections provide digitized notation for all of the sub-collections, as well as recordings for the folk songs originating from fieldwork. My research focusses exclusively on the notations of songs from the FS and ANN corpus. Notations are a reduction of the original performance to fewer musical aspects, namely the pitch and duration of notes as well as an interpretation of how notes are embedded in scale and meter; this has the advantage that the study of these musical aspects is facilitated, with the disadvantage that research questions concerning other musical aspects, as for instance *timbre*, or tone quality, cannot be answered.

I choose to work with notation as this circumvents technical problems in the audio recordings such as tempo and pitch fluctuations, band noise (i.e., the hum of some of the earlier recordings introduced from the band recorder), and artefacts such as spoken comments between verses of a song, which would complicate automatic analysis of the folk songs. The efforts of the Music Information Retrieval community of the past few years to combine approaches from the *symbolic* domain (i.e., based on music notation) and the *audio* domain (i.e., based on recordings) are an inspiring incentive to investigate the research questions posed in this dissertation based on audio recordings. Olthof et al. (2015) investigated a selection of 100 audio recordings supported by automatic analysis, but due to the aforementioned technical problems, could not fully automate extraction of pitches. As of yet, therefore, automatic analysis on the full set of audio recordings available from the Dutch song database is left for future work.

It is important to keep in mind that the notations are based on transcriptions, which are the result of human interpretation: there are points at which different human experts might disagree on the pitch or duration of a note, or the meter of a song. A remark by Bartók, a prolific folk song collector, highlights this interpretative act of the transcriber: "our eyes and ears serve as measuring apparatuses –rather imperfect apparatuses. Thus, through our imperfect senses many subjective elements will get into our transcriptions, rendering them that much less reliable." (Bartók, 1951, p. 18) He also warns that transcribers should distinguish between accidental and intended variation of a song, and not necessarily notate accidental variations. However, he also fully accepts that the distinction between accidental and deliberate variation is not easy and certainly not unambiguous (p. 16).

With the FS and ANN collections studied in this dissertation, the transcription introduces the following limitations and challenges:

1. The corpora are limited to those folk songs which have been transcribed by the time the collections were assembled.
2. All transcriptions were transposed to the keys of G major or e minor for ease of comparison between variants. This means that absolute pitches cannot be studied from the current notations.
3. As mentioned before, the transcribers divided the songs into *melodic phrases*. Even though it may be assumed that mostly these phrase divisions would not lead to

disagreement, there are cases in which, e.g., one transcriber split a song into four phrases, while a closely related variant has been split into eight phrases.

4. Different transcribers may have chosen to notate similar melodies in different octaves or meters (see Figure 1.4.a), which complicates computational comparison of melodies.
5. Timing of performances may have been represented in different ways: e.g., in the case of long pauses at phrase endings, some transcribers may have decided to notate the lengthening through fermatas, whereas others may have chosen to notate shorter durations or rests to maintain the meter of the notation (see Figure 1.4.b)
6. Pitches may be open to interpretation: if a singer did not give a very clear performance, the transcribers may have had to double-guess which pitch was intended, while it is impossible to know the intentions of the performer based on an audio recording.
7. In cases where transcribers could not arrive at a pitch interpretation at all, they may have chosen to represent a missing pitch with crossed note heads (see Figure 1.4.c). As unpitched notes could not be digitized, the pitch of the crossed note heads was entered, introducing an artefact.

The notation originating from song books can also be considered a transcription to some extent: we do not know whether the key in which a song is notated represented the pitch the song book editor, or their informant, used for recitation of the song, or whether it was chosen for convenience, e.g. with respect to possible accompanying instruments. Of the transcription choices relating to timing (5.), it can be assumed that song book editors would choose a regular meter with fermatas above extensions of bars at phrase endings: the song book renditions of folk song melodies usually do not contain meter changes.

1.5 OUTLINE OF THE DISSERTATION

The relationship between the chapters in the dissertation is graphically described in Figure 1.5. The ensuing Chapter 2 discusses how stability and variation may be quantified, based on research on musical variation in Music Information Retrieval and ethnomusicology. In particular, pattern discovery and pattern matching are suggested as possible approaches to quantify stability.

Chapter 3 presents a literature review of pattern discovery approaches to identify repeated, salient patterns in music. The literature review shows that this course would be not feasible for the problem at hand, at least not with the current methods available. Therefore, in Chapter 4 various approaches to pattern matching, or the identification of occurrences of given melodic patterns, are compared with each other, from which a method for quantifying stability is obtained.

NLB070748_01, Phrase 1

a Zo lang de boom zal bloei - en

NLB073225_01, Phrase 1

Zo - lang de boom zal bloei - en

NLB072299_01, Phrases 1 and 2

Op - eens kwam daar een jong heer - tje aan

b En sprak: zeg meis - je wat doet gij hier te staan

NLB072886_02, Phrases 1 and 2

Op - eens kwam daar een jong heer - tje aan

En sprak: zeg meis - je wat doet gij hier te staan.

NLB070326_01, Phrase 5

c Nu moet gij ra - den wie de moo - ie ring be - ko - men zal

Figure 1.4: Illustration of three ways in which transcription influences the use of musical data in the FS corpus. **a)** Choice of meter and transposition may vary per transcriber; **b)** Lengthened phrase endings may be represented by fermatas (first example), or written out (second example); **c)** Spoken word may be represented with cross notation, but are digitized as exact pitches.

Chapter 5 introduces five hypotheses on melody *memorability* which may predict which parts of melodies are retained, and which parts are lost in the course of transmission. It makes use of the pattern matching method developed in Chapter 4 to quantify stability, and tests whether it is possible to predict stability through memorability.

Finally, [Chapter 6](#) reviews the contributions of this dissertation, and its implications for related research. Moreover, it raises research questions which cannot be answered with the current approaches, and which may be the basis of future research on transmission, variation and stability.

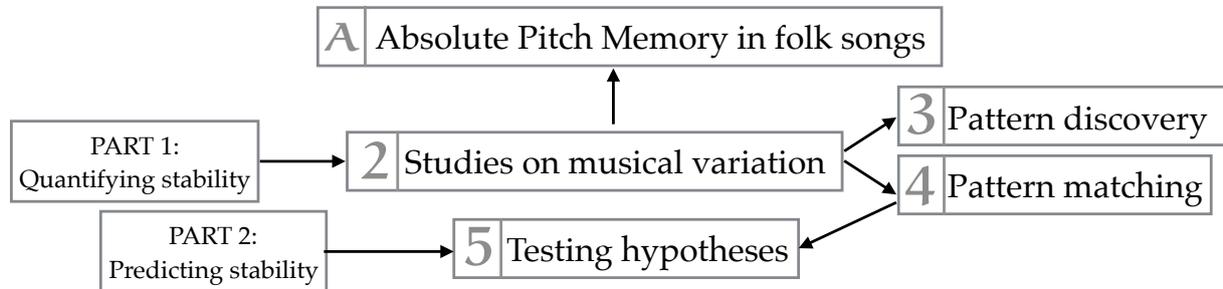


Figure 1.5: The relationships between the chapters in this dissertation, indicated by chapter number and short title, and the appendix (A).

Part I

QUANTIFYING STABILITY AND VARIATION

This part introduces approaches to quantifying stability and variation.

ANALYZING MUSICAL VARIATION

This chapter introduces possible approaches to quantifying variation and stability in folk songs. To this end, the first section gives an overview of the various musical aspects which may vary in folk songs; the second section reviews studies on musical variation, addressing the question which musical aspects would be a meaningful focus: single notes, motifs, phrases, or abstract musical aspects such as the scale or the meter of a folk song. My conclusion from this work is to focus on *note sequences*, i.e., melodic segments. The third section then discusses how the stability, or resistance to change of such a melodic segment may best be quantified. I discuss two potential approaches to quantifying stability of melodic segments in folk songs: one, a binary concept of stability, in which stable melodic segments are surrounded by unstable melodic material; the other, a graded concept of stability, in which melodic segments may be more or less stable, depending on their *frequency of occurrence*.

2.1 MUSICAL ASPECTS

The previous chapter observed that in symbolic music representation, there are fewer analyzable musical aspects than in audio representations. Yet, even for my study material of notated monophonic folk songs, there are countless musical aspects which are subject to change in transmission, and whose stability or variation may therefore be an interesting research focus.

An intriguing illustration of the many musical aspects subject to change in notations of monophonic folk song melodies can be found in Wiora's detailed inventory of variation in folk songs (Wiora, 1941). Examples (a-e) from Dutch folk songs can be found in Figure 2.1, illustrating Wiora's distinction between

- a. changes in the melodic line, or the replacement of notes affecting the width, but not the overall contour of the melody. Consider example a in Figure 2.1: of the corresponding phrases from two folk song variants, the first one has a much more condensed contour than the second.
- b. tonal changes, referring to changes in scale, the order of notes in harmonically defining structures, and changes in the underlying harmonic relationships in melodies. Observe two corresponding phrases from two folk song variants in example b, the last few notes of the first variant have a different underlying tonality (G major) than those of the second variant (D major).
- c. rhythmic changes, consisting of changed relationships between durations of notes, or a change of meter. Of the corresponding phrases from two folk song variants in example c, the second one has a punctuated, "bouncy" rhythm.

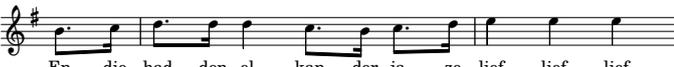
<p>NLB073404_01 - Phrase 1</p> <p>a.</p>  <p>Kom laat ons toch zo stil niet zijn,</p> <p>NLB075379_01 - Phrase 1</p>  <p>En ach moederlief, wat zal ik nu gaan maken</p>	<p>NLB073588_01 - Phrase 1</p> <p>d.</p>  <p>Een rijke heer ging eens van huis</p> <p>NLB073672_01 - Phrase 1</p>  <p>Een heer die ging zeer ver van huis</p>
<p>NLB076495_01 - Phrase 4</p> <p>b.</p>  <p>ik zag de hemel en de hel - le</p> <p>NLB074286_01 - Phrase 4</p>  <p>Ik zag den hemel en de hel - le</p>	<p>NLB075174_01</p> <p>e.</p>  <p>En er wa - ren eens twee zoe - te zoe - te lief - jes</p>  <p>En die had - den el - kan - der ja zo lief lief lief</p>  <p>En die had - den el - kan - der ja zo lief.</p> <p>NLB075040_01</p>  <p>En daar wa - ren eens twee zoe - te - lief - jes</p>  <p>En daar wa - ren eens twee zoe - te - lief - jes</p>  <p>en die had - den el - kan - der ja zo lief lief lief</p>  <p>En zij had - den el - kan - der ja zo lief.</p>
<p>NLB070134_01 - Phrase 1</p> <p>c.</p>  <p>En er waren eens twee zoete lie-fies</p> <p>NLB75018_02 - Phrase 1</p>  <p>Maar de jon-geling moest on-der de sol-da-ten.</p>	

Figure 2.1: Examples from Dutch folk songs for variation types as categorized by Wiora: **a.** changes of the melodic line, **b.** tonal changes, **c.** rhythmic changes, **d.** motivic changes, **e.** changes in form.

d. motivic changes, which entail that melodic material is either extended or contracted, differentiated or assimilated, ornamented, or less ornamented as compared to related songs. In example d, showing corresponding phrases from two folk song variants, the first one repeats the same motif twice, whereas the other uses different melodic material for the first half of the phrase.

e. changes in form, referring to changes in the order of parts, higher temporal connection or separation between parts. This is illustrated in example e: of the two shown folk song variants, the second variant repeats the first phrase, while the first one does not.

Wiora's conclusion, then, having inventorized all these alterations that singers might introduce into folk song variants, sums up to the statement that "everything in a melody is subject to change [translation: BJ]" (Wiora, 1941, p. 193). Some of the variation in Wiora's overview concerns *local musical aspects*, in that single notes, or sequences of

notes, are affected by change, i.e., changes of the melodic line, rhythmic and motivic changes. Other variation concerns *global musical aspects*, such as the scale or meter of songs. Below, I will review studies on the variation of global and local musical aspects, and based on this, motivate my choice to study sequences of local musical aspects in the form of melodic segments.

2.2 STUDIES ON MUSICAL VARIATION WITH VARIOUS MUSICAL ASPECTS

Musical variation has been studied based on a wide range of musical aspects, approaching different research goals: first, to find aspects which may help to organize folk song traditions from all over the world into styles; second, to organize collections from *one* folk song tradition such that related melodies can be found easily; and third, to analyze variation over time. Even though my research focusses on notated monophonic melodies, I also discuss research on audio recordings and music which contains *chords*, i.e., multiple pitches sounding at the same time. As these studies also make use of local and global aspects, their results underpin my choice for sequences of local aspects to research stability.

2.2.1 *Musical aspects in the comparison of musical traditions*

Studies which compare different music traditions may be seen to study the *macro-level* of cultural variation (Mesoudi, 2011), as opposed to studies at the *micro-level*, which zoom in on one specific musical tradition, or even a group of variants within such a tradition. In this category, I discuss five studies which undertake comparison of music traditions, and which mainly make use of global musical aspects to study differences between these traditions.

Lomax (1968) and his collaborators laid the foundation for large-scale comparison of music traditions, with the goal of projecting differences and similarities between music traditions onto a map of song styles. To this end, they developed a detailed style comparison system, *cantometrics*, which rates song properties such as the voice qualities of the singers, the structure of musical ensembles, the combination of different voices, and the respective focus on lyrics. In total, a catalogue of 37 rating scales was used to judge 2557 songs from 56 cultural areas. This led them to define six regions within which song styles were more similar to each other than to musical traditions from other regions. These six regions include North and South America, the Insular Pacific, Europe, Africa, and a region stretching from North Africa to Eastern Asia. The latter region is characterized by what Lomax calls the “Oriental Bardic Style” (Lomax, 1962) – a performance style in which one singer performs along to sparse accompaniment, making extensive use of vocal embellishments.

Since the work by Lomax and colleagues, comparisons of musical cultures have not been undertaken anymore until quite recently. One such example, which makes use of a very different methodology, is Juhász’ (2009) analysis of European and Asian folk songs, combining 16 different Western European and Asian folk song collections,

which each comprise 1000 to 2500 pieces. With a focus on the typical contours of folk songs from these collections, he uses *self-organizing maps* to infer contour types. To this end, neural networks are supplied with fixed-length pitch sequences from the melodies (referred to as *melody vectors*). The neural networks find similarities between the contours and group them to a map representation on a two-dimensional grid, representing the distances between the melody vectors. In a first step the contour types are learned for the 16 corpora separately, and in a second step, one self-organizing map learns all contour types for all corpora combined. Juhász concludes from the resulting overall map that there are two groups of contour types, which he calls Western (comprising Finnish, French, German, Luxembourgian, and Irish-Scottish-English folk tunes) and Eastern (comprising, among others, Hungarian, Karpatian, Anatolian, Sicilian, Karachay, Mongolian and Chinese folk tunes). He finds that the Western group shows less overlap in its contour types than the Eastern group. However, I would argue that the apparent homogeneity of the styles that Juhász observes for the Eastern contour types might be counter-balanced by great variation in other musical domains, such as vocal embellishments, which Lomax found to be characteristic for the “Oriental bardic style” in the cantometrics scheme.

Gómez, Haro, and Herrera (2009) focus on a broader selection of musical aspects. They analyze 5905 audio recordings from different regions of the world with the goal of automatically classifying them into Western or non-Western music traditions, using methods from Music Information Retrieval. Per analyzed piece, Gomez and colleagues detect the prevalent pitch classes, minima and maxima in the frequency spectrum, predominant rhythmic divisions, and the occurrence of specific drum sounds. The Western musical traditions include recordings from Europe and North America, and are subdivided into Classical, modern and traditional styles. The non-Western traditions are represented by recordings from Africa, the Arab States, Asia, Central Asia, Greenland, and the Pacific. While the musical aspects enable an accuracy of 88% in the distinction between Western and non-Western musical traditions, the classification errors for the subdivisions of the Western styles and recordings from other geographical areas reveal that tonal, rhythmic and drum descriptors lead to more classification errors than timbre descriptors. The most frequently misclassified style is Western traditional, a fact which the authors attribute to the ensemble size and recording technique of traditional music, which may be more similar to recordings from the non-Western category. This implies that while the timbre descriptors are relatively successful for classifying recordings as Western or non-Western, they may capture musical aspects which are more related to the production of an audio recording (i.e., studio vs. field recordings) than tonal or rhythmic properties, which arguably would be the aspects on which most human analysts would focus.

Like Gomez and colleagues, Panteli, Bittner, Bello, and Dixon (2017) use audio analysis techniques, investigating differences between singing styles in 2808 recordings from the Smithsonian Folkways Recordings, a record label collecting folk music from all over the world. They extract pitch contours from the recordings, and characterize the contours according to 30 descriptors which measure the pitch contours’ rate of change,

their curvature, and vibrato characteristics. After automatic classification of contours into vocal and non-vocal content, they use the machine learning method K-means to establish a pitch contour dictionary based on the 30 contour descriptors from the vocal pitch contours. Based on the prevalence of specific items from this pitch contour dictionary, Panteli and colleagues cluster the recordings. The resulting clusters correspond to groups of recordings which are from similar geographic or cultural regions, such as clusters of Eastern Mediterranean, European and Afro-Caribbean recordings. These results suggest that pitch contours may indeed be a very successful musical aspect by which to classify vocal music from different regions in the world. However, their cluster analysis is somewhat weakened by misclassified non-vocal contours from, e.g., string instruments and spoken word fragments.

While the previous studies focus on differences between musical traditions, Savage, Brown, Sakai, and Currie (2015) are interested in musical aspects which may be considered universals, as they occur in all musics in the world. To this end, Savage and colleagues rated 304 recordings from the Garland Encyclopedia of World Music according to their *CantoCore* scheme, a rating scheme based on Lomax' cantometrics. While cantometrics mixes binary and categorical musical aspects, *CantoCore* relies exclusively on binary ratings (i.e., the absence or presence of a given aspect). They grouped the recordings by continent, and tested whether any of the musical aspects considered by the *CantoCore* scheme were present in all recordings from all continents.

Out of 32 candidate features, none could be considered absolute universals, as none were present in all the investigated pieces. However, the researchers identified 18 musical aspects as "statistical universals", i.e., they were represented in all continents above chance level (Savage et al., 2015). These statistical universals state the following: pitches tend to be organized in non-equidistant scales of seven or fewer pitch classes; melodies tend to use descending or arch-shaped contours, and contain small pitch intervals; rhythms tend to be organized relative to groups of two or three isochronous beats (e.g., duple or triple meter), and form short motivic patterns with few duration values; phrases tend to be short; performance is practiced by instrumentalists as well as vocalists, predominantly in groups, and most of the performers are male. According to Trehub (2015), it is important to keep in mind that based on the rather small dataset, it may be too hasty to reject or accept musical aspects as universals, which may have just been randomly over- or underrepresented. She writes, "[t]he sampling scheme of Savage et al. was motivated by the diversity of music within cultures, but its effect was to reduce the similarities across cultures" (p. 8809).

Comparative analyses of musical traditions show that out of the vast number of musical aspects, it is difficult to select those which capture differences between musical traditions, but which may also be used to describe variation within a tradition. In some traditions, a lot of variation may be observed in terms of rhythm; in others, in tonality. Global descriptors of rhythm or tonality, such as meter or scale, may blur the fine differences found within a given musical tradition. On the other hand, local musical aspects, such as the contours used by Juhász and Sipos (2009) may be very informative for selected musical traditions, but may be less so for others. This difficulty of com-

binning comparative and detailed analyses of cultural analyses has been recognized in other domains of cultural analysis as the “micro-macro gap” (Mesoudi, 2011). Defining a vast number of musical aspects with the goal of capturing micro- as well as macro-level variation might overcome such a gap, but this would be infeasible for manual rating, as performed by Lomax (1968) and Savage et al. (2015). Automated music analysis may provide methods which can deal with large music collections as well as a large number of descriptors, but may capture differences which are not informative for humans (c.f. Gómez et al., 2009), or be deteriorated by errors from automatic classifiers (c.f. Panteli et al., 2017). Rather than finding musical aspects which may be meaningful across traditions, I conclude that the most practical approach to researching stability and variation in Dutch folk songs is to identify those musical aspects which are meaningful in Dutch and closely related musical traditions. To this end, the next section investigates research on musical variation in European folk songs.

2.2.2 *Musical aspects for organizing European folk song collections*

From the nineteenth century on, interest in folk song traditions in Western Europe has grown, resulting in large collections, in which it was hard to find specific melodies without some organizing principle. The various approaches to organizing folk song collections are informative for musical aspects which may be meaningful to analyze variation in West European folk song traditions. Krohn (1903) was the first to propose a system to organize folk songs: he suggested to order a collection of Finnish folk songs according to the number of phrases in a folk song (a global musical aspect), and according to the similarity of cadences (a local musical aspect). This suggestion was taken up by Bartók and Kodály (Bartók, 1981), who developed a catalogue of Hungarian folk song transcriptions in the early twentieth century. Kodály, following Krohn, proposed to categorize the songs according to their final notes, or cadences, whereas Bartók was in favour of describing the songs in terms of their form, number of syllables and rhythm (Járdányi, 1965). As these researchers had to perform ordering by hand, the consequences of choosing specific global or local musical aspects as criteria for organization was not easy to gauge. Computational studies on folk song categorization, in which ordering was automatized, were therefore a good way of comparing different systems, with the goal of finding criteria by which related songs could be identified successfully.

One such computational study was performed by Wolfram Steinbeck (1982), who categorized folk song melodies from a large collection of digitized German folk songs, the ESAC folk song collection (Schaffrath & Huron, 1995). Steinbeck’s categorization study relied exclusively on global musical aspects, such as the number of notes or measures of a melody, the number of distinct pitches and durations, the range of the melody, the number of changes of melodic direction, and many others. Steinbeck combined these global features to cluster melodies. The resulting clusters are hard to interpret, and therefore not many conclusions can be drawn from the study. Current state-of-the-art techniques to study similarity relationships, such as network analysis, might be

interesting to apply to Steinbeck's findings, as they may help to estimate in how far the employed global musical aspects, e.g., the number of distinct notes in a melody, are indeed meaningful musical aspects to study variation as a result of music transmission.

Barbara Jesser (1991) also analyzed melodic relationships within the ESAC folk song collection. Next to global musical aspects such as tone inventory, tonality, form, and range, she also formalized local musical aspects, in the form of rhythmic, melodic and contour types of the melodies' phrases. The contour types are automatically derived from the lowest and highest notes and turning points in the phrases, describing ascending, descending, or horizontal movement between the notes, and different combinations of these contours. She uses the global and local aspects to find relationships between melodies within two corpora: a corpus of 4178 folk songs, and a smaller corpus of 858 German language ballads.

Jesser discusses some selected cases of folk song types within the folk song corpus, but cannot meaningfully order all of the melodies in the corpus with the applied methods. For the ballad corpus, she identifies six groups, mainly by their metrical structure, which categorize about half of the songs in the corpus, but the other melodies remain unspecified. Jesser concludes that the tested musical aspects can be used successfully to find variants of a given melody, but that the identification of groups of related melodies within a folk song corpus might require different musical aspects. She sees potential in the future investigation of local musical aspects, such as motifs and their development or harmonic progressions. Yet she also raises the concern that the properties unifying groups of related melodies might be extramusical (p. 259 f.).

To compare local and global aspects of melodies systematically, van Kranenburg et al. (2013) used global aspects suggested by Steinbeck, Jesser, and McKay and Fujinaga (2004), and different substitution functions within the Needleman-Wunsch global alignment algorithm (Needleman & Wunsch, 1970), which compare local aspects such as pitch, pitch interval, duration ratio, metric weight and phrase position. They use the global and local musical aspects to categorize the ANN collection, and a larger collection of 4470 Dutch folk songs into tune families. Their classification results indicate that comparisons of local aspects are more informative than comparisons of global aspects.

Volk and van Kranenburg's (2012) surveys of folk song experts corroborate their computational results: they asked domain experts to rate the similarity of folk song variants, based on rhythm and melody, and to name other aspects which might guide the categorization of folk songs. The folk song specialists stated that their similarity judgements were not so much caused by the melodic or rhythmic similarity in general, but were highly informed by characteristic motifs – short note sequences which are highly similar between different variants of a tune family.

The various approaches to organizing folk song collections indicate that local musical aspects are likely more meaningful than global musical aspects to study variation in Western European folk songs. Experiments with global musical aspects such as ranges or scales of melodies did not lead to easily interpretable groupings of melodies (c.f. Jesser, 1991; Steinbeck, 1982). In the study by van Kranenburg et al. (2013), relationships of local aspects with the global aspects of a melody, such as the notes' metrical strength

and phrase position, did lead to clearer categorization, which means that global aspects may still be meaningful to consider in combination with local ones. The interviews by Volk and van Kranenburg (2012) suggest that melodic motifs, i.e., sequences of local aspects, may be a good way to study variation within and between tune families.

2.2.3 *Musical aspects in diachronous studies*

Recently, several publications have appeared which study variation over time in Western popular music, Western art music, and jazz, respectively. These studies are *diachronic* approaches to the development of musical styles, performing analyses on selected musical pieces from different time periods. They use isolated local musical aspects, as well as sequences of local musical aspects.

Serrà, Corral, Boguñá, Haro, and Arcos (2012) study change of Western popular music between 1955 and 2010. They make use of the million song dataset¹, a widely used resource in Music Information Retrieval, which provides music descriptions and metadata of a million pop songs. Serrà and colleagues randomly pick pieces for each year in the investigated time interval. Consequently, they analyze the sounding pitches, transposed to the same tonality, the timbre and the loudness of short segments (less than a second long) of the chosen songs, based on automatically generated music descriptors from proprietary algorithms by the Echo Nest.² They observe the distribution of so-called codewords, or categories into which they cluster the pitches and timbre of each analyzed segment. They also investigate the possible transitions between the codewords in a network analysis, i.e., for each codeword they check how often it appears before or followed by other codewords.

The results show that the same pitch codewords are favoured over time; with respect to timbre, codewords change over time, but the variety of codewords decreases; the loudness of the recordings increases over time. The authors' conclusion that this constitutes "no-evolution" of Western popular music with "no considerable changes in more than fifty years" (p. 5) is premature, however, as popular music is likely to vary in more aspects than pitches, timbre and loudness. Moreover, the analysis focuses mainly on isolated music segments of the order of one or two tones. Transition networks show how often a given collection of pitches or a given timbre is followed by other pitches or timbres, but as all pitch codewords are transposed to the same tonality, and they only report network statistics of connectedness, change in favoured chord progressions or favoured timbre successions cannot be tracked.

Another study of Western popular music by Mauch, MacCallum, Levy, and Leroi (2015) investigates a comparable time interval as Serrà and colleagues: from 1960 to 2010, the researchers randomly selected 30-second-long segments of 17094 songs, and analyzed them with respect to their pitch content and timbre. In contrast to the former study, Mauch and colleagues establish a harmony lexicon of chord bigrams, i.e., successions of two chords, which are derived from audio analysis by comparing the sounding

¹ <http://labrosa.ee.columbia.edu/millionsong/>

² <http://the.echonest.com/>

pitches to the most common chord types. They do not transpose the chords to the same tonality, so the movement of the chords as well as their quality is considered. They also establish a timbre lexicon, based on clustered features describing timbre.

To find common combinations of harmonic progressions and timbre qualities, Mauch and colleagues employ a technique called *Latent Dirichlet Allocation*, which analyzes combinations in which harmonic or timbre categories occur in song segments, and infers so-called topics. They set the algorithm to discover eight timbre and eight harmony topics. In a second step, the resulting topics are linked to semantic descriptors obtained from human listeners. In contrast to the results by Serrà and colleagues, they find great change of harmony and timbre topics over time, associated with the development of new styles. For instance, the genre of hip hop can be seen to influence popular music greatly in the early 1990s, giving prominence to the timbre topic related to “energetic, speech, bright” and a harmony topic which represents the absence of chord structure (p. 3). Moreover, the authors identify specific points in time (1964, 1983, 1991) at which topic change is rapid, which they consider revolutions related to new technologies and associated styles.

Broze and Shanahan (2013) study the development of jazz harmony. They investigate jazz chord progressions in compositions from 1924 to 1968, and find that even though the distribution of single chords does not change noticeably, chord bigrams change considerably in the observed time interval, and reflect clearly the development of jazz in the late 1950s away from functional harmony towards modal harmony.

Rodriguez Zivic, Shifres, and Cecchi (2013) investigate Western art music from 1730 to 1930, based on the Peachnote corpus, which collects *n-grams* (successions of *n* items) of chords and pitches of art music scores. They focus on pitch bigrams, from which they derive clusters that align well with the Baroque, Classical and Romantic periods: for instance, later music tends to use wider melodic intervals between consecutive pitches.

The discussed diachronic research on variation supports the merit of studying *sequences* of local musical aspects, as suggested by Volk and van Kranenburg (2012). For instance, while Serrà et al. (2012) did not find any change in popular music for isolated chords, Mauch et al. (2015) did find change in a comparable corpus for chord bigrams. Likewise, Broze and Shanahan (2013) did not observe change over time in their corpus of jazz chord progressions when considering chords in isolation, but did find it for chord bigrams. Rodriguez Zivic et al. (2013) also report change over time in their analysis of pitch bigrams.

The above studies indicate that isolated events may not provide enough insights to analyze variation: without any context, musical events may display the same statistical occurrences over the whole data set, whereas relevant changes can be observed in successions of chords or pitches. Broze and Shanahan attribute this contrast to cultural learning occurring in early childhood as opposed to exposure to specific repertoires later in life. The former engrains the variety and distribution of single events into listeners’ and musicians’ minds, forming the base of their musical perception; the latter enables them to learn successions of musical events in a given repertoire, building on the distributions of single events (Broze & Shanahan, 2013, p. 42).

For my research on stability and variation in music transmission, I therefore choose to focus on the *variation of note sequences* in the folk song melodies provided by the Meertens Tune Collections. Having established the type of musical aspect to study, the following section investigates approaches to quantify variation of note sequences in folk songs.

2.3 RESEARCH ON QUANTIFYING VARIATION OF NOTE SEQUENCES IN FOLK SONGS

I distinguish two general trends in folk song research to study variation as a result of music transmission: one, to identify *units of transmission* which remain intact in transmission (as opposed to more variable melodic material); the other, to study the *stability*, or *resistance to change*, of a given note sequence.

2.3.1 *Units of transmission*

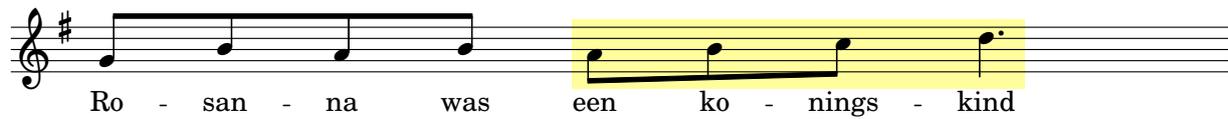
To study variation as a result of music transmission computationally, it would be a useful assumption that melodies can be reduced to discrete, immutable units, which could be isolated in a melody like proteins in a genome, and the recombination of which in related melodies would give us insights into the principles of variation within a given musical tradition. This would fit in with Dawkins' (1978) theory that as genetics, cultural artefacts may be studied as a system of discrete units or memes, a field which he dubbed *memetics*.

Nettl's (2005) discussion of *units of transmission* seems to echo such a memetic approach to folk song research: "One may think of a repertory as consisting of a vocabulary of units, perhaps melodic or rhythmic motifs, lines of music accompanying lines of poetry, cadential formulas, chords or chord sequences. We could study the process of transmission by noting how a repertory keeps these units intact, and how they are combined and recombined into larger units that are acceptable to the culture as performances. The smallest units of content may be the principal units of transmission" (p.295). Furthermore, he suggests that an oral tradition can be described both in terms of density – the degree to which units of a repertory are similar – or in terms of breadth – the "musical ground" covered by a repertory (p. 299ff.). Nettle's terminology implies that a unit of transmission can be clearly isolated within a given melody, and its relationship with other units of transmission, and its position within an oral tradition, can be quantified.

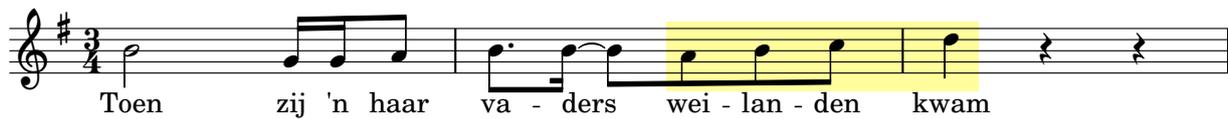
However, it is not evident how one would demarcate such smallest units of transmission in melodies from the FS collection, as illustrated by Figure 2.2, showing the first phrase of three variants from a Dutch folk song. Where would a unit of transmission start or end? The topmost example phrase is shorter than the other two example phrases – does this mean that it contains fewer units of transmission?

Another characterization that Nettle uses for units of transmission, "recurring events or signposts" (p. 297), indicates that he does not necessarily envision units of transmission as building blocks, but that a musical piece may consist of some units of trans-

NLB074452_01, Phrase 1



NLB74547_02, Phrase 1



NLB075273_01, Phrase 1



Figure 2.2: The respective **first phrase of three variants** of the Dutch folk song *Een soudaan had een dochtertje* (1), with each variant's record number in the Dutch folk song database. It would be difficult to break down the phrases into units of transmission.

mission besides other, less clearly defined musical material. Bohlman (1988) picks up this notion when he refers to units of transmission as “memory markers”, which are used to navigate through a folk song upon performance. “The density of these markers may be so great that accurate performance results in exact repetition of a song as the singer first experienced it; their musical function may be such that they encourage new phrase combinations or improvisations. [...] Taken as a whole, these memory markers become the units of transmission that make oral tradition possible” (p. 15).

But even if we were to study music as consisting of units of transmission besides other musical material, the problem remains: how would we isolate them within a melody? If they recur, does that mean that they do not change at all in the various repetitions? How should we judge whether a unit of transmission remains, in Nettl's words, “intact”? What kind of change would be acceptable for them to be still considered the same unit? As an example, refer again to Figure 2.2: the last four notes of each phrase are arguably very similar, but not completely alike. Some of these differences are introduced by transcription choices, as the phrases are notated in different meters: this might be easy to see for a human analyst, but is very difficult to disentangle computationally.

One solution might be the one proposed by Suppan (1973), who suggests describing change in terms of melody, rhythm, contour, and modes separately, such that every musical aspect of a repertory might have its own “types”, or units of transmission. This would solve some problems, as then the demand for units of transmission to recur can be very strict according to one musical aspect, but it would introduce new problems

as well: to understand mechanisms of change in oral transmission, the relationship between melodic, rhythmic, contour, and modal types would have to be defined.

In summary, units of transmission are an attractive concept in folk song research, put forward by various ethnomusicologists, but the theory is too unspecific to be falsifiable. This could be resolved by defining units of transmission according to very specific rules, yet if we failed to find such well-defined units of transmission in an oral tradition, it would be impossible to tell apart whether the assumed unit is meaningless, or if the rules defining it are wrong. Maybe because of this problem, there are no studies following up Nettl's vision of studying music transmission through units of transmission.

An alternative, data-driven approach to identifying units of transmission would be *pattern discovery*, or the inference of repeating note sequences. However, my extensive literature research, presented in [Chapter 3](#), points out some problems for this approach:

1. Most pattern discovery approaches are designed with a specific goal and genre in mind: for instance, finding themes in Classical music, or finding short motifs in jazz music. It is not possible to conclude from the success of a method for a specific goal how well it will perform for my research goal, finding repeating note sequences in folk songs.
2. While the influence of music representation and filtering has been investigated in some comparative studies, I do not feel confident to make such design choices based on the still inconclusive evidence.
3. Even though recent years have introduced standardized evaluation measures for pattern discovery, the quality of pattern discovery results is hard to assess based on these measures, especially the nature of errors that pattern discovery methods produce.

2.3.2 *Stability of note sequences*

Still, even without being able to infer units of transmission in folk songs, the basic observation of Nettl and Bohlman, that there are some parts of a melody which change less than others in oral transmission, is not meaningless. Instead of endeavouring to isolate units of transmission, which do not change, as opposed to other, variable melodic material, it is also possible to investigate given note sequences with respect to their *stability*, i.e., their *resistance to change*, which is a graded property.

Bronson, who used the concept of stability in his research on folk song evolution, stated that "there is probably no more objective test of stability than frequency of occurrence" (Bronson, 1951, p. 51). He applied this principle in his study, discussed in [Chapter 1](#), comparing stability of notes within one tune family. To study the transmission process, I adopt this suggestion to quantify stability through the number of occurrences of a given note sequence within a given tune family. For this, I depart from the tune family categorization in the ANN and FS corpora. Following Bronson's proposal,

I surmise that those note sequences which occur frequently within a given tune family possess more memorability, and therefore survive better in the transmission process, than note sequences which occur in only one variant of a tune family.

Using the tune family categorization of the ANN and FS corpora may be met with two points of criticism: first, tune families may not be a helpful concept in all musical traditions (Cowdery, 1990). Second, the tune family categorization may not always be clear, as different human analysts might associate a given melody with different tune families. Regarding the first point of criticism: tune families in the FS and ANN collection are usually quite homogenous, so the assumption that they are indeed a result of transmission from a common ancestor melody (c.f. Bayard, 1950) is reasonable, even though it is good to keep in mind that this assumption is a simplification of the real, unknown transmission processes. Regarding the second point of criticism: while there is no denying that some melodies' tune family membership may be disputable, the melodies have been checked by several domain experts independently, and the tune family categories shifted relatively little. So while there may be some ambiguities in the tune family categorization, I feel confident that this is only true for a minority of cases, such that I can quantify stability within tune families and arrive at meaningful results.

Pattern matching, the comparison of given note sequences against melodies, can deal with approximate as well as exact occurrences, and is well-researched for music. However, this approach also raises a number of problems:

1. Pattern matching can be performed with various similarity measures, which lead to different results as to which parts of a melody constitute an occurrence of a given melodic segment.
2. To find only relevant occurrences, a *similarity threshold* needs to be set which defines how much deviation is still acceptable for patterns to be considered occurrences of a given melodic segment.
3. It is not clear *a priori* which music representation is suitable for the problem at hand.

The problems related to pattern matching are addressed through quantitative research on the ANN corpus in [Chapter 4](#). Even though no computational method can solve the problems involved in pattern matching completely, this chapter does point out an approach that is suitable for finding most relevant occurrences of melodic segments: a combination of city-block distance, local alignment and structure induction, with a music representation which circumvents the problem of different transpositions in different songs, and which supplies information on the duration of notes.

2.4 CONCLUSION

This chapter summarized the types of variation which may occur in the Dutch folk song transcriptions which will be studied; moreover, based on a wide range of research on

musical variation, the potentials of local and global musical aspects for investigating variation and stability were discussed. These findings underpin my choice to investigate note sequences. As to the question of how we may quantify variation and stability of such note sequences, I introduced two possible approaches: the first approach would be to search for stable patterns through pattern discovery; the second approach would be to gauge the stability of given note sequences by observing their frequency of occurrence. In the following [Chapter 3](#), I present the state of knowledge on pattern discovery, which is still inconclusive as to which method would be successful for finding stable patterns in folk songs. [Chapter 4](#) compares various similarity measures for pattern matching, on which I base my approach of quantifying stability through frequency of occurrence.

Repetitions are a fundamental structuring principle in the majority of musical styles. They guide the listener in their experience of a musical piece, create cohesion, and facilitate the recall process (Margulis, 2014, p.22). Therefore, musical repetitions have been intensively studied in many areas of music research. Computational analysis of repetitions may be an important contribution to such studies, as it facilitates research of repetitions in large music collections, and allows to test theories on repetition and variation, potentially revealing cognitive principles underlying musical repetition. Computationally discovered repetitions may also reveal information about other musical phenomena: for instance, repetitions have been used as indicators of musical segmentation (de Haas, Volk, & Wiering, 2013), or to find related themes or choruses in large databases (Paulus, Müller, & Klapuri, 2010).

One such computational analysis of musical repetitions is musical pattern discovery, which has the goal of inferring salient repetitions in musical pieces. This chapter provides a comprehensive overview, review and discussion of the field of musical pattern discovery. I present the essence of assorted studies, and proceed to clarify the relationships between different methods, proposing a taxonomy of musical pattern discovery approaches according to the following criteria: goals, method, music representation, filtering, and evaluation. Furthermore, the chapter identifies current challenges of musical pattern discovery and suggests steps to overcome these challenges.

The focus of the survey are studies using symbolic music representations. Several of the discussed studies work with audio as well as symbolic representations (Nieto & Farbood, 2014; Pesek, Leonardis, & Marolt, 2014; Wang, Hsu, & Dubnov, 2015), but we will not discuss the problems related to pattern discovery in audio recordings here. For methods in the audio domain, we refer the reader to the overview by Klapuri (2010).

Moreover, the focus of the current chapter is on musical pattern discovery, while pattern matching approaches are addressed in the following chapter. Pattern discovery aims at the identification of motifs, themes, and other musical structures in a piece of music, or between related pieces of music (intra-, and inter-opus discovery). Typically, algorithms applied for pattern discovery do not presuppose prior knowledge of possible candidates.

An overview of all reviewed studies, classifying their various aspects according to a newly proposed taxonomy, can be found in Table 3.1. The first column of this table reports the bibliographical reference of various pattern matching approaches. Where multiple publications on the same method exist, we choose the most recent publication and discuss how it fits into the taxonomy, even though earlier papers with different goals, music representations or filtering choices may exist. For these studies, we refer the readers to the references of the most recent publications.

The columns of Table 3.1 correspond to the ensuing sections. The second column gives an overview of the various goals of pattern discovery studies, which will be reviewed in the next section. The third column categorizes the pattern discovery methods, further described in the second section. The fourth column reflects the music representations used in pattern discovery, discussed in the third section. The fifth column distinguishes different methods of filtering algorithmic results, which are analyzed in the fourth section. The sixth column gives an overview of the studies' evaluation methods used in the studies, which will be further explained in the fifth section. The final section of this chapter summarizes the insights which can be gleaned from the current state-of-the-art of pattern discovery research.

3.1 GOALS OF MUSICAL PATTERN DISCOVERY

One goal of pattern discovery may be to identify large repeating sections (SEC) within musical pieces, such as themes, chorusses, or verses. For verses or stanzas, variation of repetition may be minimal, but certainly plays a role, e.g., through ornamentation or extra internal repetitions of shorter patterns. The Johannes Kepler University Pattern Discovery Development (JKUPDD) database contains several annotations of repeated sections within pieces by Beethoven, Chopin and Mozart. Not all of the five pieces in the database have annotated sectional repetitions, however.

Another goal may be to find shorter repeated patterns of a few notes, which we will refer to here as motifs (MOT): such patterns have been annotated in monophonic jazz solos (Frieler, Pfleiderer, Zaddach, & Abeßer, 2016; Owens, 1974), in pieces by Gibbons, Bach, Mozart, Beethoven and Chopin as part of the JKUPDD database, and in monophonic Dutch folk songs, in the form of characteristic motifs, i.e., short melodic patterns by which experts identify groups of variants in Dutch folk songs (Volk & van Kranenburg, 2012). As the characteristic motifs are defined by comparisons between different variants of folk songs, finding such patterns requires inter-opus pattern discovery.

In some cases, repeating motifs may be of interest for the goal of segmentation, where motifs form meaningful subdivisions. Cambouropoulos (2006) takes this approach, and as an example discusses the well-known song *Frère Jacques*, which can be described by four repeating motifs. Another example is Réti's motivic analysis of Schumann's *Träumerei* (Réti, 1951), which segments the melody into motifs of two to six notes. Buteau and Vipperman (2009) build a computational model for Réti's analysis. While Cambouropoulos' example consists entirely of repeating segments, Réti's motivic analysis also contains some motifs which do not repeat. Pattern discovery will not be able to retrieve such non-repeating motifs. Moreover, repeating motifs may also not always be easy to discover computationally, as they may occur with considerable variation.

The last goal of studies reviewed here is to find variations in counterpoint (COU). This may be the category where variation is strongest, as counterpoint techniques involve variations which are stretched or condensed in duration, transposed in pitch, or whose contour may even be inverted or reversed. While Giraud, Groult, and Levé

Study	Goal	Method	Music rep.	Filtering	Evaluation
Buteau and Vipperman (2009)	MOT	GEOM	P ⊗ O		QUAL
Cambourpoulos (2006)	MOT	STR	PI, CON	LEN, FREQ	SEG
Collins, Arzt, Flossmann, and Widmer (2013)	SEC, MOT	GEOM	P ⊗ O	SPAC, SIM	PAT
Conklin and Anagnostopoulou (2011)	MOT	STR	PI	FREQ	QUAL
Forth (2012)	MOT	GEOM	P ⊗ O	SPAC	QUAL
Hsu, Liu, and Chen (2001)	MOT	STR	P		SPEED
Karydis, Nanopoulos, and Manolopoulos (2007)	MOT	STR	P	LEN	SPEED
Lartillot (2014)	MOT	STR	PI ⊗ DUR	LEN, FREQ, SIM	PAT
Lee and Chen (1999)		STR	PI ⊗ DUR		SPEED
Louboutin and Meredith (2016)	MOT, COU	STR	PI ⊗ DUR	SPAC	CLASS, COMP, PAT
Meek and Birmingham (2001)	SEC	STR	PI	FREQ	PAT
Meredith, Lemström, and Wiggins (2002)	MOT	GEOM	P ⊗ O		QUAL
Meredith (2015)	MOT	GEOM	P ⊗ O	SPAC	CLASS, PAT
Nieto and Farbood (2012)	MOT	TS	PI ⊗ DUR	LEN, FREQ, SPAC	PAT
Nieto and Farbood (2014)	MOT	TS	P	SIM	PAT
Pesek, Medvešek, Leonardis, and Marolt (2015)		TS	P		PAT
Ren (2016)	MOT	STR	PI ⊗ DR	LEN, FREQ	QUAL
Rolland (1999)	MOT	STR	user defined	FREQ	QUAL
Velarde and Meredith (2014)	MOT	TS	P	SIM	PAT
Wang et al. (2015)		TS	P	SIM	PAT

Table 3.1: An overview of the discussed pattern discovery studies. The first column lists the bibliographical reference; the second column defines the goal of the study: finding sections (SEC), finding salient patterns or motifs (MOT), identifying imitations in counterpoint (COU), or unspecified (empty cell); the third column categorizes the study as using a geometric (GEOM), time-series based (TS) or string-based (STR) method; the fourth column specifies the music representation as pitch (P), onset (O), duration (DUR), pitch interval (PI), contour (CON) or duration ratio (DR), where combined presentations are indicated by the tensor product; the fifth column indicates the filtering method used, which can be pattern length (LEN), frequency (FREQ), spacing (SPAC), similarity (SIM), or none (empty cell); and the rightmost column shows the evaluation method used: qualitative (QUAL), segmentation (SEG), classification (CLASS), compression (COMP), computation speed (SPEED) or by comparison against annotated patterns (PAT).

(2012) used pattern matching to identify fugue subjects in Bach fugues, their data has also been used for pattern discovery (Louboutin & Meredith, 2016). Knopke and Jürgensen (2009) investigate masses by Palestrina, for which they segment the different voices of the masses into phrases, and then compare phrases to identify variations. Hence, their work is related to pattern matching rather than pattern discovery.

This overview of different musical pattern discovery goals illustrates the variety of potentially interesting patterns. The amount of variation determines which methods and music representations may be suitable. Long repeated patterns, such as themes or chorusses, contain shorter repeated patterns, which may also be considered salient motifs. Pattern discovery methods might find different kinds of patterns at the same time, after which filtering may discard some results. Recently, several studies introduce methods without a specific musical goal, comparing their pattern discovery results to other state-of-the-art methods. While it is insightful to compare methods in this way, it is also good to keep in mind that a pattern discovery method may perform quite successfully when evaluated on given reference data, e.g., the JKUPDD database, while it may be unsuitable for other musical pattern discovery goals, e.g., finding fugue subjects.

Other computational methods for music analysis may also find musical repetitions, e.g., methods for music prediction, or for measuring motif repetivity. As these methods do not explicitly state where patterns of interest are located, but rather measure the influence of repetition on uncertainty in music prediction (Pearce & Wiggins, 2007), or summarize how many distinct motifs are in a given piece of music (Müllensiefen & Halpern, 2014), these methods are not discussed in the following.

3.2 PATTERN DISCOVERY METHODS

Some pattern discovery methods have been developed specifically for music, but most have been adapted from other disciplines, such as computational biology and natural language processing. Pattern discovery methods can be distinguished into three categories: string-based methods (STR), which assume a melody to be a sequence of discrete symbols, time-series methods (TS), which sample pitches at regular time intervals, and geometric methods (GEOM), which assume a melody to be a collection of points defining the pitch and onset of the melody's notes.

3.2.1 *String-based, time-series and geometric methods*

One approach to musical pattern discovery is to search for identical subsequences of tokens in a string representation (STR) of a melody or multiple melodies. This approach has been derived from techniques developed within Computational Biology to compare gene sequences, and is also applied in other fields, such as Natural Language Processing. Gusfield (Gusfield, 1997) provides a thorough overview of these techniques.

The simplest string-based approach to finding repeated patterns in a melody s consists of sliding all possible query patterns q past s , and recording all found matches. This approach is taken by Müllensiefen and Halpern (2014), in an exhaustive search for n -grams, and by Ren (2016) to find repeated patterns in Bach chorales, jazz standards and folk songs.

There has been research into speeding up string-based pattern discovery, through skipping comparison steps between q and s without missing any relevant patterns. One of these extensions, the algorithm by Knuth, Morris, and Pratt (1977), has been applied by Rolland (1999) for musical pattern discovery. Yet another approach is Crochemore's set partitioning method (Crochemore, 1981), which recursively splits the melody s into sets of repeating tokens. Cambouropoulos (2006) used this method to find maximally repeating patterns (i.e. repeated patterns which cannot be extended left or right and still be identical) in musical pieces. Karydis et al. (2007) refine the set-partitioning approach to find only the longest patterns for each musical piece, with the intuition that these correspond most closely to musical themes.

Meek and Birmingham (2001) transform all possible patterns up to a maximal pattern length to keys in a radix 26 system (representing 12 intervals up or down, unison, and 0 for the end of the string). After a series of transformations, which consolidate shorter into longer patterns, identical patterns are encoded by the same numerical keys. Another potentially interesting group of methods are compression algorithms, as they reduce the size of data by finding regularities in it. Louboutin and Meredith (2016) applied such general-purpose compression algorithms for a pattern discovery task.

There are a number of studies which use indexing structures to speed up the search for repeated patterns (Conklin & Anagnostopoulou, 2011; Hsu et al., 2001; Lartillot, 2014; Lee & Chen, 1999). Conklin and Anagnostopoulou (2011) use a suffix tree to represent search spaces of patterns in Cretan folk songs, which is pruned based on the patterns' frequency of occurrence. Hsu et al. (2001) compare the use of a correlational matrix tracking prefixes of repeated patterns against the use of a suffix tree for musical pattern discovery. Lartillot (2014) employs a pattern tree to model musical pieces. This is a prefix tree which allows for cyclic structures within the graph, which can capture repetitions of short patterns, forming building blocks within larger repeated structures.

In geometric methods (GEOM), the pitch, onset and other information of notes are not treated as symbols, but as values: through this approach, a melody is considered as a shape in an n -dimensional space. Repeated patterns are then identified as (near-)identical shapes. Geometric methods are especially interesting for polyphonic music, as they deal more conveniently with note events occurring at the same time (Meredith et al., 2002, p.328). For string-based methods, polyphony has been approached through, e.g., encoding distances between voice pairs (Conklin & Bergeron, 2010).

Meredith's *Structure Induction Algorithms* (SIA) (Meredith et al., 2002) order points of note pitch and onset lexicographically, and search for vectors of pitch and onset relationships which repeat elsewhere in a musical piece. This concept, used by Meredith and Collins (Collins et al., 2013; Meredith, 2015; Meredith et al., 2002) for pattern discovery, has been applied by Lemström, Mikkilä, and Mäkinen (2009) for pattern match-

ing, and is also investigated in [Chapter 4](#). The musical pattern discovery approaches by Buteau and Viperman (2009) and Szeto and Wong (2006) rely on a similar conceptualisation of music as n -dimensional shapes. In the latter study, the geometric relationships are represented as nodes and edges in graphs (Szeto & Wong, 2006).

Time-series methods (TS), like geometric measures, use values rather than symbols to represent note pitches, but treat the time axis comparable to sampling in the audio domain: an increment is chosen, for instance a sixteenth note, and for each increment, the corresponding pitch or pitches are registered in a time-series representation. Time-series methods can therefore also be used to discover patterns in polyphonic music.

A common method for discovery of repeated segments in audio is also used for symbolic time-series pattern discovery: a similarity matrix between each pair of values in the time-series is constructed, based on a distance metric to compare the values. Repeated patterns are then inferred based on diagonals of contiguous high similarity values in this matrix. This approach is taken by Nieto (2012; 2014) and Velarde and Meredith (2014). Velarde and Meredith (2014) transform the time-series representations by convolution with the Haar wavelet before constructing the similarity matrix. This transformation is meant to ensure that patterns are found even if they are not notated in the same key, as the wavelet transform registers the changes in the pitch contour rather than the absolute pitch values (see also [Section 4.2](#)).

Wang et al. (2015) use a Variable Markov Oracle, which is a memory efficient indexing structure derived from suffix trees, where notes represent states. These states are connected by links, which can point forward or backward, and which represent connections between repetitions in the sequence. As this method works with symbols rather than values, they discretize the time series with a similarity threshold determining which values are considered the same states in the Variable Markov Oracle. Pesek et al. (2015) use a time-series to build a compositional hierarchical model: a neural network in which a given layer represents combinations of units at lower layers.

3.2.2 *Exact or approximate matching*

Next to searching for exact matches, approximate matching is also of great interest to musical pattern discovery. Rhythmic, melodic and many other conceivable variations are likely to occur, such as the insertion of ornamentations during a repetition, the speeding up or slowing down of a musical sequence, deviations in pitch, or transpositions.

Several ways to define approximate matching for musical pattern discovery have been proposed, usually achieved through a distinction between approximate matches and irrelevant matches, based on a threshold on a similarity measure. For string-based methods, this can be the number of allowed mismatches (also known as Hamming distance or *k-mismatch*), or the length of the longest common subsequence (Lemström & Ukkonen, 2000). Furthermore, the threshold can also be defined as the maximum amount of edit operations in an alignment algorithm, also known as edit distance or Levenshtein distance (Levenshtein, 1966). This way, also strings of different length, or

strings containing gaps in relation to each other, can be considered as approximate matches. Rolland (1999) applied approximate matching to musical pattern discovery, using the Levenshtein distance to compare a pattern with a match candidate.

For time-series methods, values can be compared with distance metrics (for some examples of such metrics, see Section 4.2). Geometric measures, defined on points, may also employ topological distance metrics such as the Hausdorff distance. Romming and Selfridge-Field (2007) use this metric for pattern matching. Implicitly, however, approximate matching can also be achieved through more abstract music representations, as Cambouropoulos, Crochemore, Iliopoulos, Mohamed, and Sagot (2007) point out. We will address the influence of music representation in Section 3.3, but first, discuss recent developments and new challenges of musical pattern discovery research.

3.2.3 *Recent developments and new challenges*

Musical pattern discovery research has been very active in the past few years: several new methods were proposed, and systematic comparisons of methods were performed. For comparison, the Music Information Retrieval EXchange (MIREX) track *Discovery of Repeated Themes & Sections* (Collins, 2013), with its own development and test dataset (JKUPDD and JKUPDT, respectively), has been highly influential. From the comparisons so far it seems that geometric methods (Meredith, 2015) are good approaches for, especially, polyphonic music, while a time-series of wavelet transforms (Velarde & Meredith, 2014) has been successful for discovery of themes in monophonic music.

Next to the MIREX track, two recent studies compare pattern discovery methods on other datasets: Boot, Volk, and de Haas (2016) compare several pattern discovery algorithms for inter-opus and intra-opus discovery of Dutch folk song variants. They perform folk song classification based on discovered patterns, for which results from geometric methods (Meredith, 2015) lead to some of the best results, but based on parameter configurations, a time-series based method (Nieto & Farbood, 2014) performed almost equally well after intra-opus discovery, and the string-based method by Conklin and Anagnostopoulou (2011) performed even better after inter-opus discovery. None of the discovered patterns were more informative for classification than one of the study's baselines: classification based on the first few notes of each folk song melody. Louboutin and Meredith (2016) compare a geometric method (Meredith, 2015) and string-based compression algorithms for the discovery of fugue subjects, for which the geometric method performs best.

The comparisons so far show the same trend, that geometric methods seem a good approach to pattern discovery. However, one also has to keep in mind that the methods described by Meredith (2015) have been researched most, and taken along in all comparisons, while some methods may have never been tested in comparisons, based on the availability of code and research time. Moreover, filtering and music representation influence pattern discovery results, and comparisons usually test selected configurations, while the potential to improve methods through other music representations or filtering choices may not always have been investigated.

Comparison of methods on a broad palette of genres, and kinds of patterns, would be an important next step. The JKUPDD dataset has been the most frequent reference for comparison so far. It focusses on Classical music, but with five pieces, it is rather small, and is very heterogeneous, both in terms of the epochs of Classical music which are covered (from Renaissance to Romantic composers), and in terms of the kinds of patterns which are annotated (sections and motifs). Some datasets, e.g., annotations of licks in jazz solos (Frieler et al., 2016) have not been used for pattern discovery evaluation yet, while other datasets have been used rarely.

3.3 MUSIC REPRESENTATION

There are different musical aspects to be considered for comparisons of musical patterns: rhythm, pitch, but also dynamics, timbre, and many more. Symbolic methods mostly focus on pitch and rhythm, as this information is most readily available. These two musical aspects can be represented in many different ways, however: in terms of absolute values; in terms of categories or classes; in terms of contours indicating direction of change; and many others. For instance, the notes of a melody could be represented by a string of pitch names (A; G; A; D), as MIDI note numbers (57; 55; 57; 50), or as points representing both pitch name and duration of the note ((A,1.5),(G,0.5),(A,1.0),(G,1.0)). This is closely related to Conklin’s notion of musical *viewpoints* (Conklin & Witten, 1995). Conklin calls combinations of several musical dimensions, such as pitch and duration, *linked viewpoints*.

A glance at the music representation column of Table 3.1 reveals that the majority of the studies on musical pattern discovery use pitch (P) or pitch intervals (PI) as the music representation, in some of the studies this is combined with rhythmic representations such as note onset (ON) or duration (DUR). Several studies use more abstract representations describing pitch contour (CON), or relationships between consecutive durations, such as duration ratio (DR). Linked viewpoints, i.e., combinations of two music representations, following Conklin’s notation, are represented by a tensor product: for instance, the combination of pitch and onset is denoted by $P \otimes O$, and of pitch interval and duration interval by $PI \otimes DR$.

Rolland (1999) allows the users of his *FLEXPAT* software to switch between different music representations, but he does not report how this influences the results of his musical pattern discovery algorithm. Cambouropoulos et al. (2007) suggest to compare a pitch interval representation with a more abstract step-leap representation, but results of these two representations are not discussed by the authors.

An open question is how to combine multiple viewpoints for pattern discovery: they may be linked, or treated separately, which requires combining the results of multiple pattern discovery procedures. Lartillot (2014) constructs combined pattern trees for the two music representations pitch interval and onset, for which new branches are created independently. Lee and Chen (1999) find neither the use of linked viewpoints, nor of separate pattern discovery procedures satisfying, so they suggest two new indexing

structures, *Twin Suffix Trees*, and *Grid-Twin Suffix Trees*, as possible alternatives. They do not report any results produced by these different representations.

Meredith et al. (2002) suggest different ways to represent pitch: for instance, through using diatonic pitch categories rather than chromatic ones, repeated patterns which are, e.g., transformed from major to minor tonality may also be detected. These subcategories are not distinguished in Table 3.1, which just lists $P \otimes O$ to reflect that Meredith's method uses tuples of pitch and onset. Louboutin and Meredith (2016) compare how compression algorithms perform on a number of different linked viewpoints with pitch interval and duration representations.

3.3.1 *Recent developments and new challenges*

Most studies do not explicitly compare results of pattern discovery for different music representations. Louboutin and Meredith (2016) tests various music representations for compression algorithms. Their results indicate that relative distances between adjacent pitches and onsets are more informative for pattern discovery than absolute distances from the starting pitch and onset of a piece: this seems logical, as onsets from the start of the piece would never show repetitions. Likewise, the same pattern, transposed up or down, would be represented by the same relative pitch intervals, but by different pitch intervals from the starting pitch of a piece.

In general, pitch intervals are the preferred representation for the pitch domain, often in combination with duration. The linked pitch-onset viewpoints rely mostly on the strategy of geometrical methods to trace patterns through repeated relationships between pitch-onset points, rather than looking at absolute repetitions of values. More abstract music representations such as contours have been researched very little so far. The intuition with these viewpoints is that while it may be possible to find patterns which show more variation, e.g., patterns in which the sizes of pitch intervals are slightly altered upon repetition, the higher abstraction may also lead to the discovery of more irrelevant patterns. How exactly this trade-off between abstraction and precision should be judged has not yet been researched systematically, but would be very informative.

Research on music representations in musical pattern discovery may also benefit from experimental research on perception and recall of music (e.g. Dowling, 1978). Experimental research may provide theories which can be employed and tested by musical pattern discovery. The comparison of musical pattern discovery methods using different music representations informed by perceptual theories will generate insights which can feed back into research on similarity and variation in music theory and music cognition.

3.4 FILTERING

A frequently described problem in musical pattern discovery is the great amount of algorithmically discovered patterns as compared to the patterns that would be con-

sidered relevant by a human analyst. For the task of computer-aided motivic analysis, Marsden observed that “... the mathematical and computational approaches find many more motives and many more relationships between fragments than traditional motivic analysis.” (Marsden, 2012) Therefore, most of the presented studies employ a filtering step, which is supposed to separate the wheat from the chaff. Common filtering approaches judge the qualities of patterns based on their length, their frequency, their compactness or the compression that is achievable by representing a melody just in terms of the discovered patterns.

Several approaches to filtering can be distinguished, which are often combined. A common notion is that patterns should have a minimum *length* (LEN) to be interesting. This of course depends on the application of pattern discovery: for short melodic patterns such as licks in jazz solos, the frequency (FREQ) of a pattern may be more important than its length. Another approach to filtering relates to spacing (SPAC), suggesting that patterns should not contain gaps, such as rests or interposed notes, or be too close to each other. Finally, approximate matching methods may filter based on the similarity (SIM) between occurrences of discovered repeated patterns, typically optimizing a similarity threshold through training on a smaller dataset.

3.4.1 *Filtering based on length*

One commonly used filtering approach is based on the assumption that extremely short patterns may be less relevant than longer ones. Filtering may proceed in two ways: either, given multiple discovered patterns, longer patterns will be preferred over shorter ones (e.g. Cambouropoulos, 2006), or a minimum length is defined, such that shorter patterns are discarded (e.g., Nieto & Farbood, 2012). However, there may be multiple levels of repetition, which may be more or less interesting for given purposes: for instance, in the string “a b a b c a b a b c”, one could identify the patterns “a b”, “a b c” and “a b a b c”. While the last pattern is longest, the first pattern is much more frequent, and may therefore be interesting in its own right. This is why pattern discovery results are mostly filtered on criteria taking the frequency as well as the length of discovered patterns into account.

3.4.2 *Filtering based on frequency*

Another commonly used filtering approach is based on the assumption that patterns which occur more often might also be considered as more important by human analysts. As the filtering for length, it may take the form of preferring frequent patterns over less frequent patterns (e.g., Cambouropoulos, 2006; Rolland, 1999), or of discarding all patterns which occur less often than a user-defined threshold (e.g., Nieto & Farbood, 2012).

If a minimum number of o occurrences is defined, a pattern which occurs at least o times is considered a *frequent pattern*. The number of occurrences o is also known as a pattern’s *support*. A common concept in sequential pattern mining are *closed patterns*:

only patterns which, for a given support, are not contained by other patterns are considered closed. Lartillot (2014) and Ren (2016) employ the closed pattern criterion for filtering patterns.

Conklin and Anagnostopoulou (2011) are also interested in a pattern's frequency of occurrence, but weigh it against its frequency in a collection of contrasting music pieces, the *anticorpus*. This process is designed to favour patterns that are characteristic for the analyzed piece, or for a corpus. Conversely, also patterns which are underrepresented in specific pieces or genres can be interesting for music researchers (Conklin, 2013).

3.4.3 *Filtering based on spacing*

Spacing is used here to subsume two filtering concepts: compactness of a pattern and pattern distance. Compactness relates to the notion that notes belonging to patterns should be contiguous as far as possible. For instance, a pattern of three notes which includes one note at the beginning, one in the middle and one in the end of a piece is not desirable. Collins' *compactness trawling* to refine the results of Structure Induction Algorithms (Collins et al., 2013) is based on such a filtering, in which patterns of adjacent notes are preferred over patterns with many intervening notes. In a similar vein, Nieto and Farbood (2012) filter according to Gestalt rules during the search process, which means that pattern candidates containing relatively long notes or rests, or relatively large intervals will be rejected. Moreover, they define a minimum distance between patterns, with the intuition that patterns will not follow each other immediately in a musical piece. Meredith (2015) filters the results of his pattern discovery algorithm based on the requirement that no two patterns may cover the same notes in a musical piece, an algorithm which he calls COSIATEC.

3.4.4 *Filtering based on similarity*

Finally, filtering may be performed based on how much candidates for repeated patterns resemble each other: this step is only applicable for approximate matching methods. As such, Nieto and Farbood (2014) and Velarde and Meredith (2014) define a threshold of similarity, above which traces in the similarity matrix will be considered occurrences of repeated patterns. Wang's (2015) construction of a Variable Markov Oracle depends on a similarity threshold between pairs of symbols in the sequence. He optimizes this threshold on a training corpus, selecting the best model based on its *Information Rate*, a measure derived from entropy. Lartillot (2014) filters constructed pattern trees by choosing patterns which correspond both in pitch interval and duration over patterns which are only found in one music representation. Collins et al. (2013) also propose a similarity based filtering step for a geometric pattern discovery method, which within a given range of notes of exact matches, searches for notes which might be part of inexact matches.

3.4.5 *Recent developments and new challenges*

Filtering raises two questions: first, which filtering approaches lead to discovered patterns corresponding most closely to patterns which human listeners would consider salient or relevant? Second, which length, frequency, spacing or similarity thresholds should be set to maximize the suitability of methods for specific musical pattern discovery goals?

As to the first question, the geometric method first proposed by Meredith et al. (2002) is perhaps the most thoroughly researched, as the original method, SIATEC, has been filtered based with the goal of removing any overlap between discovered patterns (COSIATEC) (Meredith, 2015), or to make sure that notes of discovered patterns would be proximate (SIARCT) (Collins et al., 2013). Moreover, approximate matches with a similarity threshold have also been investigated (SIARCT-CFP) (Collins et al., 2013). The influence of frequency or length filtering on the method has not been reported yet, however, or how different filtering strategies might interact. For most other methods, different filtering strategies have not been systematically compared, while such a comparison would generate many new insights.

To answer the second question on ideal thresholds for filtering, some information can be obtained from the results of the MIREX pattern discovery challenge, where some methods were entered with various filtering settings. Moreover, Boot et al. (2016) also compare a range of filtering settings for pattern discovery methods. Likewise, Meek and Birmingham (2001) analyze different filtering settings through optimization on a training set. Regrettably, in all cases the comparison does not go beyond picking the “best” filtering parameters, i.e., the parameters which lead to the closest match with human annotations. Which criteria might underlie filtering choices for length, frequency, spacing or similarity of patterns are not explicitly discussed.

For approximate matching, Clifford and Iliopoulos (2004) draw the distinction between pairwise comparisons of values in two sequences – δ -matching – or the sum of all differences between the values in two sequences – γ -matching. To our knowledge, these two different approaches to similarity based filtering have not yet been investigated systematically for musical pattern discovery.

Moreover, much can be gained through exchange with experimental research. For instance, Margulis’ (2012) listening experiments, in which she tested how well repeated patterns were detected depending on the length of the patterns, and how many times they were presented, is informative for the influence of length and frequency on pattern salience. More listening experiments would be very enlightening, for instance to test whether pattern spacing may also play a role for salience, or which kinds of patterns will be considered similar enough by human listeners to be considered repetitions. On the other hand, it would also be an interesting next step to test whether observations from listening experiments may also be reproduced by a pattern discovery method; for instance, it would be intriguing to see if pattern discovery methods can reproduce Margulis’ results, according to which human listeners are more likely to recognize

short repeated patterns after few exposures, while recognizing long repeated patterns more readily after many exposures.

3.5 EVALUATION

In many pattern discovery studies, evaluation takes place qualitatively, i.e., selected discovered patterns are presented to the reader (QUAL). Most, if not all studies complement these qualitative findings with quantitative evaluation. Quantitative evaluation may be based on the computation speed of a given pattern discovery method (SPEED), yet this has become of less concern in recent publications. In some studies, the patterns are used to derive a meaningful segmentation and evaluated against human annotations of segmentations (SEG). In other studies, the evaluation takes place through classification: if a melody can be successfully classified by only the discovered patterns, this is taken as evidence that the patterns are meaningful (CLASS). In recent years, evaluations against musical pieces in which meaningful motifs and themes have been annotated have gained popularity (PAT).

3.5.1 *Qualitative evaluation*

The vast majority of studies present some qualitative evaluation of pattern discovery results, by showing some example patterns. While this may highlight some interesting achievements of automatic pattern discovery, such example patterns leave it unclear whether all discovered patterns are as meaningful – the presented patterns may just be the cherries picked from a large bag of potentially not very interesting patterns. Therefore it is laudable that recent studies increasingly make use of quantitative evaluation measures.

3.5.2 *Evaluation on speed*

In some studies (e.g., Lee & Chen, 1999), the researchers aim for fast solutions, which make the algorithms more interesting for practical use. Therefore, computation speed is used as an evaluation metric in these cases. This does not give an indication of the usefulness of the automatically found patterns, however. Recent studies have not focussed on speed in evaluation, since most pattern discovery methods are not aimed at realtime applications, making speed a subordinate concern.

3.5.3 *Evaluation on segmentation*

It can be argued that repetition defines structural boundaries in a musical piece. Therefore, annotations on segmentation may be used to evaluate pattern discovery methods: if discovered patterns overlap annotated structural boundaries, this indicates that these patterns would probably not be considered meaningful by human analysts. This

approach is taken by several studies (Buteau & Vipperman, 2009; Cambouropoulos, 2006).

3.5.4 *Evaluation on classification*

Boot et al. (2016) and Louboutin and Meredith (2016) evaluate the success of pattern discovery method based on their success at classifying folk song melodies. These melodies are reduced to the discovered patterns, and the authors investigate whether the patterns provide enough information to assign the melodies to tune families correctly, as defined in the Annotated Corpus of the Meertens Tune Collections.

3.5.5 *Evaluation on compression*

Compression algorithms make use of repetitions in data to reduce data size: if a repeated pattern needs to be stored only once with pointers to its location in the data, this saves storage space over storing the repeated patterns explicitly. Therefore, compression rate has been used in several studies as a measure of how effectively repeating patterns in musical pieces are revealed through pattern discovery (Louboutin & Meredith, 2016). Similarly, Boot et al. (2016) complement their analysis of pattern discovery methods for folk song classifications by reporting the coverage, i.e., the percentage of melody notes belonging to discovered patterns.

3.5.6 *Evaluation on annotated patterns*

Some of the presented studies use annotations of motifs, themes or other meaningful patterns to evaluate the results of musical pattern discovery. Such annotations range from overviews of frequently used licks in jazz improvisation (Owens, 1974) to themes in Western art music (Barlow & Morgenstern, 1948), and are typically created by domain specialists, who annotate what they consider the most relevant patterns of the analyzed music collection.

The first comparisons with such reference annotations were performed through counting exact correspondences between annotated and automatically discovered patterns (Collins et al., 2013; Meek & Birmingham, 2001; Nieto & Farbood, 2012). This approach does not take into account that pattern discovery methods may find patterns in a slightly shifted position with respect to annotated patterns.

Recent comparisons of pattern discovery algorithms therefore made use of cardinality scores, based on the number of shared notes between automatically discovered and annotated patterns (Collins, 2013). Two goals may be considered when evaluating pattern discovery results on annotations: one, to correctly identify all distinct patterns annotated in musical pieces; another, to correctly identify all occurrences of the distinct patterns. For the identification of all distinct annotated patterns, Collins (2013) suggests the measures establishment precision, recall and F1-score; for the correct identification of all occurrences, occurrence precision, recall and F1-score.

Task	$F1_{Est}$	$F1_{Occ}$	3LF1	Method
polyphonic	[.42, .66]	[.42, .77]	[.32, .58]	Meredith (2015)
monophonic	[.55, .93]	[.50, .74]	[.32, .68]	Velarde and Meredith (2014)

Table 3.2: The ranges of establishment F1-score, occurrence F1-score and Three-Layer F1-score for the best-performing methods in the polyphonic and monophonic MIREX pattern discovery tasks.

Meredith (2015) suggests three-layer precision, recall and F1-score as alternative measures. These measures are also based on the number of shared notes between annotated and discovered patterns, which forms the first layer of the evaluation. From this, a second layer is derived, which compares distinct patterns, and is therefore comparable to Collins’ establishment measures. The third layer evaluates whether the occurrences of patterns are found correctly, comparable to Collins’ occurrence measures.

To give an impression of typical results with relation to commonly used reference annotations, Table 3.2 presents the ranges of the establishment F1-score, occurrence F1-score and three-layer F1-score for the pattern discovery methods which overall scored highest for the five pieces of the 2016 musical pattern discovery MIREX evaluation, for the polyphonic and monophonic task. The range of values shows that the success of methods depends very much on the piece on which evaluation takes place. This is further illustrated by Louboutin and Meredith’s (2016) evaluation of COSIATEC on fugue subject discovery in a collection of Bach fugues, for which they report three-layer F1-score of 0.123, which is substantially lower than the scores of the algorithm in the MIREX task.

3.5.7 *Recent developments and new challenges*

Recent studies introduced many valuable approaches to quantitative evaluation, either implicitly on classification or compression, or explicitly on pattern annotations. These evaluations give a good first impression as to which methods might be good candidates to discover salient repeated patterns. However, as evaluation is performed with specific pattern discovery goals, and in specific musical styles, replicating evaluation of established methods on more datasets would be an important extension of the current knowledge.

Pattern annotations may be open to the criticism that different annotators might consider different patterns as relevant, or disagree on where a relevant pattern starts or ends. Therefore, multiple annotator judgements on relevant repeating patterns might be an interesting extension of current annotated datasets, as this would make the nature of annotator disagreement explicit. Potentially, pattern discovery algorithms could be evaluated on different annotators’ judgements separately, or annotators’ judgements might be pooled through a majority vote.

The quantitative evaluation approaches of the past few years are more informative than evaluations in the pioneering musical pattern discovery studies, which provided qualitative evaluations of a few selected patterns and did not give much insight into the overall success of a method. Yet without any qualitative evaluation, classification accuracy, compression rate, or precision and recall measures do not give any real insights into the problems of musical pattern discovery: which cases are handled successfully, and where do the various methods fall short? Without qualitative analysis of the errors produced by methods, it is hard to pinpoint potential areas of improvement.

One evaluation strategy has not been applied for pattern discovery so far: evaluation through prediction. As repeated patterns enable human listeners to predict the next events in a musical piece successfully, which may be one of the reasons we derive pleasure from listening to music (Huron, 2007), musical pattern discovery may be evaluated based on how well it succeeds in predicting musical events. One disadvantage of this strategy, as with classification or compression, is that the location of important repeated patterns would not necessarily be revealed by a prediction task.

The best route to gain more knowledge in musical pattern discovery is to evaluate as broadly as possible: through implicit quantitative evaluation methods, such as classification, compression and prediction; through explicit quantitative evaluation by comparison with pattern annotations; and through qualitative evaluation of errors.

3.6 CONCLUSION

Our literature overview has highlighted the different kinds of patterns which studies have aimed at so far: sections, motifs, and fugue subjects. These patterns differ in length, and in the variation that is admissible for the patterns to be recognized as repetitions. Pattern discovery may be approached through string-based, time-series or geometric methods. Multiple music representations have been applied. So far, pitch or pitch interval representations, often combined with onset or duration information, are the most frequently used music representations. Approaches to filtering are based on the length, frequency, spacing and similarity of patterns, strategies which have been all broadly applied, and which are sometimes traded off against each other, as in the case of length and frequency of patterns. For evaluation, qualitative evaluation of discovered patterns has been complemented with various quantitative measures, of which speed is the least frequently reported in recent studies. Other quantitative evaluation methods include implicit strategies, such as segmentation, classification and compression, and explicit evaluation by comparison to annotated patterns.

Recent comparisons of musical pattern discovery methods for different evaluation scenarios have increased insights into the relative success of the state-of-the-art methods; however, as evaluation often takes place with relation to selected pattern discovery goals, or in specific music genres, it is still hard to gauge how well the various methods generalize to other goals, or to other genres. Some pattern annotations on which evaluation might take place, as well as alternative evaluation strategies, may have still been overlooked so far.

The most difficult challenge remains to understand how music representation and filtering interact with each other, and with a given pattern discovery method. A desirable outcome of the presented, mostly retrieval-oriented research would be a dialogue with experimental research on music perception and recall: pattern discovery methods may be improved by incorporating knowledge from experimental research; in turn, insights into if and how pattern discovery methods behave differently than human analysts can benefit music cognition.

Generally, musical pattern discovery would benefit from stating the underlying assumptions leading to choices of music representation and filtering as explicitly as possible, and to shift the focus away from optimizing performance and onto testing conceivable assumptions: for example, is contour more important for repetition recognition than pitch intervals (Dowling, 1978)? This may lead to less success in terms of evaluation measures, but may eventually yield more knowledge. The value of errors has also been under-appreciated so far: what can we learn from annotated patterns which pattern discovery methods cannot find? Which annotated patterns are discovered readily by different pattern discovery methods? Previous comparative studies may still provide a wealth of such error data which has not yet been investigated.

The wealth of musical pattern discovery studies of the past few years, including recent attempts to bridge pattern discovery in the symbolic and audio domain, and systematic comparisons of methods, give all reason to look forward to more exciting explorations of the seemingly simple but hard to model aptitude of humans to hear repetitions in music.

FINDING OCCURRENCES OF MELODIC SEGMENTS IN FOLK SONGS

A large body of computational music research has been devoted to the study of variation of folk songs, in order to understand what characterizes a specific folk style (e.g., Conklin & Anagnostopoulou, 2011; Juhász, 2006), or to study change in an oral tradition (e.g., Bronson, 1950; Louhivuori, 1990; Olthof et al., 2015). In particular, a very active area of research is the automatic comparison of folk song melodies, with the aim of reproducing human judgments of relationships between songs (e.g., Bade, Nürnberger, Stober, Garbers, & Wiering, 2009; Boot et al., 2016; Eerola, Jäärvinen, Louhivuori, & Toiviainen, 2001; Garbers et al., 2009; Hillewaere, Manderick, & Conklin, 2009; Müllensiefen & Frieler, 2007). Recent evidence shows that human listeners do not so much recognize folk songs by virtue of their global structure, but instead focus on the presence or absence of short melodic segments, such as motifs or phrases (Volk & van Kranenburg, 2012).

This chapter compares a number of similarity measures as potential computational approaches to locate melodic segments in symbolic representations of folk song variants. We investigate six existing similarity measures suggested by studies in ethnomusicology and Music Information Retrieval as promising approaches to find occurrences.

In computational ethnomusicology, various measures for comparing folk song melodies have been proposed: as such, correlation distance (Scherrer & Scherrer, 1971), city-block distance and Euclidean distance (Steinbeck, 1982) have been considered promising. Research on melodic similarity in folk songs also showed that alignment measures can be used to find related melodies in a large corpus of folk songs (van Kranenburg et al., 2013).

As this chapter focusses on similarity of melodic segments rather than whole melodies, recent research in musical pattern discovery is also of particular interest. Two well-performing measures in the associated MIREX challenge of 2014 (Meredith, 2014; Velarde & Meredith, 2014) have shown success when evaluated on the Johannes Kepler University Patterns Test Database (JKUPTD).¹ We test whether the underlying similarity measures of the pattern discovery methods also perform well in finding occurrences of melodic segments.

The six measures investigated in this chapter were also used in an earlier study (Janssen et al., 2015) and evaluated against binary labels of occurrence and non-occurrence. Here, we evaluate not only whether occurrences are detected correctly, but also whether they are found in the correct position. Moreover, we evaluate on a bigger data set, namely the Annotated Corpus of the Meertens Tune Collections, MTC-ANN 2.0 (van Kranenburg et al., 2016).

¹ http://www.music-ir.org/mirex/wiki/2014:Discovery_of_Repeated_Themes_%26_Sections_Results

Two measures compared in our previous study (Janssen et al., 2015) – B-spline alignment (Urbano, Lloréns, Morato, & Sánchez-Cuadrado, 2011) and Implication-Realization structure alignment (Grachten, Arcos, & López de Mántaras, 2005) – are not evaluated here as in their current implementation, they do not allow determining the positions of occurrences in a melody.

We present an overview of the compared similarity measures in Table 4.1, with their abbreviation used throughout the chapter, and bibliographical references to the relevant papers.

Abbreviation	Similarity measure	Authors
CD	Correlation distance	(Scherrer & Scherrer, 1971)
CBD	City-block distance	(Steinbeck, 1982)
ED	Euclidean distance	(Steinbeck, 1982)
LA	Local alignment	(van Kranenburg et al., 2013)
SIAM	Structure induction	(Meredith, 2014)
WT	Wavelet transform	(Velarde & Meredith, 2014)

Table 4.1: An overview of the measures for music similarity compared in this research, with information on the authors and year of the related publication.

We evaluate the measures by comparison to phrase annotations by three domain experts on a selection of folk songs, produced specifically for this purpose. We employ the similarity measures and the annotations to address four research questions:

- Q1. Which of the proposed similarity measures performs best at finding occurrences of melodic segments in folk songs?
- Q2. Folk songs are often notated in different octaves or keys, or in different meters, as exemplified by two variants displayed in Figure 4.1. How can the resulting transposition and time dilation differences best be resolved? Does a different music representation improve the performance of similarity measures?
- Q3. Can a combination of the best-performing measures improve agreement with human annotations?
- Q4. Our folk song corpus contains distinct groups of variants. How robust are the best-performing measures to such subgroups within the data set?

The remainder of this chapter is organised as follows: first, we describe our corpus of folk songs, which has annotations of phrase occurrences. Next, we give details on the compared similarity measures, and the methods used to implement the similarity measures, and to evaluate them. In Section 4.4, we perform an overall comparison of the six similarity measures (Q1). Section 4.5 addresses the influence of transposition and time

NLB072664_01 - Phrase 1



NLB075074_01 - Phrase 1



Figure 4.1: The **first phrase of two variants of a folk song**, notated at different octaves and in different meters. Similarity comparison of the pitches and durations might lead to no agreement between the two variants, even though they are clearly very related.

dilation on the results (Q2). Section 4.6 introduces a combined measure based on the best-performing similarity measures and music representations (Q3), and Section 4.7 investigates the robustness of the best measures towards variation in the data set (Q4). The evidence from our results leads to a number of concluding remarks and incentives for future research.

4.1 MATERIAL

We evaluate the similarity measures on the MTC-ANN 2.0 corpus of Dutch folk songs. We parse the `**kern` files as provided by MTC-ANN 2.0 and transform the melodies and segments into the required music representations using `music21` (Cuthbert & Ariza, 2010). Even though MTC-ANN 2.0 comprises very well documented data, there are some difficulties to overcome when comparing the digitized melodies computationally. Most importantly, the transcription choices between variants may be different: where one melody may have been notated in 3/4, and with a melodic range from D₄ to G₄, another transcriber may have chosen a 6/8 meter, and a melodic range from D₃ to G₃, as shown in Figure 4.1. This means that notes which are perceptually very similar might be hard to match based on the digitized transcriptions. Musical similarity measures might be sensitive to these differences, unless they are transposition or time dilation invariant, i.e., work equally well under different pitch transpositions or meters.

For the corpus of 360 melodies categorized into 26 tune families, we asked three Dutch folk song experts to annotate similarity relationships between phrases within tune families. The annotators all have a musicological background, and had worked with the folk song collection for at least some months previous to the annotation procedure. They annotated 1891 phrases in total. The phrases contain, on average, nine notes, with a standard deviation of two notes. The data set with its numerous annotations is publicly available.²

² <http://www.liederenbank.nl/mtc/>

For each tune family, the annotators compared all the phrases within the tune family with each other, and gave each phrase a label consisting of a letter and a number. If two phrases were considered “almost identical”, they received exactly the same label; if they were considered “related but varied”, they received the same letter, but different numbers; and if two phrases were considered “different”, they received different letters (cf. an annotation example in Figure 4.2).

The three domain experts worked independently on the same data, annotating each tune family separately, in an order that they could choose themselves. To investigate the subjectivity of similarity judgements, we measure the agreement between the three annotators on pairwise phrase similarity using Fleiss’ Kappa, which yields $\kappa = 0.76$, constituting substantial agreement.

The annotation was organized in this way to guarantee that the task was feasible: checking for instances of each phrase in a tune family in all its variants (27,182 comparisons) would have been much more time consuming than assigning labels to the 1891 phrases, based on their similarity. Moreover, the three levels of annotation facilitate evaluation for two goals: finding only almost identical occurrences, and finding also varied occurrences. These two goals might require quite different approaches. In the present study, we focus on finding “almost identical” occurrences.

4.2 COMPARED SIMILARITY MEASURES

In this section, we present the six compared similarity measures, describing the music representations used for each measure. We describe the measures in three subgroups: first, measures comparing equal-length note sequences; second, measures comparing variable-length note sequences; third, measures comparing more abstract representations of the melody.

Some measures use note duration next to pitch information, whereas others discard the note duration, which is the easiest way of dealing with time dilation differences. Therefore, we distinguish between music representation as *pitch sequences*, which discard the durations of notes, and *duration weighted pitch sequences*, which repeat a given pitch depending on the length of the notes. We represent a crotchet or quarter note by 16 pitch values, a quaver or eighth note by 8 pitch values, and so on. Onsets of small duration units, especially triplets, may fall between these sampling points, which shifts their onset slightly in the representation. Structure induction requires a music representation in *onset, pitch* pairs.

In order to deal with transposition differences in folk songs, van Kranenburg et al. (2013) transpose melodies to the same key using pitch histogram intersection. We take a similar approach. For each melody, a pitch histogram is computed with MIDI note numbers as bins, with the count of each note number weighted by its total duration in a melody. The pitch histogram intersection of two histograms h_s and h_t , with shift σ is defined as

$$PHI(h_s, h_t, \sigma) = \sum_{k=1}^r \min(h_{s,k+\sigma}, h_{t,k}), \quad (4.1)$$

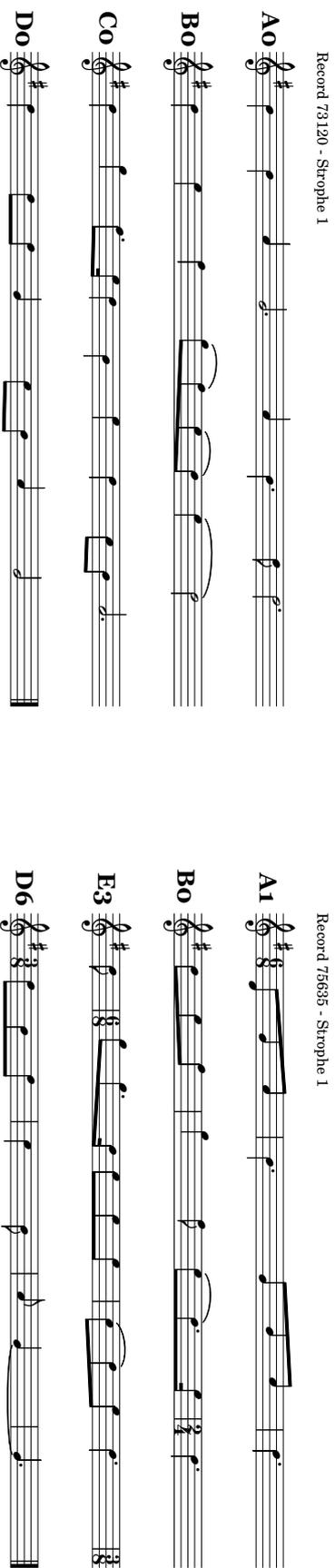


Figure 4.2: An **example for two melodies** from the same tune family with annotations. The first phrase of each melody is labeled with the same letter (A), but different numbers, indicating that the phrases are “related but varied”, the second phrase is labeled B0 in both melodies, indicating that the phrases are “almost identical”.

where k denotes the index of the bin, and r the total number of bins. We define a non-existing bin to have value zero. For each tune family, we randomly pick one reference melody and for each other melody in the tune family we compute the σ that yields a maximum value for the histogram intersection, and transpose that melody by σ semitones. This process results in *pitch-adjusted sequences*.

To test how the choice of reference melody affects the results of pitch histogram intersection, we performed the procedure 100 times, with a randomly picked reference melody per tune family in every iteration. We compare the resulting pitch differences between tune family variants with pitch differences as a result of manually adjusted pitches, available through the MTC-ANN-2.0 dataset. We compare all 2822 pairs of tune family variants. On average, pitch histogram intersection adjusts 93.3% of the melody pairs correctly, so the procedure succeeds in the vast majority of cases. The standard deviation of the success rate is 2.4%, which is low enough to conclude that it does not matter greatly which melody is picked as a reference melody for the pitch histogram intersection procedure.

4.2.1 Similarity Measures Comparing Equal-Length Note Sequences

To describe the following three measures, we refer to two melodic segments q and p of length n , which have elements q_i and p_i . The measures described in this section are distance measures, such that lower values of $dist(q, p)$ indicate higher similarity. Finding an occurrence of a melodic segment within a melody with a fixed-length similarity measure is achieved through the comparison of the query segment against all possible segments of the same length in the melody. The candidate segments with maximal similarity to the query segment are retained as matches, and the positions of these matches within the match melody are saved along with the achieved similarity. The implementation of the fixed-length similarity measures in Python is available online.³ It uses the *spatial.distance* library of *scipy* (Oliphant, 2007).

Scherrer and Scherrer (1971) suggest correlation distance to compare folk song melodies, represented as duration weighted pitch sequences. Correlation distance is independent of the transposition and melodic range of a melody, but in the current music representation, it is affected by time dilation differences.

$$dist(q, p) = 1 - \frac{\sum_{i=1}^n (q_i - \bar{q})(p_i - \bar{p})}{\sqrt{\sum_{i=1}^n (q_i - \bar{q})^2} \sqrt{\sum_{i=1}^n (p_i - \bar{p})^2}} \quad (4.2)$$

Steinbeck (1982) proposes two similarity metrics for the classification of folk song melodies: city-block distance (Equation 4.3) and Euclidean distance (Equation 4.4). He suggests to compare pitch sequences with these similarity measures, next to various other features of melodies such as their range, or the number of notes in a melody (p. 251f.). As we are interested in finding occurrences of segments rather than comparing

³ <https://github.com/BeritJanssen/MelodicOccurrences>

whole melodies, we compare pitch sequences, based on the pitch distances between each note in the sequence.

$$\text{dist}(q, p) = \sum_{i=1}^n |q_i - p_i| \quad (4.3)$$

$$\text{dist}(q, p) = \sqrt{\sum_{i=1}^n (q_i - p_i)^2} \quad (4.4)$$

City-block distance and Euclidean distance are not transposition invariant, but as they are applied to pitch sequences, time dilation differences have minor influence. All the equal-length measures in this section will be influenced by variations introducing more notes into a melodic segment, such as melodic ornamentation. Variable-length similarity measures, discussed in the following section, can deal with such variations more effectively.

4.2.2 Similarity Measures Comparing Variable-Length Note Sequences

To formalize the following two measures, we refer to a melodic segment q of length n and a melody s of length m , with elements q_i and s_j . The measures described in this section are similarity measures, such that higher values of $\text{sim}(q, s)$ indicate higher similarity. The implementation of these methods in Python is available online.³

Mongeau and Sankoff (1990) suggest the use of alignment methods for measuring music similarity, and they have been proven to work well for folk songs (van Kranenburg et al., 2013). We apply local alignment (T. Smith & Waterman, 1981), which returns the similarity of the segments within a given melody which match the query best.

To compute the optimal local alignment, a matrix A is recursively filled according to equation 4.5. The matrix is initialized as $A(i, 0) = 0, i \in \{0, \dots, n\}$, and $A(0, j) = 0, j \in \{0, \dots, m\}$. $W_{\text{insertion}}$ and W_{deletion} define the weights for inserting an element from melody s into segment q , and for deleting an element from segment q , respectively. $\text{subs}(q_i, s_j)$ is the substitution function, which gives a weight depending on the similarity of the notes q_i and s_j .

$$A(i, j) = \max \begin{cases} A(i-1, j-1) + \text{subs}(q_i, s_j) \\ A(i, j-1) + W_{\text{insertion}} \\ A(i-1, j) + W_{\text{deletion}} \\ 0 \end{cases} \quad (4.5)$$

We apply local alignment to pitch adjusted sequences. In this representation, local alignment is not affected by transposition differences, and it should be robust with

respect to time dilation. For the insertion and deletion weights, we use $W_{insertion} = W_{deletion} = -0.5$, and we define the substitution score as

$$subs(q_i, s_j) = \begin{cases} 1 & \text{if } q_i = s_j \\ -1 & \text{otherwise} \end{cases}. \quad (4.6)$$

The insertion and deletion weight are chosen to be equal, and to be smaller than the weight of a substitution with a different pitch; substitution with the same pitch is rewarded. Effectively, this means that the alignment matrix will have non-zero values only if substitutions with the same pitch occur.

The local alignment score is the maximum value in the alignment matrix A . This maximum value can appear in more than one cell of the alignment matrix, due to phrase repetition. This means that several matches can be associated with a given local alignment score. To determine the positions of the matches associated with the maximum alignment score, we register for each cell of the alignment matrix whether its value was caused by insertion, deletion or substitution. We backtrace the alignment from every cell containing the maximal alignment score, which we take as the end position of a match, continuing until encountering a cell containing zero, which is taken as the beginning of a match.

We normalize the maximal alignment score by the number of notes n in the query segment, which gives us the similarity of the detected match with the query segment.

$$sim(q, s) = \frac{1}{n} \max_{i,j} (A(i, j)) \quad (4.7)$$

Structure induction algorithms (Meredith, 2006) formalize a melody as a set of points in a space defined by note onset and pitch, and perform well for musical pattern discovery (Meredith, 2014). They measure the difference between melodic segments through so-called translation vectors. The translation vector \mathbf{T} between points in two melodic segments can be seen as the difference between the points q_i and s_j in onset, pitch space. As such, it is transposition invariant, but will be influenced by time dilation differences.

$$\mathbf{T} = \begin{pmatrix} q_{i,onset} \\ q_{i,pitch} \end{pmatrix} - \begin{pmatrix} s_{j,onset} \\ s_{j,pitch} \end{pmatrix} \quad (4.8)$$

The maximally translatable pattern (MTP) of a translation vector \mathbf{T} for two melodies q and s is then defined as the set of melody points q_i which can be transformed to melody points s_j with the translation vector \mathbf{T} .

$$MTP(q, s, \mathbf{T}) = \{q_i | q_i \in q \wedge q_i + \mathbf{T} \in s\} \quad (4.9)$$

We use the pattern matching method SIAM, defining the similarity of two melodies as the largest set match achievable through translation with any vector, normalized by the length n of the query melody:

$$sim(q, s) = \frac{1}{n} \max_{\mathbf{T}} |MTP(q, s, \mathbf{T})| \quad (4.10)$$

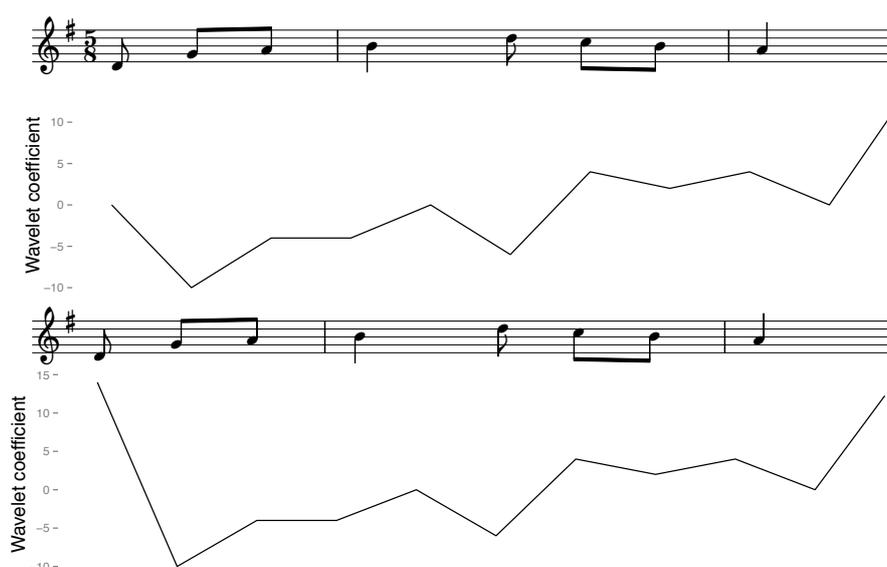


Figure 4.3: The **first two phrases of a melody** from the tune family “Daar ging een heer 1”, with the values of the Haar wavelet coefficient underneath.

The maximally translatable patterns leading to highest similarity are selected as matches, and their positions are determined through checking the onsets of the first and last note of the MTPs.

4.2.3 Similarity Measures Comparing Abstract Representations

Wavelet transform converts a pitch sequence into a more abstract representation prior to comparison. We apply wavelet transform to each query segment q and melody s in the data set prior to searching for matches.

Velarde, Weyde, and Meredith (2013) use wavelet coefficients to compare melodies: melodic segments are transformed with the Haar wavelet, at the scale of quarter notes. The wavelet coefficients indicate whether there is a contour change at a given moment in the melody, and similarity between two melodies is computed through city-block distance of their wavelet coefficients. The method achieved considerable success for pattern discovery (Velarde & Meredith, 2014).

We use the authors’ Matlab implementation to compute wavelet coefficients of duration weighted pitch sequences. An example for an excerpt from a melody and the associated wavelet coefficients can be found in Figure 4.3. In accordance with Velarde and Meredith’s procedure, we use city-block distance to compare wavelet coefficients of query segment and match candidates, retaining similarity and position information of matches as described in Section 4.2.1.

Through the choice of music representation and comparison of the wavelet coefficients, this is an equal-length similarity measure sensitive to time dilation; however, it is transposition invariant.

4.3 EVALUATION

For the evaluation, we distinguish three concepts: *match*, *instance*, and *occurrence*. A *match* is a note sequence in a melody s at which maximum similarity with the query segment q is achieved, as detected by one of the similarity measures. An *occurrence* is a match whose similarity score exceeds a given threshold. An *instance* of a query phrase in a melody is given if the annotators indicate that a query phrase q is found within a given melody s . There can be multiple matches, occurrences and instances of a query phrase in a given melody, due to phrase repetitions.

We evaluate each of the 1890 phrases in the data set as query segments. Using the various similarity measures, we detect for each query segment q , per tune family, its matches in every melody s , excluding the melody from which the query segment was taken. As we are interested in the positions of the matches, we then determine which notes belong to the match. We assign to each note in a melody belonging to a match the similarity score of that match; the other notes receive an arbitrary score which for each measure exceeds the largest (CD, CBD, ED, WT) or smallest (LA, SIAM) similarity values of all matches.

Different thresholds on the similarity measures determine which notes are selected as constituting occurrences. Notes from matches with similarity values below (for the distance measures CD, CBD, ED, and WT) or above (for LA and SIAM) are considered as belonging to occurrences. We vary the similarity threshold for each measure step-wise from the matches' minimum similarity to maximum similarity, and for each step, compare the retained occurrences to the human annotations.

We evaluate the occurrences against the annotations of "almost identical" instances of the query segments in all melodies from the same tune family. As we would like to know which instances of query phrases most annotators agree on, we combine the three annotators' judgements into a *majority vote*: if for a given query segment q in one melody t , two or more annotators agree that a phrase p with exactly the same label (letter and number) appears in another melody s of the same tune family, we consider phrase p 's notes to constitute an instance of query segment q in s .

Conversely, if there is no such phrase in melody s to which two or more annotators have assigned exactly the same label as q , the notes of melody s do not represent any instances of that phrase. This means that the phrases considered "related but varied" are not treated as instances of the query segment for the purpose of this study. The query phrases are compared with a total of 1,264,752 notes, of which 169,615 constitute instances of the query phrases.

All the notes which annotators consider to constitute instances of a query phrase are positive cases (P), all other notes are negative cases (N). The notes that a similarity measure with a given threshold detects as part of an occurrence are the positive predictions (PP), all other notes are negative predictions (NP). We define the intersection of P and PP, i.e., the notes which constitute an occurrence according to both a similarity measure with a given threshold and the majority of the annotators, as true positives (TP). True negatives (TN) are the notes which both annotators and similarity measures

do not find to constitute an occurrence, i.e., the intersection of N and NP. False positives (FP) are defined as the intersection of N and PP, and false negatives (FN) as the intersection of P and NP.

We summarize the relationship between true positives and false positives for each measure in a receiver-operating characteristic (ROC) curve with the threshold as parameter and the axes defined by true positive rate (tpr) and false positive rate (fpr). The greater the area under the ROC curve (AUC), the better positive cases are separable from negative cases.

We would like to know the optimal similarity threshold for each measure, to retrieve as many as possible notes annotated as instances correctly (high recall), and retrieving as few as possible irrelevant notes (high precision). A common approach to strike this balance is the F1-score, the harmonic mean of precision and recall. However, as our data has a strong bias (86.6%) towards negative cases, the F1-score is not an adequate criterion, as it focusses on true positives only. Therefore, we evaluate both positive and negative cases with sensitivity, specificity, positive and negative predictive values, and optimize the similarity threshold with respect to all these values through Matthews' correlation coefficient (Matthews, 1975).

Sensitivity, or recall, is equal to the true positive rate. It is defined as the number of true positives, divided by all positive cases, i.e., the number of notes correctly detected as part of occurrences, divided by all notes considered by annotators to constitute instances of query phrases.

$$SEN = \frac{TP}{P} \quad (4.11)$$

Specificity, or true negative rate, is defined as the number of true negatives, divided by all negative cases, i.e., the number of notes which are correctly labeled as not belonging to an occurrence, divided by all notes considered by annotators to not belong to any occurrences.

$$SPC = \frac{TN}{N} = 1 - fpr \quad (4.12)$$

The positive predictive value, or precision, is defined as the number of true positives, divided by all positive predicted cases, i.e., the number of all relevant notes labelled as part of an occurrence, divided by all notes detected to constitute occurrences by the similarity measure.

$$PPV = \frac{TP}{PP} \quad (4.13)$$

The negative predictive value is defined as the number of true negatives, divided by all negative predicted cases, i.e., the number of notes correctly labelled as not belonging to an occurrence, divided by all notes not constituting an occurrence according to the similarity measure.

$$NPV = \frac{TN}{NP} \quad (4.14)$$

To maximize both true positive and true negative rate, i.e., sensitivity and specificity, their sum should be as large as possible. The same goes for the positive and negative predictive values, the sum of which should be as large as possible. Powers (2007) suggests the measures informedness and markedness, which are zero for random performance, and one for perfect performance:

$$INF = SEN + SPC - 1 \quad (4.15)$$

$$MRK = PPV + NPV - 1 \quad (4.16)$$

Moreover, informedness and markedness are the component regression coefficients of Matthews' correlation coefficient ϕ , which is a good way of describing the overall agreement between a predictor and the ground truth (Powers, 2007). $\phi = 1.0$ for perfect agreement between ground truth and predictors, $\phi = 0.0$ for random performance, and $\phi = -1.0$ if there is a complete disagreement between ground truth and predictors, such that every positive case is a negative prediction, and vice versa.

$$\phi = \sqrt{INF \cdot MRK} \quad (4.17)$$

4.3.1 *Glass ceiling*

As our ground truth is defined as the majority vote of three annotators, we analyze the agreement of the three annotators with the majority vote. This gives us an indication of the "glass ceiling" of the task, or how much agreement with the ground truth is maximally achievable. If the annotators do not perfectly agree on occurrences in our data set, it is not realistic to expect that a similarity measure can achieve perfect agreement with the current ground truth (Flexer & Grill, 2016).

Table 4.2 shows that all annotators show similar agreement (measured by Matthews' correlation coefficient) with the annotators' majority vote. There are individual differences, however: for example, annotator 3 shows lower sensitivity, which is counterbalanced by a higher positive predictive value. This means that this annotator misses some of the occurrences on which the two other annotators agree, but finds almost no spurious occurrences.

The closer the compared similarity measures get to the annotators' agreement with the majority vote of $\phi \simeq 0.86$, the better we take them to be at finding occurrences of melodic segments in folk song melodies.

4.3.2 *Baselines*

Next to the best possible performance, we would like to know what a very naive approach would do, and introduce two baselines: one which considers every note as part

of an occurrence (*always*), leading to perfect sensitivity, and a baseline which considers no note as part of an occurrence (*never*), leading to perfect specificity. The positive predictive value of *always* and the negative predictive value of *never* reflect the aforementioned bias towards negative cases; the respective other predictive values are zero as there are no negative predictions for *always*, and no positive predictions for *never*. As informedness is 0.0 in both cases, Matthews' correlation coefficient also leads to $\phi = 0.0$, meaning both have random agreement with the ground truth.

Annotator	ϕ	SEN	SPC	PPV	NPV
Annotator ₁	0.877	0.900	0.982	0.887	0.985
Annotator ₂	0.865	0.913	0.976	0.855	0.986
Annotator ₃	0.861	0.815	0.993	0.947	0.972
Baseline	ϕ	SEN	SPC	PPV	NPV
<i>always</i>	0.0	1.0	0.0	0.134	0.0
<i>never</i>	0.0	0.0	1.0	0.0	0.866

Table 4.2: The glass ceiling (top), or the annotators' agreement with the majority vote, and the majority vote agreement of the baselines (bottom), assuming every note (*always*) or no note (*never*) to be an occurrence. We report Matthews' correlation coefficient (ϕ) for the overall agreement, and the associated sensitivity (SEN), specificity (SPC), positive and negative predictive values (PPV, NPV)

4.4 COMPARISON OF SIMILARITY MEASURES

Presently, we compare the previously described six similarity measures, applied to the music representations for which they were proposed. The results suggest some answers to our first research question (Q₁), i.e., which of the measures best serves the purpose of finding correct occurrences of melodic segments in folk songs.

4.4.1 Results

Figure 4.4 shows the ROC curves of the six compared measures, which reflect the true positive rate versus the false positive rate of the measures over a range of similarity thresholds. The higher and sharper the "elbow" in the upper left corner, the better a measure can separate between positive and negative cases. Chance level performance would be on the diagonal connecting zero true and false positive rate to full true and false positive rate.

The straightness of the curves on the right is caused by the fact that a considerable amount of the notes annotated as instances are not found by the measures. The ROC

curve interpolates between considering all matches found by a given measure as occurrences, and considering all notes in the data set as constituting occurrences, leading to $tpr = fpr = 1.0$.

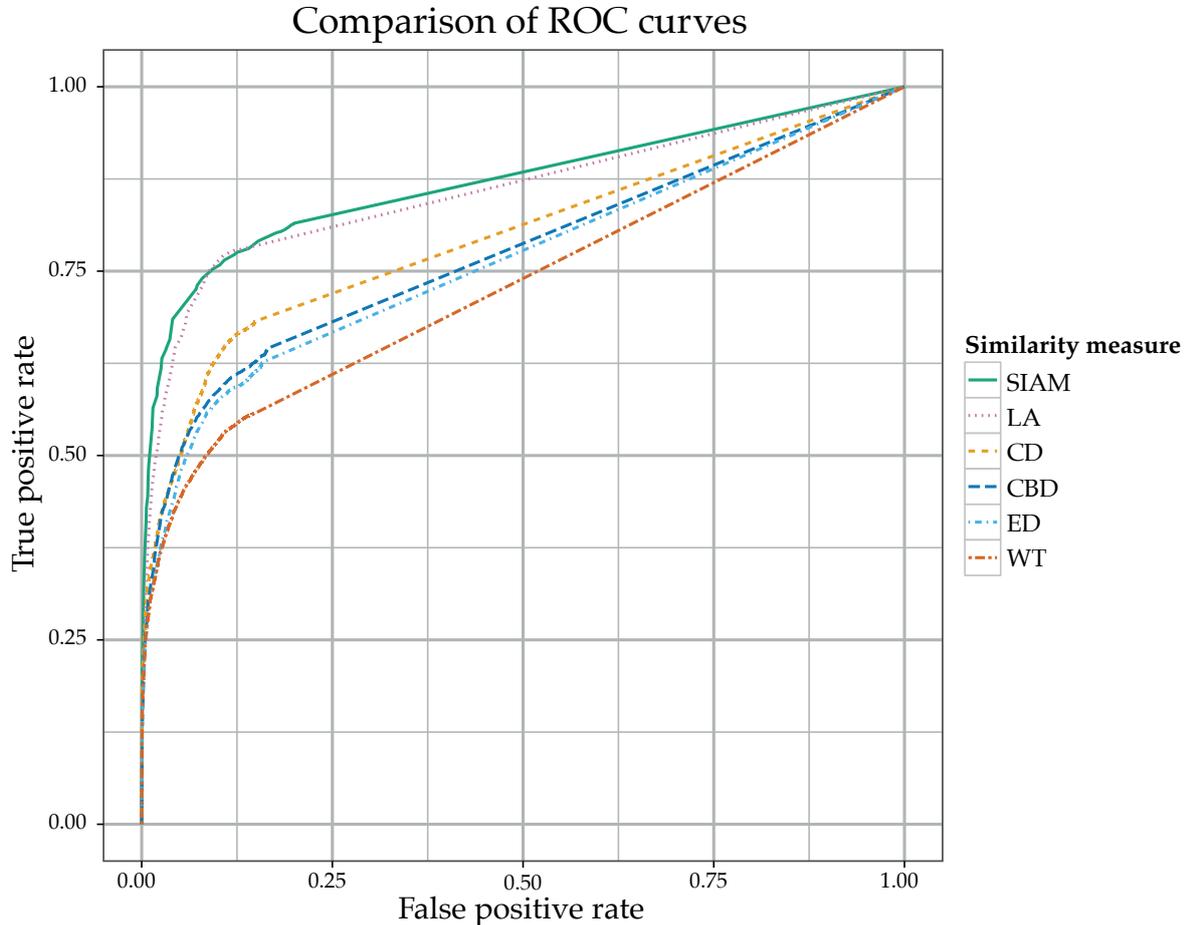


Figure 4.4: The ROC curves for the various similarity measures, showing the increase of false positive rate against the increase of the true positive rate, with the threshold as parameter.

For each measure, we report the area under the ROC curve, to numerically represent the difference between the curves in Figure 4.4. Moreover, we select the similarity threshold which maximizes Matthews' correlation coefficient, and report the associated ϕ , sensitivity, specificity, positive and negative predictive values. These measures are summarized in Table 4.3.

Table 4.3 shows that all of the compared measures agree much better with the ground truth than the baselines (*always* and *never*), but do not reach the level of the annotator agreement with the majority vote (cf. Table 4.2). Of the six measures, wavelet transform (WT) achieves least agreement with the annotators, followed by the distance measures suggested in the field of ethnomusicology (ED, CBD and CD). Local alignment (LA) and structure induction (SIAM) agree best with the majority vote and achieve

Measure	AUC	ϕ	SEN	SPC	PPV	NPV
WT	0.731	0.459	0.367	0.976	0.703	0.909
ED	0.764	0.468	0.482	0.948	0.589	0.922
CBD	0.774	0.499	0.425	0.973	0.708	0.916
CD	0.797	0.503	0.414	0.977	0.732	0.915
LA	0.859	0.621	0.646	0.956	0.695	0.946
SIAM	0.870	0.665	0.632	0.973	0.787	0.945

Table 4.3: Results of the compared similarity measures: area under the ROC curve (AUC), maximal ϕ correlation coefficient with associated sensitivity (SEN), specificity (SPC), positive and negative predictive values (PPV, NPV).

Matthews’ correlation coefficients of around $\phi = 0.621$ and $\phi = 0.665$, respectively. This is still much lower than the annotator agreement, but shows that the measures find most relevant occurrences, while producing less spurious than relevant results.

4.4.2 Discussion

With the present results, the distance measures Euclidean distance and city-block distance (ED, CBD) do not seem to be promising candidates for finding occurrences of melodic segments in melodies. Still, while they do not achieve high agreement as measured in ϕ , they perform widely above the baselines. The relatively higher success of correlation distance (CD) is most likely to be attributed to the more fine-grained music representation in the form of duration weighted pitch sequences, which reflect the duration of the notes.

It is surprising that the performance of wavelet transform (WT) lies below the other compared similarity measures, as in our previous study (Janssen et al., 2015) which evaluated occurrences without taking their positions into account, it performed better than the distance measures. The low sensitivity, mainly responsible for the low maximal ϕ , is caused to a large extent by undetected phrase repetitions. As wavelet coefficients represent contour change in the pitch sequence, identical phrases with the same pitch sequence representation may have different wavelet transforms, depending on notes preceding the first note of a phrase, as illustrated in Figure 4.3. Therefore, in only 10% of the melodies with more than one instance of a given query phrase, wavelet finds more than one occurrence.

Local alignment (LA) and structure induction (SIAM) perform better than the before-mentioned measures. One reason for this might be that they are both variable-length similarity measures, and therefore deal with slight rhythmic variation and ornamentation more effectively. Moreover, both are transposition invariant, local alignment due

to the pitch adjustment performed on the pitch sequence, structure induction due to the fact that it finds transpositions between pitches by definition.

From the present results it is not possible to differentiate whether the best-performing measures do well because their comparison method is effective, or because of the music representations they use. It also seems that duration information might improve performance, as SIAM and CD, with duration information, perform comparatively well. Moreover, in respect to duration, time dilation differences might still affect the results negatively, and a music representation which attempts to correct these differences might improve results of the best measures even further.

The next section therefore compares different music representations for the compared measures, which gives clearer insights as to which of the observed differences in the present comparison are due to the measures themselves, and which differences can be overcome with different music representations. This also allows us to perform another comparison of the similarity measures with optimized music representations.

4.5 DEALING WITH TRANSPOSITION AND TIME DILATION DIFFERENCES

The automatic comparison of folk song melodies is impeded by transposition and time dilation differences of the transcriptions, as illustrated in Figure 4.1. It remains an open question which music representation can best resolve these differences (research question Q2 in the introduction). Therefore, we compare seven different music representations here, applied to each of the similarity measures as appropriate.

4.5.1 *Music representations*

In the previous section, four similarity measures used a pitch sequence (P) as music representation, which does not resolve transposition differences, and does not take the duration of notes into account. To solve the problem of transposition differences, two approaches are conceivable: a music representation consisting of sequences of pitch intervals (PI), i.e., sequences of differences between successive pitches, and pitch adjusted sequences (PA), as described and used for local alignment in the previous section.

With respect to the representation of duration, we have already seen the use of pitch and onset tuples (PO) for structure induction, and duration weighted pitch sequences (DW) for correlation distance and wavelet transform in the previous section. The latter representation can of course also be combined with pitch adjustment, and the resulting representation (PADW) will be compared, too.

To solve the problem of time dilation differences, we test whether time dilation differences can be corrected through automatic comparison of duration value frequencies, analogous to pitch adjustment. To this end, we calculate duration histograms, in which seven duration bins are filled with the count of each duration. Only durations which are in 2:1 integer ratios are considered, as other durations, such as punctuated rhythms or triplets, would not allow easy scaling. The smallest considered duration is a hemidemisemiquaver, or 64th note, and all doublings of this duration are considered

up to a semibreve, or whole note. Analogous to Equation 4.1, we define the duration histogram intersection of two duration histograms h_t and h_s , with a total number of r duration bins k :

$$\text{DHI}(h_t, h_s, \sigma) = \sum_{k=1}^r \min(h_{t,k+\sigma}, h_{s,k}), \quad (4.18)$$

For each tune family, we randomly pick one reference melody and for each other melody in the tune family we compute the shift σ that yields a maximum value for the histogram intersection, and use that σ to calculate the multiplier of the onsets of melody t with relation to melody s :

$$\text{Mult}(t, s) = 2^\sigma \quad (4.19)$$

We also tested the influence of the randomly picked reference melodies on the results of duration histogram intersection by running the procedure 100 times, and comparing with annotated duration adjustments. Of the 2822 pairs of tune family variants, 66.5% were adjusted in the same way as annotated. This means that a third of the pairs are adjusted incorrectly, so it is an open question whether duration adjustment improves results, in spite of its rather high error rate. At any rate, the low standard deviation of 1.3% of the success rate means that it does not matter greatly which melodies are picked as reference melodies.

The result of this procedure leads us to a music representation which is pitch and duration adjusted (DA). We also make use of the metadata of the Annotated Corpus to find out the hand-adjusted (HA) optimal transposition and time dilation of each melody. Hand-adjustment is not feasible for a large collection of folk songs, but is a useful reference for comparison with the automatically adjusted music representations.

Wavelet transform and structure induction (WT, SIAM) are defined for specific representations, namely a duration weighted pitch sequence (DW) and pitch/onset tuples (PO), respectively. As such, not all music representations are applicable for these measures. For WT, only duration weighted pitch sequences and adjustments thereof are tested (DW, PADW, DA, HA). For SIAM, the duration adjustment and hand adjustment (DA, HA) are applied to the pitch/onset tuples, which differs slightly from the DA and HA representations in the other measures, in which duration weighed pitch sequences are adjusted.

4.5.2 Results

From Figure 4.5 it can be seen that music representation has considerable influence on the success of the similarity measures. Overall, most music representations show better performance than the pitch sequence representation (P). The only exception is the pitch interval representation (PI): attempting to resolve transposition differences between songs through pitch intervals deteriorates performance.

Duration information (DW) improves the performance of some distance measures and local alignment (LA, CD, CBD, ED), as does pitch adjustment (PA). A combination

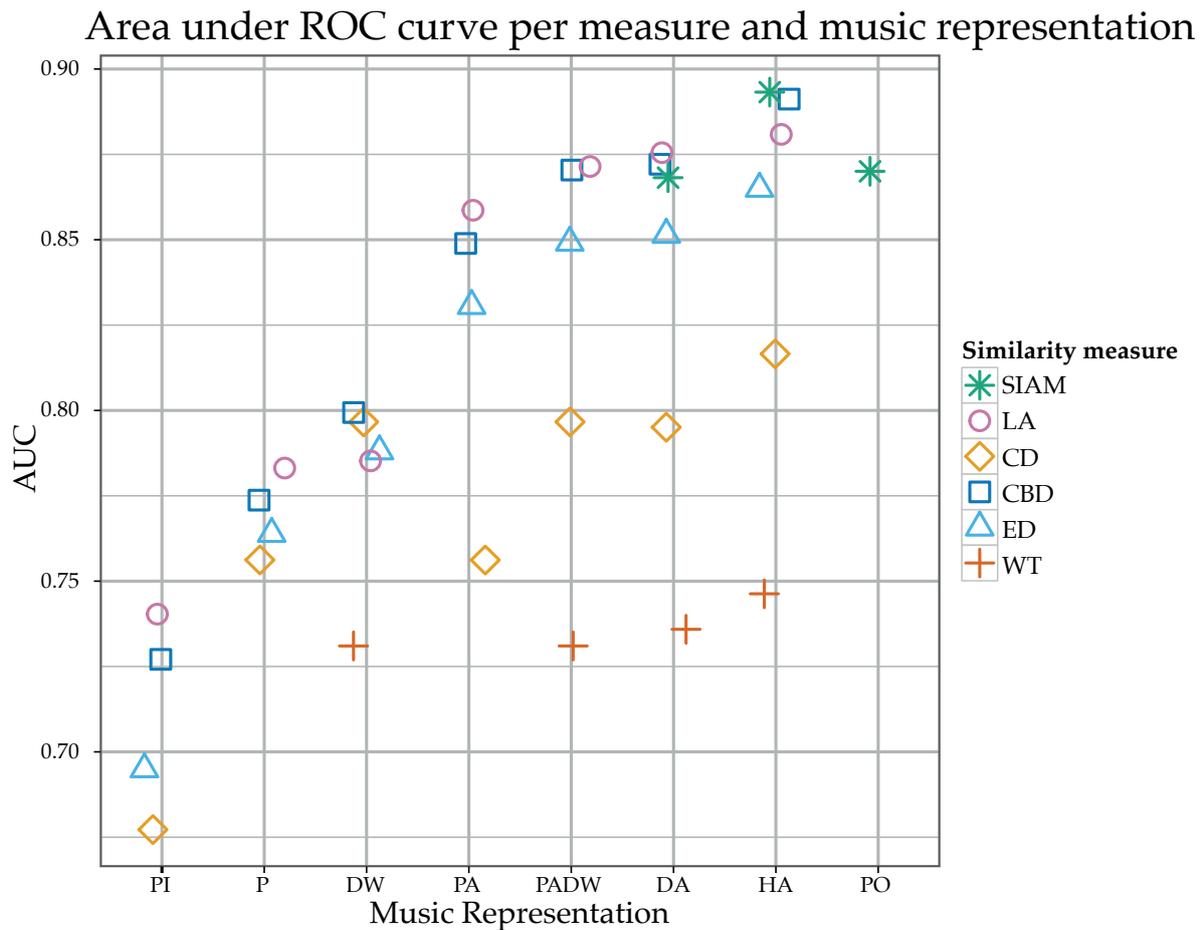


Figure 4.5: The Area under the ROC curves (AUC) of the similarity measures for different music representations: pitch interval (PI), pitch (P), duration weighted (DW), pitch adjusted (PA), pitch adjusted and duration weighted (PADW), metrically adjusted (DA), hand-adjusted (HA), and pitch/onset (PO). For wavelet transform (WT) and structure induction (SIAM), not all music representations are applicable, and only SIAM uses the pitch/onset representation.

of the two (PADW) improves these measures even further. Duration adjustment (DA) of the duration weighted sequences gives a slight advantage for some measures (CBD, LA), but does not seem to affect the other measures much (ED, CD, WT, SIAM).

The difference with the hand-adjusted (HA) representation, resulting in the best performance for all measures, shows that automatic adjustment is not completely able to resolve transposition and time dilation differences. A full overview of all music representations and measures, with the resulting AUC as well as maximal ϕ with associated retrieval measures can be found in Table B.2 in the Appendix.

Figure 4.6 shows another comparison of ROC curves for the six similarity measures, with optimized music representations. We pick for each measure the music representation which results in the highest AUC. As we could not improve some measures (CD,

SIAM) through other music representations, their curves are identical to those in Figure 4.4. We find that a number of measures (ED-DA, CBD-DA) perform much better than before as a result of the corrections for transposition and time dilation differences. Local alignment (LA-DA) and city-block distance (CBD-DA) even outperform SIAM with these adjustments.

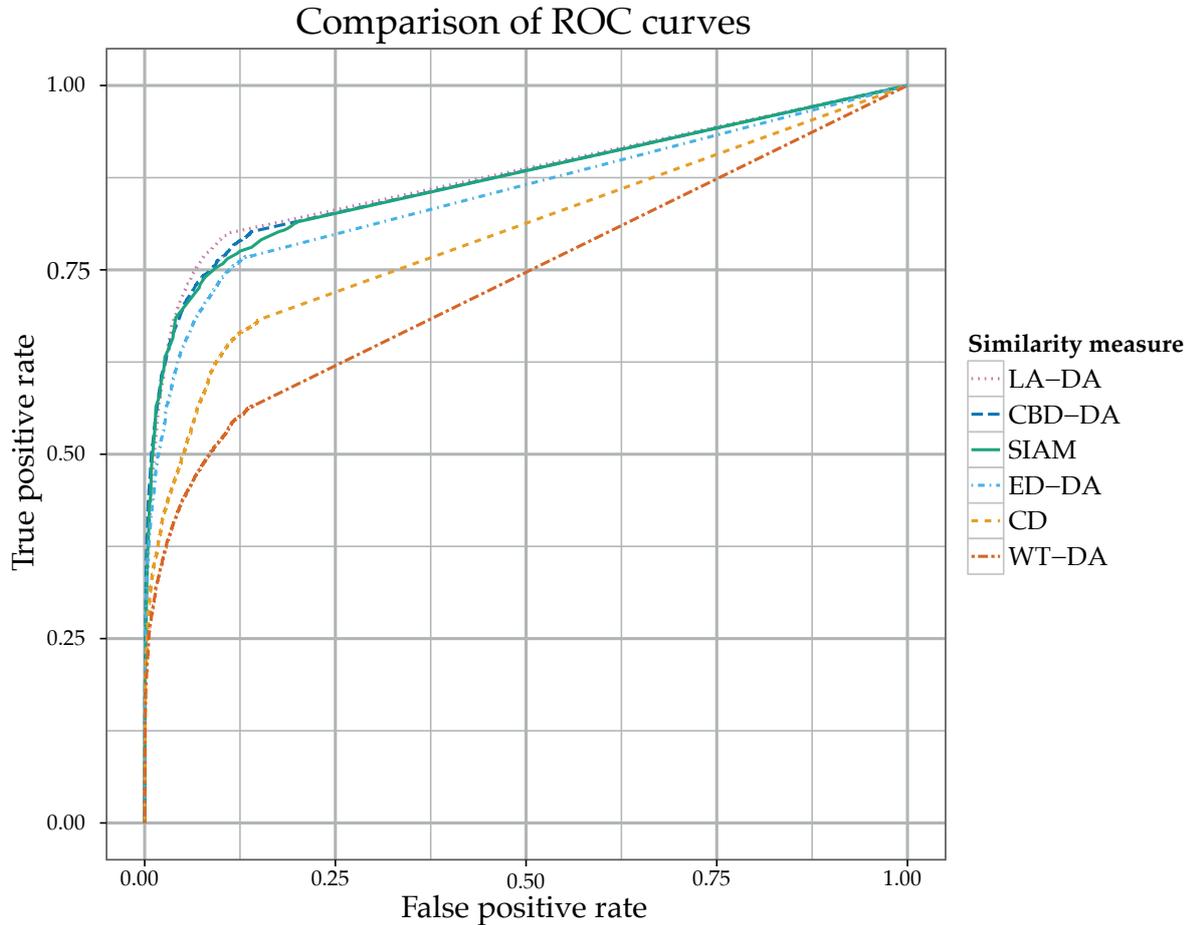


Figure 4.6: The ROC curves for the various similarity measures with optimized music representations, showing the increase of false positive rate against the increase of the true positive rate, with the threshold as parameter.

In Table 4.4, we report the area under the ROC curve for all measures with optimized music representations, as well as the maximized ϕ correlation coefficient with associated sensitivity, specificity, positive and negative predictive values.

With optimized music representation, local alignment and city-block distance achieve values for ϕ close to that of structure induction (SIAM). The differences among these three measures can mainly be found in their sensitivity and positive predictive value, as SIAM and CBD-DA achieve lower sensitivity than LA-DA, but compensate by higher positive predictive values.

Measure	AUC	ϕ	SEN	SPC	PPV	NPV
WT-DA	0.736	0.454	0.320	0.985	0.772	0.903
CD	0.797	0.503	0.414	0.977	0.732	0.915
ED-DA	0.851	0.610	0.627	0.957	0.693	0.943
SIAM	0.870	0.665	0.632	0.973	0.787	0.945
CBD-DA	0.872	0.663	0.608	0.978	0.808	0.942
LA-DA	0.875	0.668	0.675	0.965	0.748	0.950

Table 4.4: Results of the similarity measures with optimized music representations: area under the ROC curve (AUC), maximal ϕ correlation coefficient with associated sensitivity (SEN), specificity (SPC), positive and negative predictive values (PPV, NPV).

Euclidean distance is also improved considerably through duration and pitch adjustment; however, its ϕ is somewhat lower than that of the aforementioned measures. Correlation distance and wavelet transform could not be much improved through any of the tested music representations, and remain at relatively low ϕ values.

4.5.3 Discussion

The present section shows that transposition and time dilation differences have considerable influence on the results of several of the compared measures (CBD, ED, LA). We conclude that the relative success of local alignment in the previous section was caused by its pitch adjusted music representation, and that city-block distance and Euclidean distance perform much better on a pitch adjusted representation, too. However, local alignment achieves slightly higher AUC than the distance measures for each representation, showing that it is the most effective overall.

As wavelet transform, correlation distance and structure induction (WT, CD, SIAM) are already defined as transposition invariant, they cannot be improved through pitch adjustment. Wavelet transform is improved through duration adjustment to some extent. The similarity threshold associated with maximal agreement ϕ is stricter for the duration adjusted case, i.e., fewer matches are considered occurrences, accounting for higher positive predictive value but lower sensitivity (cf. Table B.2). This leads us to the conclusion that the drawback of wavelet transform observed in the previous section, i.e., that it may miss phrase repetitions within a melody, cannot be resolved through our strategy for duration adjustment.

Correlation distance and structure induction perform slightly worse with duration adjustment as compared to their original music representation (cf. Table B.2). For both measures, the similarity threshold associated with maximal agreement ϕ is not affected by duration adjustment. Duration adjustment increases the number of occurrences for both measures. As some of these occurrences are true positives, this leads to higher

sensitivity. Inversely, we have seen that about a third of the automatic adjustments are incorrect, and these mis-adjustments produce false positives, decreasing the positive predictive value.

In summary, we can observe that transposition differences can be adequately resolved through pitch histogram intersection, while a better way of adjusting duration is needed, as the present approach of duration histogram intersection leads to many errors, and improves the performance only slightly or even not at all.

Based on our comparison of similarity measures with optimized music representations, city-block distance and local alignment with pitch and duration adjustment, and structure induction (CBD-DA, LA-DA, SIAM) are the best approaches to finding occurrences of melodic segments in folk song melodies. None of them reach the level of agreement with the majority vote as the human annotators (cf. Table 4.2), however.

This leads to the question whether a combination of the best-performing measures might show better performance than the individual measures. This question will be investigated in the following section.

4.6 COMBINATION OF THE BEST-PERFORMING MEASURES

We combine the three best-performing measures (CBD-DA, LA-DA, SIAM), observing whether this combination improves performance, addressing Q₃ from the introduction.

4.6.1 Method

For each measure, we retain only those matches which exceed the best similarity threshold, obtained from optimization with respect to ϕ . For CBD-DA, matches with $\text{dist}(q, p) \leq 0.98$, for LA-DA, matches with $\text{sim}(q, s) \geq 0.55$, and for SIAM, matches with $\text{sim}(q, s) \geq 0.58$ are retained.

We combine the three best similarity measures in the same way as we combine the annotators' judgements to a majority vote. To this end, we redefine the notion of occurrence: we consider only those notes to constitute an occurrence which two or more measures detect as part of a match, given the respective measures' optimal similarity thresholds. We investigate how well this *combined measure* agrees with the annotators' majority vote.

4.6.2 Results

Table 4.5 presents Matthews' correlation coefficient, sensitivity, specificity, positive and negative predictive value of the combined measure. The agreement $\phi = 0.703$ is higher than that of the individual measures, and outperforms the hand-adjusted music representations of all individual measures.

ϕ	SEN	SPC	PPV	NPV
0.703	0.648	0.981	0.84	0.947

Table 4.5: Results of a combined similarity measure from SIAM, CBD-DA and LA-DA, represented by the maximal ϕ correlation coefficient with associated sensitivity (SEN), specificity (SPC), positive and negative predictive values (PPV, NPV).

4.6.3 Discussion

The combined measure's increased performance is mainly caused by its positive predictive value ($PPV = 0.84$), which is considerably higher than the values achieved by any individual measure, and close to the values of two of the annotators. The sensitivity $SEN = 0.648$ is comparable to that of the individual measures, so it is still a lot lower than the annotators' sensitivity, meaning that the combined measure still misses more instances of melodic segments than human experts.

Based on our study, we find that the combined measure is the best currently achievable method for detecting occurrences of melodic segments automatically. However, we assume the same optimal threshold of the individual similarity measures over the whole data set. This would be inappropriate if there were subgroups of the tested melodies which necessitate higher or lower thresholds to achieve optimal agreement with the annotations. Moreover, the agreement is also likely to vary in different subgroups of melodies, leading to different error rates, depending on the selection of melodies tested.

Therefore, in the next section we proceed to test how leaving out tune families from the data set affects the optimized similarity threshold of the three best-performing measures, and how much the agreement with the ground truth varies depending on the evaluated tune family.

4.7 OPTIMIZATION AND PERFORMANCE OF SIMILARITY MEASURES FOR DATA SUBSETS

In the present section, we investigate whether subgroups of our data set affect the optimized threshold of the three best-performing similarity measures (LA-DA, CBD-DA and SIAM) to such an extent that it is inappropriate to assume one optimal threshold for the whole data set. Moreover, we observe the variation of the agreement ϕ with the ground truth, depending on the evaluated subset. This analysis addresses research question Q4 from the introduction.

As the tune families form relatively homogenous subgroups of melodies within the Annotated Corpus, we use the 26 tune families as subsets. This has the disadvantage that the subsets have different sizes, but the advantage of knowing a priori that the subsets are different by human definition.

4.7.1 Method

For each of the 26 tune families, we optimize the similarity threshold for each measure, leaving that tune family out of the data set. For this “leave one tune family out” procedure, we remove the matches from the tune family under consideration from the data set, and vary the similarity threshold in this reduced data set, selecting the threshold that maximizes Matthews’ correlation coefficient ϕ with the ground truth.

Next, we use this “leave one tune family out” optimized threshold to detect occurrences in the considered tune family, and observe the resulting agreement (ϕ_{tf}) with the ground truth of this tune family. This gives us a different value ϕ_{tf} for the 26 tune families. Ideally, we would like ϕ_{tf} to be high on average, and show small variance.

For comparison of the optimized thresholds $thres$ after leaving out one tune family, we standardize them, using the arithmetic mean and standard deviation of all similarity scores produced by a given measure.

$$thres_{std} = \frac{thres - \overline{sim}}{SD(sim)} \quad (4.20)$$

As a result, the standardized threshold $thres_{std}$ is mapped into a space centered on 0, representing the average similarity score, and in which each unit represents one standard deviation of the similarity scores $SD(sim)$. As city-block distance has similarity values ranging from $0 \leq dist \leq 5.29$, while local alignment and structure induction are bounded by the interval $sim = (0, 1]$, the standardization allows better interpretation of the differences between optimized thresholds.

To compare the variation in agreement ϕ_{tf} of the individual measures, the combined measure, and the annotators with the ground truth, we use a Tukey box and whiskers plot (Tukey, 1977), in which the median is indicated by a horizontal line, and the first (25%) and third (75%) quartile of the data by the horizontal edges of the box. All data exceeding the first and third quartile by no more than 1.5 times the inter-quartile-range are represented by vertical lines. All data outside this range are considered outliers and plotted as individual dots.

4.7.2 Similarity thresholds

The thresholds vary very little when specific tune families are left out of the optimization procedure: most “leave one tune family out” optimizations result in the same optimal threshold as the optimizations on the full data set in the previous section, indicated by black stripes in Figure 4.7. There are some minor deviations, but none larger than 0.3 standard deviations, noticeable in SIAM’s thresholds.

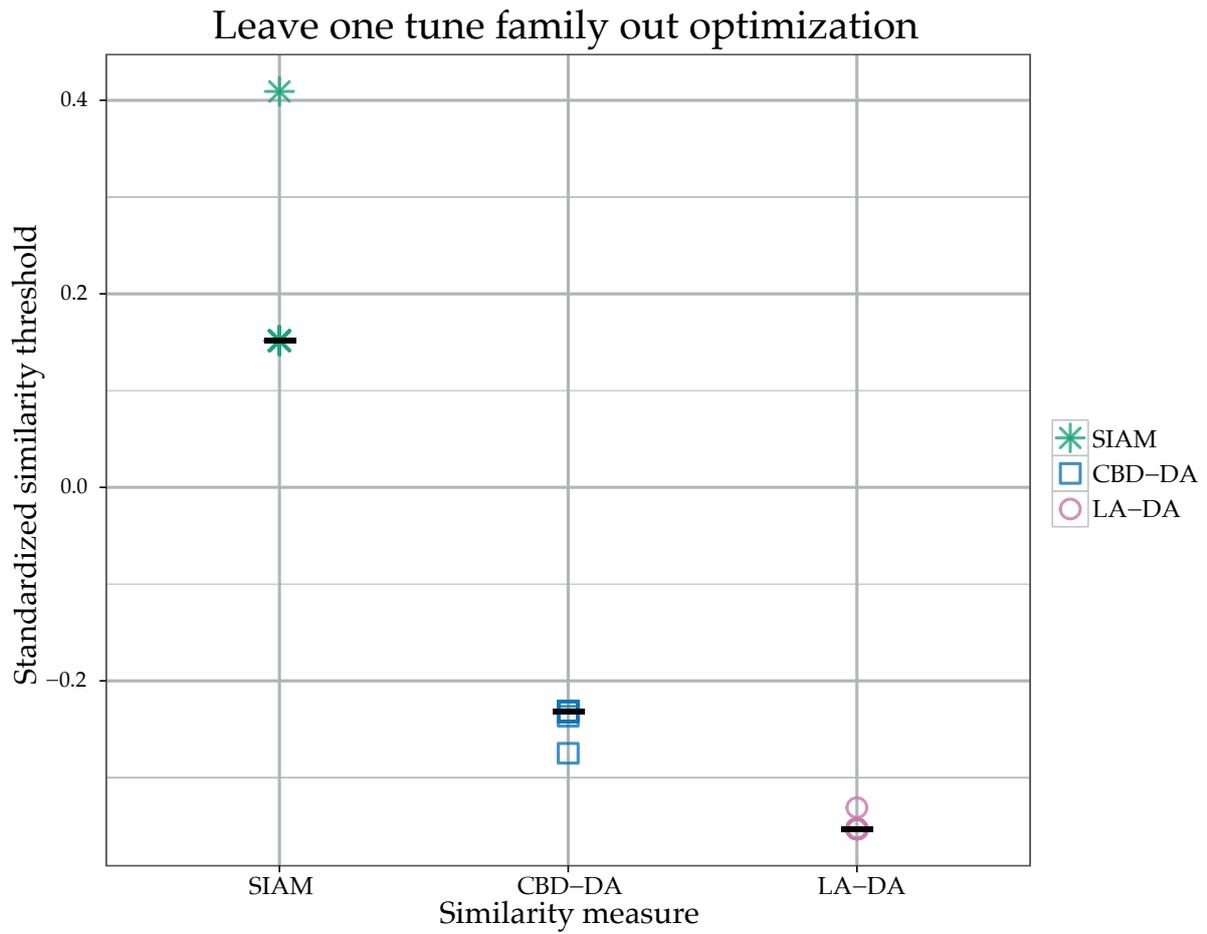


Figure 4.7: The thresholds resulting from “leave one tune family out” optimization. The black stripes indicate the threshold of the optimization of the full data set. All of the measures’ thresholds are close to each other.

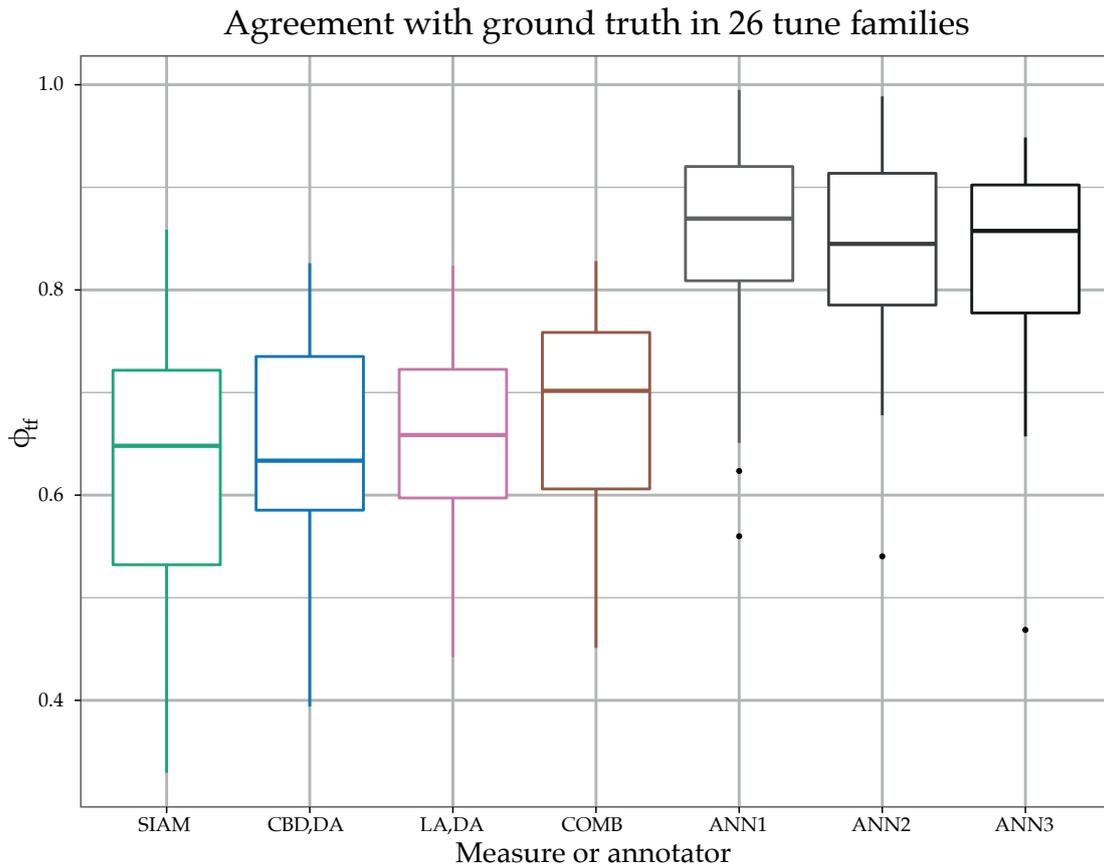


Figure 4.8: The agreement of the three similarity measures and the annotators with the majority vote, evaluated separately for each tune family (in ϕ_{tf}). The similarity measures show more variation than the annotators, even though there are also some remarkable low outliers for the annotators.

4.7.3 Agreement with ground truth

The agreement with the ground truth, measured in the tune-family dependent Matthews' correlation coefficient ϕ_{tf} , depends greatly on the considered tune family, as can be seen in Figure 4.8. This is true especially for the similarity measures SIAM and CBD-DA, which result in a wide range of values for ϕ_{tf} , while LA-DA shows less variation in ϕ_{tf} .

The combined measure (COMB) achieves consistently higher agreement with the ground truth than the measures of which it is composed, as can be seen in its higher mean (indicated by a horizontal line in the box plot), though its variation between $0.45 < \phi_{tf} < 0.83$, depending on the evaluated tune family, is considerable.

The annotators are more consistent than the similarity measures overall, but there are some remarkable outliers for all of them, some as low as $\phi_{tf} = 0.47$, which is comparable to some of poorest algorithmic results.

4.7.4 Discussion

The thresholds vary little when leaving out tune families from the optimization procedure (cf. Figure 4.7), indicating that it is reasonable to assume the same optimal similarity threshold throughout the whole data set. This means that the combined measure can also be applied with one similarity threshold per measure to the whole data set.

The variation in agreement when evaluated against the tune families separately (cf. Figure 4.8) indicates that SIAM and CBD-DA are less robust towards differences between tune families than LA-DA and the combined measure.

Less variation in ϕ_{if} means that a measure is more consistent with respect to the number of errors it produces, regardless of the tune family under consideration. Neither any of the individual measures, nor the combined measure shows enough consistency that a computational folk song study using them should consider the error constant over all subsets of a folk song corpus.

As the annotators also show considerable variation in their agreement with the majority vote, it is unlikely that a computational method can find occurrences in this folk song corpus without producing variable amounts of errors, depending on the evaluated tune family.

4.8 CONCLUSION

We have investigated the success of six similarity measures at finding occurrences of melodic segments in folk songs. We tested how well the similarity measures would find occurrences of phrases, evaluating their results against the majority vote of three annotators' judgements of phrase occurrences in Dutch folk songs. We summarize the answers to the four research questions posed in the introduction, and conclude with some steps for future work.

Regarding the question of which similarity measure is best suited for finding occurrences (Q1), our results of Section 4.4 indicate that structure induction and local alignment are the most successful approaches for this task given the music representation for which they were defined. However, when duration as well as pitch information is supplied, and time dilation and transposition are corrected, city-block distance performs even slightly better than structure induction, and the results of local alignment can be improved, as shown in Section 4.5.

We show that the performance of all similarity measures can be improved when time dilation and transposition differences between folk songs are adjusted (Q2, Section 4.5). The best way to adjust pitch differences automatically is histogram intersection, leading to much improved results. Providing information on the duration as well as pitch of compared notes improves the success of all measures considerably, but time dilation differences remain a problem. Our approach to adjust durations automatically through histogram intersection led to slight improvement for some measures, but no improvement for others.

A combination of the best-performing measures (SIAM, CBD-DA, LA-DA) does indeed perform better than each measure individually (Q3), and is the best measure arising from our comparison. It produces about 16% spurious results, close to the values of human annotators. However, the combined measure misses about a third of the relevant instances of query segments, whereas the annotators miss only around 10%. In consequence, the combined measure is not a replacement for human judgements on melodic occurrences, but to our knowledge produces the best results with the current similarity measures and music representations.

In Section 4.7, we show that the agreement of the three best-performing similarity measures with the ground truth differs depending on the evaluated tune family (Q4). However, we also show that human annotators show almost as much variation. Our optimization of the similarity threshold on subsets of the full data set also leads to almost no change in the selected similarity thresholds of SIAM, CBD-DA and LA-DA, meaning that it is appropriate to assume the same threshold for the whole data set. Yet in statistical analyses of occurrences detected by these measures or the combined measure, it would be inappropriate to assume the same error rate throughout the whole data set. When categories within a music collection are defined, as is the case with tune families in the Meertens Tune Collections, it is therefore advisable to make use of these categories and to assume different error terms for each of them.

Further research into alternative similarity measures and better ways of representing musical information is needed to improve the success of computational detection of melodic occurrences. Our research on music representation indicates that better methods to adjust time dilation differences will lead to much improved results. Moreover, other weighting schemes for local alignment still need to be explored. Another area of improvement is the combination of the judgements from different similarity measures into one combined measure, for which more successful ways than the currently used majority vote approach may be found.

The annotations used in this study distinguish between two levels of instances, those which are “related but varied” and those which are “almost identical”. We have focussed on the latter category in the current study; it would be interesting to see whether the best-performing similarity measures of this study and their combination would also work best for the “related but varied” category, and if so, in how much the optimal similarity thresholds would be affected.

It is also important to validate our findings in different folk song corpora, and in different genres. Unfortunately, no comparable annotations on occurrences in folk songs exist to our knowledge. Annotations in works of Classical music, used as validation sets for pattern discovery, might be an interesting ground of comparison. More annotation data and comparative research is needed to overcome some of the challenges we have presented in finding occurrences of melodic segments in folk songs, and in melodies from other genres, and to ascertain the robustness of computational methods.

Part II

PREDICTING STABILITY

This part shows how stability and variation in folk song melodies may be predicted.

PREDICTING STABILITY IN FOLK SONG TRANSMISSION

Songs and instrumental pieces in a musical tradition are subject to change: as they are adopted by a new generation of listeners and musicians, they evolve into something new while retaining some of their original characteristics. This chapter investigates to what extent this change of melodies may be explained by hypotheses based on the memorability of melodies.

To address this question, we investigate a corpus of folk songs collected in the second half of the twentieth century, in which we can identify groups of variants. The variants are results of real-life melody transmission, something which would be difficult to study in an experimental setting, but for which the present folk song collection possesses high ecological validity. In folk song research, there is a long-standing interest in those melodic segments which resist change during melody transmission. This resistance to change is also referred to as *stability* (Bronson, 1951).

According to models of cultural evolution, the relative frequency of cultural artefacts can be explained based on *drift* alone: certain phrases might have been copied more frequently than others purely based on chance, and the relative stability of a given phrase in a collection of folk songs would be random (Henrich & Boyd, 2002). We hypothesize, instead, that stability can be predicted through the memorability of melodies.

To quantify stability, or the amount of variation a folk song segment undergoes through oral transmission, we follow Bronson's notion that "there is probably no more objective test of stability than frequency of occurrence." (Bronson, 1951, p. 51). We formalize the relative stability of a melodic segment as its frequency of occurrence across variants of the same folk song. We focus on melodic phrases from the folk songs and employ pattern matching to determine whether or not a match for a given phrase may be found in a given folk song variant, based on similarity measures tested in Music Information Retrieval, and evaluated on a subset of folk songs in [Chapter 4](#). We then test whether there is a statistical relationship between a given phrase's matches in variants, and the same phrase's memorability, i.e., properties which might facilitate its recall.

5.1 HYPOTHESIZED PREDICTORS FOR STABILITY

This section gives an overview of literature on which we base our five hypotheses on the stability of folk song phrases. Our first hypothesis states that *phrase length* predicts the variation of a folk song phrase, supported by evidence from serial recall experiments. Our second hypothesis states that the number of *repetitions* of a phrase in a folk song melody predicts its variation, also supported by evidence from serial recall studies. Our third hypothesis states the *phrase position* predicts the variation of a folk song phrase, supported by evidence from computational musicology studies, artificial

transmission chains and serial recall experiments. Our fourth hypothesis states that *predictability* predicts the variation of a folk song phrase, a hypothesis based on concepts in music theory, and supported by evidence from music cognition. Our fifth hypothesis states that *motif repetitivity* predicts the variation of a folk song phrase, supported by evidence from musical corpus analysis.

5.1.1 *Phrase length*

It is reasonable to expect that the length of a phrase influences how much it will vary in transmission: a phrase with many notes contains more items that need to be correctly reproduced, and will therefore be harder to remember than a phrase with few notes. While this notion has not yet been experimentally tested for the recall of melodies, there is supporting evidence from serial recall experiments. Serial recall experiments typically test how well participants in studies remember word lists – presented visually or auditorily – or purely visual or spatial cues. Such recall experiments with lists of different lengths have shown that increasing the length of a memorized list decreases the proportion of correctly recalled items (Ward, 2002). This leads me to the hypothesis that shorter phrases will be more stable in music transmission. This hypothesis does not take into account that the memory load of long phrases may be still reduced by chunking (Miller, 1956), which corresponds to the fifth hypothesis, that motif repetitivity influences phrase variation.

5.1.2 *Phrase repetition*

Moreover, rehearsal in the form of phrase repetitions might play a role: a phrase that is repeated several times within a melody might be memorized more faithfully than a phrase that only occurs once in each verse. The repetition can be considered rehearsal, which has been shown to increase retention of items (Murdock & Metcalfe, 1978).

5.1.3 *Phrase position*

Besides, the position of a melodic phrase within a piece might influence its memorability: in serial recall experiments, these effects are known as *serial position effects* (Deese & Kaufman, 1957). When the start of lists is remembered better, this is considered a *primacy effect* (Murdock, 1962). When words were presented auditorily, Crowder and Morton (1969) found that the end of lists were remembered better, which might lead one to expect that melodies, also auditory in nature, exhibit a *recency effect*. The studies by Bronson (1951) and Louhivuori (1990) on comparisons of folk song variants (see [Chapter 2](#)) suggest that both the first few and the last few notes of a phrase were most stable in tune families, so potentially both primacy and recency effects play a role at the same time. However, it is hard to estimate whether effects found for notes would also hold for whole melodic phrases.

Rubin's (1977) experiments on long-term retention of well-known spoken word passages (the Preamble to the constitution of the United States, Psalm 23, and Hamlet's monologue from the eponymous Shakespeare play), are maybe closest to the situation we are interested in, namely the recitation of folk song phrases from memory. According to Rubin (1977), words at the start of spoken word passages are recalled better than items in the middle or at the end. Therefore, we assume that phrases at the start of melodies may also be more stable. Of course, serial position effects may be caused by an individual's more frequent exposure to items early or late in a melody (Ward, 2002), in which case we would expect that rehearsal is more important than serial position to explain the stability of melodic segments.

5.1.4 *Melodic expectancy*

Another set of theories is related to expectancy in melodies: according to Kleeman's (1985) discussion of selection criteria for music transmission, only meaningful music, and hence, only music which can be processed by listeners based on their musical expectations, will be selected for transmission (p. 17).

In this vein, Schmuckler (1997) found a relationship between expectancy ratings and melody recall in an experimental study on folk song melodies. To this end, 16 participants were instructed to rate how well artificial variants of 14 folk songs confirmed their expectancy. The variants of the folk songs were generated by scrambling the notes at the end of each song, maintaining the rhythmical structure and the end note. Afterwards, participants had to identify the melodies they had encountered in the first part of the experiment, presented along with previously unheard melodies. The hit rates were positively correlated with the expectancy rating, indicating that those melodies which conform best to melodic expectations of listeners are also most reliably recalled.

An alternative prediction would be that more unexpected items will actually be easier to remember. This is corroborated by evidence from free recall, where items which are unusual are usually better remembered (von Restorff, 1933). For music, Müllensiefen and Halpern (2014) found that memorability of melodies was increased if they contained a large amount of unique motifs, i.e., melodic material which is unusual and therefore unexpected. This means that expectancy may influence variation of melodies in opposing ways, which we both adopt as hypotheses (see hypotheses 4a and 4b in the list of hypotheses below).

Meyer (1956) was one of the first researchers who linked melodic expectancy to music perception, to explain aesthetic and emotional responses to music. According to Meyer, expectancy may be caused by general tendencies of perception, as also formulated in *Gestalt* theory, for instance the notion that entities which are close to each other will be perceived as connected (p. 86ff.). This inspired his student Narmour (1990) to postulate the *implication-realization* theory, in which the distance, or *pitch interval*, between two notes implies the direction and size of the next pitch interval. According to Narmour, the different possibilities of the expectancy generated by a given pitch inter-

val, and its realization or violation in the ensuing pitch interval, can be summarized in eight categories, shown in Figure 5.1. He uses these categories to explain the aesthetic impact of well-known examples from Western art music.

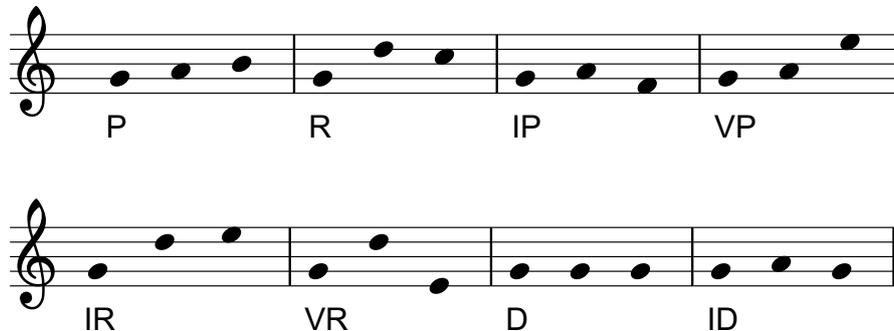


Figure 5.1: The eight basic structures generating melodic expectations in music according to Narmour's (1990) implication-realization theory.

Narmour's implication-realization theory does not make any quantifiable predictions about how expected or surprising a given note is; this was the incentive for Schellenberg (1996) to formalize Narmour's model, such that for a given implicative and realized pitch interval, there is a value predicting the expectancy associated with the realization. Schellenberg summarized Narmour's theory in five principles – registral direction, intervallic difference, registral return, proximity and closure – and found that they corresponded well with experimental data on listener expectancies (Schellenberg, 1996). A principal-components analysis revealed that the model could be reduced even further to two factors, *pitch-proximity* and *pitch-reversal*, without significant loss in explanatory power (Schellenberg, 1997).

Another theory of Meyer states that expectancy is generated by learned probabilities of given events: listeners expect musical events they have heard frequently before, and will be surprised by musical events they hear for the first time (Meyer, 1957). Conklin and Witten (1995) adopted this theory for their own expectancy model, based on frequency of musical events and their successions, as observed in a given music collection. For instance, if the model were to learn from Mozart's variants of *A vous dirais je maman*, it could be used to predict melodies of children's songs, and given the notes accompanying "Twinkle, twinkle", it could with high probability predict the notes completing the melody of *Twinkle, twinkle, little star*. Conklin and Witten furthermore combine short and long term models, i.e., one model which is trained on the frequencies of events in a music collection of many pieces, meant to model a listener's expectations based on lifelong exposure to music (long term model), and one which is trained incrementally on the piece in which prediction is going to take place, meant to model expectations formed during listening to a new piece (short term model). A modified implementation of Conklin and Witten's model by Pearce and Wiggins (2004), IDyOM (Information Dynamics of Music), is publicly available and can therefore be used to quantify expectancy in melodies.

Next to expectancy, one might also hypothesize that participants in a musical tradition make an aesthetic choice on which musical pieces they will pass on. However, there is some evidence that preference for music is also related to frequent listening, the so-called *mere exposure effect* (Zajonc, 1980): listeners will find a musical piece they have heard multiple times most pleasant, even if they are not aware that they have heard it before (Szpunar, Schellenberg, & Pliner, 2004). This means that aesthetic selection might also be implicitly informed by previous exposure, which would make it highly correlated with expectancy. In fact, Huron (2007) postulates that the pleasure we experience when listening to music is caused by expectations, fulfilled or unfulfilled.

5.1.5 Repeating motifs

Müllensiefen and Halpern (2014) investigated a large number of musical features derived from music notation of 80 Western pop songs, to see which of them would best predict the memorability of 80 pop song excerpts. The memorability was determined in a recall experiment with 34 participants, who listened to half of the excerpts. After, the participants were presented with all 80 excerpts, having to indicate whether they had heard a given melody before, and how pleasant they considered it. The researchers considered responses on the pleasantness to represent *implicit* memory for the music, through the above-mentioned mere exposure effect.

The ratings of explicit and implicit memory were then related to musical features. Next to some global aspects of music, these mainly consisted of values based on the frequency of short note sequences, which they referred to as *M-types*. They observed how often M-types occurred in a melodic excerpt in isolation, and how unique their occurrence was, compared to their frequency in a background corpus of different pop song melodies. They called the musical aspects of a given musical piece its *first-order features*, and the musical aspects arising from comparison against the background corpus *second-order features*.

Müllensiefen and Halpern's results indicate that a melody is more easily remembered explicitly if it consists of motifs which are unusual when compared to other songs, but are repeated within the melody. Unique motifs are also an important condition for implicit memory of melodies, even though then the motifs should not repeat too much within a given melody, and their pitch intervals should be small, their contour simple, while there should be a wide range of note durations. The different results for explicit and implicit memory are puzzling, but motifs seem to play an important role for both.

That global musical aspects such as contour and interval size of a melody can successfully predict implicit memory points to another interpretation of the results: explicit and implicit memory may be related to schematic and veridical expectations (Huron, 2007). Listeners build up schematic expectations while listening to a musical piece: for example, if a given motif has been repeated several times before, a listener might expect it to occur again. Veridical expectations, on the other hand, are rule-like expectations formed through lifelong exposure to music. Based on veridical expectations, listeners

may expect that there are no big interval leaps in melodies, and that their contours tend to be arch-formed (Huron, 1996). So if listeners find melodies which have small pitch intervals and simple contours more pleasant, these melodies may conform better to their veridical expectations. It is questionable, however, if such melodies would also be those which would be most memorable in a music tradition. Rather, they might be unspecific material that fits in any melody, and is therefore present throughout a tradition, but not associated with a specific musical piece.

Van Balen, Burgoyne, Bountouridis, Müllensiefen, and Veltkamp (2015) approached memorability of melodies differently: they registered the reaction times in a game. The goal of the game was to indicate whether or not the player recognized a given song segment (cf. Burgoyne, Bountouridis, Van Balen, & Honing, 2013). If the player's response was fast, Van Balen and colleagues surmized that the song segment in question was very memorable, or catchy. They used a range of features to predict the memorability of the melodies, among which the features used by Müllensiefen and Halpern (2014), to which they added audio features, i.e., features extracted from audio recordings rather than score representations.

Analogous to Müllensiefen and Halpern's M-types, they investigated the frequency of pitch trigrams and pitch interval bigrams derived from the audio signal: in the song segment itself, in the song from which the segment was taken, and in a background corpus of pop songs. They analyzed first-order features, i.e., features based on the segment itself, and second-order features, i.e., features based on how the segment compared to the complete song or the background corpus. They summarized a total of 44 features by means of a principal components analysis, which determined which of the features were correlated.

One of Van Balen and colleagues' strongest predictors of memorability turned out to be motif repetitivity, which is in line with Müllensiefen and Halpern's findings on explicit melody recall. As our study focusses on melodies which were explicitly remembered by their singers, rather than pleasantness ratings of these melodies, we therefore adopt the prediction that motif repetitivity will increase a phrase's stability. Motif repetitivity can also be seen as related to chunking, as repeating motifs would provide meaningful subdivisions within a phrase. Chunking has been shown to facilitate learning in various domains (Gobet et al., 2001).

Based on the above observations, we investigate the following five hypotheses of how variation of folk songs may be predicted through theories on melody recall:

- H1. Shorter phrases show less variation.
- H2. Phrases which repeat within their source melody show less variation.
- H3. Phrases which occur early in their source melody show less variation.
- H4. A phrase's expectancy is related to its variation.
 - a) Phrases which contain highly expected melodic material show less variation.
 - b) Phrases which contain highly surprising melodic material show less variation.

H5. Phrases composed of repeating motifs show less variation.

5.2 MATERIAL

Our research was carried out using the folk song corpus (MTC-FS-1.0) from the Meertens Tune Collections.¹ We use the tune family categories in this corpus to investigate stability between song variants from the same tune family. Each variant is considered to represent the variation imposed by a particular singer or song book editor to a given melody. Consequently, we analyze which phrases of the songs belonging to a tune family vary more, or vary less between different variants: if a phrase occurs in many variants, this means that this phrase is less subject to change, or more stable.

A subset of the FS corpus, the annotated corpus (MTC-ANN-2.0) was used in [Chapter 4](#) to optimize the pattern matching method, and will not be used in this study. The remaining 3760 songs of the MTC-FS-1.0 corpus were separated into sub-corpora as follows: 1) a test corpus of 1695 melodies with tune families comprising at least five variants, but excluding tune families from the training corpus; 2) a background corpus of 1000 melodies with tune families comprising very few variants.

The background corpus was used to train information theoretical models, and the test corpus was used to test the relationship between variation of the folk song phrases and their hypothesized memorability. All melodies which could potentially be related to melodies from the test corpus – because they might be hitherto unrecognized variants of a tune family in the test corpus (tune family membership undefined), or because they were subtypes of a tune family in the test corpus – were excluded from the background corpus. This means that 1065 melodies were not used for this study.

5.3 FORMALIZING HYPOTHESES

This section describes the formalization of hypotheses on memorability of melodies.² For illustration purposes, we present a running example in [Figure 5.2](#), a folk song melody from the tune family *Van Peer en Lijn* (1), part of the test corpus. This melody has ten phrases and shows how under the current formalizations, different hypotheses arrive at different predictions of stability for each phrase. Throughout this section, we refer to a query phrase as q , which is taken from its source melody, s . The source melody's notes are referred to as s_j . The query phrase starts at index $j = a$ and has a length of n notes.

¹ www.liederenbank.nl/mtc

² The implementations of the hypotheses can be found at <https://github.com/BeritJanssen/Stability>

NLB074521_01

1 Zeg vrien - den luis - ter naar het lied

2 toen Peer en Lijn ging trou - - wen

3 Het huw - lijk is zo al ge - schied

4 het zal hun nog be - rou - - wen.

5 De eer - ste dag was 't al maar lach

6 en men deed er niets dan slem - pen

7 want Peer en Lijn moes - ten vro - lijk zijn

8 met bas - sen en trom - pet - - ten

9 Tra - la lie - e ti ral - la - la tra - la lie - e ti ra - la - la

10 Tra - la lie - e ti ral - la - la tra - lie - a ra - la - la.

Figure 5.2: An **example melody** from the test corpus, belonging to the tune family *Van Peer en Lijn* (1), which comprises six variants. This melody is used to illustrate the formalizations of the hypotheses. The number on top of the sheet music shows the record number in the Dutch folk song database, the numbers left of the staves show the sequential phrase indices.

5.3.1 Influence of phrase length

We test whether the length of the phrases influences their stability by defining the *phrase length* as the number of notes n of which a given phrase is composed.

$$\text{Len}(q) = n \quad (5.1)$$

In the example melody, the shortest phrases (phrase 2 and 4) have a length of seven notes, the longest phrase (phrase 9) has 16 notes. According to the prediction of the list length effect, we would expect the second and fourth phrases to be more stable than the ninth phrase. Over the whole dataset, phrase length takes values in the range $[3, 26]$, with a mean of $\overline{Len} = 9.11$ and a standard deviation of $\mathbf{SD}(Len) = 2.23$.

5.3.2 Influence of rehearsal

Rehearsal is modelled based on phrase repetitions: if a phrase is repeated multiple times within a melody, it is subject to more rehearsal, hence it may be expected to be more stable. The resulting predictor, *phrase repetition*, is measured by counting the occurrences of a phrase in its source melody. All phrases in a melody s are defined as sets of notes P_{id} . id refers to the sequential index of the phrase P in the melody. Each phrase's notes are represented by their onset from the start of the phrase and their pitch. The query phrase is a set of notes Q with the same representation. For every phrase P_{id} we determine its equality score to Q as follows:

$$Eq(P_{id}, Q) = \begin{cases} 1 & \text{if } P_{id} = Q \\ 0 & \text{otherwise} \end{cases} \quad (5.2)$$

Then we measure the number of phrase repetitions Rep of the query phrase q by summing the equality scores of all g phrases P_{id} in the melody.

$$Rep(q) = \sum_{id=1}^g Eq(P_{id}, Q) \quad (5.3)$$

In the example melody, the first and second phrase repeat exactly as the third and fourth phrase, respectively. The other phrases do not repeat anywhere in the melody. This means that phrase repetition is $Rep = 2$ for the first four phrases, $Rep = 1$ for the other six phrases. This would lead to the prediction that the first four phrases are more stable than the last six phrases. Phrase repetition takes values in the range of $[1, 4]$ in the dataset, with a mean of $\overline{Rep} = 1.17$ and a standard deviation of $\mathbf{SD}(Rep) = 0.39$.

5.3.3 Influence of the primacy effect

We test the primacy effect based on the position of a phrase in its source melody. We formalize the *phrase position* as a given phrase's sequential index, q_{id} , from $q_{id} = 1$ to $q_{id} = g$ for all g phrases in the source melody.

$$Pos(q) = q_{id} \quad (5.4)$$

In the example melody, the first phrase has a value of $Pos = 1$, and the last phrase a value of $Pos = 10$. Phrase position takes values in the range of $[1, 22]$ in the dataset, with a mean of $\overline{Pos} = 3.44$ and a standard deviation of $\mathbf{SD}(Pos) = 2.06$.

5.3.4 Influence of expectancy

To quantify expectancy, we make use of two formalizations: one by Schellenberg (1997), which is based on observations from music theory, and one by Pearce and Wiggins (2004), which is based on statistical analysis of a background corpus.

We base both models on pitch intervals between consecutive notes. The pitch of a given note $pitch(s_j)$, or its height in the human hearing range, is represented by its MIDI note number. This entails that pitches are represented by integers, and that a semitone difference between two pitches is indicated by a difference of one. The pitch interval between a note s_j and its predecessor s_{j-1} is defined by $pInt(s_j) = pitch(s_j) - pitch(s_{j-1})$, where a positive sign indicates that the preceding note is lower, and a negative sign that the preceding note is higher. Both models make predictions for single *notes*, rather than whole phrases. We derive predictions for whole phrases through averaging the note values over the length of the phrase.

For the first note of any melody's first phrase, there are no expectations yet, as there is no previous melodic material on which such expectations could be based. One might choose to treat the first note of later phrases analogously, and hold that there are no expectations at the beginning of each phrase. We choose the alternative: the first note of a phrase represents expectations with relation to previous phrases, which we consider to be more realistic, as singers and listeners of songs will probably not treat phrases in folk songs as completely isolated, but in relation to melodic material that preceded a phrase.

EXPECTANCY: MUSIC THEORY The first component of Schellenberg's model, *pitch proximity*, states that listeners expect small steps between melody tones. The further a given note is away from its predecessor, the more unexpected it is. The model does not make any predictions for pitch intervals equal to or larger than an octave.

$$PitchProx(s_j) = \begin{cases} |pInt(s_j)| & \text{if } |pInt(s_j)| < 12 \\ \text{undefined} & \text{otherwise} \end{cases} \quad (5.5)$$

Pitch-reversal is the linear combination of two other principles, registral direction and registral return. The principle of registral direction states that after large implicative intervals, a change of direction is more expected than a continuation of the direction. The tritone, i.e., a pitch interval of six semitones, is not defined in this principle.

$$PitchRev_{dir}(s_j) = \begin{cases} 0 & \text{if } |pInt(s_{j-1})| < 6 \\ 1 & \text{if } 6 < |pInt(s_{j-1})| < 12 \text{ and } pInt(s_j) \cdot pInt(s_{j-1}) < 0 \\ -1 & \text{if } 6 < |pInt(s_{j-1})| < 12 \text{ and } pInt(s_j) \cdot pInt(s_{j-1}) > 0 \\ \text{undefined} & \text{otherwise} \end{cases}$$

(5.6)

The other component of pitch-reversal, registral return, states that if the realized interval has a different direction than the implicative interval, the size of the intervals is expected to be similar, i.e., they should not differ by more than two semitones. If the implicative interval describes a tone repetition, or if the difference between two consecutive pitch intervals of opposite direction is too large, pitch-reversal is zero, otherwise it has the value of 1.5.

$$\text{PitchRev}_{\text{ret}}(s_j) = \begin{cases} 1.5 & \text{if } |\text{pInt}(s_j) > 0| \text{ and} \\ & \text{pInt}(s_j) \cdot \text{pInt}(s_{j-1}) < 0 \text{ and} \\ & \text{pInt}(s_{j-1}) + \text{pInt}(s_j) \leq 2 \\ 0 & \text{otherwise} \end{cases} \quad (5.7)$$

Combined, registral direction and registral return form the pitch-reversal principle.

$$\text{PitchRev}(s_j) = \text{PitchRev}_{\text{dir}}(s_j) + \text{PitchRev}_{\text{ret}}(s_j) \quad (5.8)$$

Figure 5.3, drawn after a figure by Schellenberg, shows a schematic overview for the different values pitch reversal can take under different conditions.

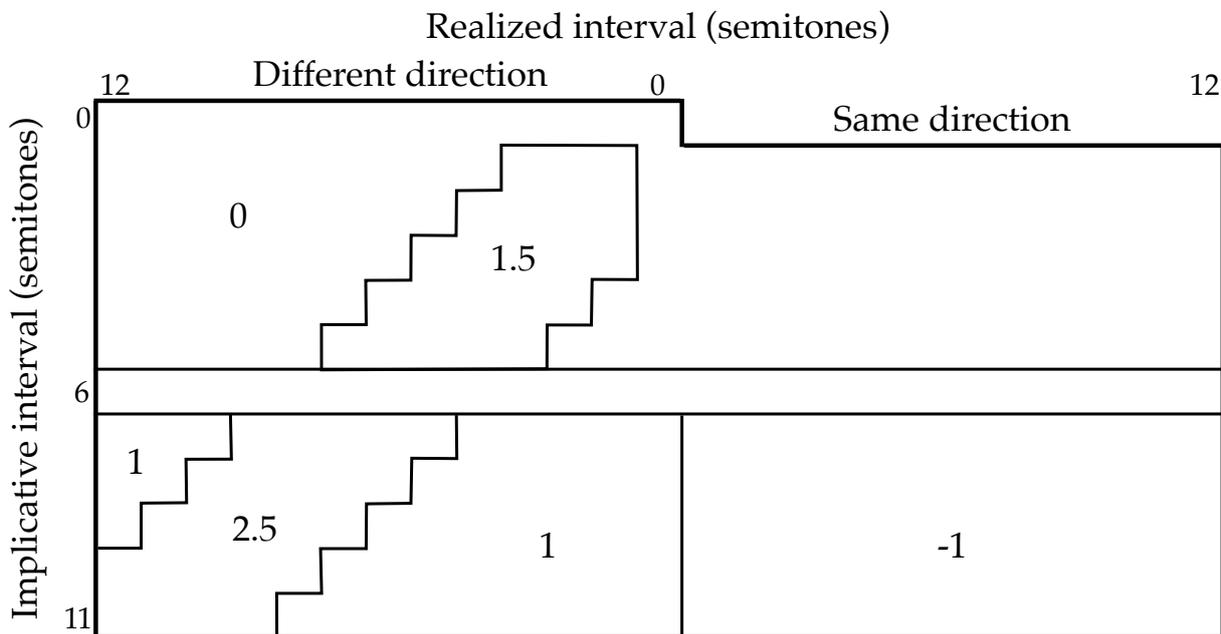


Figure 5.3: A visualization of the pitch reversal principle as defined by Schellenberg (1997), drawn after his figure. The vertical axis represents the size of the implicative interval, from 0 to 11, and the horizontal axis the size of the realized interval, from 0 to 12, which can have either the same direction (right side of the panel) or a different direction (left side of the panel).

In Figure 5.4.a we show the first phrase of the example melody, with the pitch proximity values printed underneath each note, referring to the pitch interval to its preceding note. Note that the pitch interval, and therefore pitch proximity, is not defined for the first note of a melody, as there is no previous pitch from which a pitch interval could be measured.

To calculate the average proximity of a phrase, the pitch proximity values of the notes s_j belonging to a given phrase are averaged over the whole phrase, and the negative value of this average is used for easier interpretation, such that if a phrase has a high value of pitch proximity, its pitches are close to each other, while lower values indicate larger pitch intervals. Notes for which pitch proximity is not defined are discarded from the averaging procedure.

$$Prox(q) = -\frac{1}{n} \sum_{j=a}^{a+n} PitchProx(s_j) \quad (5.9)$$

We show the pitch proximity values for the seventh and eighth phrase of the example melody in Figure 5.4.a. The average proximity of the two phrases amounts to $Prox = -13/9 = -1.44$ and $Prox = -20/7 = -2.85$, respectively, which means that we would expect the seventh phrase to be more stable than the eighth phrase. Average proximity takes values in the range of $[-6.0, 0.0]$ in the whole data set, with a mean of $\overline{Prox} = -2.01$ and a standard deviation of $\mathbf{SD}(Prox) = 0.69$.

The other factor in Schellenberg's model is *pitch reversal*, which summarizes the long-standing observation from music theory that if leaps between melody notes do occur, they tend to be followed by stepwise motion in the opposite direction (Meyer, 1956). For a given melody note, this principle results in values ranging from $PitchRev(s_j) = -1$, or least expected, to $PitchRev(s_j) = 2.5$, or most expected. As with pitch proximity, we calculate the average reversal of a phrase through calculating the arithmetic mean of its constituent notes. As pitch reversal makes predictions based on two pitch intervals, it is not defined for the first two notes of a melody. Notes for which pitch reversal is not defined are discarded from the averaging procedure.

$$Rev(q) = \frac{1}{n} \sum_{j=a}^{a+n} PitchRev(s_j) \quad (5.10)$$

We show the pitch reversal values for the seventh and eighth phrase of the example melody in Figure 5.4.b. The average reversal of the two example phrases amounts to $Rev = 3/9 = 0.33$ and $Rev = 1/7 = 0.14$, respectively, which means that we would expect the seventh phrase to be more stable than the eighth phrase. Average reversal takes values in the range of $[-0.5, 1.5]$ in the whole data set, with a mean of $\overline{Rev} = 0.30$ and a standard deviation of $\mathbf{SD}(Rev) = 0.24$.

EXPECTANCY: INFORMATION THEORY The IDyOM (Information Dynamics of Music) model by Pearce analyzes the frequencies of *n-grams* in a music collection, and

a. Pitch proximity

NLB074521_01, Phrases 7 and 8

2 0 2 2 1 0 3 2 1

0 1 10 4 3 0 2

b. Pitch reversal

NLB074521_01, Phrases 7 and 8

1.5 0 0 0 0 0 1.5 0

0 0 0 1 0 0 0

c. Information content

NLB074521_01, Phrases 7 and 8

4.71 (5) 2.27 (3) 1.97 (3) 2.01 (4) 2.16 (5) 2.87 (6) 3.59 (6) 1.61 (6) 0.54 (7)

3.44 (8) 3.15 (5) 10.75 (5) 2.89 (2) 0.51 (3) 2.44 (4) 2.72 (3)

Figure 5.4: Phrase 7 and 8 of the example melody, showing the expectancy values for each note resulting from different theories: a. expectancy values according to Schellenberg’s pitch-proximity principle; b. expectancy values according to Schellenberg’s pitch-reversal principle; c. Information Content, calculated with IDyOM, based on a background corpus. The numbers in brackets indicate how much context is considered to calculate information content, which in this case ranges from 2 (two previous notes considered) to 8 (eight previous notes considered).

based on these observed frequencies, assigns probabilities to notes in unseen melodies, given the notes preceding it. The preceding notes are also referred to as *context*. The length of the context can be set by the user. If the model cannot find a relevant *n-gram*

of the context length specified by the user, it backtracks to shorter melodic contexts, and uses those frequencies to return the probability of a given note.

We let the model analyze our background corpus, with the melodies represented as pitch intervals. As we are interested in contexts of phrase length, we limit the *n-gram* length to the average phrase length of nine, meaning the model will never look for longer contexts than eight preceding notes, and backtrack for shorter contexts if necessary. We use IDyOM's long-term model, i.e., the model does not update itself while observing the query phrases, and we apply the interpolation weighting scheme *C*, which balances longer and shorter melodic contexts evenly. This was proven to be the best performing weighting scheme in experiments by Pearce (2005).

We express the expectancy of a given melodic segment through its average information content. Information content is the natural logarithm of the inverse probability $\mathbb{P}(s_j)$ of a note to occur given the previous melodic context, based on the probabilities of the background corpus. We choose information content rather than probability, as the logarithmic representation makes it possible to compare the typically small probability values in a more meaningful way. Information content is often also referred to as *Surprisal*, as its values increase as events get *less* expected.

We average the information content of all notes in a query phrase by their arithmetic mean, which is equivalent to a geometric mean of the probabilities. We call this average information content surprisal in the following, to indicate that higher values denote less expected phrases.

$$Sur(q) = \frac{1}{n} \sum_{j=a}^{a+n} \log\left(\frac{1}{\mathbb{P}(s_j)}\right) \quad (5.11)$$

We show the information content for the seventh and eighth phrase of the example melody in Figure 5.4.c. The context used to generate the information content is shown in brackets. The surprisal of the two example phrases amounts to $Sur = 21.74/9 = 2.42$ and $Sur = 25.88/7 = 3.7$, respectively, which means that we would expect the seventh phrase to be more stable than the eighth phrase. Surprisal takes values in the range of $[1.15, 6.86]$ in the whole data set, with a mean of $\overline{Sur} = 2.68$ and a standard deviation of $\mathbf{SD}(Sur) = 0.53$.

5.3.5 The influence of repeating motifs

As Müllensiefen and Halpern (2014) and Van Balen et al. (2015) found a relationship between repetitiveness of short motifs and the recall of a melody, we follow their procedure and use the FANTASTIC toolbox (Müllensiefen, 2009) to compute a frequency table of such short motifs *t* for each phrase, and to measure motif repetitiveness through normalized entropy. The motifs are *n-grams* of character sequences representing the pitch and duration relationships between notes. The lengths of motifs to be investigated can be determined by the user. For each investigated motif length *l*, the frequency of unique motifs $v_{z,l}$ is counted, and compared to the total number of motifs of

that length $N_{t,l}$ covering the phrase. The normalized entropy H_l is then calculated from each unique motif's relative frequency $f_{z,l}$, i.e., how often a given motif $v_{z,l}$ occurs in a phrase with relation to all motifs of that length in the phrase.

The relative frequencies of all unique motifs are multiplied with their binary logarithm, summed, and divided by the binary logarithm of the number of all motifs of that length in the phrase ($N_{u,l}$) for normalization. A value of $H = 1.0$ then indicates maximal normalized entropy, and minimal repetitiveness: there are no repeating motifs of length l at all in the phrase; a lower value indicates that there are some repeating motifs.

$$H_l = - \frac{\sum_{z=1}^{N_{t,l}} f_{z,l} \cdot \log_2 f_{z,l}}{\log_2 N_{u,l}} \quad (5.12)$$

The mean entropy of the motifs is then the average over all possible motif lengths. We analyze, in accordance with earlier work, motifs from two notes to six notes in length. We take the negative value of this average to define motif repetitivity: the higher the mean entropy, or the more distinct motifs in the phrase, the lower the repetitivity.

$$MR(q) = - \frac{\sum_{l=2}^6 H_l}{5} \quad (5.13)$$

To illustrate the concept, refer to Figure 5.5, showing the second (also fourth) and sixth phrase of the example melody. The second phrase consists of repeated steps up by a third, and can be subdivided into three identical sequences of two notes each (as the representation of the FANTASTIC toolbox does not distinguish between minor and major intervals): this would mean that this phrase has higher motif repetitivity than the sixth phrase.

NLB074521_01, Phrases 2/4 and 6

The figure shows two musical staves in G major (one sharp) and 2/4 time. The top staff represents the second phrase, consisting of six notes: G4 (quarter), A4 (quarter), B4 (quarter), A4 (quarter), G4 (quarter), and F#4 (quarter). Below the notes are interval symbols: s1e (same pitch, equal duration), u3e (up a third, equal duration), u3e (up a third, equal duration), u3e (up a third, equal duration), d2l (down a second, longer duration), and d3q (down a third, quicker duration). The bottom staff represents the sixth phrase, consisting of seven notes: G4 (quarter), A4 (quarter), B4 (quarter), A4 (quarter), G4 (quarter), F#4 (quarter), and E4 (quarter). Below the notes are interval symbols: s1e (same pitch, equal duration), u4l (up a fourth, longer duration), d4e (down a fourth, equal duration), u4e (up a fourth, equal duration), u3e (up a third, equal duration), u4l (up a fourth, longer duration), and d2q (down a second, quicker duration).

Figure 5.5: The second, also fourth, and sixth phrase of the example melody with symbols representing the pitch and duration relationships between adjacent notes. Notes can either stay at the same pitch (s1), or move up or down by a diatonic pitch interval (e.g., u4, d2). Durations can either be equal (e), quicker (q), or longer (l).

To give an example calculation for the second, also fourth, and sixth phrase of the example melody, shown in Figure 5.5, first observe the string representations underneath the notes. FANTASTIC (Müllensiefen, 2009) represents relationships between adjacent

notes as follows: pitches can either stay the same (s_1), move up or down by a diatonic pitch interval (e.g., u_4 , d_2). In this representation, it does not matter whether, e.g., a step down (d_2) contains one or two semitones. Durations either stay equal (e), get quicker (q) or longer (l).

We will refer to the string representations of the motifs, in accordance with Müllensiefen and Halpern (2014), as *M-Types*. The second/fourth phrase of the example melody consists of six two-note M-types, of which one repeats three times, so there are four unique M-types. This leads to the entropy of two-note M-Types:

$$H_2 = -\frac{3/6 \log_2 3/6 + 3 \cdot (1/6 \log_2 1/6)}{\log_2 6} = -\frac{-0.5 - 3 \cdot 0.43}{2.58} = 0.69 \quad (5.14)$$

These M-types can be combined to the following three-note M-Types:

$s_1q_u_3e$
 $u_3e_u_3e$
 $u_3e_d_2l$
 $d_2l_d_3q$

One M-Type, “ $u_3e_u_3e$ ”, is repeated. In total, there are five M-Types in the melody, and four unique M-Types. This means that the entropy for length $l = 3$ for this phrase is

$$H_3 = -\frac{2/5 \log_2 2/5 + 3 \cdot (1/5 \log_2 1/5)}{\log_2 5} = -\frac{-0.53 - 3 \cdot 0.46}{2.32} = 0.82 \quad (5.15)$$

There are no repeated four-note M-Types, so the entropy of M-Types of length $l = 4$ for this phrase is maximal, $H(4) = 1.0$. This means that all longer M-types also have maximal entropy, and the motif repetivity, the average of the entropies for all lengths of *mtypes*, is $MR = -(0.69 + 0.82 + 3 \cdot 1.0) / 5 = -0.90$.

The sixth phrase of the example melody (the second phrase shown in the figure) consists of seven M-Types, of which only one (u_4l) appears twice. This leads to the entropy of two-note M-Types:

$$H_2 = -\frac{2/7 \log_2 2/7 + 5 \cdot (1/7 \log_2 1/7)}{\log_2 7} = -\frac{-0.52 - 5 \cdot 0.40}{2.81} = 0.89 \quad (5.16)$$

For the longer M-Types, there are no repetitions, hence the entropy is maximal at $H_{3,4,5,6} = 1.0$. This leads to a motif repetivity of $MR = -(0.89 + 4 \cdot 1.0 / 5) = -0.98$.

In summary, the motif repetivity of the second/fourth phrase amounts to $MR = -0.90$, and of the sixth phrase to $MR = -0.98$, so we would expect the second and fourth phrase to be more stable than the sixth phrase. Motif repetivity takes values in the range of $[-1.0, 0.0]$ in the whole data set, with a mean of $\overline{MR} = -0.92$ and a standard deviation of $SD(MR) = 0.09$.

5.4 RESEARCH METHOD

This section describes the research method developed to study stability of folk song phrases. As laid out in the previous chapter, the stability of phrases is determined through pattern matching: each phrase in the corpus is compared to all variants belonging to the tune family from which the query phrase is taken, and if there is a note sequence in a given variant which is very similar to the query phrase, this is rated as an *occurrence* of that phrase in that variant. Conversely, if there is no note sequence which is similar to the query phrase in a given variant, this constitutes a *non-occurrence*. A phrase which occurs in many variants of its tune family has a higher frequency of occurrence, or in other words, it is more stable. For more details on the pattern matching method, see [Chapter 4](#).

As an example, consider the query phrase q and variants s_1 and s_2 , as shown in [Figure 5.6](#). The phrase and variants are part of the tune family *Vrienden kom hier en luister naar mijn lied 2* from the MTC-FS-1.0 corpus, which has five variants, providing 22 query phrases in total. The shown phrase occurs in one of the two shown variants, and not in the other, as determined by the pattern matching method. The example phrase occurs in three of the four variants against which it is matched. As there are two possible outcomes of the pattern matching procedure – occurrence or non-occurrence – the chance frequency of occurrence is 50%. The shown query phrase therefore exceeds the chance frequency of occurrence.

5.4.1 *Logistic regression*

It could of course be purely random that some query phrases occur more frequently than others. Conversely, if there is any statistical relationship between frequency of occurrence and the hypothesized memorability of a phrase, this points towards a tendency of phrases with specific properties to be more stable than others. We determine the presence of such a statistical relationship through *logistic regression*. While the outcome of pattern matching is binary (occurrence or non-occurrence), the statistical model used to predict the outcome is continuous: logistic regression, as the name suggests, uses a logistic probability function, the logit, which is an s-shaped curve reflecting the *probability of occurrence* \mathbb{P} according to the model.

$$\text{logit}(\mathbb{P}) = \log\left(\frac{\mathbb{P}}{1 - \mathbb{P}}\right) \quad (5.17)$$

The goal of logistic regression is to find a curve that best separates the true events from the false events. In our case, this means that we want to predict the probability \mathbb{P} that a given query phrase q has a match in a given melody s , based on the vector \mathbf{F} of the independent variables hypothesized to contribute to long-term memorability of melodies.

$$\text{logit}(\mathbb{P}) = \beta \mathbf{F} + \epsilon \quad (5.18)$$

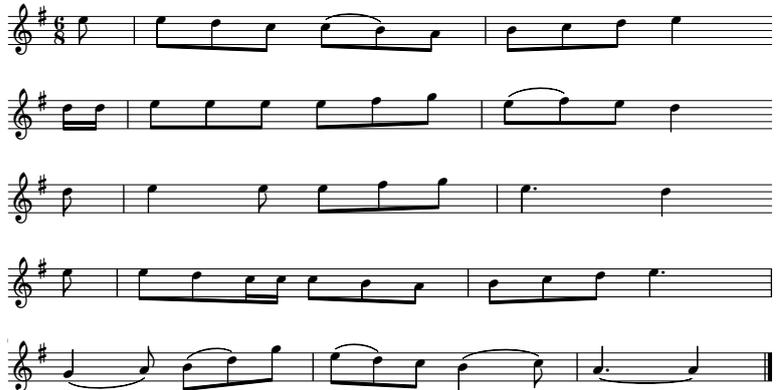
Non-occurrence in variant s_1
NLB073030_01



Query phrase q
NLB072093_01, Phrase 1



Occurrence in variant s_2
NLB073030_03



The figure displays musical notation for a query phrase and two variants of a tune family. The query phrase 'q' is shown on the left. To the right, variant s_1 (top) does not contain the query phrase, while variant s_2 (bottom) does. The notation includes treble clefs, a key signature of one sharp (F#), and a 6/8 time signature. The query phrase consists of a single melodic line. Variant s_1 is shown in two staves, and variant s_2 is shown in four staves.

Figure 5.6: An example for the matching of a **query phrase** q (shown on the left) in two variants s_1, s_2 (shown on the right) of the same tune family. The pattern matching method determines that there is an occurrence of q in variant s_2 (shown at the bottom), but that there is no occurrence in variant s_1 (shown on top).

In this equation, β represents the slope of the prediction function, ϵ represents the random effects of the model, i.e., the random error for each melodic segment, assumed to be normally distributed. If the prediction of the probability of occurrence (i.e., the inverse logit of the prediction function) were perfect, this would lead to a curve separating the occurrences clearly from the non-occurrences. For instance, if the model were to predict that query phrase q has a higher probability of occurrence than the chance level $\mathbb{P}(q, s) > 0.5$ to appear in any variant s of its tune family, we would like to find that indeed q occurs in most variants s of the associated tune family.

For the same tune family *Vrienden kom hier en luister naar mijn lied 2* from which the query phrase is taken, Figure 5.7 shows the logistic regression model of predicting the probability of occurrence of all phrases in this tune family in all its variants. The logistic regression model is indicated by the black, s-shaped line, and the grey area around it represents the standard error of the model. For this example model, I used average information content as a predictor, rescaled to a range of $[0, 1]$, where a low value indicates low memorability, and a high value high memorability. Query phrase

q has a memorability of $M(q) = 0.896$ on this scale, and according to the model, it has a higher probability of occurrence than chance, corresponding to around $\mathbb{P}(q) = 0.8$.

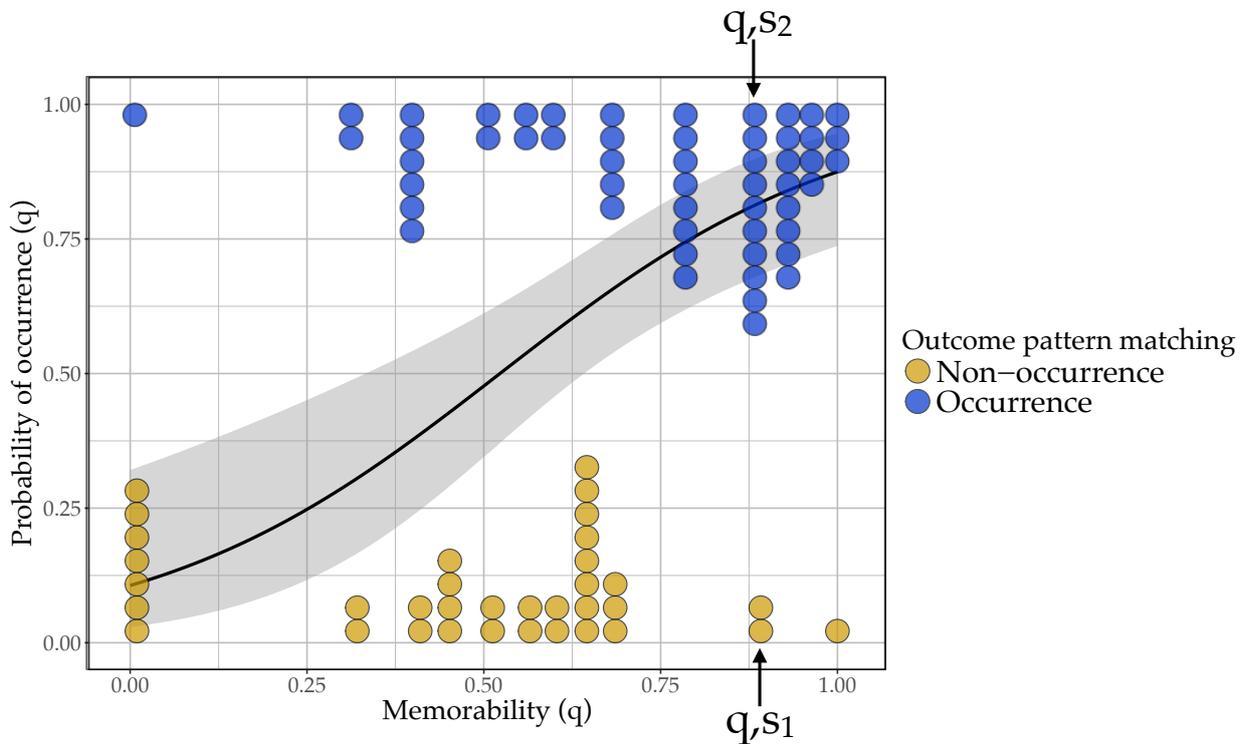


Figure 5.7: An example logistic regression model which predicts occurrences of melodic phrases in variants of the tune family *Vrienden kom hier en luister naar mijn lied 2*. The x axis shows the hypothesized memorability of the various query phrases, the y axis the model's prediction of their probability of occurrence. The logistic regression model, predicting the probability of occurrence of a given phrase in a given variant based on its memorability, is indicated by the black line, the standard error of the model by the grey area. The dots above and below indicate the actual occurrences (blue) and non-occurrences (yellow) of all query phrases in all variants in the tune family.

The dots above and below the figure indicate the actual occurrences (blue, top) and non-occurrences (yellow, bottom) of the query phrases. There are 88 data points in total, because there are 22 query phrases which are each matched against four variants. The data points corresponding to the query phrase and variants from Figure 5.6 are located at the right side of the figure, meaning their memorability and predicted probability of occurrence are higher than average. Next to the four data points corresponding to query phrase q , other points corresponding to other query phrases are shown at the same location on the x axis, as they have very similar memorability values. Overall, more occurrences are located on the right side of the figure, and more non-occurrences on the left side of the figure, which means that the model represents the data reasonably well, even though it does certainly not separate occurrences and non-occurrences perfectly. The fit between the data and the model is measured as R^2 , the coefficient of

determination which describes the percentage of the variation in the data which can be explained by the model, of which more below.

5.4.2 *Generalized Linear Mixed Model*

Chapter 4 pointed out that the pattern matching method may lead to different percentages of errors in different tune families. This makes it problematic to treat all tune families in one logistic regression model: it essentially means that the chance level of occurrence may be raised for some tune families, in which many spurious occurrences are found, or lowered in others, in which relevant occurrences are missed. This problem could be addressed by analyzing each tune family with a separate logistic regression model; yet this would mean that we could not globally estimate how well a specific hypothesis accounts for probability of occurrence. We therefore choose another solution to model the relationship between phrase properties and occurrence: a generalized linear mixed model (GLMM) which can model the variation of all data at the same time.

Generalized linear models are a framework in which relationships between independent variables and dependent variables of binomial, multinomial, ordinal and continuous distributions can be investigated. A special case of this framework are mixed models, in which not only a general random effect (ϵ), but also random effects for subgroups of the data can be taken into account. This way, we can model the tune family dependent error of the computational method. We assume that every tune family has a different intercept term in the model. The intercept describes the chance level of occurrence of any phrase within a tune family. Depending on the tune family intercept, the logistic decision function between occurrence vs. non-occurrence of the model is shifted.

We again assume \mathbf{F} as the vector representing the independent variables of the query phrases, β as the slope of the prediction function, ϵ as the random error, but now also take into account the random effect μ , based on the individual error of each tune family, summarized in the vector \mathbf{tf} . Then the model can be formalized as follows:

$$\text{logit}(\mathbb{P}) = \beta \mathbf{F} + \mu \mathbf{tf} + \epsilon \quad (5.19)$$

One could also think of the fixed effects, expressed by $\mu \mathbf{tf}$ as the between-tune-family variance, and the random effects, expressed by ϵ , as the within-tune-family variance. Using this model, we test our hypotheses on possible predictors for stability.

To be able to compare the independent variables derived from our hypotheses, we standardize all variables x of the predictor vector by subtracting the arithmetic mean \bar{x} , and dividing by the standard deviation $\mathbf{SD}(x)$ of a given variable.

$$F_x = \frac{x - \bar{x}}{\mathbf{SD}(x)} \quad (5.20)$$

This leads to the overall model for all phrase occurrences, in which units can be compared with each other. We apply a Generalized Linear Mixed Model with fixed

Parameter estimate	3df	4df	5df	6df	7df	8df	9df
Surprisal	-0.27	-0.29	-0.30	-0.30	-0.29	-0.24	-0.24
Phrase length		-0.32	-0.32	-0.33	-0.30	-0.31	-0.30
Phrase position			-0.10	-0.12	-0.12	-0.10	-0.10
Phrase repetition				0.08	0.09	0.09	0.09
Motif repetivity					0.08	0.08	0.08
Average proximity						0.09	0.10
Average reversal							0.05
AIC_c	209159.8	206889.7	206584.4	206355.5	206157.6	206012.1	205941.5

Table 5.1: The best models for different degrees of freedom, from 3df with one parameter, to 9df with seven parameters. For each model, the second order Akaike information criterion (AIC_c) is shown, with lower values indicating better model fit. Surprisal is the parameter which leads to the best model with only one predictor; the other parameters are listed in the order by which they are added, leading to the best model fit when all parameters are used.

slopes and random intercepts for each tune family to the test corpus of the dataset containing 9,639 phrases from 147 tune families, using the R package LME4.³

5.4.3 Model selection

We select the independent variables contributing to the strongest model predicting stability of folk song phrases, using the R library MuMIn.⁴ This model selection compares all possible combinations of independent variables and ranks them based on their second-order Akaike information criterion (AIC_c) (Hurvich & Tsai, 1989). The second-order Akaike information criterion penalizes the addition of extra parameters to a model, such that it strikes a balance between model fit and parsimony (Burnham & Anderson, 2004). Furthermore, we estimate the effect size of the best model with a technique to determine R^2 of mixed models by Nakagawa and Schielzeth (2013).

5.5 RESULTS

We show the best models selected from three degrees of freedom (3df), with one model parameter, to nine degrees of freedom (9df), with seven model parameters, in Table 5.1. The models' second-order Akaike information criteria decrease as more parameters get added, indicating better model fit. Our results show that the strongest model for the stability of melodic phrases is the full model with all independent variables: phrase

³ <https://CRAN.R-project.org/package=lme4>

⁴ <https://CRAN.R-project.org/package=MUMIn>

length, phrase repetition, phrase position, pitch proximity, pitch reversal, surprisal and motif repetivity. This model yields an AIC_c lower by 70.65 than the second best model. Table 5.2 shows the estimated prediction coefficients, the variances of the tune family dependent error and the residual error for the full model, as well as the model fit in R^2 . The fixed effects alone, marginalized, explain $R^2_{marginal} = 0.05$, or about 5% of the variance, which is a mid-sized effect for mixed models (Cohen, 1992; Kirk, 1996). When the tune family dependent random effects are considered along with the fixed effects ($R^2_{conditional}$), 22% of the variation in the data is explained.

Parameter	$\hat{\beta}$	95% CI
Intercept	-0.22	[-0.35, -0.08]
Surprisal	-0.24	[-0.25, -0.22]
Phrase length	-0.30	[-0.32, -0.29]
Phrase position	-0.10	[-0.11, -0.09]
Phrase repetition	0.09	[0.08, 0.10]
Motif repetivity	0.08	[0.07, 0.09]
Average proximity	0.10	[0.08, 0.11]
Average reversal	0.05	[0.04, 0.06]
σ_{tf}	0.84	[0.74, 0.95]
$R^2_{marginal}$.05
$R^2_{conditional}$.22

Table 5.2: The parameters of the best model of the model selection: estimated regression coefficient $\hat{\beta}$ and 95% confidence interval for *phrase length*, *phrase repetitions* within the source melody, *phrase position* in the source melody, *pitch proximity* and *pitch reversal* as defined by Schellenberg (1997), *expectancy*, as defined by IDyOM (Pearce & Wiggins, 2004), and *motif repetivity*, as defined by Müllensiefen (2009). At the bottom of the table we report the standard deviation of the random effect (tune family), as well as the marginalized and conditional R^2 calculated according to Nakagawa and Schielzeth (2013).

The prediction coefficients show that phrase length and surprisal possess most predictive power: with increase of a given query phrase's length, its stability decreases. Higher expectancy leads to increased stability. Furthermore, the coefficients also indicate that earlier phrases tend to be more stable, as with an increase in the phrase index, the odds that a query phrase occurs in a given melody are decreased. Moreover, an increase in pitch proximity, or a decrease in the average size of the pitch intervals in a phrase, leads to a higher chance of an occurrence. More repetitions of a query phrase also result in the increased odds of occurrence. Pitch reversal and motif repetivity contribute least strongly to the model, but the signs of the parameters are as expected: if

a phrase confirms expectations of pitch reversal, its odds of occurrence are increased, and likewise, if a phrase contains many repeating motifs, its odds of occurrence are increased.

We also tested the model for multicollinearity, confirming that the approximate correlations of parameter estimates do not exceed 0.6, which justifies our treatment of the model parameters as independent predictors.

To illustrate the predictions of the model, we show the predicted as well as the observed frequency of occurrence for the ten phrases of the example melody in Figure 5.8. According to the model, the first four phrases have the highest probability of occurrence, and indeed these phrases also have the highest observed frequency of occurrence (i.e., stability). The predictions do differ from many of the observed values, as for instance the higher stability of phrase 1 and 3 as compared to phrase 2 and 4 is not captured by the model.

5.6 DISCUSSION

The current research shows that folk song collections are a valuable resource for studying the relationship between melody variation and memorability. All proposed hypotheses relating to recall in general and music recall in particular contribute to prediction of folk song variation, as model selection among all combinations of parameters leads to a model with all hypotheses as predictors.

Of course, the variation that is explained with the current model is still rather low at $R^2 = 0.05$. This might mean that there are potentially more, and stronger predictors for melody variation that have not been tested in this study. It is also good to keep in mind that the phrase occurrences in folk songs do not represent “clean” experimental data in which all aspects but melody recall are controlled. The ecological validity comes at the cost of potential noise. Some aspects that might deteriorate the observed variation are a) the computational method to detect occurrences; b) the inherent ambiguity of phrase occurrences, i.e., humans do not agree on occurrences perfectly (Janssen, van Kranenburg, & Volk, 2017); c) a bias in the corpus towards specific regions and demographic groups (Grijp, 2008).

Alternatively, one could assume that a large proportion of melody variation is a result of drift, and therefore random (Henrich & Boyd, 2002). Therefore, it is enlightening that the hypotheses *do* contribute to explaining variation in the dataset, in spite of potential noise in the data. Memorability predicts the amount of melodic variation, or stability, as follows: phrases which resist change should be short (list length effect, H1) and contain little surprising melodic material (i.e., low surprisal, a formalization of expectancy, H4a). Moreover, it is beneficial if a phrase occurs relatively early in a melody (primacy effect, H3), and has mostly small pitch intervals (i.e., high average proximity, a formalization of expectancy, H4a). The repetition of a phrase in its source melody also contributes to its memorability (rehearsal effect, H2), even though this effect is somewhat weaker in our analysis than other predictors. Average reversal, or the tendency for a melody to adhere to the gap fill principle, i.e., following a leap by step-

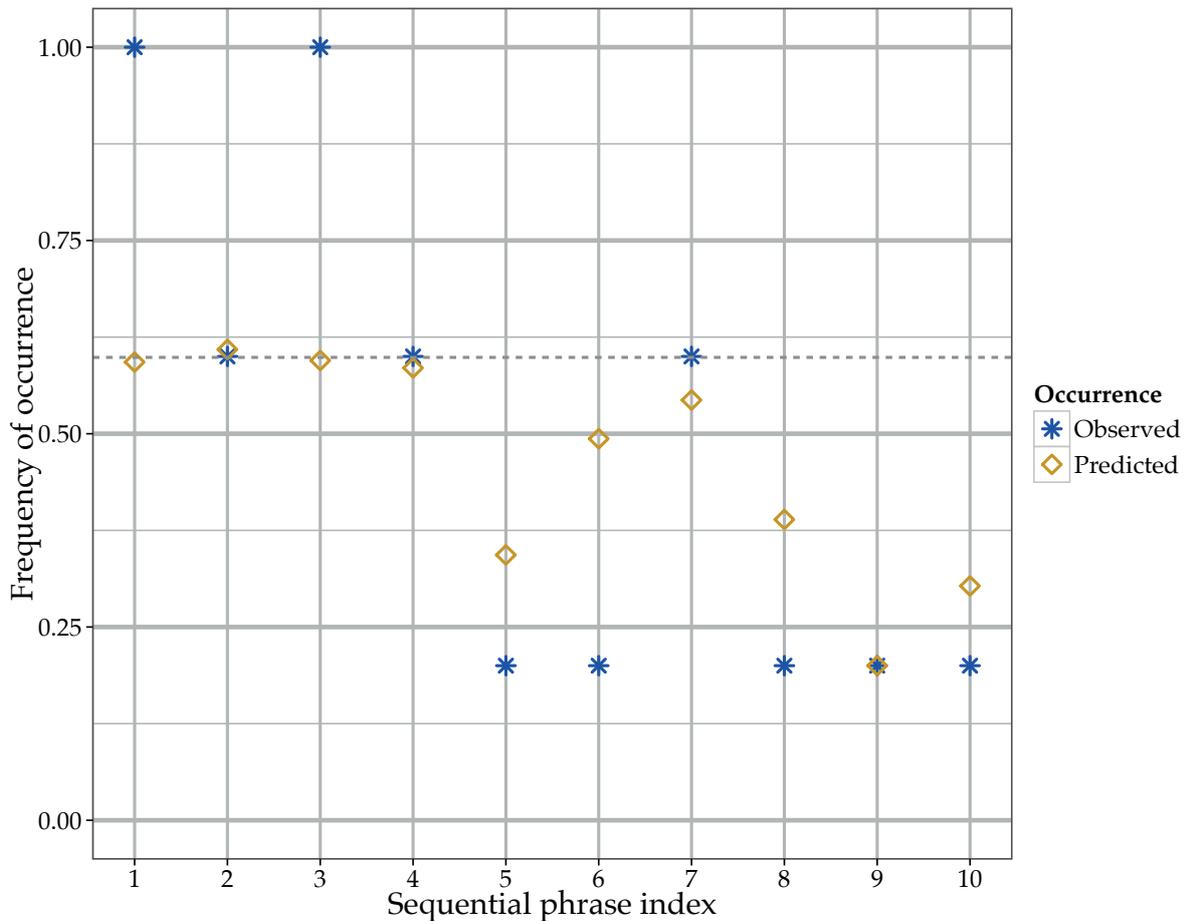


Figure 5.8: The predicted (yellow diamonds) and observed (blue stars) frequency of occurrence, in percent, for the ten phrases of the example melody. The predictions are generated by the generalized linear mixed model, for the model parameters see Table 5.2. The observed frequency of occurrence is based on how many of the five variants, other than the example melody, contain a given phrase from the example melody. The dashed line shows the model's intercept for frequency of occurrence for this tune family, which is at 58%, meaning that is slightly more likely for the phrases of this tune family to occur in the respective variants than not.

wise motion in the opposite direction (expectancy, H4a) and motif repetitivity within the phrase (H5) seem to account for long-term memorability to a more limited extent. All predictors related to expectancy indicate that more expected melodic material increases stability, leading us to reject H4b.

As for possible drawbacks of the presented study, the three predictors related to expectancy (average proximity, average reversal and surprisal) share the disadvantage that for the first few notes of a melody, no or little information on expectancy is available. This means that there is a potential imbalance between the initial and later phrases of a melody, as the predictor values of initial phrases are based on less information. The alternative, treating every phrase as isolated, so that no context from previous phrases

is used for creating expectancy values, seemed unrealistic, however, as the recall of phrases is cued by previous melodic material (cf. Rubin, 1995, p. 190). For the current folk song collection, in which the same melody is sung multiple times with different verses, it may be interesting to investigate in how far considering the end of a given melody as the melodic context for the start of this melody influences expectancy predictions.

The expectancy predictors defined by Schellenberg (1997), average proximity and average reversal, may be comparatively unsuccessful model parameters as they were not necessarily designed to be averaged for a longer melodic context: they were defined to quantify the fulfilment of listener expectations at a given note. However, these predictors still contribute to a better model, which shows that they capture some information on memorability which may predict variation of melodies in this corpus.

The relatively low contribution of motif repetitivity as a predictor for melodic variation may be partly ascribed to the fact that the phrases are very short melodic material, and as such rarely contain repeated motifs. It would be interesting to investigate if motif repetitivity increases stability of longer melodic contexts, e.g., full folk song melodies. For the current analysis of phrases with an average length of nine notes, which are unlikely to contain repeated motifs longer than four notes, it may be sufficient to limit the maximal *n-gram* length to four notes for future research on motif repetitivity in phrases. To hold our use of the method comparable to earlier research, we decided to analyze motifs of the same lengths as previous authors. Moreover, there is no disadvantage to considering longer *n-grams* other than longer computation time, as the FANTASTIC toolbox automatically disregards *n-grams* which are longer than the length of a phrase.

With the current approach, we cannot address the influence of other memory effects on melody variation, such as fill-in effects, spacing effects or confusion errors. Fill-in effects (Conrad & Hull, 1964), which lead to the later inclusion of an item that was skipped earlier in serial recall, may also play a role in melody recall. This might be observed, for instance, if melodic material within a phrase or melody is rearranged, such that a motif which usually starts a melody appears later instead. With the current method, these effects would be missed, as only the amount of melodic variation, but not the kind of melodic variation, is investigated. In the same vein, the spacing effect from free recall (c.f. Hintzman, 1969; Madigan, 1969), which relates to the space between rehearsals of items, cannot be studied on the basis of phrases, which do not necessarily repeat within a melody, and if they do, usually are not spaced far apart. Instead, shorter melodic contexts might be interesting to study this effect.

Furthermore, confusion errors (Page & Norris, 1998), which in serial recall of words lead to the erroneous recall of acoustically similar words, might also be interesting to study for melody variation. This might occur if instead of a melodic phrase in a given folk song, a similar phrase from another folk song is recalled. Such confusion errors were found in a melody recall experiment by Sloboda and Parker (1985), in which participants had to memorize fragments of several folk songs: melodies recalled earlier affected the recall of melodies learned later in the experiment. As our study analyzes

variation per tune family and not across different tune families, melodic material that might correspond between different folk songs is not identified as such.

As our analysis of an existing folk song corpus highlighted some mechanisms behind melodic variation which may be tied to memorability of melodies, this shows that it would certainly be fruitful to perform more studies based on computational music analysis: such research could be performed on the present folk song corpus to investigate other potential effects of recall, or our methods could be applied to other music collections, to see whether our findings can be replicated with respect to melodic variation in other musical traditions. As such, the present collection, and other folk song collections, might be an overlooked resource to study recall and long term memory for melodies.

CONCLUSIONS AND FUTURE WORK

This dissertation provides what is, to my knowledge, the first extensive computational study investigating music transmission, through quantifying the stability of folk song phrases, and predicting this stability based on their hypothesized memorability. Through a number of choices, assumptions and formalizations, of course this study is very specific, and much more may be learned on stability, variation, and music transmission in the future. To conclude my dissertation, I provide an overview of my contributions to the three general topics of my dissertation – transmission, quantifying stability and variation, and predicting stability – as well as how the current knowledge may be extended in future work.

6.1 TRANSMISSION

As we have seen in [Chapter 1](#), music transmission has been studied in the laboratory through artificial transmission chains, while the aim of the present dissertation was to learn about stability in the transmission of folk song melodies from an existing collection of Dutch folk songs. My computational study links stability of folk song phrases to hypotheses on memorability, providing the insight that stability may be influenced by the length of a given phrase; where and how often it occurs in a melody; how much expected material such a phrase contains; and whether it consists of repeating motifs.

It would be interesting to see whether these predictors for stability would also hold in an artificial transmission chain experiment. Could the amount of variation when melodies are learned in an experimental setting also be predicted through the hypotheses tested on the folk song corpus? For instance, would highly expected melodic material be transmitted with less variation than less expected melodic material? As my computational study was performed on melodies which were recorded or notated long after the singers or editors had learned the melodies, it would also be interesting to investigate whether immediate recall of melodies in a laboratory setting leads to different kinds of variation than if melodies are recalled weeks or months later.

One application of the artificial transmission chain paradigm would be to use the same folk song phrases which were tested in the corpus study of [Chapter 5](#); another, to use whole melodies, either from the Dutch folk song database, or from other folk song collections, and perform an experiment more comparable to Klusen's (1978) work, as reported in [chapter Chapter 1](#).

A different interesting application of the artificial transmission chain paradigm would be to study how randomly constructed melodies would behave in transmission: this would be comparable to recent work on the evolution of language. Such studies include the transmission of an artificial language of nonsense syllables (Kirby, Cornish, & Smith, 2008), or of pitch trajectories (Verhoef, 2013) along an artificial transmission

chain. Such research revealed a tendency of human participants to regularize: starting from random associations between moving shapes and syllables or pitch trajectories, systematic associations between a shape and a specific syllable or pitch trajectory emerge over the course of the transmission chain.

This human tendency to impose structure unto random sequences has also recently been confirmed for music in two studies investigating random drum patterns in artificial transmission chains (Jacoby & McDermott, 2017; Ravignani & Delgado, 2016): these studies reveal that humans tend to modify a drum pattern of random time intervals such that it contains only few rhythmic categories. This finding is especially interesting in the light of the musical universals studied by Savage et al. (2015), according to which music world-wide is constructed of a limited set of rhythmic categories.

Next to study music transmission in artificial transmission chains with human participants, another interesting line of research would be the simulation of music transmission. Such work has been performed by Miranda (2008), who studied transmission of tone trajectories between robot agents, and found that a stable, shared lexicon emerged from a wide range of different trajectories. The research by MacCallum, Mauch, Burt, and Leroi (2012) may also be considered simulation of transmission. They investigated the variation of short, repeating musical sequences through transmission. These sequences, dubbed *Darwin Tunes*, were generated by algorithmic instructions on note placement, instrumentation and performance parameters, and rated by human participants in a web experiment. The pieces which were rated as most appealing by participants were used to generate new musical pieces. The researchers show that the perceived aesthetic value of the generated music increases over the generations, up until a certain plateau.

These approaches are inspiring for future experiments simulating real-world music transmission. An artificial transmission chain with agents would not need to model vocal production or auditory perception explicitly, as Miranda (2008) did for his robot agents. What could be improved over Miranda's work is to use more agents: Miranda only used five agents, which leaves open the possibility that the stable lexicon emerged through the small sample size, rather than an affordance of such an agent system in general. As for more experiments with music transmitted based on selection, rather than performance of human participants, as MacCallum et al. (2012) have done, such an experimental setup circumvents one problem that may emerge in experimental setups requiring participants to sing or tap: namely that participants may hesitate to perform music due to their insecurity about their performing abilities. If participants merely have to select a variant between several musical pieces, this may reduce such performance related artefacts. One possible improvement over MacCallum and colleagues' approach would be to not use two musical pieces to generate new musical pieces, but to combine musical material in more flexible ways, as it is unlikely that musical pieces have two "parents" in the same way that many biological organisms derive their genetic material from two ancestors.

Another strategy to study transmission is through construction of transmission paths. A recent study by Le Bomin, Lecointre, and Heyer (2016) investigated possible mu-

sic transmission paths in 58 Bantu and Pygmy societies in Gabon. To this end, they coded the presence or absence of 322 musical traits, relating to repertoire, musical instruments, rhythmical motifs, meter, scale, and performance characteristics. Through observing in how far different societies shared traits, they constructed a *phylogenetic tree*. A phylogenetic tree (or cladogram) is a tree structure with which the relationships between different species, societies or cultural artefacts can be visualized. The resulting tree had two distinct clusters which corresponded to matrilineal and patrilineal societies in Gabon; clusters at a lower level corresponded to geographical regions. The study is a great example of how music transmission can be studied in lieu of the impossibility to observe how exactly transmission progresses in a musical culture. Similar work in other musical traditions could contribute immensely to knowledge on music transmission.

6.2 QUANTIFYING STABILITY AND VARIATION

Chapter 2 examines a wide range of studies on musical variation based on the quantification of diverse musical aspects. Based on the success of earlier studies investigating sequences of local musical aspects such as pitches or chords, I conclude that note sequences are a worthwhile musical aspect to study. However, this does not mean that other musical aspects would not be interesting to investigate in terms of stability.

One important focus for future research may be single notes: this was how stability was first investigated by Bronson (1951). Such an approach would require a good technique for finding which notes in different variants of a tune family correspond to each other. This could be achieved automatically through multiple sequence alignment. While optimal multiple sequence alignment is not tractable with current algorithms, heuristic algorithms have been used for comparison of musical works (e.g. Bountouridis, Koops, Wiering, & Veltkamp, 2016). It would be interesting to see whether automatically aligned melodies indeed manifest that notes at the start and end of phrases are more stable than notes in other positions (c.f. Bronson, 1951; Klusen et al., 1978; Louhivuori, 1990).

While my overview of various studies seems to indicate that global musical aspects may not capture variation within musical traditions very well, this also does not mean that global musical aspects should be disregarded altogether. They may carry interesting information when musical traditions are compared, as geographically close traditions may share scales or meters with each other, which might be used to construct networks of shared aspects.

If variation *within* a musical tradition is to be studied, global musical aspects may also provide insights when they are considered in relation to local musical aspects. For instance, one could investigate what relationship notes which resist change hold with the underlying meter or scale. Interestingly, in meters and scales, there is also a notion of stability: notes which occur in metrically strong positions, or which define a given scale, are considered stable (Lerdahl & Jackendoff, 1983). It would be interesting to see whether stability in transmission corresponds to metric or tonal stability. A meter or

scale could then be perceived as a framework which constrains the choice of a note's onset or pitch at specific locations in a melody, comparable to Rubin's (1995) observation that rhyme and semantics form constraints for variation at specific locations in orally transmitted poetry. This can also be seen in relation to Jones' *joint accent theory* (1987), according to which melodies in which metrical and melodic accents coincide are easier to remember.

While my analysis of stability in Dutch folk songs is based on notations of folk songs, there is still a wealth of information in the many folk song recordings in the Dutch folk song database which is not available from the transcriptions. One example, the analysis of the tonic pitch chroma at which folk songs are recited (see [Appendix A](#)), reveals that the tonic pitch chroma may also be stable, at least for some tune families, as singers may have absolute pitch memory for the tonic pitch chroma in which they learned the melodies.

As with the symbolic analyses, audio based studies suggest that meaningful variation is most likely to be found in sequences of local musical aspects (c.f. Mauch et al., 2015). Specific research questions which may still be addressed with recordings from the Dutch folk song database may relate to the timing of the singers: as we have shown in [Chapter 1](#), some performances may lead to very different interpretations of timing by transcribers, which raises the question whether in these cases, the singers use very diverse rhythmic categories. Quantifying the durations between onsets might be an interesting way of investigating in how far the singers' performance may use more rhythmic categories than would be expected based on the observations by Savage et al. (2015) and Ravignani and Delgado (2016), and whether these rhythmic categories have small integer relationships with each other (c.f. Jacoby & McDermott, 2017). As automatic onset detection of singing voices is still a difficult problem for current state-of-the-art methods (Gómez & Bonada, 2013), such an analysis may not be feasible for big datasets at this stage, however.

To quantify stability of note sequences, I discussed the state of knowledge in musical pattern discovery in [Chapter 3](#). I provided a thorough overview of the field, but with the current state of knowledge it is not yet clear how successful the various methods may be for finding the stable patterns in tune families. There is ongoing research in this direction on the Dutch folk song database (van Kranenburg & Conklin, 2016), which tries to establish a connection between patterns which are unique to a given tune family, and annotated *characteristic motifs*, which according to domain experts are important melodic material in a given tune family. If future research in this vein reveals a method to reliably extract stable melodic patterns from folk songs, this would mean a comparison of such stable patterns against other melodic material in a tune family could reveal even more predictors for stability.

While my discussion of pattern discovery methods focussed on approaches for symbolically represented music, improved methods for pattern discovery from audio representations would also contribute immensely to research on transmission, variation and stability. Symbolic music analysis limits research to such music which has been notated, which is mostly Western art, folk and popular music, while improved methods

for finding patterns in audio presentations would facilitate research on repeated pitch, chord or rhythmic patterns in a much wider range of music.

Another interesting application of pattern discovery would be to discover patterns not only within, but also between tune families. In bioinformatics, common patterns between different amino acid sequences, discovered through multiple sequence alignment, are used to construct phylogenetic trees (Castresana, 2000). If we were to find common melodic patterns in a whole corpus of melodies, these patterns might also reveal relationships between melodies across tune family boundaries, and show how melodies from different tune families are related. For instance, in the Dutch song database, melodies with the same lyrics may still have been categorized into different tune families, such as *Daar was laatst een meisje loos 1* and *Daar was laatst een meisje loos 2*. Pattern correspondences might provide insights as to whether such tune families were once united, or whether the alternative melodies are closer related to other tune families in the corpus (cf. also Nettl, 2005, p. 298f.).

Pattern discovery may also be interesting to model memorability of melodies: meaningful musical patterns may provide a way to compress a musical piece such it can be described to a great extent through its patterns and their repetitions (Meredith, 2015). The resulting data compression was mentioned as a possible method of evaluating pattern discovery. Such compression could also be seen as a measure of how well a human listener may be able to recall a melody as it can be chunked into shorter repeating patterns. Hence, musical pieces which can be compressed effectively through pattern discovery might also be more stable, which links back to the motif repetitivity hypothesis of Chapter 5. By revealing repeating structures in a musical piece, pattern discovery may also be a way of modelling expectancy, as a musical pattern that occurs several times in a piece may help listeners predict the end of such a pattern when they hear the first few notes of it (c.f. Widmer, 2016).

Chapter 4 contributed a detailed investigation of various similarity measures for the problem of finding melodic phrases in folk songs. While the results presented here may provide some evidence as to which similarity measures may be successful for finding note sequences in longer melodic contexts, my conclusions are still very specific for the particular problem of finding similar phrases within tune families of Dutch folk songs. Research replicating the comparison of similarity measures for other research goals, and in other music corpora, would contribute to learning more about how human judgements on occurrences of note sequences might be modelled by computational comparison. Possible future applications of the compared similarity measures would be to analyze stability of shorter note sequences, e.g., licks in jazz improvisation, or to analyze stability of phrases in other folk song corpora. As most of the compared similarity measures are defined for monophonic music, other applications for similarity measures, such as finding musical quotations in Classical works, might still require some research into successful measures for polyphonic music.

Another application of the compared similarity measures could be to investigate the annotations of “related but varied” phrase occurrences. These annotations in MTC-ANN2.0 have not been studied so far, but may reveal much more about stability and

variation: namely what type of variation introduced through transmission is perceived as similar enough to be still considered a phrase variant, rather than a different phrase altogether. This question connects to Nettl's recommendation to investigate how a musical tradition keeps musical patterns "intact" (Nettl, 2005, p.295), and relates to ongoing research on musical similarity (Volk & van Kranenburg, 2012).

One of the outcomes of our evaluation of similarity measures is the considerable variance in error rate, depending on the analyzed tune family. While this was taken into account in the regression model in Chapter 5, the implication of this outcome still need to be investigated. The variance in the error rate suggests that in some tune families, occurrences perceived by humans are easier to capture through computational comparison than in others. What underlying factors account for the difficulty with which computational similarity measures detect occurrences in different tune families is still an open question. Addressing this question may reveal that stability behaves differently in different tune families, as some important similarity relationships may not be possible to detect through comparison of pitch and duration alone. Stability might then be perceived in other musical aspects, to which the annotators, but not the similarity measures, were sensitive.

6.3 PREDICTING STABILITY

Chapter 5 showed a number of factors which determine memorability of folk song phrases, and therefore predict their stability to some extent. With this result, we can reject the possibility that stability of Dutch folk song phrases is entirely random; this said, it is also notable that only a small percentage of stability or variability of folk song phrases can be predicted through the proposed hypotheses – a large amount of variation in the data still eludes such prediction. More studies along the lines of the previous chapter are therefore needed, potentially testing more hypotheses based on recall studies, or on theories from music cognition.

However, for understanding stability and variation in music transmission, one might also look at models from cultural evolution. Henrich and Boyd (2002) propose several mechanisms by which specific cultural traits are favoured over others. They call such tendencies of individuals or groups to prefer a given trait over another a *bias* which influences the distribution of traits throughout transmission. Of the proposed bias categories, the hypotheses tested in Chapter 5 all belong to the category of *content bias*, i.e., the content of a melodic phrase itself determines whether it is stable. Another bias category proposed by Henrich and Boyd (2002), *prestige bias*, relates to the individuals participating in transmission: some may be seen as more successful, and therefore other individuals may be more likely to copy from them. We can imagine such a scenario in Classical music, where recordings by famous performers of a given composition are studied in detail by aspiring performers, meaning that the most prestigious performers' choices for expressive timing and dynamics would be more likely to be imitated than those by less renowned musicians.

Another bias category, *conformity bias*, predicts that the most frequent cultural trait in a given society will be copied by individuals. Of course, there is already a tendency of more frequent traits to be copied as it is more likely an individual is familiar with them; conformity bias means that individuals, being familiar with several alternatives, pick the one that they observe most frequently, exceeding the effects which can be purely explained by drift, i.e., change introduced through random sampling. In music, conformity bias and prestige bias may not be easy to disentangle: musical pieces which are successful often get more airplay, and renowned musicians give more concerts. Nevertheless, there may be some situations in which the two effects might be separated, for instance when music lovers reject successful music or musicians in favour of music or musicians they perceive as more skilful.

To study alternative predictions for stability of cultural evolutionary models, one would need to study different music collections, however: to study conformity bias, one would need to perform a diachronic analysis, by counting different variants of a folk song or phrase at a given time, and observing whether the more frequent variants are reproduced at a later stage more often than would be expected purely through drift. As the Dutch folk song database mostly contains songs recorded between 1950 and 1970, with few earlier versions from song books, it is much more suitable for synchronic analysis. Salganik, Dodds, and Watts (2006) provide some evidence for the conformity bias in music sales: in a simulated music market, individuals were more likely to download songs that had been previously downloaded by peers. The authors acknowledge that this simulated market does not necessarily model real music markets, as individuals had no other information about songs than which songs had been downloaded before, while real consumers will likely learn about songs and musicians via other channels than mere download statistics. Still, conformity bias may play a role in the choice of specific musical pieces, or variations therein, over others.

An interesting way of studying conformity bias in real music markets would be to use data from the Billboard Hot 100 charts, which reflect the success of popular music singles based on airplay, sales, downloads and streaming. One might then investigate whether songs which are demonstrably widespread are more likely to be covered by other musicians. Such a research would also require information on cover songs, which could be retrieved from crowd-sourced data bases such as *secondhandsongs*. To my knowledge, such a research has not been performed yet, but is also not entirely straightforward, as music consumption has changed over the past few decades, and therefore the way success of particular singles is determined by the Billboard Hot 100 has changed as well. As an alternative, artificial transmission chain experiments could be used to determine the influence of conformity bias: if participants hear different versions of a song, their choice to sing a particular variant over another might reveal whether they adhere to conformity bias.

To investigate prestige bias, one would need information on what musicians are perceived as most successful by their peers. Even though the Dutch folk song database contains biographies of some singers, which might give occasional insights into the prestige of the singers, this information is by no means complete, and not structured

enough for computational analysis. Analysis of cover songs in popular music might also reveal prestige bias, if fairly unknown but acclaimed artists were chosen as models over other, more successful artists.

My study in [Chapter 5](#) focusses on one musical tradition, and the methods and findings from this dissertation cannot be necessarily applied for other musical traditions. Testing the hypotheses I used to predict stability in Dutch folk songs in other music traditions would be an invaluable step towards closing the “micro-macro gap” (Mesoudi, 2011), i.e., to understand which phenomena of transmission are tradition-specific, and which phenomena may play a role in all known musical traditions. Another approach to narrowing the micro-macro gap would be to investigate in how far the melodic phrases I investigated can be found across musical traditions. For example, would a highly stable melodic phrase within one tune family also occur in other tune families, or in melodies from other folk song traditions?

Next to the micro-macro gap, I would also postulate the presence of another gap in research on music transmission: the gap between trends of variation – which I researched in this dissertation – as opposed to the boundaries of variation, which rely on the cognitive capabilities of human listeners to perceive, remember and perform music. These capabilities may be researched through cross-cultural comparison and iterated learning, as discussed before. Moreover, comparative research on music and language, as well as comparative research across species may contribute to understand *musicality* – the ability to perceive music – and how it shapes musical traditions (c.f. Honing, ten Cate, Peretz, & Trehub, 2015; Rohrmeier, Zuidema, Wiggins, & Scharff, 2015). Musicality determines the boundaries of variation: some forms of musical variation may be too difficult to remember or perceive, such that they are never circulated in transmission. Raffman (2003) makes such a case for twelve-tone music, claiming that human musicality does not afford perception of serial structure, and scales of so many distinct pitches, such that it never caught on with the wider public. While this work on understanding the boundaries of variation, of pondering the possibility of impossible music (c.f. Hauser, 2009), is ongoing and fascinating, mapping out the trajectories of transmission, and tendencies of variation within the space of possible music is another invaluable course of research.

My dissertation has been a contribution to this last endeavour. Some of my findings, such as that stable melodies tend to consist of small pitch intervals, and of repeating motifs, may form the link to musicality: small pitch intervals and repeating motifs have also attested as statistical universals in the world’s musical traditions by Savage et al. (2015). Other predictors of stability, such as melodic expectancy, also relate to musicality in general, in which prediction plays an integral role (Huron, 2007). It would be fascinating to find that the same properties which afford stability, constituting trends of variation, also are part and parcel of human musicality, defining the boundaries of variation.

Part III

APPENDIX



THE ROLE OF ABSOLUTE PITCH MEMORY IN THE ORAL TRANSMISSION OF FOLKSONGS

While the ability to instantly identify and label an isolated tone as being a particular note in the tonal system is very rare (Takeuchi & Hulse, 1993), research suggests that memory for absolute pitch information is in fact widespread (e.g., Levitin, 1994; Schellenberg & Trehub, 2003). Expanding on these earlier studies, in this study, pitches of Dutch folksong recordings from the *Onder de Groene Linde* collection that are available via the Meertens Tune Collections¹ were analyzed.

The goal of the study was to determine whether there is consistency in sung tonic pitch chroma across individuals when singing the same folksong independently of each other, across place and time. The results show that there is indeed some *inter-recording tonic pitch consistency* in the recordings of a small collection of folksongs. Inter-recording tonic pitch consistency is consistency in sung tonic pitch chroma across individuals when singing the same folk song independently of each other. As such, this is the first study to suggest that Absolute Pitch Memory (APM) plays a role in oral transmission of folksongs.

Our working hypothesis is that all of the tune families should show some level of inter-recording tonic pitch consistency as a sign that the melodies were memorized and transmitted on the basis of absolute pitch height, instead of just melodic contour (Dowling & Fujitani, 1971). If the empirical data supports this hypothesis, this would imply a role for APM in the oral transmission of folksongs, as suggested by earlier studies on the subject (e.g., Halpern, 1989; Levitin, 1994). The alternative hypothesis predicts that none of the investigated songs show inter-recording tonic pitch consistency, and instead the sung tonic pitches can be expected to be uniformly distributed over recordings. Interpretations of these possible outcomes will be discussed later on in this paper.

Below, we will first provide some background information on the topic of absolute pitch. After elaborating on the differences between two types of absolute pitch, AP as opposed to APM, we will present two lines of research on absolute pitch, followed by different theories on auditory memory, and an overview of the musical material used in this study. Subsequently, we will present our methods and results, followed by a discussion. Finally, we will present our conclusions and provide some suggestions for future research.

¹ www.liederenbank.nl/mtc

A.1 BACKGROUND

A.1.1 *Traditional Absolute Pitch Versus Absolute Pitch Memory*

Traditional Absolute Pitch (AP) – the ability to instantly identify and produce a certain tone without any reference note – is extremely rare; only 1 in 10000 individuals have this ability (Takeuchi & Hulse, 1993), and as a result it has been termed a gift by some researchers (e.g., Athos et al., 2007; Bachem, 1940; Gregersen, Kowalsky, Kohn, & Marvin, 2001). Others insist that anyone has the potential to acquire this traditional sense of AP, but that training in a critical period is needed to fully acquire the skill (e.g., Vitouch, 2003). Interestingly, all infants process pitch information in an absolute fashion, and some researchers have suggested that humans gradually shift to relative pitch processing when they get older and, as a result, lose their absolute pitch processing abilities (Saffran & Griepentrog, 2001). More recent evidence suggests that infants are in fact capable of relative pitch processing from early on as well, if a slightly different task is used (Plantinga & Trainor, 2005). If infants possess both relative and absolute pitch processing capabilities, and use them both depending on task requirements, why would adults lose either of the two?

Perhaps adults do not lose either of the processing capabilities mentioned above. The past 25 years or so, researchers have partly shifted their attention to what some experts call “absolute tonality” (Vitouch, 2003) or “residual AP” (Deutsch, 2002). This has resulted in an increasing body of evidence in support of the notion that the capability to store absolute pitch information is in fact retained for melodies, even for adults with little formal musical training (Levitin, 1994; Schellenberg & Trehub, 2003; Terhardt & Ward, 1982; Vitouch & Gaugusch, 2000). Levitin (1994) has argued that AP possessors in the traditional sense – as opposed to the non-possessors – are able to label isolated pitches verbally, but that anyone possesses absolute memory for pitch (APM) in melodies. This claim – that APM is in fact widespread – is supported by evidence from two types of research. One of these lines of research uses identification tasks to address the issue, whereas the other uses production tasks. Some related work of both types of research will be discussed in the next section.

A.1.2 *Related Work on Absolute Pitch Memory*

Identification experiments use an experimental design in which participants are presented with two versions of an excerpt of a musical piece. One of these two versions is played to them in the correct key, while the other is transposed (“shifted”) upward or downward by a certain number of semitones. The participants then have to judge which version is the correct one.

An example of such an experiment was performed by Schellenberg and Trehub (2003). Even those participants with little formal musical training (and no reported AP capabilities) were able to identify the correct version of familiar TV program tunes highly above chance. When the difference in tonic pitch between the correct and the

incorrect version was one semitone, the participants chose the correct version 58% of times. If there was a two semitone difference between the excerpts, the participants even identified the correct version 70% of times. These results show that at least some APM seems to be retained into adulthood.

The second type of experiment investigates to what degree participants are able to produce or reproduce a melody in the correct key. Levitin (1994) showed that musically untrained participants achieved above chance pitch accuracy in a production task. In two trials, students with little formal musical training were asked to select a familiar song from CDs that Levitin provided and asked the participants to sing a self-selected part of this song on the correct pitch chroma (correct being: the pitch as it was recorded on the CD). About 40% of participants sang familiar rock tunes on the correct pitch in at least one of two trials, suggesting that even singers with little to no formal musical training are to some extent able to produce or reproduce these pitches from memory. These results again indicate that APM is widespread.

More recently, Frieler et al. (2013) were able to replicate Levitin's experiment in six European labs, though with effect sizes that were slightly smaller than those reported in the original study. Findings varied widely between different laboratories, and as a result the authors stressed the importance of replication in music psychology research.

Additional evidence for APM using a production task design comes from within-subject experiments done by Bergeson and Trehub (2002) and Halpern (1989). In the first study, mothers were asked to sing a song to their infant that they would also sing to their infant in a non-experimental setting. They were instructed to sing the same song again a week later. Tonic pitch chroma was measured on both occasions. The mean tonic pitch deviation of the second, as compared to the first performance was less than a semitone. In Halpern's (1989) experiment, participants had to sing the opening tones of holiday and children songs in two trials. The opening tones of their second performance only deviated two semitones on average, compared to their first performance.

These last two studies by Bergeson and Trehub (2002) and Halpern (1989) are especially interesting for our study, because the type of songs that the participants had to sing in their studies closely resemble the type of songs used in the current study, in that there was no standardized version available to the participants. In other words, participants did not have an external reference such as sheet music or a recording. Instead, the absolute pitch information was retrieved from memory only, and therefore this memory of earlier performances, among which the participants' performance in the first trial, served as an internal reference for their second performance.

A.1.3 *Song Memory*

How do singers recall melodies from long term memory? Some researchers have proposed that attributes such as pitch, tempo, and timbre are stored in a multiple-trace memory system (Levitin & Rogers, 2005) that is connected to a perceptual system (Dalla Bella, Peretz, & Aronoff, 2003). Levitin and Rogers suggest that memory

for AP is a low-level feature and co-exists with an abstract memory for other features such as the sequence of relative pitches in the melody, and most likely the emotion associated with the piece as well (Eschrich, Münte, & Altenmüller, 2008).

Furthermore, it has been suggested that imagery plays a role in the retrieval of a memory. Before and while singing a melody, one is actively imagining the melody. When trying to recall and reproduce it, one is combining all features, such as absolute and relative pitch, tempo, and timbre into the end product, the melody itself, while getting immediate perceptual feedback from one's own voice. The more often a melody has been rehearsed, the more accurate the representations become (Keller, Cowan, & Saults, 1995). This in turn results in better recall and thus reproduction of the melody. This could be a reason why musicians often perform better in song (re)production experiments. All this is consistent with neurological evidence from Zatorre and Halpern (1993), demonstrating that once auditory cortical areas have been damaged, for example due to a focal auditory cortex lesion, perceptual and imagery deficits co-exist, resulting in worse retrieval of memories that include auditory features.

Finally, procedural memory has been suggested to be involved in the retrieval and (re)production of a melody representation (Brown & Palmer, 2012). Located in the motor cortex, a specific area representing the vocal tract could potentially be connected to the several representations of the melodies stored in memory as well as the vocal tract itself. In this scenario, it is as though the vocal tract "remembers" the exact tension associated with each pitch in the representation. However, this possibility thus far remains mostly theoretical, as little evidence of such procedural memory has been presented in the vocal domain (see, e.g., Brown & Palmer, 2012). Also, as Levitin (2013) has recently noted, procedural memory cannot solely account for the accurate pitch reproduction adults are capable of.

A.1.4 *Material*

All recordings were part of the *Onder de Groene Linde* (OGL) collection as recorded by Will Scheepers and Ate Doornbosch from the 1950s and onwards are available online via the Meertens Tune Collections. These recordings feature mostly older adults with little to no formal musical training, singing Dutch folk songs from memory only. Like in some of the production studies mentioned earlier (e.g., Bergeson & Trehub, 2002), there was no standardized version available for these songs.

Experts have grouped the recordings in the Meertens Tune Collections into tune families of closely related variants, which enabled us to compare the tonic pitch of different variants to each other, and thereby explore the potential role of APM in the memory of songs transmitted in oral traditions.

The oral transmission of folksongs is, besides factors such as perception, performance and creativity, highly influenced by memory (van Kranenburg, 2010). It is thus important to show that there is at least some pitch consistency on a between-subject scale. Once one is able to define the components of auditory memory that play a cen-

tral role in oral transmission, one can study these components and use the resulting knowledge for a theory of oral transmission of folksongs.

Analyses on two partly different datasets were done to investigate the role of APM in folksongs. Dataset A was used for a between tune family analysis examining the inter-recording tonic pitch consistency of variants from seven tune families. Dataset B consists of a larger set of recordings of two tune families, allowing for a within analysis of subsets grouped according to three factors that might influence the songs' inter-recording tonic pitch consistency.

The first factor that could possibly influence our results is the lyrics of the song. Different textual versions of one melody may belong to the same tune family. Contextual cues and affect (Bergeson & Trehub, 2002) have been proposed to facilitate retrieval of performance details of musical melodies. One of these contextual cues could be the lyrics accompanying the melody (cf. Levitin & Rogers, 2005). Secondly, gender may influence the pitch on which a folksong is usually sung, because of the different typical ranges of men's and women's voices (Titze, 1989). Thirdly, if there is a role for APM in oral transmission of folksongs, the tonic pitch of folksongs might develop independently in different geographical regions, predicting possible different mean tonic pitches in recordings from different regions. In the ensuing section, the methods and the results of the between tune family analysis will be described.

A.2 DATASET A: BETWEEN TUNE FAMILY ANALYSIS

A.2.1 *Method*

To investigate the role of APM in folksongs, we analyzed the tonic pitches sung in the recordings available via the Meertens Tune Collections post hoc, using comparable methods to those used in the production studies mentioned earlier in this paper (e.g., Bergeson & Trehub, 2002; Halpern, 1989). The recordings in the Meertens Tune Collections that were used for analysis are of various lengths, ranging from one to ten minutes. We decided to compare the first verses of each recording, as other factors such as pitch chroma fluctuations (i.e., drift) might have intervened if first verses had been compared with later verses. Moreover, there is evidence that the first verse best reflects the representation of the tonic pitch retrieved from memory (Halpern, 1989; Klinger, Campbell, & Goolsby, 1998; D. S. Smith, 1991).

In our between tune family analysis, we examined sung tonic pitches in recordings of 7 tune families that were chosen from the Meertens Tune Collections. The tune families selected for the between tune family analysis were: *Daar was laatst een meisje loos* (N = 20), *Er reed er eens een ruiter* (N = 20), *Het was laatst op een zomerdag* (N = 20), *Al is ons prinsje nog zo klein* (N = 19), *Het vrouwtje van Stavoren* (N = 20), *Mijn vader zei laatst tegen mij* (N = 20), *Wat hoor ik hier in het midden van de nacht* (N = 20). The particular tune families used were chosen based on the criterion that there were at least 20 recordings of these tunes available [5]. One recording of *Al was ons prinsje nog zo klein* was removed from the data set, because it contained an instrument playing the

tune rather than someone singing it, leaving 19 recordings of this tune family for data analysis.

Yin, a pitch detection algorithm (De Cheveigné & Kawahara, 2002), was used to detect the fundamental frequencies of various pitches sung in the recordings. From the output of this algorithm, a density estimation of the distribution of pitches can be computed using a Gaussian Kernel following the method of Biró, van Kranenburg, Ness, Tzanetakis, and Volk (2012). The first author then checked the frequencies achieving the highest density with the help of a tone generator², to determine which of these frequencies represented the tonic frequency.

A.2.2 *Quantitative analysis with circular statistics*

Pitch can be represented in several ways. We will adopt the representation as proposed by Shepard (1982) who distinguished between pitch height and pitch chroma. The latter is a common grouping principle (also referred to as pitch class), where a tone of, e.g., 440 Hz (orchestral A₄) is considered the same as 220 Hz (A₃) or 880Hz (A₅). Since there is also some evidence that the brain represents pitch along these two dimensions (Warren et al., 2003), we used pitch chroma (or octave normalization) as a way to group the pitches analyzed.

Because of the circular nature of octave normalized pitch (Levitin, 1994) a circular test such as the Rayleigh test was needed to determine or reject uniformity of the data (Fisher, 1995). We assured that the investigated pitch distributions fit a unimodal distribution through Hartigan's dip test (Hartigan & Hartigan, 1985) before performing the Rayleigh test, which presupposes unimodal distribution. If Rayleigh's test returns a result equal to or below $p \leq .05$, the distribution of the sung pitches significantly deviates from a uniform distribution, in favour of a unimodal distribution.

To convert our measurements (fundamental frequency of the tonic) to a scale suitable for circular statistics, we first converted our frequencies *freq* to MIDI note numbers, an equivalent of pitch on a linear scale. Because an octave can be divided in twelve semitones, or in this case, twelve MIDI note numbers, every MIDI note number can be thought of as 30 degrees. Therefore the MIDI note numbers were then in turn converted to angles θ (in degrees) by multiplying them by 30, ensuring successful octave generalization for our obtained results.

$$\theta = 12 \cdot (\log(\text{freq}/220)/\log(2)) + 57) \cdot 30, \quad (\text{A.1})$$

where n = MIDI tones and f = measured tonic frequency in Hertz.

For the analysis we used Oriana³, a tool specially developed for this type of statistics.

² <http://www.audionotch.com/app/tune/>

³ Windows version 4.01, Kovach Computing Services, Pentraeth, Anglesey, Wales, U.K.

A.2.3 *Baseline*

To determine how much of the tonic pitch consistency might be caused by the vocal range of the singers, and the melodic range of the folk songs, we ran a Monte Carlo simulation of each tune family. Vocal ranges used for this simulation were based on Moore's (1991) analysis of singing ranges of geriatric persons of the ages between 60 and 110, as such roughly matching the singers recorded in the Meertens database.

The average vocal range for male singers was set to G₂ to D₄, the average female range was set to F₃ to C₅. As the female to male ratio in our dataset was 4:1, we gave higher probability ($\mathbb{P} = .8$) to a female vocal range. The vocal range was represented in MIDI note numbers, on a continuous scale.

The melodic ranges of the simulated songs were sampled from the available transcriptions for the tune families used in this study. The pitches of the first five notes of each transcription were examined to determine the highest and lowest starting pitch. The melodic range was represented as semitone steps between the highest and lowest starting pitch.

The simulation of the tonic pitch departed from the lowest starting pitch, such that the melodic range fit into the vocal range. For instance, if the vocal range sampled was that of a female singer ([53.0, 72.0] in continuous MIDI note numbers) and the melodic range was seven semitones, the lowest starting pitch would be sampled from values between 53.0 and 65.0. Finally, the position of the nearest tonic pitch above the lowest starting pitch – information that was also derived from the transcriptions – was determined, providing the sampled tonic pitch.

We ran 100 simulations of 20 tonic pitches, and on each simulation, we performed Hartigan's dip test. If the test implied a non-unimodal distribution, we excluded the simulation. Otherwise, we performed the Rayleigh test, of which the resulting p values were collected. The mean p value of the Rayleigh tests on the Monte Carlo simulations indicate whether there is a bias for an unimodal distribution, based on the vocal range of the singers, and melodic range of the songs.

A.2.4 *Results*

The results of the between tune family analysis are shown in A.1. The null hypothesis of bimodal or multimodal distribution could be rejected for all tune families ($p > 0.6$ throughout). The null hypothesis of uniformity was rejected for the tune family *Daar was laatst een meisje loos*, $r = .49$, $p < .01$. The null hypothesis of uniformity could not be rejected however for *Er reed er eens een ruiter*, $r = .35$, $p = .08$, *Het was laatst op een zomerdag*, $r = .16$, $p = .61$, *Het vrouwtje van Stavoren*, $r = .19$, $p = .51$, *Al was ons prinsje nog zo klein*, $r = .24$, $p = .34$, *Mijn vader zei laatst tegen mij*, $r = .28$, $p = .21$ and *Wat hoor ik hier in het midden van de nacht*, $r = .12$, $p = .74$.

The significant results of *Loos* indicate that there may be some APM involved in the recall and reproduction of this particular folksong. The baseline simulation of the vocal and melodic range yielded an average p-value of $p=0.39$ on the Rayleigh tests. Three

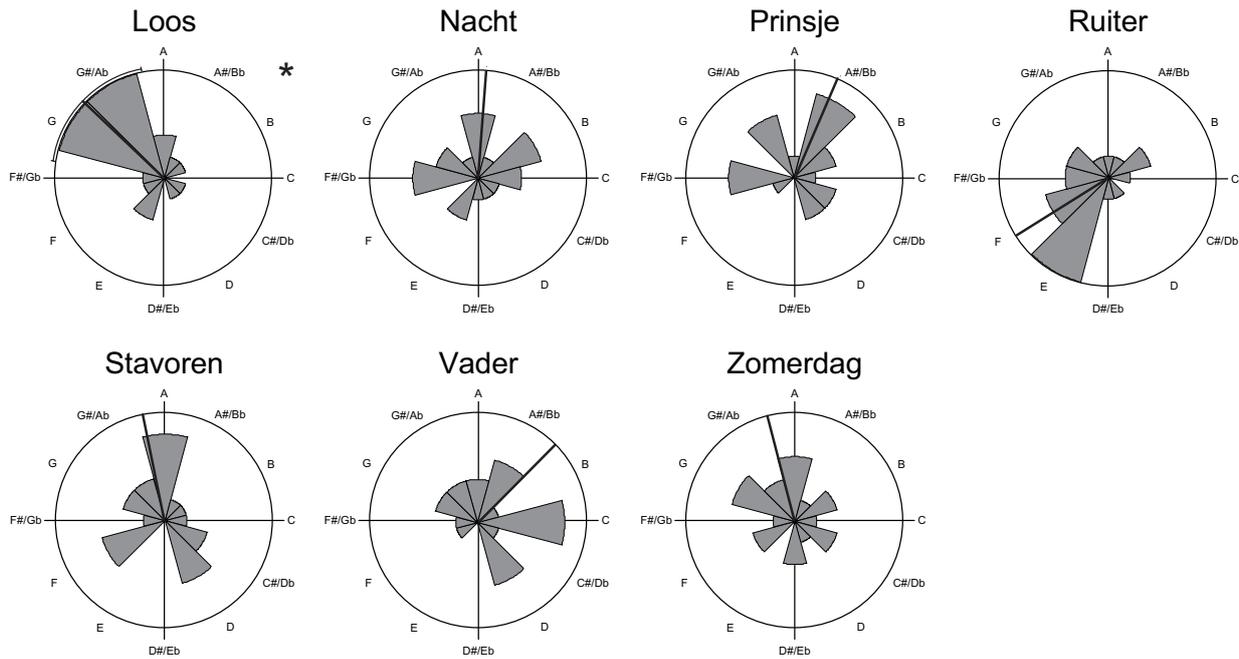


Figure A.1: Between tune family results. Circular histograms representing the distribution of pitch chroma for seven tune families. Arrows indicate the mean vector. Significant results are marked with an asterisk and confidence interval. The legend shows how pitch chroma is mapped on degrees in the diagram.

simulations were excluded due to a significant result of Hartigan's dip test. This implies that the evidence for unimodal distribution is not likely caused by the restrictions of the singers' vocal range and the song's melodic range.

Therefore, the significant results of *Loos* indicate that there may be some APM involved in the recall and reproduction of this particular folksong. The remaining tune families may not have yielded significant results for various possible reasons: among others, the dataset might simply not be large enough to observe significant deviations from the uniform distribution, because of possible factors of variance such as lyrics, gender and geographical origin involved. To test these assumptions, two of the tune families of which there were more recordings available were again used in our within tune family analysis.

A.3 DATASET B: BETWEEN AND WITHIN TUNE FAMILY ANALYSIS

A.3.1 Method

Dataset B consisted of two tune families, *Wat hoor ik hier in het midden van de nacht* ($N = 53$) and *Mijn vader zei laatst tegen mij* ($N = 67$). The procedure used to determine the fundamental frequencies of the first verse of these recordings was similar to the one used for dataset A. This time, gender of the singer(s), geographical origin, and difference in lyrics of the recordings were noted for additional analysis.

A.3.2 *Baseline*

The full tune families and the subsets of female and male singers, of geographical origin and lyrics were also subjected to a simulation of starting pitches, corresponding to the simulations for dataset A. The sample sizes in the simulations were set to correspond to the sizes of the investigated datasets.

A.3.3 *Results*

The results of the between and within tune family analysis performed on dataset B are shown in A.2, this time comparing a greater number of variants than reported in A.1. Following a Hartigan's dip test, the null hypothesis of bi- or multimodality was rejected for all subsets of the tune families, with the exception of the geographic subset of the province Drenthe, for which Hartigan's dip test yielded $D = 0.1191$ and $p < 0.03$. In consequence, we performed Rao's spacing test (Jammalamadaka & SenGupta, 2001) instead of the Rayleigh test to investigate the distribution.

The null hypothesis of uniformity was rejected for *Mijn vader zei laatst tegen mij* ($N = 67$), $r = .26$, $p < .01$. It could not be rejected for *Wat hoor ik hier in het midden van de nacht* ($N = 53$), $r = .35$, $p = .06$. In the simulation of starting pitches of *Vader*, the Rayleigh tests achieved an average p -value of $p = 0.36$, indicating that there is no tendency towards uniform distribution based on vocal range and melodic range alone.

Next, additional analyses were done for *Mijn vader zei laatst tegen mij* to determine whether there was any effect of the lyrics that were sung. There were two major textual versions among the recordings, *Text Vader* ($N = 33$) and *Text Boerenzoons* ($N = 32$). A Rayleigh test was conducted to determine uniformity. The null hypothesis of uniformity was rejected for both *Text Vader*, $r = .36$, $p < .02$ and *Text Boerenzoons*, $r = .34$, $p < .02$. Interestingly, the means (17 and 111 degrees, respectively, see A.2) for both textual versions were as much as 3 semitones apart, even though range and contour of both versions were highly similar. The simulation of vocal range did not show a bias for unimodal distribution, with the mean p value at $p = 0.47$ for *Text Vader*, and at $p = 0.55$ for *Text Boerenzoons*. For both subsets, the dip test indicated unimodality for all simulations.

Gender might also be a possible factor of variance in the sung pitches, as discussed in the introduction. Based on the Rayleigh test, the null hypothesis of uniformity was rejected for recordings of *Mijn vader zei laatst tegen mij* that were sung by female singers, $r = .29$, $p < .02$ ($N = 49$) but not for recordings the same song that were sung by male singers, $r = .24$, $p = .49$ ($N = 12$). Similarly, the null hypothesis of uniformity was rejected for recordings of *Wat hoor ik hier in het midden van de nacht* that were sung by female singers, $r = .39$, $p < .01$ ($N = 33$) but not for recordings of *Wat hoor ik hier in het midden van de nacht* that were sung by male singers, $r = .20$, $p = .56$ ($N = 14$). This implies that the pitch consistency in female recordings is remarkably higher than the consistency in male recordings. However, there were relatively few male singers, which could have contributed to the non-significant results. The baseline simulations of the

male and female subgroups showed no tendency towards significant p-values on the Rayleigh tests, with means of $p=0.36$ for the female subgroup of *Vader* (two excluded simulations due to significant dip test), and $p = 0.15$ for the female subgroup of *Nacht* (no excluded simulations).

One last factor that we included in our analysis was geographical origin. In the Meertens Tune Collections, the origin of the recordings is also included in the description. There are twelve provinces in the Netherlands (see A.3), most of which have their very own identity and culture. Most provinces did not have enough data points to do a statistical analysis based on geographical origin, even in our within tune family dataset. Two provinces however did have a reasonable number of recordings to look for a trend in the data. This was the case for particular tune families in the second dataset: Groningen for *Mijn vader zei laatst tegen mij* ($N = 22$) and Drenthe for *Wat hoor ik hier in het midden van de nacht* ($N = 16$) (see A.3). A Rayleigh test was conducted for recordings from both provinces to determine uniformity when geographical origin was included as a factor. The null hypothesis of uniformity was rejected for recordings of *Mijn vader zei laatst tegen mij* originating from Groningen, $r = .38$, $p < .05$, but could not be rejected for recordings of *Wat hoor ik hier in het midden van de nacht* from Drenthe, $p > .1$ (see A.2). For both cases, the simulations of the geographical subsets resulted in mostly non-significant p values. The mean p-value for the Rayleigh tests on the Groningen subset for *Vader* were $p = 0.5$, and for the Drenthe subset of *Nacht* $p = 0.62$. No simulations were excluded for the geographical subsets because of non-unimodal distributions. This supports our assumption that higher pitch consistencies are caused by geographical factors, rather than by the melodic ranges of the songs.

A.4 DISCUSSION

A.4.1 *A Role for Absolute Pitch Memory in Oral Transmission of Folk Songs*

The results indicate that absolute pitch memory has a role to play in oral transmission of folksongs. As there was no standardized version of these folksongs available to the singers, these singers reproduced the songs solely from memory. However, not all tune families showed inter-recording tonic pitch consistency. In this discussion, several potential explanations for these mixed results are discussed.

If there is no standard version, one might expect that subjects sing the tune on a different tonic pitch every time, and the absolute pitch information is not reinforced and stored as part of the mental representation of the song. In that case a standardized version of the song might be needed as a reference to establish absolute pitch information in memory, as Levitin (1994) already hypothesized. This reference could be the original version, played by musical instruments, or a written standard version of the song.

However, a standard version often does not exist for folksongs, and especially not for those in the *Onder de Groene Linde* collection, because its collectors, Will Scheepers and Ate Doornbosch, had a preference for folksongs which were not explicitly learned,

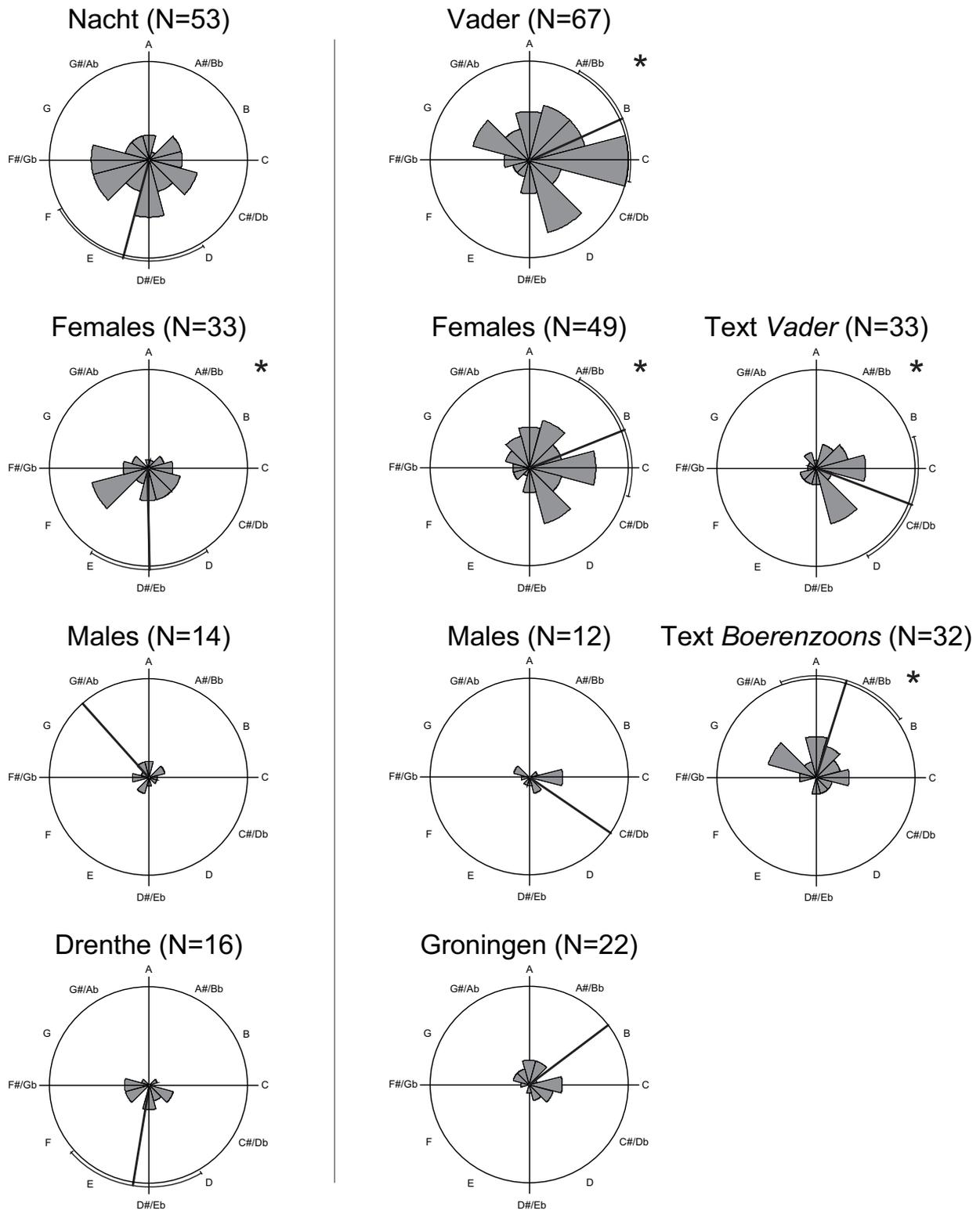


Figure A.2: Within tune family results. Circular histograms representing the distribution of pitch chroma for the tune families *Nacht* (left panel) and *Vader* (right panel). Arrows indicate the mean vector. Significant results are marked with an asterisk and confidence interval. The legend shows how pitch chroma is mapped on degrees in the diagram. N.B. Subsets of textual versions and gender do not add up to the complete dataset because there were some recordings with other texts, and unspecified gender of the singers.

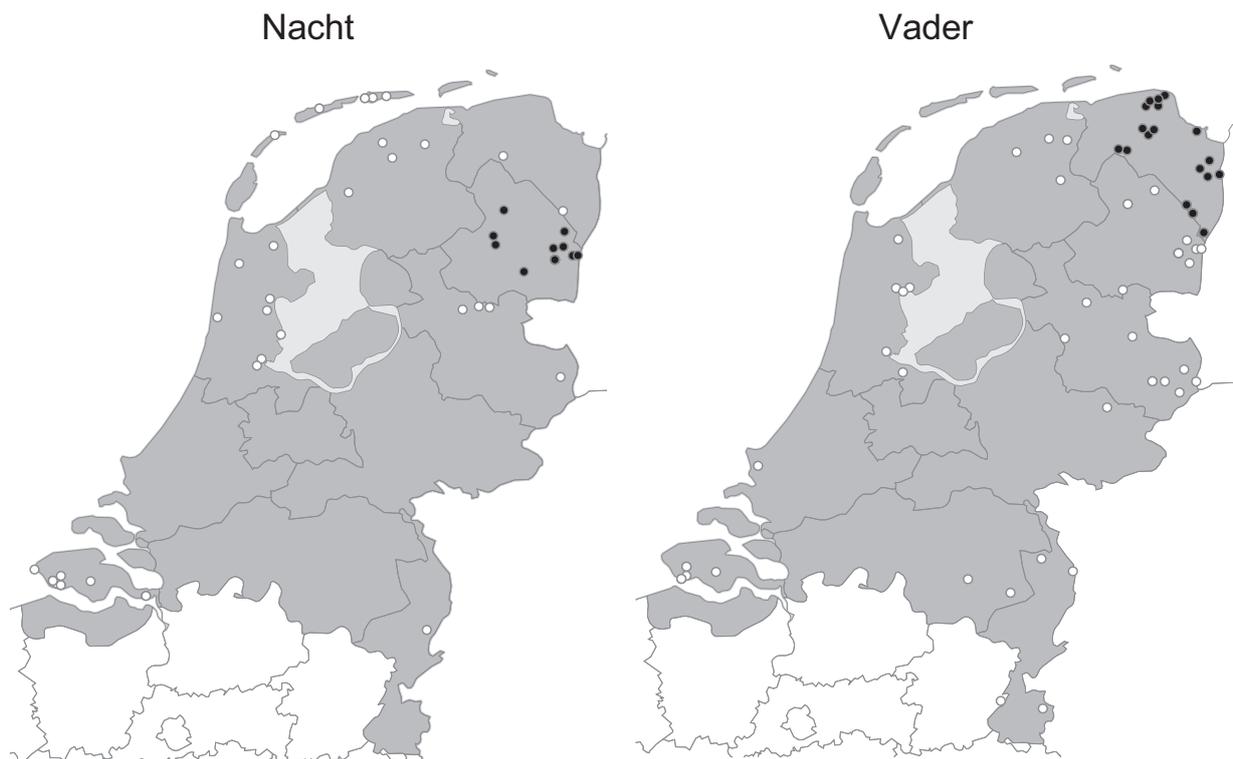


Figure A.3: Geographical distribution of the recordings used for *Nacht* (left panel) and *Vader* (right panel) over the Netherlands. Black dots mark recordings in a province that have a reasonable number of recordings for analysis (i.e., Drenthe for *Nacht* and Groningen for *Vader*), white dots show the remaining recording locations from the respective tune families. N.B. A single dot can represent multiple recordings.

for instance at school or from church song books. *Daar was laatst een meisje loos* might be an exception, as it did appear in a standardized fashion in songbooks from the early 20th century onwards. Interestingly, this was the only song showing significant inter-recording tonic pitch consistency in our between tune family analysis. However, this does not explain why the two larger tune families, which both did not have a standardized version, did show inter-recording tonic pitch consistency after controlling for certain variables such as gender, which we will discuss later.

Another explanation for the mixed results could be that the singers were not able to produce the tone they wanted to produce Levitin (1994). In that case, the mental representation of the melody might have been correct, but the singers were unsuccessful in reproducing the tone they had in mind. These unsuccessful attempts could be due to physical tone production problems (Levitin, 1994), or a lack of musical training. In general, pitch memory and the perception of pitch relations are a function of musical experience (Krumhansl, 2000; Schellenberg & Trehub, 2003). The singers in the recordings of the *Onder de Groene Linde* collection had mostly no formal musical training, and therefore might have been less precise in reproducing the tone they had in mind.

Through our baseline study we show that the range of the starting pitches of a melody do not seem to constrain the choice of pitch chroma significantly. However, one aspect to investigate in the future is whether the contour of the starting pitches has an influence on the chosen pitch height. For example, a pitch jump of a fifth following the starting pitch might constrain the singer's choices in a different way than an ascending scale spanning a fifth.

There are several other reasons for a possible lack of pitch consistency in the recordings, more related to the particular materials used in this study. For example, in most of their recording sessions, Doornbosch and Scheepers asked the singers to sing more than one folksong. If there is no pitch consistency to be found, this might be due to influences of performances of other songs during the same recording session. According to Levitin (1994), the tonal center of the first song that was sung during a session might simply determine the tonal center of the second song as well.

We cannot compare our results in terms of effect sizes to Levitin's (1994) and Frieler's (2013) results. Frieler and colleagues estimated the power and effect sizes of their results by testing the number of 'hits' in their trials by means of a Goodness-of-Fit χ^2 -test. However, due to the fact that there is no standardized version available of the folk songs sung in the recordings used in our study, we do not have 'hits', and our analysis relies purely on the distribution of tonic pitches.

A.4.2 Gender, Lyrics and Geographical Origins

In the introduction, we suggested three factors that might influence the results. We have tested for these three factors in our second study, but the results are tentative. To truly control for these variables, one would require a larger dataset than the one available in this study. In such a larger dataset, if one would control for the factors gender, geography, contextual factors, according to the "APM is widespread" hypothesis one

might find inter-recording tonic pitch consistency in every subset of variants from the same geographic region, with the same lyrics, or sung by singers with the same gender.

Different means in tonic pitch for women and men for the same song were obtained, perhaps due to the different fundamental frequencies men and women prefer for speech and song. These differences can in turn be attributed to physiological differences, for example the size of the larynx (Titze, 1989). There was a remarkably higher inter-recording tonic pitch consistency between female recordings for both the songs in the second dataset. This may have been due to puberty voice changes for male subjects (Harries, Walker, Williams, Hawkins, & Hughes, 1997). As these songs are often learned at young age, male singers might have learned these songs before puberty. The voice changes associated with male puberty might have distorted their reproduction of the melody as they had learned it. It was thus harder to reproduce the initial learned pitches. However, this remains mostly speculative, because relatively few recordings of male singers were available.

Regional variance may also explain the lack of pitch consistency in the smaller datasets. If geographical origin is indeed a factor of pitch variance, as has been proposed in the introduction, this might lead to disparities in the tonic pitch, especially for the smaller tune families, because the song might have been sung on a different tonic pitch in different parts of the Netherlands (see A.3). Those provinces that did have a reasonable number of recordings of particular tune families for statistical analysis showed inter-recording tonic pitch consistency, indicating that the same songs may be sung on different tonic pitches in different geographical areas. However, again a bigger sample is needed to draw more definite conclusions about this matter.

Interestingly, the means for two textual versions of one of the tune families in the second dataset were as much as 3 semitones apart, suggesting that tune families with different lyrics, but highly similar relative melodies may have two separate absolute pitch representations in memory. This is in accordance with the proposals by Bergeson and Trehub (2002) and Levitin and Rogers (2005), stating that lyrics accompanying the melody can facilitate retrieval of performance details of musical melodies. However, a larger dataset is needed to be able to test these interpretations.

A.5 CONCLUSIONS

In the current study, we tested whether there was any inter-recording tonic pitch consistency in recordings of folksongs available via in the Dutch Song database. In dataset A, one tune family, *Daar was laatst een meisje loos*, showed significant inter-recording tonic pitch consistency based on a collection of twenty recordings. In dataset B, one tune family, *Mijn vader zei laatst tegen mij*, showed tonic pitch consistency over the whole dataset, but also when grouped based on lyrics, suggesting a possible role for contextual information such as the lyrics. The second tune family in this dataset, *Wat hoor ik hier in het midden van de nacht*, showed near-significant inter-recording tonic pitch consistency. When controlled for gender, both these tune families showed tonic pitch consistency for female, but not for male singers. Also, for both these tune families,

there was significant consistency for tonic pitch when grouped on geographical origin of the recordings.

Nevertheless, to really identify the influence of these factors, a larger dataset is needed. For this the Meertens Tune Collections continues to be an excellent source. The recordings in the database are well documented and therefore easily retrievable. They have been grouped by tune family, which is important when looking at phenomena related to oral transmission of folksongs. Further research in this direction would be a valuable addition to our knowledge on the role of absolute pitch memory in oral transmission. Such research could be performed on recordings from other song collections, and use similar methods as were used in our research.

SIMILARITY MEASURES AND MUSIC REPRESENTATIONS

Measure	AUC	ϕ	SEN	SPC	PPV	NPV
WT-DW	0.731	0.459	0.368	0.976	0.703	0.909
WT-PADW	0.731	0.459	0.368	0.976	0.703	0.909
WT-DA	0.736	0.454	0.320	0.985	0.772	0.903
WT-HA	0.746	0.460	0.324	0.986	0.778	0.904
ED-PI	0.695	0.373	0.243	0.985	0.718	0.894
ED-P	0.764	0.468	0.482	0.948	0.589	0.922
ED-DW	0.788	0.540	0.441	0.980	0.774	0.919
ED-PA	0.831	0.554	0.616	0.940	0.612	0.940
ED-PADW	0.849	0.618	0.619	0.962	0.716	0.942
ED-DA	0.851	0.610	0.627	0.957	0.693	0.943
ED-HA	0.865	0.612	0.610	0.962	0.714	0.941
CBD-PI	0.727	0.424	0.365	0.967	0.634	0.908
CBD-P	0.774	0.499	0.425	0.973	0.708	0.916
CBD-DW	0.799	0.581	0.468	0.985	0.824	0.923
CBD-PA	0.849	0.589	0.564	0.966	0.720	0.935
CBD-PADW	0.870	0.663	0.601	0.979	0.818	0.941
CBD-DA	0.872	0.663	0.608	0.978	0.808	0.942
CBD-HA	0.891	0.696	0.651	0.978	0.822	0.948

Table B.1: Area under ROC curve, maximal ϕ correlation coefficient with associated sensitivity (SEN), specificity (SPC), positive and negative predictive values (PPV, NPV) for wavelet transform (WT), Euclidean distance (ED) and city-block distance (CBD) in all applicable music representations.

Measure	AUC	ϕ	SEN	SPC	PPV	NPV
CD-PI	0.677	0.313	0.214	0.979	0.617	0.889
CD-P	0.756	0.426	0.266	0.990	0.810	0.897
CD-PA	0.849	0.589	0.564	0.966	0.720	0.935
CD-DW	0.797	0.503	0.414	0.977	0.732	0.915
CD-PADW	0.797	0.503	0.414	0.977	0.733	0.915
CD-DA	0.795	0.501	0.420	0.975	0.720	0.916
CD-HA	0.817	0.525	0.448	0.975	0.734	0.919
LA-PI	0.740	0.470	0.416	0.967	0.662	0.915
LA-P	0.783	0.533	0.491	0.967	0.695	0.925
LA-DW	0.785	0.573	0.510	0.974	0.750	0.928
LA-PA	0.859	0.621	0.646	0.956	0.695	0.946
LA-PADW	0.871	0.665	0.658	0.968	0.759	0.948
LA-DA	0.875	0.668	0.675	0.965	0.748	0.950
LA-HA	0.881	0.682	0.695	0.965	0.753	0.953
SIAM-PO	0.870	0.665	0.632	0.973	0.787	0.945
SIAM-DA	0.868	0.663	0.641	0.971	0.772	0.946
SIAM-HA	0.893	0.696	0.688	0.970	0.783	0.953

Table B.2: Area under ROC curve, maximal ϕ correlation coefficient with associated sensitivity (SEN), specificity (SPC), positive and negative predictive values (PPV, NPV) for correlation distance (CD), local alignment (LA) and structure induction (SIAM) in all applicable music representations.

BIBLIOGRAPHY

- Athos, A., E, Levinson, B., Kistler, A., Zemansky, J., Bostrom, A., Freimer, N., & Gitschier, J. (2007). Dichotomy and perceptual distortions in absolute pitch ability. *Proceedings of the National Academy of Sciences*, 104(37), 14795–14800.
- Bachem, A. (1940). The genesis of absolute pitch. *Journal of the Acoustical Society of America*, 434–439.
- Bade, K., Nürnberger, A., Stober, S., Garbers, J., & Wiering, F. (2009). Supporting Folk-Song Research by Automatic Metric Learning and Ranking. In *Proceedings of the 10th International Society for Music Information Retrieval Conference (ISMIR 2009)* (pp. 741–746). Kobe, Japan.
- Barlow, H., & Morgenstern, S. (1948). *A dictionary of musical themes*. New York, NY: Crown Publishers.
- Bartlett, F. C. (1920). Some Experiments on the Reproduction of Folk-Stories. *Folklore*, 31(1), 30–47. doi: [10.1080/0015587X.1920.9719143](https://doi.org/10.1080/0015587X.1920.9719143)
- Bartók, B. (1951). Introduction to Part One. In *Serbo-Croatian Folk Songs. Texts and Transcriptions of Seventy-five Folk Songs from the Milman Perry Collection and a Morphology of Serbo-Croatian Folk Melodies* (pp. 3–20). New York, NY: Columbia University Press.
- Bartók, B. (1981). *The Hungarian Folk Song* (B. Suchoff, Ed.). London, United Kingdom: Oxford University Press.
- Bayard, S. P. (1950). Prolegomena to a Study of the Principal Melodic Families of British-American Folk Song. *The Journal of American Folklore*, 63(247), 1–44.
- Bergeson, T. R., & Trehub, S. E. (2002). Absolute pitch and tempo in mothers' songs to infants. *Psychological Science*, 13(1), 72–75. doi: [10.1111/1467-9280.00413](https://doi.org/10.1111/1467-9280.00413)
- Biró, D. P., van Kranenburg, P., Ness, S., Tzanetakis, G., & Volk, A. (2012). Stability and variation in cadence formulas in oral and semi-oral chant traditions—a computational approach. In *Proceedings of the 12th International Conference on Music Perception and Cognition and the 8th Triennial Conference of the European Society for Cognitive Sciences of Music (ICMPC/ESCOM 2012)* (pp. 98–105). Tsaloniki, Greece.
- Bohlman, P. V. (1988). Folk Music and Oral Tradition. In *The study of folk music in the modern world* (pp. 14–32). Bloomington, IN: Indiana University Press.
- Boot, P., Volk, A., & de Haas, W. B. (2016). Evaluating the Role of Repeated Patterns in Folk Song Classification and Compression. *Journal of New Music Research*, 45(3), 223–238.
- Bountouridis, D., Koops, H. V., Wiering, F., & Veltkamp, R. C. (2016). Music Outlier Detection Using Multiple Sequence Alignment and Independent Ensembles. In M. E. Houle & E. Schubert (Eds.), *Similarity Search and Applications: 9th International Conference (SISAP 2016)* (pp. 286–300). Tokyo, Japan: Springer. doi:

- 10.1007/978-3-319-46759-7
- Bronson, B. H. (1950). Some Observations About Melodic Variation in British-American Folk Tunes. *Journal of the American Musicological Society*, 3(2), 120–134.
- Bronson, B. H. (1951). Melodic Stability in Oral Transmission. *Journal of the International Folk Music Council*, 3, 50–55.
- Brown, R. M., & Palmer, C. (2012). Auditory–motor learning influences auditory memory for music. *Memory & Cognition*, 40(4), 567–578.
- Broze, Y., & Shanahan, D. (2013). Diachronic Changes in Jazz Harmony. *Music Perception: An Interdisciplinary Journal*, 31(1), 32–45. doi: 10.1525/mp.2013.31.1.32
- Burdick, A. (2012). Humanities to Digital Humanities. In A. Burdick, J. Drucker, P. Lunenfeld, T. Presner, & J. Schnapp (Eds.), *Digital Humanities* (pp. 3–26). Cambridge, MA: MIT Press.
- Burgoyne, J. A., Bountouridis, D., Van Balen, J., & Honing, H. (2013). Hooked: A Game for Discovering what Makes Music Catchy. In *Proceedings of the 14th International Society for Music Information Retrieval Conference (ISMIR 2013)* (pp. 245–250). Curitiba, Brazil.
- Burnham, K. P., & Anderson, R. (2004). Multimodel Inference: Understanding AIC and BIC in Model Selection. *Sociological Methods & Research*, 33(2), 261–304. doi: 10.1177/0049124104268644
- Buteau, C., & Vipperman, J. (2009). Melodic Clustering Within Motivic Spaces: Visualization in OpenMusic and Application to Schumann’s *Träumerei*. In T. Klouche & T. Noll (Eds.), *Mathematics and Computation in Music* (pp. 59–66). Berlin, Germany: Springer.
- Cambouropoulos, E. (2006). Musical Parallelism and Melodic Segmentation: A Computational Approach. *Music Perception: An Interdisciplinary Journal*, 23(3), 249–268.
- Cambouropoulos, E., Crochemore, M., Iliopoulos, C. S., Mohamed, M., & Sagot, M.-F. (2007). All maximal-pairs in step-leap representation of melodic sequence. *Information Sciences*, 177(9), 1954–1962. doi: 10.1016/j.ins.2006.11.012
- Castresana, J. (2000). Selection of Conserved Blocks from Multiple Alignments for Their Use in Phylogenetic Analysis. *Molecular Biology and Evolution*, 17(4), 540. doi: 10.1093/oxfordjournals.molbev.a026334
- De Cheveigné, A., & Kawahara, H. (2002). YIN, a fundamental frequency estimator for speech and music. *The Journal of the Acoustical Society of America*, 111(4), 1917–1930.
- Clifford, R., & Iliopoulos, C. S. (2004). Approximate string matching for music analysis. *Soft Computing*, 8, 597–603. doi: 10.1007/s00500-004-0384-5
- Cohen, J. (1992). A power primer. *Psychological bulletin*, 112(1), 155–159.
- Collins, T. (2013). *Discovery of repeated themes and sections*. http://www.music-ir.org/mirex/wiki/2013:Discovery_of_Repeated_Themes_%26_Sections. (Section 4, accessed: 2013-05-04)
- Collins, T., Arzt, A., Flossmann, S., & Widmer, G. (2013). SIARCT-CFP: Improving Precision and the Discovery of Inexact Musical Patterns in Point-set Representations. In *Proceedings of the 14th International Society for Music Information Retrieval*

- Conference (ISMIR 2013) (pp. 549–554). Curitiba, Brazil.
- Conklin, D. (2013). Antipattern Discovery in Folk Tunes. *Journal of New Music Research*, 42(2), 161–169. doi: [10.1080/09298215.2013.809125](https://doi.org/10.1080/09298215.2013.809125)
- Conklin, D., & Anagnostopoulou, C. (2011). Comparative Pattern Analysis of Cretan Folk Songs. *Journal of New Music Research*, 40(2), 119–125. doi: [10.1080/09298215.2011.573562](https://doi.org/10.1080/09298215.2011.573562)
- Conklin, D., & Bergeron, M. (2010). Discovery of contrapuntal patterns. In *Proceedings of the 11th International Society for Music Information Retrieval Conference (ISMIR 2010)* (pp. 201–206). Utrecht, the Netherlands.
- Conklin, D., & Witten, I. H. (1995). Multiple Viewpoint Systems for Music Prediction. *Journal of New Music Research*, 24(1), 51–73.
- Conrad, R., & Hull, A. (1964). Information, Acoustic Confusion and Memory Span. *British Journal of Psychology*, 55(4), 429–432.
- Cowdery, J. R. (1990). *The Melodic Tradition of Ireland*. Kent, OH: Kent State University Press.
- Crochemore, M. (1981). An Optimal Algorithm for Computing the Repetitions in a Word. *Information Processing Letters*, 12(5), 244–250.
- Crowder, R. G., & Morton, J. (1969). Precategorical acoustic storage (PAS). *Perception & Psychophysics*, 5(6), 365–373. doi: [10.3758/BF03210660](https://doi.org/10.3758/BF03210660)
- Cuthbert, M. S., & Ariza, C. (2010). music21 : A Toolkit for Computer-Aided Musicology and Symbolic Music Data. In *Proceedings of the 11th International Society for Music Information Retrieval Conference (ISMIR 2010)* (pp. 637–642). Utrecht, the Netherlands.
- Dalla Bella, S., Peretz, I., & Aronoff, N. (2003). Time course of melody recognition: A gating paradigm study. *Perception & Psychophysics*, 65(7), 1019–1028.
- Dawkins, R. (1978). *The Selfish Gene*. London, United Kingdom: Granada Publishing.
- Deese, J., & Kaufman, R. A. (1957). Serial effects in recall of unorganized and sequentially organized verbal material. *Journal of Experimental Psychology*, 54(3), 180–187. doi: [10.1037/h0040536](https://doi.org/10.1037/h0040536)
- Deutsch, D. (2002). The puzzle of absolute pitch. *Current Directions in Psychological Science*, 11(6), 200–204.
- Dowling, W. J. (1978). Scale and Contour: Two Components of a Theory of Memory for Melodies. *Psychological Review*, 85(4), 341–354.
- Dowling, W. J., & Fujitani, D. S. (1971). Contour, interval, and pitch recognition in memory for melodies. *The Journal of the Acoustical Society of America*, 49(2B), 524–531.
- Eerola, T., Jäärvinen, T., Louhivuori, J., & Toiviainen, P. (2001). Statistical features and perceived similarity of folk melodies. *Music Perception: An Interdisciplinary Journal*, 18(3), 275–296.
- Eschrich, S., Münte, T. F., & Altenmüller, E. O. (2008). Unforgettable film music: the role of emotion in episodic long-term memory for music. *BMC Neuroscience*, 9, 48.
- Fisher, N. I. (1995). *Statistical analysis of circular data*. Cambridge, United Kingdom:

Cambridge University Press.

- Flexer, A., & Grill, T. (2016). The Problem of Limited Inter-rater Agreement in Modelling Music Similarity. *Journal of New Music Research*, 8215(August), 1–13. doi: [10.1080/09298215.2016.1200631](https://doi.org/10.1080/09298215.2016.1200631)
- Forth, J. (2012). *Cognitively-Motivated Geometric Methods of Pattern Discovery and Models of Similarity in Music* (Ph.D. Thesis). Goldsmiths University, London.
- Frieler, K., Fischinger, T., Schlemmer, K., Lothwesen, K., Jakubowski, K., & Müllensiefen, D. (2013). Absolute memory for pitch: A comparative replication of Levitin's 1994 study in six European labs. *Musicae Scientiae*, 17(3), 73–92. doi: [10.1177/102986490200600104](https://doi.org/10.1177/102986490200600104)
- Frieler, K., Pfeleiderer, M., Zaddach, W.-G., & Abeßer, J. (2016). Midlevel analysis of monophonic jazz solos: A new approach to the study of improvisation. *Musicae Scientiae*, 20(2), 143–162. doi: [10.1177/1029864916636440](https://doi.org/10.1177/1029864916636440)
- Garbers, J., Volk, A., van Kranenburg, P., Wiering, F., Grijp, L. P., & Veltkamp, R. C. (2009). On Pitch and Chord Stability in Folk Song Variation Retrieval. In T. Klouche & T. Noll (Eds.), *Mathematics and Computation in Music: First International Conference, MCM 2007. Communications in Computer and Information Science* (Vol. 37, pp. 97–106). Berlin, Germany: Springer.
- Giraud, M., Groult, R., & Levé, F. (2012). Subject and counter-subject detection for analysis of the Well-Tempered Clavier fugues. In *9th International Symposium on Computer Music Modelling and Retrieval (CMMR 2012)*. London, United Kingdom.
- Gobet, F., Lane, P., Croker, S., Cheng, P., Jones, G., Oliver, I., & Pine, J. (2001). Chunking mechanisms in human learning. *Trends in Cognitive Sciences*, 5(6), 236–243. doi: [10.1016/S1364-6613\(00\)01662-4](https://doi.org/10.1016/S1364-6613(00)01662-4)
- Gómez, E., & Bonada, J. (2013). Experimental Comparison of Automatic Transcription Towards Computer-Assisted Flamenco Transcription: An Experimental Comparison of Automatic Transcription Algorithms as Applied to A Cappella Singing. *Computer Music Journal*, 37(2), 73–90. doi: [10.1162/COMJ](https://doi.org/10.1162/COMJ)
- Gómez, E., Haro, M., & Herrera, P. (2009). Music and Geography: Content Description of Musical Audio from Different Parts of the World. In *Proceedings of the 10th International Society for Music Information Retrieval Conference (ISMIR 2009)* (pp. 753–758). Kobe, Japan.
- Grachten, M., Arcos, J. L., & López de Mántaras, R. (2005). Melody Retrieval using the Implication / Realization Model. In *2nd Annual Music Information Retrieval eXchange (MIREX2005)*. London, United Kingdom.
- Gregersen, P. K., Kowalsky, E., Kohn, N., & Marvin, E. W. (2001). Early childhood music education and predisposition to absolute pitch: Teasing apart genes and environment. *American Journal of Medical Genetics*, 98(3), 280–282.
- Grijp, L. P. (2008). Introduction. In L. P. Grijp & I. van Beersum (Eds.), *Under the Green Linden. 163 Dutch ballads from the oral tradition* (pp. 18–27). Hilversum, the Netherlands: Music & Words.
- Grijp, L. P., & Roodenburg, H. (2005). *Blues en Balladen*. Amsterdam, the Netherlands: Amsterdam University Press.

- Gusfield, D. (1997). *Algorithms on strings, trees and sequences: computer science and computational biology*. New York, NY: Cambridge University Press.
- de Haas, W. B., Volk, A., & Wiering, F. (2013). Structural Segmentation of Music Based on Repeated Harmonies. In *2013 IEEE International Symposium on Multimedia* (pp. 255–258). Anaheim, CA. doi: [10.1109/ISM.2013.48](https://doi.org/10.1109/ISM.2013.48)
- Halpern, A. R. (1989). Memory for the absolute pitch of familiar songs. *Memory & Cognition*, *17*(5), 572–581.
- Harries, M. L. L., Walker, J. M., Williams, D. M., Hawkins, S., & Hughes, I. A. (1997). Changes in the male voice at puberty. *Archives of disease in childhood*, *77*(5), 445–447.
- Hartigan, J., & Hartigan, P. (1985). The Dip Test of Unimodality. *Annals of Statistics*, *13*(1), 70–84.
- Hauser, M. D. (2009). The possibility of impossible cultures. *Nature*, *460*(7252), 190–6. doi: [10.1038/460190a](https://doi.org/10.1038/460190a)
- Henrich, J., & Boyd, R. (2002). On Modeling Cognition and Culture representations. *Journal of Cognition and Culture*, *2*(2), 87–112.
- Hillewaere, R., Manderick, B., & Conklin, D. (2009). Global Feature Versus Event Models for Folk Song Classification. In *Proceedings of the 10th International Society for Music Information Retrieval Conference (ISMIR 2009)* (pp. 729–734). Kobe, Japan.
- Hintzman, D. L. (1969). Apparent frequency as a function of frequency and the spacing of repetitions. *Journal of Experimental Psychology*, *80*(1), 139–145. doi: [10.1037/h0027133](https://doi.org/10.1037/h0027133)
- Honing, H., ten Cate, C., Peretz, I., & Trehub, S. E. (2015). Without it no music: cognition, biology and evolution of musicality. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, *370*(1664), 20140088. doi: [10.1098/rstb.2014.0088](https://doi.org/10.1098/rstb.2014.0088)
- Hsu, J. L., Liu, C. C., & Chen, A. L. P. (2001). Discovering nontrivial repeating patterns in music data. *IEEE Transactions on Multimedia*, *3*(3), 311–325. doi: [10.1109/6046.944475](https://doi.org/10.1109/6046.944475)
- Huron, D. (1996). The Melodic Arch in Western Folksongs. *Computing in Musicology*, *10*, 3–23.
- Huron, D. (2007). *Sweet Anticipation. Music and the Psychology of Expectation*. Cambridge, MA: MIT Press.
- Hurvich, C. M., & Tsai, C.-L. (1989). Regression and Time Series Model Selection in Small Samples. *Biometrika*, *76*(2), 297–307.
- Jacoby, N., & McDermott, J. H. (2017). Integer Ratio Priors on Musical Rhythm Revealed Cross-culturally by Iterated Reproduction Article Integer Ratio Priors on Musical Rhythm Revealed Cross-culturally by Iterated Reproduction. *Current Biology*, 1–12. doi: [10.1016/j.cub.2016.12.031](https://doi.org/10.1016/j.cub.2016.12.031)
- Jammalamadaka, S. R., & SenGupta, A. (2001). *Topics in Circular Statistics*. Singapore: World Scientific Press.
- Janssen, B., Burgoyne, J. A., & Honing, H. (2017). Predicting Variation of Folk Songs: A Corpus Analysis Study on the Memorability of Melodies. *Frontiers in Psychology*,

- 8, 621. doi: [10.3389/fpsyg.2017.00621](https://doi.org/10.3389/fpsyg.2017.00621)
- Janssen, B., de Haas, W. B., Volk, A., & van Kranenburg, P. (2014). Finding Repeated Patterns in Music: State of Knowledge, Challenges, Perspectives. In M. Aramaki, O. Derrien, R. Kronland-Martinet, & S. Ystad (Eds.), *Sound, Music, and Motion: 10th International Symposium, CMMR 2013, Revised Selected Papers (LNCS 8905)* (pp. 277–297). Heidelberg, Germany: Springer. doi: [10.1007/978-3-319-12976-1_8](https://doi.org/10.1007/978-3-319-12976-1_8)
- Janssen, B., de Haas, W. B., Volk, A., & van Kranenburg, P. (2013). Discovering repeated patterns in music: state of knowledge, challenges, perspectives. In *Proceedings of the 10th International Symposium on Computer Music Interdisciplinary Research (CMMR 2013)* (pp. 225–240). Marseille, France.
- Janssen, B., van Kranenburg, P., & Volk, A. (2015). A Comparison of Symbolic Similarity Measures for Finding Occurrences of Melodic Segments. In *Proceedings of the 16th International Society for Music Information Retrieval Conference (ISMIR 2015)*. Málaga, Spain.
- Janssen, B., van Kranenburg, P., & Volk, A. (2017). Finding Occurrences of Melodic Segments in Folk Songs: a Comparison of Symbolic Similarity Measures. *Journal of New Music Research*, 46(2), 118–134. doi: [10.1080/09298215.2017.1316292](https://doi.org/10.1080/09298215.2017.1316292)
- Járdányi, P. (1965). Experiences and Results in Systematizing Hungarian Folk-Songs. *Studia Musicologica Academiae Scientiarum Hungaricae*, 7, 287–291.
- Jesser, B. (1991). *Interaktive Melodieanalyse. Methodik und Anwendung computergestützter Analyseverfahren in Musikethnologie und Volksliedforschung: typologische Untersuchung der Balladensammlung des DVA*. Bern, Switzerland: Peter Lang.
- Jones, M. R. (1987). Dynamic pattern structure in music: recent theory and research. *Perception & Psychophysics*, 41(6), 621–634.
- Juhász, Z. (2006). A systematic comparison of different European folk music traditions using self-organizing maps. *Journal of New Music Research*, 35(2), 95–112. doi: [10.1080/09298210600834912](https://doi.org/10.1080/09298210600834912)
- Juhász, Z., & Sipos, J. (2009). A Comparative Analysis of Eurasian Folksong Corpora, using Self Organising Maps. *Journal of Interdisciplinary Music Studies*, 1–16. doi: [10.4407/jims.2009.11.005](https://doi.org/10.4407/jims.2009.11.005)
- Karydis, I., Nanopoulos, A., & Manolopoulos, Y. (2007). Finding maximum-length repeating patterns in music databases. *Multimedia Tools and Applications*, 32, 49–71. doi: [10.1007/s11042-006-0068-5](https://doi.org/10.1007/s11042-006-0068-5)
- Keller, T. A., Cowan, N., & Sauls, J. S. (1995). Can auditory memory for tone pitch be rehearsed? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 21(3), 635–645.
- Kirby, S., Cornish, H., & Smith, K. (2008). Cumulative cultural evolution in the laboratory: an experimental approach to the origins of structure in human language. *Proceedings of the National Academy of Sciences of the United States of America*, 105(31), 10681–6. doi: [10.1073/pnas.0707835105](https://doi.org/10.1073/pnas.0707835105)
- Kirk, R. E. (1996). Practical significance: A concept whose time has come. *Educational and psychological measurement*, 56(5), 746–759.
- Klapuri, A. (2010). Pattern induction and matching in music signals. In *Exploring Music*

- Contents: Proceedings of the 7th International Symposium on Computer Modeling and Retrieval (CMMR 2010)* (pp. 188–204). Málaga, Spain.
- Kleeman, J. E. (1985). The Parameters of Musical Transmission. *The Journal of Musicology*, 4(1), 1–22.
- Klinger, R., Campbell, P. S., & Goolsby, T. (1998). Approaches to children's song acquisition: Immersion and phrase-by-phrase. *Journal of Research in Music Education*, 46(1), 24–34.
- Klusen, E., Moog, H., & Piel, W. (1978). Experimente zur mündlichen Tradition von Melodien. In *Jahrbuch für Volksliedforschung* (pp. 11–23).
- Knopke, I., & Jürgensen, F. (2009). A System for Identifying Common Melodic Phrases in the Masses of Palestrina. *Journal of New Music Research*, 38(2), 171–181. doi: [10.1080/09298210903288329](https://doi.org/10.1080/09298210903288329)
- Knuth, D. E., Morris, J. H., & Pratt, V. R. (1977). Fast Pattern Matching in Strings. *SIAM Journal of Computing*, 6(2), 323–350.
- van Kranenburg, P. (2010). *A Computational Approach to Content-Based Retrieval of Folk Song Melodies* (Ph.D. Thesis). University of Utrecht.
- van Kranenburg, P., & Conklin, D. (2016). A Pattern Mining Approach to Study a Collection of Dutch Folk-Songs. In *Proceedings of the 6th International Workshop on Folk Music Analysis* (pp. 71–73). Dublin, Ireland.
- van Kranenburg, P., Janssen, B., & Volk, A. (2016). *Meertens Tune Collections: Annotated Corpus 2.0* (Tech. Rep.). Amsterdam, the Netherlands: Meertens Online Reports.
- van Kranenburg, P., Volk, A., & Wiering, F. (2013). A Comparison between Global and Local Features for Computational Classification of Folk Song Melodies. *Journal of New Music Research*, 42(1), 1–18. doi: [10.1080/09298215.2012.718790](https://doi.org/10.1080/09298215.2012.718790)
- Krohn, I. (1903). Welche ist die beste Methode, um Volks- und volksmäßige Lieder nach ihrer melodischen (nicht textlichen) Beschaffenheit lexikalisch zu ordnen? *Sammelbände der Internationalen Musikgesellschaft*, 4(4), 643–660.
- Krumhansl, C. L. (2000). Rhythm and pitch in music cognition. *Psychological Bulletin*, 126(1), 159–179.
- Lartillot, O. (2014). In-depth Motivic Analysis Based on Multiparametric Closed Pattern and Cyclic Sequence Mining. In *Proceedings of the 15th International Society for Music Information Retrieval Conference (ISMIR 2014)* (pp. 361–366). Taipei, Taiwan.
- Le Bomin, S., Lecointre, G., & Heyer, E. (2016). The evolution of musical diversity: The key role of vertical transmission. *PLoS ONE*, 11(3), 1–17. doi: [10.1371/journal.pone.0151570](https://doi.org/10.1371/journal.pone.0151570)
- Lee, W., & Chen, A. L. P. (1999). Efficient multifeature index structures for music data retrieval. In *Proceedings of SPIE: Storage and Retrieval for Media Databases 2000* (Vol. 3972, pp. 177–188). San José, CA. doi: [10.1117/12.373547](https://doi.org/10.1117/12.373547)
- Lemström, K., Mikkilä, N., & Mäkinen, V. (2009). Filtering methods for content-based retrieval on indexed symbolic music databases. *Information Retrieval*, 13(1), 1–21. doi: [10.1007/s10791-009-9097-9](https://doi.org/10.1007/s10791-009-9097-9)
- Lemström, K., & Ukkonen, E. (2000). Including Interval Encoding into Edit Distance Based Music Comparison and Retrieval. In *Proceedings of the AISB'2000 Symposium*

- on *Creative & Cultural Aspects and Applications of AI & Cognitive Science*. Birmingham, United Kingdom.
- Lerdahl, F., & Jackendoff, R. (1983). *A Generative Theory of Tonal Music*. Cambridge, MA: MIT Press.
- Levenshtein, V. I. (1966). Binary Codecs Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics - Doklady*, 10(8), 707–710.
- Levitin, D. J. (1994). Absolute memory for musical pitch: evidence from the production of learned melodies. *Perception & Psychophysics*, 56(4), 414–23.
- Levitin, D. J. (2013). Commentary on “Absolute memory for pitch: A comparative replication of Levitin’s 1994 study in six European labs”. *Musicae Scientiae*, 17(3), 350–355.
- Levitin, D. J., & Rogers, S. E. (2005). Absolute pitch: perception, coding, and controversies. *Trends in Cognitive Sciences*, 9(1), 26–33.
- Lomax, A. (1962). Song Structure and Social Structure. *Ethnology*, 1(4), 425–451.
- Lomax, A. (1968). *Folk Song Style and Culture*. Washington D.C.: American Association for the Advancement of Science.
- Louboutin, C., & Meredith, D. (2016). Using general-purpose compression algorithms for music analysis. *Journal of New Music Research*, 45(1), 1–16. doi: [10.1080/09298215.2015.1133656](https://doi.org/10.1080/09298215.2015.1133656)
- Louhivuori, J. (1990). Computer Aided Analysis of Finnish Spiritual Folk Melodies. In H. Braun (Ed.), *Probleme der Volksmusikforschung* (pp. 312–323). Bern, Switzerland: Peter Lang.
- MacCallum, R. M., Mauch, M., Burt, A., & Leroi, A. M. (2012). Evolution of music by public choice. *Proceedings of the National Academy of Sciences of the United States of America*(7), 1–6. doi: [10.5061/dryad.h0228](https://doi.org/10.5061/dryad.h0228)
- Madigan, S. A. (1969). Intraserial repetition and coding processes in free recall. *Journal Of Verbal Learning And Verbal Behavior*, 8(6), 828–835. doi: [10.1016/S0022-5371\(69\)80050-2](https://doi.org/10.1016/S0022-5371(69)80050-2)
- Margulis, E. H. (2012). Musical Repetition Detection Across Multiple Exposures. *Music Perception: An Interdisciplinary Journal*, 29(4), 377–385.
- Margulis, E. H. (2014). *On Repeat: How Music Plays the Mind*. Oxford, United Kingdom: Oxford University Press.
- Marsden, A. (2012). Counselling a better relationship between mathematics and musicology. *Journal of Mathematics and Music: Mathematical and Computational Approaches to Music Theory, Analysis, Composition and Performance*, 6(2), 145–153.
- Matthews, B. W. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *BBA - Protein Structure*, 405(2), 442–451. doi: [10.1016/0005-2795\(75\)90109-9](https://doi.org/10.1016/0005-2795(75)90109-9)
- Mauch, M., MacCallum, R. M., Levy, M., & Leroi, A. M. (2015). The evolution of popular music: USA 1960-2010. *Royal Society Open Science*, 2(5), 150081. doi: [10.1098/rsos.150081](https://doi.org/10.1098/rsos.150081)
- Mckay, C., & Fujinaga, I. (2004). Automatic Genre Classification Using High-Level Musical Feature Sets. In *ISMIR 2004, 5th International Conference on Music Information*

- Retrieval* (pp. 525–530). Barcelona, Spain.
- Meek, C., & Birmingham, W. P. (2001). Thematic Extractor. In *ISMIR 2001, 2nd International Symposium on Music Information Retrieval* (pp. 119–128). Bloomington, IN.
- Meredith, D. (2006). Point-set algorithms for pattern discovery and pattern matching in music. In T. Crawford & R. C. Veltkamp (Eds.), *Content-Based Retrieval. Dagstuhl Seminar Proceedings 06171*. Dagstuhl, Germany.
- Meredith, D. (2014). COSIATEC and SIATECCompress: Pattern Discovery by Geometric Compression. In *Music Information Retrieval Evaluation eXchange*.
- Meredith, D. (2015). Music Analysis and Point-Set Compression. *Journal of New Music Research*, 44(3), 245–270. doi: [10.1080/09298215.2015.1045003](https://doi.org/10.1080/09298215.2015.1045003)
- Meredith, D., Lemström, K., & Wiggins, G. A. (2002). Algorithms for discovering repeated patterns in multidimensional representations of polyphonic music. *Journal of New Music Research*, 31(4), 321–345.
- Mesoudi, A. (2011). *Cultural Evolution: How Darwinian Theory Can Explain Human Culture & Synthesize the Social Sciences*. Chicago, IL: University of Chicago Press.
- Meyer, L. B. (1956). *Emotion and Meaning in Music*. Chicago, IL: The University of Chicago Press.
- Meyer, L. B. (1957). Meaning in Music and Information Theory. *The Journal of Aesthetics and Art Criticism*, 15(4), 412–424.
- Miller, G. E. (1956). The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological Review*, 63(2), 81–97.
- Miranda, E. R. (2008). Emergent songs by social robots. *Journal of Experimental & Theoretical Artificial Intelligence*, 20(4), 319–334. doi: [10.1080/09528130701664640](https://doi.org/10.1080/09528130701664640)
- Mongeau, M., & Sankoff, D. (1990). Comparison of Musical Sequences. *Computers and the Humanities*, 24, 161–175.
- Moore, R. S. (1991). Comparison of Children's and Adults' Vocal Ranges and Preferred Tessituras in Singing Familiar Songs. *Bulletin of the Council for Research in Music Education*, 107(107), 13–22. doi: [10.2307/40318417](https://doi.org/10.2307/40318417)
- Müllensiefen, D. (2009). *FANTASTIC: Feature ANalysis Technology Accessing STatistics (In a Corpus): Technical Report* (Tech. Rep.). London, United Kingdom: Goldsmiths University.
- Müllensiefen, D., & Frieler, K. (2007). Modelling experts' notions of melodic similarity. *Musicae Scientiae*, 11(183), 183–210. doi: [10.1177/102986490701100108](https://doi.org/10.1177/102986490701100108)
- Müllensiefen, D., & Halpern, A. R. (2014). The Role of Features and Context in Recognition of Novel Melodies. *Music Perception: An Interdisciplinary Journal*, 31(5), 418–435.
- Murdock, B. (1962). The serial position effect of free recall. *Journal of Experimental Psychology*, 64(5), 482–488. doi: [10.1037/h0045106](https://doi.org/10.1037/h0045106)
- Murdock, B., & Metcalfe, J. (1978). Controlled Rehearsal in Single-Trial Free Recall. *Journal of Verbal Learning and Verbal Behavior*, 17, 309–324.
- Nakagawa, S., & Schielzeth, H. (2013). A general and simple method for obtaining R² from generalized linear mixed-effects models. *Methods in Ecology and Evolution*,

- 4(2), 133–142. doi: [10.1111/j.2041-210X.2012.00261.x](https://doi.org/10.1111/j.2041-210X.2012.00261.x)
- Narmour, E. (1990). *The Analysis and Cognition of Basic Melodic Structures. The Implication-Realization Model*. Chicago, IL: University of Chicago Press.
- Needleman, S. B., & Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48(3), 443–453. doi: [10.1016/0022-2836\(70\)90057-4](https://doi.org/10.1016/0022-2836(70)90057-4)
- Nettl, B. (2005). *The Study of Ethnomusicology. Thirty-one Issues and Concepts*. Champaign, IL: University of Illinois Press.
- Nieto, O., & Farbood, M. M. (2012). Perceptual Evaluation of Automatically Extracted Musical Motives. In *Proceedings of the 12th International Conference on Music Perception and Cognition and the 8th Triennial Conference of the European Society for Cognitive Sciences of Music (ICMPC/ESCOM 2012)* (pp. 723–727). Tessaloniki, Greece.
- Nieto, O., & Farbood, M. M. (2014). Identifying Polyphonic Patterns From Audio Recordings Using Music Segmentation Techniques. In *Proceedings of the 15th International Society for Music Information Retrieval Conference (ISMIR 2014)* (pp. 411–416). Taipei, Taiwan.
- Oliphant, T. E. (2007). Python for Scientific Computing. *Computing in Science and Engineering*, 9(3), 10–20. doi: <http://dx.doi.org/10.1109/MCSE.2007.58>
- Olthof, M., Janssen, B., & Honing, H. (2015). The Role Of Absolute Pitch Memory In The Oral Transmission of Folksongs. *Empirical Musicology Review*, 10(3), 161–174. doi: [10.18061/emr.v10i3.4435](https://doi.org/10.18061/emr.v10i3.4435)
- Owens, T. (1974). *Charlie Parker : techniques of improvisation* (Ph.D. Thesis). Kent State University.
- Page, M. P. A., & Norris, D. (1998). The Primacy Model: A New Model of Immediate Serial Recall. *Psychological Review*, 105(4), 761–781. doi: [10.1037/0033-295X.105.4.761-781](https://doi.org/10.1037/0033-295X.105.4.761-781)
- Panteli, M., Bittner, R., Bello, J. P., & Dixon, S. (2017). Towards the Characterization of Singing Styles in World Music. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 636–640). New Orleans, LA.
- Paulus, J., Müller, M., & Klapuri, A. (2010). Audio-based music structure analysis. In *Proceedings of the 11th International Society for Music Information Retrieval Conference (ISMIR 2010)* (pp. 625–636). Utrecht, the Netherlands.
- Pearce, M. T. (2005). *The Construction and Evaluation of Statistical Models of Melodic Structure in Music Perception and Composition* (Ph.D. Thesis). City University, London.
- Pearce, M. T., & Wiggins, G. A. (2004). Improved Methods for Statistical Modelling of Monophonic Music. *Journal of New Music Research*, 33(4), 367–385. doi: [10.1080/0929821052000343840](https://doi.org/10.1080/0929821052000343840)
- Pearce, M. T., & Wiggins, G. A. (2007). Evaluating Cognitive Models of Musical Composition. In *Proceedings of the 4th International Joint Workshop on Computational Creativity* (pp. 73–80). London, United Kingdom.
- Pesek, M., Leonardis, A., & Marolt, M. (2014). A compositional hierarchical model for music information retrieval. In *Proceedings of the 15th International Conference on Music Information Retrieval (ISMIR 2014)* (pp. 131–136). Taipei, Taiwan.

- Pesek, M., Medvešek, U., Leonardis, A., & Marolt, M. (2015). SymCHM: a compositional hierarchical model for pattern discovery in symbolic music representations. In *11th Annual Music Information Retrieval eXchange (MIREX2015)* (pp. 1–3). Málaga, Spain. doi: [10.13140/RG.2.1.4128.2965](https://doi.org/10.13140/RG.2.1.4128.2965)
- Plantinga, J., & Trainor, L. J. (2005). Memory for melody: Infants use a relative pitch code. *Cognition*, *98*(1), 1–11.
- Powers, D. M. W. (2007). Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation. *Journal of Machine Learning Technologies*, *2*(1), 24. doi: [10.1.1.214.9232](https://doi.org/10.1.1.214.9232)
- Raffman, D. (2003). Is Twelve-Tone Music Artistically Defective? *Midwest Studies in Philosophy*, *27*, 68–87.
- Ravignani, A., & Delgado, T. (2016). Musical evolution in the lab exhibits rhythmic universals. *Nature*, *0007*(December), 1–7. doi: [10.1038/s41562-016-0007](https://doi.org/10.1038/s41562-016-0007)
- Ren, I. Y. (2016). Closed Patterns in Folk Music and Other Genres. In *Proceedings of the 6th International Workshop on Folk Music Analysis* (pp. 56–58). Dublin, Ireland.
- von Restorff, H. (1933). Über die Wirkung von Bereichsbildungen im Spurenfeld. *Psychologische Forschung*, *18*(1), 299–342. doi: [10.1007/BFo2409636](https://doi.org/10.1007/BFo2409636)
- Réti, R. (1951). *The thematic process in music*. New York, NY: Macmillan.
- Rodriguez Zivic, P. H., Shifres, F., & Cecchi, G. A. (2013). Perceptual basis of evolving Western musical styles. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(24), 10034–8.
- Rohrmeier, M., Zuidema, W., Wiggins, G. A., & Scharff, C. (2015). Principles of structure building in music, language and animal song. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *370*(1664). doi: [10.1098/rstb.2014.0097](https://doi.org/10.1098/rstb.2014.0097)
- Rolland, P.-Y. (1999). Discovering Patterns in Musical Sequences. *Journal of New Music Research*, *28*(4), 334–351.
- Romming, C. A., & Selfridge-Field, E. (2007). Algorithms for Polyphonic Music Retrieval: the Hausdorff Metric and Geometric Hashing. In *Proceedings of the 8th International Society for Music Information Retrieval Conference (ISMIR 2007)* (pp. 457–462). Vienna, Austria.
- Rubin, D. C. (1977). Very Long-Term Memory for Prose and Verse. *Journal of Verbal Learning and Verbal Behaviour*(16), 611–621.
- Rubin, D. C. (1995). *Memory in Oral Traditions. The Cognitive Psychology of Epic, Ballads, and Counting-out Rhymes*. New York, NY: Oxford University Press.
- Saffran, J. R., & Griepentrog, G. J. (2001). Absolute pitch in infant auditory learning: evidence for developmental reorganization. *Developmental Psychology*, *37*(1), 74.
- Salganik, M. J., Dodds, P. S., & Watts, D. J. (2006). Experimental Study of Inequality and Unpredictability in an Artificial Cultural Market. *Science*, *311*(5762), 854–856. doi: [10.1126/science.1121066](https://doi.org/10.1126/science.1121066)
- Savage, P. E., Brown, S., Sakai, E., & Currie, T. E. (2015). Statistical universals reveal the structures and functions of human music. *Proceedings of the National Academy of Sciences*, *112*(29), 8987–8992. doi: [10.1073/pnas.1414495112](https://doi.org/10.1073/pnas.1414495112)

- Schaffrath, H., & Huron, D. (1995). *The Essen folksong collection in the humdrum kern format*. Menlo Park, CA: Center for Computer Assisted Research in the Humanities.
- Schellenberg, E. G. (1996). Expectancy in melody: Tests of the implication-realization model. *Cognition*, 58(1), 75–125.
- Schellenberg, E. G. (1997). Simplifying the Implication-Realization Model of Melodic Expectancy. *Music Perception: An Interdisciplinary Journal*, 14(3), 295–318.
- Schellenberg, E. G., & Trehub, S. E. (2003). Good Pitch Memory Is Widespread. *Psychological Science*, 14(3), 262–266. doi: [10.1111/1467-9280.03432](https://doi.org/10.1111/1467-9280.03432)
- Scherrer, D. K., & Scherrer, P. H. (1971). An Experiment in the Computer Measurement of Melodic Variation in Folksong. *The Journal of American Folklore*, 84(332), 230–241.
- Schmuckler, M. A. (1997). Expectancy Effects in Memory for Melodies. *Canadian Journal of Experimental Psychology*, 51(4), 292–306.
- Serrà, J., Corral, A., Boguñá, M., Haro, M., & Arcos, J. L. (2012). Measuring the evolution of contemporary western popular music. *Scientific Reports*, 2, 521. doi: [10.1038/srep00521](https://doi.org/10.1038/srep00521)
- Shepard, R. N. (1982). Geometrical approximations to the structure of musical pitch. *Psychological Review*, 89(4), 305–333.
- Sloboda, J., & Parker, D. (1985). Immediate recall of melodies. In P. Howell, I. Cross, & R. West (Eds.), *Musical Structure and Cognition* (pp. 143–167). London, United Kingdom: Academic Press.
- Smith, D. S. (1991). A comparison of group performance and song familiarity on cued recall tasks with older adults. *Journal of Music Therapy*, 28(1), 2–13.
- Smith, T., & Waterman, M. (1981). Identification of common molecular subsequences. *Journal of Molecular Biology*, 147(1), 195–197. doi: [10.1016/0022-2836\(81\)90087-5](https://doi.org/10.1016/0022-2836(81)90087-5)
- Steinbeck, W. (1982). *Struktur und Ähnlichkeit. Methoden automatisierter Melodienanalyse*. Kassel, Germany: Bärenreiter.
- Suppan, W. (1973). Zur Verwendung der Begriffe Gestalt, Struktur, Modell und Typus in der Musikethnologie. In D. Stockmann & J. Steszewski (Eds.), *Analyse und Klassifikation von Volksmelodien* (pp. 41–52). Krakow, Poland: Polskie Wydawnictwo Muzyczne.
- Szeto, W. M., & Wong, M. H. (2006). A graph-theoretical approach for pattern matching in post-tonal music analysis. *Journal of New Music Research*, 35(4), 307–321. doi: [10.1080/09298210701535749](https://doi.org/10.1080/09298210701535749)
- Szpunar, K. K., Schellenberg, E. G., & Pliner, P. (2004). Liking and memory for musical stimuli as a function of exposure. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 30(2), 370–81. doi: [10.1037/0278-7393.30.2.370](https://doi.org/10.1037/0278-7393.30.2.370)
- Takeuchi, A. H., & Hulse, S. H. (1993). Absolute Pitch. *Psychological Bulletin*, 113(2), 345–61.
- Tappert, W. (1890/1965). *Wandernde Melodien. Eine musikalische Studie*. Oosterhout, the Netherlands: Anthropological Publications.
- Terhardt, E., & Ward, W. D. (1982). Recognition of musical key: Exploratory study. *The*

- Journal of the Acoustical Society of America*, 72(1), 26–33.
- Titze, I. R. (1989). Physiologic and acoustic differences between male and female voices. *The Journal of the Acoustical Society of America*, 85(4), 1699–1707.
- Trehub, S. E. (2015). Cross-cultural convergence of musical features. *Proceedings of the National Academy of Sciences of the United States of America*, 112(29), 8809–8810. doi: [10.1073/pnas.1510724112](https://doi.org/10.1073/pnas.1510724112)
- Tukey, J. W. (1977). *Exploratory Data Analysis*. Reading, MA: Addison-Wesley.
- Urbano, J., Lloréns, J., Morato, J., & Sánchez-Cuadrado, S. (2011). Melodic Similarity through Shape Similarity. In S. Ystad, M. Aramaki, R. Kronland-Martinet, & K. Jensen (Eds.), *Exploring Music Contents: 7th International Symposium, CMMR 2010, Málaga, Spain, June 21-24, 2010. Revised Papers (LNCS 6684)* (pp. 338–355). Berlin, Germany: Springer.
- Van Balen, J., Burgoyne, J. A., Bountouridis, D., Müllensiefen, D., & Veltkamp, R. (2015). Corpus Analysis Tools for Computational Hook Discovery. In *Proceedings of the 16th International Society for Music Information Retrieval Conference (ISMIR 2015)* (pp. 227–233). Málaga, Spain.
- Velarde, G., & Meredith, D. (2014). A Wavelet-Based Approach to the Discovery of Themes and Sections in Monophonic Melodies. In *Music Information Retrieval Evaluation eXchange*.
- Velarde, G., Weyde, T., & Meredith, D. (2013). An approach to melodic segmentation and classification based on filtering with the Haar-wavelet. *Journal of New Music Research*, 42(4), 325–345. doi: [10.1080/09298215.2013.841713](https://doi.org/10.1080/09298215.2013.841713)
- Verhoef, T. (2013). *Efficient coding in speech sounds. Cultural evolution and the emergence of structure in artificial languages* (Ph.D. Thesis). University of Amsterdam.
- Vitouch, O. (2003). Absolutist models of absolute pitch are absolutely misleading. *Music Perception: An Interdisciplinary Journal*, 21(1), 111–117.
- Vitouch, O., & Gaugusch, A. (2000). Absolute recognition of musical keys in non-absolute-pitch-possessors. In *Proceedings of the 6th International Conference on Music Perception and Cognition (ICMPC 2000)*. Keele, United Kingdom.
- Volk, A., & van Kranenburg, P. (2012). Melodic similarity among folk songs: An annotation study on similarity-based categorization in music. *Musicae Scientiae*, 16(3), 317–339. doi: [10.1177/1029864912448329](https://doi.org/10.1177/1029864912448329)
- Volk, A., Wiering, F., & van Kranenburg, P. (2011). Unfolding the Potential of Computational Musicology. In *Proceedings of the Thirteenth International Conference on Informatics and Semiotics in Organisations: Problems and Possibilities of Computational Humanities (ICISO 2011)* (pp. 137–144). Leeuwarden, the Netherlands.
- Wang, C.-i., Hsu, J., & Dubnov, S. (2015). Music Pattern Discovery with Variable Markov Oracle: A Unified Approach to Symbolic and Audio Representations. In *Proceedings of the 16th International Society for Music Information Retrieval Conference (ISMIR 2015)* (pp. 176–182). Málaga, Spain.
- Ward, G. (2002). A recency-based account of the list length effect in free recall. *Memory & Cognition*, 30(6), 885–892. doi: [10.3758/BF03195774](https://doi.org/10.3758/BF03195774)
- Widmer, G. (2016). Getting Closer to the Essence of Music: The Con Espressione

- Manifesto. *ACM Transactions on Intelligent Systems and Technology*, 8(2), 19:1–19:13. doi: [10.1145/2899004](https://doi.org/10.1145/2899004)
- Wiora, W. (1941). Systematik der musikalischen Erscheinungen des Umsingens. In *Jahrbuch für Volksliedforschung* (pp. 128–195). Freiburg, Germany: Zentrum für Populäre Kultur und Musik.
- Zajonc, R. B. (1980). Feeling and thinking: Preferences need no inferences. *American Psychologist*, 35(2), 151–175. doi: [10.1634/theoncologist.9-90005-10](https://doi.org/10.1634/theoncologist.9-90005-10)
- Zatorre, R. J., & Halpern, A. R. (1993). Effect of unilateral temporal-lobe excision on perception and imagery of songs. *Neuropsychologia*, 31(3), 221–232.

GLOSSARY

<i>Alignment</i>	A computational method which, in the context of this dissertation, compares two pitch sequences, and finds the optimal correspondences between pitches, which may result in gaps in one sequence in relation to the other sequence.
<i>Anacrusis</i>	The first note of a melody or phrase if it appears before the first accent defined by the meter.
<i>Area Under Curve (AUC)</i>	A measure describing retrieval success, describing the area under the Receiver Operating Characteristic curve.
<i>Audio</i>	In Music Information Retrieval: analysis of music recordings.
<i>Background corpus</i>	A collection of musical pieces on which training of computational model (e.g., models for expectancy) takes place.
<i>Bar</i>	A subdivision of a musical piece, typically containing two to three accented notes, defined by a meter.
<i>Bias</i>	In cultural evolution, a bias refers to the tendency of cultural items to be transmitted in a certain way. See also prestige bias, conformity bias, content bias.
<i>Binary</i>	A variable with two outcomes, e.g., true vs. false, or occurrence vs. non-occurrence.
<i>Cadence</i>	A melodic or harmonic formula marking the end of a phrase or melody.
<i>Chord</i>	Several pitches sounding at the same time.
<i>Chord progression</i>	A sequence of chords.
<i>Chroma</i>	See pitch chroma.
<i>City-block distance</i>	A similarity measure based, in the context of this dissertation, on the difference of pitches in two compared pitch sequences.

<i>Classification</i>	The procedure of sorting items into categories, e.g., to assign melodies to tune families.
<i>Clustering</i>	The procedure of grouping items which are similar, e.g., to group melodies without prior categories such as tune families.
<i>Combined measure</i>	The pattern matching method developed in Chapter 4 , combining city-block distance, local alignment and structure induction to find occurrences of melodic phrases.
<i>Content bias</i>	A tendency to adopt a specific version of a melody due to the inherent properties of that melody.
<i>Contour</i>	The melodic contour is the perceived a melody, the way pitches move up or down.
<i>Correlation distance</i>	A similarity measure which compares, in the context of this dissertation, in how far two pitch sequences have a similar contour.
<i>Dependent variable</i>	In a statistical model, observations that are predicted by the model.
<i>Diachronic</i>	A diachronic analysis investigates change of cultural items over time.
<i>Drift</i>	A concept from evolutionary theory according to which variation is random, based on sampling artefacts in small populations.
<i>Duration</i>	In this dissertation, the length of a note in score time, e.g., quarter or half notes. Synonymous with inter-onset interval
<i>Euclidean distance</i>	A similarity measure which compares, in the context of this dissertation, the overall distance between values in two pitch sequences.
<i>Features</i>	In machine learning, another word for variables of models. In this dissertation, also synonymous to musical aspects.
<i>Fermata</i>	A prolonged note, mostly at the end of a phrase or musical piece.

<i>Fixed-length comparison</i>	The comparison of two pitch sequences of the same length.
<i>Frequency of occurrence</i>	The formalization of stability used in this dissertation: the more often a note sequence occurs, the more stable it is.
<i>Global musical aspects</i>	Musical aspects which describe a whole melody in one value or category.
<i>Harmony</i>	The presence of chord progressions in a musical piece. Chord progressions can also be implied, e.g., through an imagined accompaniment of a monophonic musical piece.
<i>Histogram intersection</i>	Used in this dissertation to compare in how far pitches or durations in two musical pieces may be results of transposition or time dilation.
<i>Informedness</i>	A measure which describes how many positive cases and negative cases from an annotation are correctly retrieved by a computational method.
<i>Instance</i>	In this dissertation, a given melodic phrase has an instance in a melody if two out of three annotators agree that the phrase occurs in that melody.
<i>Independent variable</i>	In a statistical model, the property that is used to predict observations.
<i>Inter-onset interval</i>	The space between two note onsets.
<i>Key</i>	In Western music notation, key describes in relation to which root note a musical piece is notated.
<i>Logistic regression</i>	A regression model which captures the relationship between continuous independent variables and binary dependent variables.
<i>Local alignment</i>	alignment which, in the context of this dissertation finds the optimal correspondence between a short pitch sequence within a melody.
<i>Local musical aspects</i>	Musical aspects which describe single notes or chords in a melody.

<i>Match</i>	In the context of this dissertation, a correspondence between a phrase with a melodic segment within a melody detected by a similarity measure.
<i>Majority vote</i>	The combination of multiple judgements through considering, e.g., a binary variable true only if a majority of judges has marked this variable as true. In the context of this dissertation, majority vote is used to combine annotations on phrase occurrences, and to accumulate the best similarity measures' results into the combined measure.
<i>Markedness</i>	A measure which describes the percentage of relevant results of the total number of items retrieved by a computational method.
<i>Maximally translatable pattern (MTP)</i>	In structure induction, the translation vector that leads to the most correspondences between two point sets representing melody notes.
<i>Meter</i>	A pattern of accented and unaccented beats underlying rhythms.
<i>Melody</i>	In the context of this dissertation, a melody is described by the pitches and durations of its constituent notes.
<i>Melos</i>	The pitches of a melody.
<i>Monophonic</i>	A musical piece in which only one pitch at a time is heard.
<i>Music Information Retrieval (MIR)</i>	A research field which strives to obtain information from music recordings or scores through computational analysis.
<i>Node</i>	A connected unit in a network model.
<i>Negative predictive value</i>	The percentage of relevant negative cases out of all negative cases retrieved by a computational method.
<i>Neural network</i>	A network model which is inspired by connections of neurons in the human brain, consisting of layers of nodes.
<i>Octave</i>	Two periodic sounds of which one waveform's period is double the other waveform's period are perceived as an octave apart.

<i>Onset</i>	The start of a note, measured from the start of the piece.
<i>Occurrence</i>	In the context of this dissertation, a match found by a similarity measure that exceeds a given similarity threshold. Of the combined measure: a match that exceeds the optimized similarity thresholds of city-block distance, local alignment and structure induction.
<i>Phylogenetic tree</i>	A computational method to analyze potential ancestor relationships in data assumed to arise from an evolutionary process.
<i>Pattern discovery</i>	In the context of this dissertation, a computational method to infer repeating note sequences within a melody.
<i>Pattern matching</i>	In the context of this dissertation, a computational method which departs from a given note sequence and observes whether or not it occurs in a melody.
<i>Pitch</i>	Musical sounds are often perceived as possessing a specific pitch, caused by their waveforms' periodic structure.
<i>Pitch chroma</i>	Pitch chroma describes the different categories of pitches within an octave.
<i>Pitch height</i>	Pitch height describes whether a pitch is on the low or high end of the human hearing range.
<i>Pitch interval</i>	The difference between two pitches.
<i>Positive predictive value</i>	The percentage of relevant positive cases out of all positive cases retrieved by a computational method. Also called precision.
<i>Primacy effect</i>	In serial recall studies, the primacy effect describes the phenomenon that items at the start of a list are recalled better than later items.
<i>Receiver-Operating Characteristic (ROC) curve</i>	A curve which visualizes the relationship between relevant and irrelevant cases retrieved by a similarity measure over the whole range of possible similarity thresholds.

<i>Recency effect</i>	In serial recall studies, the recency effect describes the phenomenon that items at the end of a list are recalled better than earlier items.
<i>Recursion</i>	A structure which is described in terms of repetitions of itself. See also recursion.
<i>Regression</i>	A statistical model which infers whether the development of a given property (independent variable) is statistically related to a given observation (dependent variable).
<i>Rhythm</i>	Rhythm describes the structure of a melody with respect to the duration of its notes.
<i>Scale</i>	The set of distinct pitch chromas in a musical piece.
<i>Self-organizing maps</i>	A neural network which projects similarity relationships between items onto a map.
<i>Serial position effect</i>	In serial recall studies, the serial position effect describes that the order of items may have an effect on human abilities to recall the items. More specific effects related to serial position are the primacy and recency effect.
<i>Sensitivity</i>	The percentage of positive cases, out of all positive cases, retrieved by a computational method. Also known as recall.
<i>Similarity measure</i>	In the context of this dissertation, a similarity measure quantifies the correspondence between two note sequences.
<i>Specificity</i>	The percentage of negative cases, out of all negative cases, retrieved by a computational method.
<i>Structure induction</i>	In the context of this dissertation, a similarity measure which compares relationships of notes represented as pitch-onset points.
<i>Suffix tree</i>	An indexing structure used to speed up pattern discovery.
<i>Symbolic</i>	in Music Information Retrieval: analysis of music notation

<i>Synchronic</i>	A synchronic analysis investigates change of cultural items from the same time.
<i>Timbre</i>	The colour or texture of a sound, determined by its waveform.
<i>Tonic</i>	The prominent root note of a scale.
<i>Tonality</i>	The presence of harmony in a musical piece.
<i>Tune family</i>	A tune family is a group of related melodies, which share some, but not necessarily all melodic traits.
<i>Variable-length comparison</i>	The comparison of two pitch sequences which may differ in length.
<i>Viewpoint</i>	A different word for a local musical aspect, such as the pitch or duration of a note.
<i>Waveform</i>	A measurement of the air pressure fluctuations perceived as a sound.
<i>Wavelet transform</i>	In the context of this dissertation, a way of measuring contour changes within a melody.

SAMENVATTING

BEHOUDEN OF VERLOREN DOOR OVERLEVERING? Het analyseren en voorspellen van stabiliteit in Nederlandse volksliederen

Mijn proefschrift onderzoekt de variatie van Nederlandse volksliederen met computationele methoden, met specifieke aandacht voor stabiliteit, of de mate waarmee zich een melodie verzet tegen verandering in de muzikale overlevering. Het eerste hoofdstuk introduceert terminologie, geeft een overzicht van gerelateerd onderzoek van muzikale overlevering, bespreekt de Nederlandse liederenbank, en in bijzonder de *Meertens Tune Collections*, de basis van mijn analyses. Deel I van de dissertatie omvat het tweede, derde en vierde hoofdstuk, en focust op het kwantificeren van stabiliteit en variatie. Het tweede hoofdstuk onderzoekt welke muzikale aspecten belangrijk zijn om stabiliteit en variatie te onderzoeken. Hiervoor bespreek ik relevant onderzoek uit ethnomusicologie, muziekcognitie en Music Information Retrieval over muzikale variatie. De volgende twee hoofdstukken lichten twee methodes uit waarmee stabiliteit gekwantificeerd kan worden. Het derde hoofdstuk vat de bevindingen uit onderzoek over pattern discovery samen. Pattern discovery zou gebruikt kunnen worden om stabiele muzikale patronen te vinden die vaak in melodische varianten voorkomen. Het vierde hoofdstuk introduceert een nieuwe pattern matching methode, gebaseerd op een combinatie van goed presterende gelijkheidsmaten. Deze methode maakt het mogelijk om de stabiliteit van melodische frases in volksliederen te meten, gebaseerd op hun voorkomendheid in varianten. Deel II van mijn dissertatie omvat het vijfde en zesde hoofdstuk, laat zien hoe stabiliteit gemeten en voorspeld kan worden, en plaatst de constateerde bevindingen in een bredere context van culturele en muzikale evolutie. Het vijfde hoofdstuk gebruikt de speciaal ontwikkelde pattern matching methode om stabiliteit van volksliedfrases te meten. Het toetst een aantal hypotheses waarmee deze stabiliteit voorspeld kan worden. De resultaten laten zien dat een deel van stabiliteit en variatie voorspeld kan worden door de lengte, positie en aantal herhalingen van een frase binnen een melodie, evenals door verwacht melodisch materiaal en herhalende motieven binnen een frase. Terwijl een aanzienlijk gedeelte van variatie en stabiliteit hierdoor niet voorspeld kan worden, suggereert de aanwezigheid van relevante voorspellende factoren wel dat stabiliteit en variatie in muzikale transmissie niet geheel toevallig is. Het zesde hoofdstuk vat de bijdrages van dit proefschrift voor het bestuderen van muzikale transmissie, het kwantificeren en voorspellen van stabiliteit

samen. Het hoofdstuk discussieert ook de beperkingen van mijn onderzoek, en hoe deze beperkingen overkomen kunnen worden door verder onderzoek, geïnspireerd door bevindingen uit muziekcognitie, muzikaliteit en culturele evolutie.

SUMMARY

RETAINED OR LOST IN TRANSMISSION? Analyzing and Predicting Stability in Dutch Folk Songs

My dissertation investigates the variation of Dutch folk songs with computational methods, with a special interest in stability, or a melody's resistance to change in transmission. The first chapter introduces terminology, gives an overview of related work on music transmission, and reviews the Dutch folk song database and in particular the *Meertens Tune Collections*, which were studied in this dissertation. Part I of the dissertation, comprising the second, third and fourth chapter, establishes possible ways to quantify stability and variation. The second chapter investigates which musical aspects would be a good focus to study stability and variation, reviewing relevant research from ethnomusicology, music cognition and Music Information Retrieval on musical variation. The two ensuing chapters review computational methods by which stability might be quantified. The third chapter summarizes findings from pattern discovery research, which might be used for inferring stable melodic patterns occurring frequently in variants. The fourth chapter introduces a novel pattern matching method based on a combination of well-performing similarity measures, which may be used to quantify stability of folk song phrases based on their frequency of occurrence. Part II of the dissertation, comprising the fifth and sixth chapter, shows how stability can be measured and predicted, and places the contributions of the current research into the wider context of cultural and music evolution. The fifth chapter uses the specifically developed pattern matching method to determine stability of melodic phrases, and tests a number of hypotheses by which such stability may be predicted. The results show that a moderate amount of stability and variation can be explained through the length, position and number of repetitions of a phrase within a melody, as well as whether it consists of expected melodic material or repeating motifs. While a considerable amount of stability and variation cannot be explained through these hypotheses, the detection of relevant predictors suggests that stability and variation in music transmission are not random. The sixth chapter summarizes the contributions of my dissertation to the study of music transmission, as well as to the quantification and prediction of stability. The chapter also discusses the limitations of the presented approaches, and how they may be overcome by future research, inspired by concepts from music cognition, musicality, and cultural evolution.

ZUSAMMENFASSUNG

ERHALTEN ODER VERLOREN DURCH ÜBERLIEFERUNG? Analyse und Prognose von Stabilität in niederländischen Volksliedern

Meine Doktorarbeit untersucht die Variation von niederländischen Volksliedern, mit besonderem Fokus auf Stabilität, worunter ich den Widerstand einer Melodie gegen Veränderung in der musikalischen Überlieferung verstehe. Das erste Kapitel führt Begriffe ein, gibt einen Überblick über den Forschungsstand zur musikalischen Überlieferung, und bespricht die *Dutch folk song database* und im besonderen die *Meertens Tune Collections*, die die Grundlage für meine Studien bilden. Teil I der Dissertation umfasst das zweite, dritte und vierte Kapitel, und erörtert Möglichkeiten, Stabilität und Variation zu quantifizieren. Das zweite Kapitel untersucht, welche musikalischen Aspekte ein geeigneter Fokus für die Erforschung von Variation und Stabilität wären, und fasst hierzu Erkenntnisse aus relevanten Studien in der vergleichenden Musikwissenschaft, Musikkognition und Music Information Retrieval zusammen. Die zwei folgenden Kapitel behandeln rechnergestützte Methoden um Stabilität zu quantifizieren. Das dritte Kapitel fasst Ergebnisse aus der Pattern Discovery-Forschung zusammen, welche genutzt werden könnten, um stabile melodische Fragmente zu bestimmen. Das vierte Kapitel führt eine neuartige Pattern Matching-Methode ein, die auf einer Kombination von erfolgreichen Ähnlichkeitsmaßen beruht. Diese Methode kann eingesetzt werden, um Stabilität von melodischen Phrasen auf Basis ihrer Häufigkeit in Volksliedvarianten zu bestimmen. Teil II der Dissertation umfasst das fünfte und sechste Kapitel, und zeigt Methoden für die Messung und Prognose von Stabilität auf, sowie den Zusammenhang zwischen Studien zu musikalischer Stabilität und kultureller und musikalischer Evolution. Das fünfte Kapitel benutzt die speziell entwickelte Pattern Matching-Methode um Stabilität von melodischen Phrasen zu bestimmen, und untersucht eine Reihe von Hypothesen, mit Hilfe derer Stabilität prognostiziert werden kann. Die Ergebnisse zeigen, dass Stabilität bis zu einem gewissen Grad mit Länge, Position und der Anzahl Wiederholungen einer Phrase in einem Volkslied in Zusammenhang steht. Zusätzliche Einflussfaktoren sind die Hörerwartungen die durch melodisches Material geschaffen werden, und wiederholende Motive in den Phrasen. Ein erheblicher Anteil von Stabilität und Variation kann nicht durch diese Hypothesen erklärt werden, aber die Anwesenheit von relevanten Prediktoren legt nahe, dass Stabilität und Variation in musikalischer Überlieferung nicht zufällig sind. Das sechste Kapitel erörtert den Beitrag meiner Dissertation zur Erforschung von musikalischer Über-

lieferung und zur Quantifizierung und Prognose von Stabilität. Das Kapitel zeigt auch die Einschränkungen meiner Methodik auf, und erbringt Vorschläge, wie diese Einschränkungen unter Zuhilfenahme von Erkenntnissen aus Musikkognition, Musikalität und kultureller Evolution anhand zukünftiger Studien behoben werden können.

BIOGRAPHY

Berit Janssen was born in Brockel, Germany, on May 30, 1983. She started her studies of Systematic Musicology and English Literature at the University of Hamburg in 2003. From 2005 to 2006, she studied at Anglia Ruskin University, Cambridge, where she took classes in Composition, Music Software and English Literature. In 2009, she obtained the degree Magistra Artium (MA) from the University of Hamburg. Her Master thesis investigated the influence of reverberation on sound and music. She worked as a researcher and programmer at the Studio for Electro-Instrumental Music (STEIM), Amsterdam, between 2009 and 2010, and as a workshop designer for applying digital technology to the arts at the Centrum voor Kunst en Cultuur (CKC), Zoetermeer, between 2010 and 2011. In 2012, she started her Ph.D. at the Meertens Institute as part of the Computational Humanities project *Tunes & Tales*, funded by the Royal Netherlands Academy of Arts and Sciences. Since 2017, she has been working as a developer at the Digital Humanities Lab of the University of Utrecht.

TITLES IN THE ILLC DISSERTATION SERIES

- ILLC DS-2009-01: **Jakub Szymanik**
Quantifiers in TIME and SPACE. Computational Complexity of Generalized Quantifiers in Natural Language
- ILLC DS-2009-02: **Hartmut Fitz**
Neural Syntax
- ILLC DS-2009-03: **Brian Thomas Semmes**
A Game for the Borel Functions
- ILLC DS-2009-04: **Sara L. Uckelman**
Modalities in Medieval Logic
- ILLC DS-2009-05: **Andreas Witzel**
Knowledge and Games: Theory and Implementation
- ILLC DS-2009-06: **Chantal Bax**
Subjectivity after Wittgenstein. Wittgenstein's embodied and embedded subject and the debate about the death of man.
- ILLC DS-2009-07: **Kata Balogh**
Theme with Variations. A Context-based Analysis of Focus
- ILLC DS-2009-08: **Tomohiro Hoshi**
Epistemic Dynamics and Protocol Information
- ILLC DS-2009-09: **Olivia Ladinig**
Temporal expectations and their violations
- ILLC DS-2009-10: **Tikitu de Jager**
"Now that you mention it, I wonder...": Awareness, Attention, Assumption
- ILLC DS-2009-11: **Michael Franke**
Signal to Act: Game Theory in Pragmatics
- ILLC DS-2009-12: **Joel Uckelman**
More Than the Sum of Its Parts: Compact Preference Representation Over Combinatorial Domains
- ILLC DS-2009-13: **Stefan Bold**
Cardinals as Ultrapowers. A Canonical Measure Analysis under the Axiom of Determinacy.
- ILLC DS-2010-01: **Reut Tsarfaty**
Relational-Realizational Parsing

- ILLC DS-2010-02: **Jonathan Zvesper**
Playing with Information
- ILLC DS-2010-03: **Cédric Dégrement**
The Temporal Mind. Observations on the logic of belief change in interactive systems
- ILLC DS-2010-04: **Daisuke Ikegami**
Games in Set Theory and Logic
- ILLC DS-2010-05: **Jarmo Kontinen**
Coherence and Complexity in Fragments of Dependence Logic
- ILLC DS-2010-06: **Yanjing Wang**
Epistemic Modelling and Protocol Dynamics
- ILLC DS-2010-07: **Marc Staudacher**
Use theories of meaning between conventions and social norms
- ILLC DS-2010-08: **Amélie Gheerbrant**
Fixed-Point Logics on Trees
- ILLC DS-2010-09: **Gaëlle Fontaine**
Modal Fixpoint Logic: Some Model Theoretic Questions
- ILLC DS-2010-10: **Jacob Vosmaer**
Logic, Algebra and Topology. Investigations into canonical extensions, duality theory and point-free topology.
- ILLC DS-2010-11: **Nina Gierasimczuk**
Knowing One's Limits. Logical Analysis of Inductive Inference
- ILLC DS-2010-12: **Martin Mose Bentzen**
Stit, lit, and Deontic Logic for Action Types
- ILLC DS-2011-01: **Wouter M. Koolen**
Combining Strategies Efficiently: High-Quality Decisions from Conflicting Advice
- ILLC DS-2011-02: **Fernando Raymundo Velazquez-Quesada**
Small steps in dynamics of information
- ILLC DS-2011-03: **Marijn Koolen**
The Meaning of Structure: the Value of Link Evidence for Information Retrieval
- ILLC DS-2011-04: **Junte Zhang**
System Evaluation of Archival Description and Access
- ILLC DS-2011-05: **Lauri Keskinen**
Characterizing All Models in Infinite Cardinalities

- ILLC DS-2011-06: **Rianne Kaptein**
Effective Focused Retrieval by Exploiting Query Context and Document Structure
- ILLC DS-2011-07: **Jop Briët**
Grothendieck Inequalities, Nonlocal Games and Optimization
- ILLC DS-2011-08: **Stefan Minica**
Dynamic Logic of Questions
- ILLC DS-2011-09: **Raul Andres Leal**
Modalities Through the Looking Glass: A study on coalgebraic modal logic and their applications
- ILLC DS-2011-10: **Lena Kurzen**
Complexity in Interaction
- ILLC DS-2011-11: **Gideon Borensztajn**
The neural basis of structure in language
- ILLC DS-2012-01: **Federico Sangati**
Decomposing and Regenerating Syntactic Trees
- ILLC DS-2012-02: **Markos Mylonakis**
Learning the Latent Structure of Translation
- ILLC DS-2012-03: **Edgar José Andrade Lotero**
Models of Language: Towards a practice-based account of information in natural language
- ILLC DS-2012-04: **Yurii Khomskii**
Regularity Properties and Definability in the Real Number Continuum: idealized forcing, polarized partitions, Hausdorff gaps and mad families in the projective hierarchy.
- ILLC DS-2012-05: **David García Soriano**
Query-Efficient Computation in Property Testing and Learning Theory
- ILLC DS-2012-06: **Dimitris Gakis**
Contextual Metaphilosophy - The Case of Wittgenstein
- ILLC DS-2012-07: **Pietro Galliani**
The Dynamics of Imperfect Information
- ILLC DS-2012-08: **Umberto Grandi**
Binary Aggregation with Integrity Constraints
- ILLC DS-2012-09: **Wesley Halcrow Holliday**
Knowing What Follows: Epistemic Closure and Epistemic Logic
- ILLC DS-2012-10: **Jeremy Meyers**
Locations, Bodies, and Sets: A model theoretic investigation into nominalistic mereologies

- ILLC DS-2012-11: **Floor Sietsma**
Logics of Communication and Knowledge
- ILLC DS-2012-12: **Joris Dormans**
Engineering emergence: applied theory for game design
- ILLC DS-2013-01: **Simon Pauw**
Size Matters: Grounding Quantifiers in Spatial Perception
- ILLC DS-2013-02: **Virginie Fiutek**
Playing with Knowledge and Belief
- ILLC DS-2013-03: **Giannicola Scarpa**
Quantum entanglement in non-local games, graph parameters and zero-error information theory
- ILLC DS-2014-01: **Machiel Keestra**
Sculpting the Space of Actions. Explaining Human Action by Integrating Intentions and Mechanisms
- ILLC DS-2014-02: **Thomas Icard**
The Algorithmic Mind: A Study of Inference in Action
- ILLC DS-2014-03: **Harald A. Bastiaanse**
Very, Many, Small, Penguins
- ILLC DS-2014-04: **Ben Rodenhäuser**
A Matter of Trust: Dynamic Attitudes in Epistemic Logic
- ILLC DS-2015-01: **María Inés Crespo**
Affecting Meaning. Subjectivity and evaluativity in gradable adjectives.
- ILLC DS-2015-02: **Mathias Winther Madsen**
The Kid, the Clerk, and the Gambler - Critical Studies in Statistics and Cognitive Science
- ILLC DS-2015-03: **Shengyang Zhong**
Orthogonality and Quantum Geometry: Towards a Relational Reconstruction of Quantum Theory
- ILLC DS-2015-04: **Sumit Sourabh**
Correspondence and Canonicity in Non-Classical Logic
- ILLC DS-2015-05: **Facundo Carreiro**
Fragments of Fixpoint Logics: Automata and Expressiveness
- ILLC DS-2016-01: **Ivano A. Ciardelli**
Questions in Logic

- ILLC DS-2016-02: **Zoé Christoff**
Dynamic Logics of Networks: Information Flow and the Spread of Opinion
- ILLC DS-2016-03: **Fleur Leonie Bouwer**
What do we need to hear a beat? The influence of attention, musical abilities, and accents on the perception of metrical rhythm
- ILLC DS-2016-04: **Johannes Marti**
Interpreting Linguistic Behavior with Possible World Models
- ILLC DS-2016-05: **Phong Lê**
Learning Vector Representations for Sentences - The Recursive Deep Learning Approach
- ILLC DS-2016-06: **Gideon Maillette de Buy Wenniger**
Aligning the Foundations of Hierarchical Statistical Machine Translation
- ILLC DS-2016-07: **Andreas van Cranenburgh**
Rich Statistical Parsing and Literary Language
- ILLC DS-2016-08: **Florian Speelman**
Position-based Quantum Cryptography and Catalytic Computation
- ILLC DS-2016-09: **Teresa Piovesan**
Quantum entanglement: insights via graph parameters and conic optimization
- ILLC DS-2016-10: **Paula Henk**
Nonstandard Provability for Peano Arithmetic. A Modal Perspective
- ILLC DS-2017-01: **Paolo Galeazzi**
Play Without Regret
- ILLC DS-2017-02: **Riccardo Pinosio**
The Logic of Kant's Temporal Continuum
- ILLC DS-2017-03: **Matthijs Westera**
Exhaustivity and intonation: a unified theory
- ILLC DS-2017-04: **Giovanni Cinà**
Categories for the working modal logician
- ILLC DS-2017-05: **Shane Noah Steinert-Threlkeld**
Communication and Computation: New Questions About Compositionality
- ILLC DS-2017-06: **Peter Hawke**
The Problem of Epistemic Relevance
- ILLC DS-2017-07: **Aybüke Özgün**
Evidence in Epistemic Logic: A Topological Perspective

ILLC DS-2017-08: **Raquel Garrido Alhama**

Computational Modelling of Artificial Language Learning: Retention, Recognition & Recurrence

ILLC DS-2017-09: **Miloš Stanojević**

Permutation Forests for Modeling Word Order in Machine Translation

