

# ULRR

## Melodic similarity algorithms for scores – a comparative evaluation of contrasting approaches

Item Type	Thesis
Authors	Cahill, Margaret
Download date	2026-04-16 14:54:06
Item License	<a href="https://creativecommons.org/licenses/by-nc-sa/1.0/">https://creativecommons.org/licenses/by-nc-sa/1.0/</a>
Link to Item	<a href="https://hdl.handle.net/10344/11188">https://hdl.handle.net/10344/11188</a>

# **Melodic Similarity Algorithms for Scores – A Comparative Evaluation of Contrasting Approaches**

Margaret Cahill

Submitted in partial fulfillment of the requirements for the degree of Ph.D.

The University of Limerick.

Supervised by Dr. Donncha Ó Maidín

Submitted to the University of Limerick, November 2008.

# **Abstract**

## **Melodic Similarity Algorithms for Scores – A Comparative Evaluation of Contrasting Approaches**

**Margaret Cahill**

This thesis is concerned with melodic similarity algorithms for musical scores. The performance of two contrasting approaches is explored and the chosen algorithms fine-tuned and evaluated using human observations of similarity. The most relevant musical features for assessing melodic similarity are identified from music perception research. Two contrasting algorithmic approaches are selected – a geometric algorithm and the string-matching edit distance approach. A number of different versions of both algorithms are implemented to assess the success of the musical features used. The internal weights of the algorithms are fine-tuned using a testbed of melodies for which human judgements of similarity have been gathered. These melodies are extracted from a piece of music in Theme and Variation form. While focusing on perceptual accuracy of the human similarity judgements, the best performing algorithms are identified and discussed. The internal algorithm weights are verified using additional extracts from the set of Theme and Variations. The ability of the algorithms to successfully generalise to a broader range of music is explored using two further collections of melodies in contrasting musical styles for which human observations of similarity exist.

## **Declaration**

**Title:** Melodic Similarity Algorithms for Scores – A Comparative Evaluation of Contrasting Approaches

**Author:** Margaret Cahill

**Award:** Ph.D.

**Supervisor:** Dr. Donncha Ó Maidín

I hereby declare that this thesis is entirely my own work, and does not contain material previously published by any other author, except where due reference or acknowledgement has been made. Furthermore, I declare that it has not previously been submitted for any other academic award.

Margaret Cahill, November 2008.

## **Acknowledgements**

My heartfelt thanks goes to my supervisor Dr. Donncha Ó Maidín for his patience, words of wisdom and insight in the long years it has taken to complete this thesis and for his generosity in seeing it through to the end despite having retired. Thanks to Deirdre also for allowing me to regularly invade her home during the last year of this work and for always making me feel welcome.

Thanks are also due to my long-suffering friends and family who have listened to the many moments of frustration I had over the course of this research and who never questioned the time it was taking.

Most of all I am incredibly grateful to Kieran, who believed that I could finish this work when I very much doubted my own ability to see it through. The completion of this thesis is due in no small part to the endless encouragement he gave me from afar.

# Table of Contents

<b>Chapter 1 Introduction.....</b>	<b>1</b>
1.1 Melodic Similarity Algorithms for Music Scores .....	1
1.2 The Choice of Algorithms.....	2
1.3 Music Representations and the Level of Information Available.....	3
1.4 Method .....	5
1.5 Contributions .....	6
1.6 Thesis Layout .....	7
<b>Chapter 2 Score Features for use with Melodic Similarity Algorithms.....</b>	<b>9</b>
2.1 Musical Features of a Score .....	9
2.2 Melodic Similarity in Music Perception Literature.....	10
2.2.1 Overview of Melodic Memory .....	10
2.2.2 Pitch .....	11
2.2.3 Transposition of Melodies .....	12
2.2.4 Rhythm and memory for melody .....	13
2.2.5 Tempo .....	15
2.2.6 Melody Identification and Music Features .....	16
2.2.7 Accents .....	16
2.2.8 Summary of Accents and Relevance to Melodic Similarity Algorithms .....	21
2.3 Features Used by Other Researchers.....	23
2.3.1 Features Other than Pitch and Duration .....	24
2.4 Musical Features Chosen for Implementation .....	26
<b>Chapter 3 Gathering Human Similarity Judgements for a Testbed of Melodies ..</b>	<b>31</b>
3.1 Design Considerations for the Listening Experiment .....	31
3.2 Related Listening Experiments .....	32

3.3 Design Considerations.....	33
3.3.1 Music Material Used in Related Experiments.....	33
3.3.2 The Choice of Testbed Melodies.....	35
3.3.3 The Type of Similarity Judgements.....	37
3.3.4 The Rating Scale Used .....	38
3.3.5 Pilot Tests .....	38
3.3.6 The Structure of the Listening Experiment .....	39
3.3.7 Practice runs, demonstrations etc. ....	40
3.3.8 Comments and Feedback.....	41
3.4 The Listening Experiment – Implementation and Method .....	42
3.4.1 Subjects.....	42
3.4.2 Generation of MIDI Files .....	42
3.4.3 Procedure .....	43
3.5 Results .....	43
3.5.1 Part A (bars 1-4 of the variations) .....	43
3.5.2 Part B (bars 5-9 of the variations) .....	44
3.5.3 Consistency of Subjects.....	45
3.5.4 The Reduced Data Set – Part A .....	46
3.5.5 The Reduced Data Set – Part B .....	47
3.5.6 Consistency and Reliability of the Human Similarity Judgements Gathered.....	47
3.5.7 Pooling the Ratings.....	50
<b>Chapter 4 The Algorithms.....</b>	<b>52</b>
4.1 Implementation Environment.....	52
4.2 The Geometric Melodic Similarity Algorithm.....	52
4.2.1 Eliminating the Effect of Melody Length.....	54

4.2.2 Transposition to different keys .....	54
4.3 Implementing the Geometric Melodic Similarity Algorithm.....	55
4.3.1 Pitch difference only.....	55
4.3.2 Pitch difference weighted by duration.....	56
4.3.3 Musical Implications of the Four Duration Methods .....	57
4.3.4 Metrical accents .....	60
4.3.5 Combining pitch, duration, and metrical accent.....	61
4.3.6 Rests .....	61
4.4 Edit Distance Algorithms – An Introduction .....	62
4.5 Smith McNab and Witten – A Basic Edit Distance Algorithm.....	63
4.5.1 Filling the Matrix.....	64
4.6 Mongeau and Sankoff – Extensions to the Basic Algorithm .....	64
4.7 Tracing the Edits Operations and Checking Note Alignments .....	67
4.8 Implementing the Edit Distance Algorithms.....	68
4.8.1 Expected Results .....	70
<b>Chapter 5 Fine-tuning the Algorithms and Results .....</b>	<b>72</b>
5.1 Related Research .....	72
5.2 Assessing the Performance of the Algorithms Using the Human Similarity Judgements .....	74
5.2.1 Normalising the data.....	75
5.2.2 Comparing the Algorithm Output with the Human Similarity Judgements..	75
5.3 Using a Hill-Climbing Algorithm to Fine-tune the Metrical Accent Weights.....	77
5.4 Results for the Geometric Algorithm .....	79
5.4.1 Issues Encountered when Using All Nine Variation Melodies .....	81
5.4.2 Transpositions.....	82
5.4.3 Different Time Signatures .....	82

5.4.4 Compression/Expansion in the Time Domain.....	83
5.4.5 The Time Ratio Technique for Melodies in Different Time Signatures .....	84
5.4.6 Metrical Accent Weights in Melodies with Different Time Signatures.....	86
5.4.7 The Variation Technique used in Variation VII.....	86
5.4.8 Results when Variations IV and VI are included. ....	88
5.4.9 Discussion of Results.....	89
5.5 Fine-tuning the Edit Distance Algorithms .....	90
5.5.1 Checking the Trace .....	90
5.6 Results for Smith et al.'s Edit Distance Algorithms .....	92
5.7 Results of Mongeau and Sankoff's Edit Distance Algorithms .....	93
5.7.1 Mongeau and Sankoff – pitch difference without fragmentation and consolidation.....	94
5.7.2 Mongeau and Sankoff – pitch consonance without fragmentation and consolidation.....	94
5.7.3 Discussion of the Pitch Consonance Results. ....	95
5.7.4 Mongeau and Sankoff - pitch difference with fragmentation and consolidation.....	98
5.7.5 Discussion of the fragmentation and consolidation results .....	98
5.7.6 Mongeau and Sankoff - pitch consonance with fragmentation and consolidation.....	100
5.8 Combined Results for the Geometric and Edit Distance Algorithms. ....	102
5.9 Results when Variations IV and VI are Included.....	105
5.10 Verification of the Algorithmic Results Using the Part B Melodies from the Testbed .....	110
5.10.1 Assessing the Performance of the Algorithms .....	110
5.10.2 The Geometric Algorithms.....	111
5.10.3 The Edit Distance Algorithms .....	114

5.10.4 The Edit Distance Algorithms with the Fine-tuned Values from the set of Eight Variations .....	117
5.10.5 Conclusion of Verification of Results using the Part B Melodies.....	118

**Chapter 6 Exploring the Ability of the Algorithms to Generalise to other Music 120**

6.1 The Additional Collections of Melodies .....	121
6.2 Exploring the Fine-tuned Algorithms using MIREX ground truth Data .....	121
6.2.1 MIREX 2005 ground truth Data.....	121
6.2.2 The Edit Distance Algorithms and Transposed Melodies .....	123
6.2.3 The Construction of the ground truth .....	124
6.2.4 Assessing the Performance of the Algorithms using the MIREX Data .....	125
6.3 Query Melody 000.111.706 .....	126
6.3.1 Results .....	127
6.3.2 The Geometric Algorithms.....	130
6.3.3 The Edit Distance Algorithms .....	132
6.3.4 Which Melody is More Similar to the Query Melody?.....	134
6.4 Query Melody 800.000.193 .....	136
6.4.1 Results .....	138
6.4.2 The Geometric Algorithms.....	140
6.4.3 The Edit Distance Algorithms .....	140
6.5 Query Melody 230.005.489 .....	142
6.5.1 Results .....	145
6.5.2 Differences between the Algorithms and the ground truth.....	145
6.5.3 Further Discussion of the Results .....	145
6.5.4 Fine-tuned Values for the Set of Eight Variations .....	148
6.5.5 Summary of Findings .....	149

6.6 Exploring the performance of the Algorithms Using A Collection of Irish Folk Music .....	150
6.6.1 Overview of Ceol Rince na hÉireann and Irish Folk Music .....	150
6.6.2 The Tunes Chosen from Ceol Rince na hÉireann, Vol 1. ....	150
6.6.3 Breathnach’s Annotations .....	152
6.6.4 The Algorithms Implemented.....	152
6.6.5 Results .....	153
<b>Chapter 7 Conclusion .....</b>	<b>158</b>
7.1 Summary of Findings .....	158
7.1.1 Fine-tuning the Geometric Algorithms.....	159
7.1.2 Fine-tuning the Edit Distance Algorithms.....	161
7.1.3 Conclusions of the Fine-Tuning Stage .....	162
7.1.4 Verifying the Results .....	163
7.1.5 The Ability of the Algorithms to Generalise to Other Melodies.....	163
7.1.6 Issues Encountered with the MIREX Ground Truth .....	164
7.2 Limitations of the Algorithms and Potential Solutions.....	165
7.2.1 Alterations in the Time Domain and Transposition of Melodies .....	165
7.2.2 Rests .....	167
7.2.3 The Length of Melodies .....	168
7.3 Future Work .....	169
7.3.1 Extensions to the Edit Distance Algorithms.....	169
7.3.2 Implementation and Investigation of Melodic Accents.....	169
7.3.3 Melodic Search Systems and Segmentation .....	169
Appendix A The Listening Experiment/Testbed Melodies .....	172
Appendix B Results of the Listening Experiment .....	175
Appendix C Graphed Results for a Range of Metrical Accent Values.....	196

Appendix D Sample Melodies from Ceol Rince na hÉireann .....	198
Appendix E Publications.....	200
References .....	201

## List of figures

Figure 2.1: The possible locations of the pitch contour accent is shown by the arrow. ....	28
Figure 2.2: The first two bars of Variation IX containing many contour changes. ....	28
Figure 3.1: The first two bars of the Theme and Variations IV and VII. ....	35
Figure 3.2: The first four bars of the Theme, Variation III and Variation V ....	36
Figure 4.1: The division of the score into time window. ....	53
Figure 4.2: An example of the duration weights when Method 1 is used. ....	58
Figure 4.3: An example of the duration weights when Method 2 is used. ....	58
Figure 4.4: The implications of using Method 2 for calculating the duration weights. ....	59
Figure 4.5: An example of the duration weights when Method 3 is used. ....	59
Figure 4.6: The implications of using Method 3 for calculating the duration weights. ....	59
Figure 4.7: An example of the duration weights when Method 4 is used. ....	60
Figure 4.8: An example of the fragmentation of a note. ....	66
Figure 4.9: An example of the consolidation of a group of notes. ....	66
Figure 4.10: Excerpt from an example source and target melody. ....	67
Figure 4.11 The first two bars of the Theme and Variation VIII and Variation IX. ....	70
Figure 5.1: The time windows used to process the first bar of the Theme and Variation I. ....	82
Figure 5.2: The first two bars of the Theme and Variations IV and VI. ....	83
Figure 5.3: Sample melodies showing expansion (or stretching) in the time domain. ....	83
Figure 5.4: The adjusted durations and time windows of Melody 2 from Figure 5.3. ....	84
Figure 5.5: Each bar of the Theme is matched against a bar of Variation IV. ....	84
Figure 5.6: Each bar of the Theme is matched against two bars of Variation IV. ....	87
Figure 5.7: An example source and target melody with one possibility for the alignment of notes. ....	91
Figure 5.8: Five sample melodies, illustrating Mongeau and Sankoff's pitch consonance weights. ....	97
Figure 5.9: Mongeau and Sankoff's example of fragmentation ....	99
Figure 5.10: Examples of the fragmentation of a quarter-note from the Theme. ....	99
Figure 5.11: An example of a questionable fragmentation identified by the algorithm. ....	99
Figure 5.12: Example of fragmentation from bar 4 of Variation VIII. ....	100
Figure 6.1: Query melody 000.111.706 and five candidate melodies. ....	126
Figure 6.2: An example of an Appoggiatura. ....	127

Figure 6.3: An example of an Acciaccatura .....	127
Figure 6.4: The pitch differences between the query melody and candidate melody 000.116.073. ....	131
Figure 6.5: Candidate melody 000.113.932. ....	131
Figure 6.6: Candidate melody 000.113.932 as processed by the geometric algorithm .....	131
Figure 6.7: The pitch differences between the query melody and candidate melody 000.113.932. ....	132
Figure 6.8: The query melody aligned with candidate melody 000.116.073. ....	132
Figure 6.9: The edit operations reported for candidate melody 000.116.073. ....	133
Figure 6.10: Candidate melody 00.113.932. ....	133
Figure 6.11: The comparison of notes between the query melody and candidate melody 000.113.032. .	134
Figure 6.12: Four melodies featuring the same pitches and the same relative durations. ....	134
Figure 6.13: Melody A from Figure 42 in $\frac{3}{4}$ time. ....	135
Figure 6.14: Query melody 800.000.193 and six candidate melodies. ....	137
Figure 6.15: Sample melodies in different keys with common note pitches in terms of distance from the tonic. ....	141
Figure 6.16: The query melody. ....	141
Figure 6.17: Candidate melody 000.109.445. ....	142
Figure 6.18: Query melody 230.005.489 and seventeen candidate melodies. ....	144
Figure 6.19: Candidate melodies 451.504.065 - a time-stretched transposed version of melody 230.004.687. ....	147
Figure 6.20: The common pitches between the query melody and candidate melody 700.001.741. ....	148
Figure 7.1: Two melodies that differ only by the presence of a rest in melody 2. ....	168

## List of Tables

Table 2.1: The main musical features of a score. ....	9
Table 2.2: Accent categories defined by Parncutt (1997). ....	19
Table 2.3: Features used for determining similarity in symbolic music. ....	23
Table 3.1: Examples of other rating scales for melodic similarity. ....	38
Table 3.2: The melodies used in each part of the listening experiment. ....	39
Table 3.3: The layout of the listening experiment. ....	42
Table 3.4: Inter-subject correlation calculations for the reduced data sets. ....	48
Table 3.5: An example of the inter-subject correlation using two subjects. ....	49
Table 3.6: An example of the inter-subject correlation using two subjects. ....	49
Table 3.7: The Cronbach’s alpha reliability values. ....	50
Table 3.8: The median ratings gathered in the listening experiment. ....	51
Table 4.1: Mongeau and Sankoff’s (1990) weights based on the consonance of the intervals. ....	65
Table 4.2: Mongeau and Sankoff’s (1990) weights when either note is not in a major or minor key. ....	65
Table 4.3: The versions of the edit distance algorithms implemented. ....	70
Table 5.1: The range of the human and algorithmic similarity measures. ....	74
Table 5.2: The meaning of the extreme values of the human and algorithmic similarity measures. ....	74
Table 5.3: An example of the calculation of the SumAbsDiff value. ....	76
Table 5.4: Results for the six 4/4 variation melodies, sorted by the SumAbsDiff result. ....	80
Table 5.5: The duration/adjusted duration of the melodies from Figure 5.3. ....	84
Table 5.6: The calculation of the first time window adjusted by the time ratio. ....	85
Table 5.7: The calculation of the time windows of the first bar adjusted by the time ratio. ....	85
Table 5.8: An example of metrical accent weights for a range of time signatures based on the 4/4 time signature weights. ....	87
Table 5.9: Results for the eight variation melodies sorted by the SumAbsDiff result. ....	89
Table 5.10: The results for the versions of Smith et al.’s (1998) edit distance algorithms. ....	93
Table 5.11: The results for the pitch difference versions of Mongeau and Sankoff’s (1990) edit distance algorithms. ....	94
Table 5.12: The results for the pitch consonance versions of Mongeau and Sankoff’s (1990) edit distance algorithms. ....	95

Table 5.13: The results for the pitch difference with fragmentation/consolidation versions of Mongeau and Sankoff's (1990) edit distance algorithms. ....	98
Table 5.14: The results for the pitch consonance with fragmentation/consolidation versions of Mongeau and Sankoff's (1990) edit distance algorithms. ....	101
Table 5.15: The combined results for the geometric and edit distance algorithms for the six Part A 4/4 variation melodies, sorted by the SumAbsDiff result. ....	104
Table 5.16: The combined results for the geometric and edit distance algorithms for the eight Part A variation melodies, sorted by the SumAbsDiff result. ....	107
Table 5.17: The sum of the difference in ranks. ....	110
Table 5.18: An example of the actual algorithm results before the ranks are calculated. ....	111
Table 5.19: The comparison of the ranks between the human and algorithmic similarity measures. ....	112
Table 5.20: The results of the geometric algorithms for the six 4/4 Part B variations melodies. ....	113
Table 5.21: The results of the edit distance algorithms for the six 4/4 Part B variations melodies. ....	115
Table 5.22: An example of the normalised results from the edit distance algorithms. ....	116
Table 5.23: The calculation of the sum of the difference in ranks for the edit distance algorithm results. ....	116
Table 5.24: The combined geometric and edit distance results for the eight Part B variation melodies. .	118
Table 6.1: The combined results of the geometric and edit distance algorithms for RISM melody 000.111.706. ....	130
Table 6.2: The combined results of the geometric and edit distance algorithms for RISM melody 800.000.193. ....	139
Table 6.3: The similar melodies from Ceol Rince na hÉireann identified by Breathnach and the algorithms. ....	157
Table 7.1: The top three geometric algorithms (6 variation melodies). ....	159
Table 7.2: The overall top five algorithms (6 variation melodies). ....	160
Table 7.3: The top three geometric algorithms (8 variation melodies). ....	160
Table 7.4: The top three edit distance algorithms (6 variation melodies). ....	161
Table 7.5: The overall top five algorithms (8 variation melodies). ....	162

# Chapter 1

## Introduction

### 1.1 Melodic Similarity Algorithms for Music Scores

A melodic similarity algorithm is one that compares a pair of melodies and produces a value that reflects the degree of similarity/dissimilarity between them. This research is concerned with identifying suitable melodic similarity algorithms for use with monophonic musical scores. Such an algorithm could be used to search collections of scores for exact and similar matches, an endeavour which forms part of the field of Music Information Retrieval (MIR), so called after its text equivalent, Information Retrieval (IR). In musical terms, melodic similarity algorithms can be used to automate the process of comparing various occurrences of a melody, theme or motif in a single piece of music or across an entire database of scores. Examples of applications of such algorithms include an investigation of recurring melodic themes throughout a composer's entire body of work, the comparison of the similarity of themes and motifs used by different composers, or simply searching a database for a tune whose name has been forgotten. The nature of the relationship between similar melodies can take a number of forms. In some instances, such as in Theme and Variation form, there are marked purposeful similarities between melodies. Similarity between melodies also exists due to borrowings from melodies, such as the re-use of motifs in a work. Melodies from oral traditions such as folk music also feature many similarities as the melodies are often slightly incrementally modified over time as they are passed on between musicians. Finally, coincidental similarities also occur where an unplanned relationship exists between melodies.

While melodic similarity algorithms are useful for identifying purposefully related/similar melodies they are also potentially useful for uncovering new unknown relationships between melodies and as such are useful for music analysis and musicology research. Algorithms can be deemed successful when their results approximate human perceptions of similarity.

Applying a computer to the issue of similarity, through the use of algorithms and collections of scores in digital representation formats, makes possible in a matter of seconds and minutes what might previously have taken a music researcher days, weeks, months or even longer to complete and enables search tasks that would otherwise have been infeasible. A number of collections of digitized scores (sometimes referred to as corpora) that are suitable for processing with such algorithms are currently available as a result of large digitizing projects (MuseData; Kern; EsAC).

## **1.2 The Choice of Algorithms**

In the past, researchers often encoded scores in a variety of representation formats which are stored as text files on computer for processing. Currently, many commercial music notation applications and software for scanning musical scores allow the user to easily save music notation in score representation formats, thus providing a document processable by computer. Comparing documents of text for Information Retrieval (IR) purposes is a well-established field and as a result, many of the algorithms used to measure the similarity between melodies are based on approximate string-matching algorithms (Mongeau and Sankoff 1990; Orpen and Huron 1992; McNab et al. 1997; Smith et al. 1998; Crawford et al. 1998; Uitdenbogerd and Zobel 1998, 1999; Downie 1999; Lemström 2000, Robine et al. 2007). Methods based on exact string-matching algorithms such as Knuth-Morris-Pratt, Boyer-Moore and Shift-Or algorithms are not considered as the degree of similarity is in question and not simply the identification of exact matches only. Also, since most score representation formats are text-based, this was a natural starting point for work in the area of music similarity. The most commonly used string-matching algorithm in this context is the edit-distance algorithm. Chen remarks that “In the field of MIR, approximation is measured mainly by the edit distance...” (Chen 2003). Edit distance algorithms assign a cost value for deleting, replacing, and inserting notes in order to change one melodic string into another (Smith et al. 1998; Uitdenbogerd and Zobel 1999; Mongeau and Sankoff 1990; Orpen and Huron 1992). However, melodies are multi-dimensional and operations that are suited to processing text are not necessarily suitable for use in a musical context. In order to evaluate the suitability of this approach, Mongeau and Sankoff’s (1990) and Smith et al.’s (1998) edit distance algorithms are implemented and the internal algorithm parameters fine-tuned using a testbed of melodies for which human judgments of

similarity have been collected. The results are compared to those of a contrasting approach using a geometric algorithm for melodic similarity (Ó Maidín 1982 and 1998a).

In a geometric algorithm, a melody is represented as a two-dimensional collection of music objects, rather than as a string. The algorithm implemented here was shown to be successful for comparing segments of folk music pieces in Ó Maidín (1998a). This algorithm is loosely based on music perception principles and the research on melodic memory and recognition presented in Chapter 2 informs aspects of this particular implementation of the algorithm. The internal algorithm parameters are also fine-tuned using the same testbed and human observations as the edit distance algorithms.

The results of both algorithms are compared using the human similarity observations as a benchmark. The fine-tuned internal parameters are then further verified on a related set of testbed melodies with human similarity judgements. Finally, the success of the algorithms and their ability to generalise to a wider range of music is explored.

The testbeds used here consists of pre-segmented melodies. This research is not concerned with the internal structure of these melodic segments, which is the focus of melodic pattern extraction research (Lartillot 2007; Lartillot 2005; Conklin and Anagnostopoulou 2006).

In this research the issue of melodic similarity algorithms is approached primarily as a musical problem, and secondly as a computer processing problem. This is reflected in the focus on accuracy relative to human perceptions of similarity, rather than on the speed of processing attained by the algorithm on the computer, and in the investigation and implementation of ideas that arise from the perceptual literature related to melodic memory and recognition.

### **1.3 Music Representations and the Level of Information Available**

There are many researchers working on retrieval, search and similarity tasks for music in such disparate fields of interest as musicology and music theory, engineering, computer science, library and information science and cognitive science and psychology (Downie and Futrelle 2002). Researchers use music in audio, MIDI, score, and other formats, with most concentrating on a single one of these representations. Each

representation contains a particular level of information about a piece of music. MIDI representations of scores and score notation formats are often described as symbolic representations, as they provide prescriptive information about how a piece of music should sound or be played, as opposed to audio formats, which store the actual music itself (in sampled form). It is noted that MIDI is not always used for such prescriptive purposes but can also be used to store the recording of a performance. The music score generally represents the composer's intentions for the piece of music and includes not only pitch and duration information about notes, but instructions on how individual notes and passages of music should be played. Score databases in MIDI format store the pitch, duration, and velocity of the notes of a score as commands to be executed over time and is a much-reduced representation when compared to the detail contained in a musical score. In some search projects pre-processing of the score database, and/or the melody to be searched for, is carried out to convert them to an abstracted representation. Examples of such representations include schemes that express a piece of music as a series of pitches with inter-onset-interval (the distance from the start of one note to the start of the next) and representing pitch in terms of interval or contour classes. String-matching algorithms often represent a piece of music as a series of pitches (intervals, contour, pitch height etc.), as the nature of such an algorithm is to reduce the musical content to a set of strings that can be processed as text. An audio recording of a piece of music, on the other hand, consists of samples of sound from performance of the piece and stores enough information that allows that performance to be replayed exactly as was recorded. Recording of different performances and different performers are aurally different, as no instantiations of the piece are ever exactly the same.

The nature of the musical information represented by each of these music formats is pertinent to the design, use and implementation of melodic similarity algorithms. Information available in one format is often not readily accessible when using another.

The score is the most detailed and comprehensive of the symbolic representations and is the music representation format used in this research. It contains detailed instructions on how notes and passages of music should be played, including implicit details such as articulation, phrasing, key and time signatures and barlines. Further details not explicitly notated in the score are also understood by musicians to be inherent in the music. Such musical elements can be difficult to extract from audio files and many of them

are not present in MIDI files. One of the main questions of this research is to determine which musical features available from a score representation are most useful for melodic similarity algorithms. A full discussion of this issue is presented in the next chapter.

## **1.4 Method**

The research was broken down into the following main steps:

1. The identification of potentially useful melodic features from music perception research.
2. The use of a listening experiment to collect similarity ratings for a small testbed of music that consists of melodies that are known to be similar to one another.
3. The implementation of some of the relevant features identified by the music perception literature in modified versions of an existing geometric melodic similarity algorithm.
4. The implementation of multiple versions of two edit distance algorithms as examples of a common string-matching method.
5. The fine-tuning of weights (internal algorithm parameters) used in both algorithms using the human perceptual judgements of similarity (see step 2).
6. The identification of the most successful features and algorithms in comparison to the human similarity judgements.
7. The verification of the adjusted weight values using new but related melodies for which similarity judgements were also gathered (see step 2).
8. The exploration of the success of the algorithm, the features used and the adjusted weight values in a broader musical context, using two stylistically different collections of music for which human observations on the similarity of the melodies are available.

Among the questions to be answered over the course of this research are the suitability and usefulness of not only the chosen algorithms, but of the weight values identified from the fine-tuning stage and the features suggested by the music perception literature. The appropriateness of the use of the initial testbed of melodies and the related

similarity judgements are also under scrutiny, as is the ability of the fine-tuned algorithms to generalise to other melodies.

## 1.5 Contributions

The process of algorithm design and evaluation adopted here is a methodical and objective approach to the problem. In much of the literature relating to melodic similarity algorithms, the researchers themselves arbitrarily pick the musical features to be used and the values for internal components of the algorithm, for example, Ó Maidín's metrical stress weights (1998a) and Smith et al's insert and delete costs (1998). In other cases, the success of the algorithm is judged purely by the researcher, for example in the case of Mongeau and Sankoff (1990), where they adjust the internal algorithm weights so that they get the closest similarity results to their own subjective judgement. The use of reliable and objective similarity judgements to fine-tune the internal algorithm values, rather than arbitrarily picking these values, is a distinctive characteristic of this research.

A review of music perception literature on melodic memory and identification is a key feature of this work. Theories and experimental findings relevant to melodic similarity are identified and discussed. These findings are used to inform choices made in the implementation of the algorithms, rather than relying solely on the intuition and musical knowledge of the researcher. A survey of the musical features used in some of the most commonly cited melodic similarity research is carried out and research that uses features other than pitch and duration alone is identified.

The set of human similarity judgements for the chosen testbed can also be seen as a contribution to the field. (There are no commonly used test and evaluation collections such as those available for text-retrieval.) The choice of testbed material (melodic segments from real music), the carefully constructed listening experiment and the analysis of the results for consistency combine to provide reliable perceptual judgements.

The fine-tuned weights and values are themselves an important contribution to this field of research. The metrical accent values that result in the best performance for each geometric algorithm are noted, as are the cost and weight values that provide the best performance for each version of the edit distance algorithms.

A further contribution is the comparative evaluation of two contrasting methods for algorithmically assessing melodic similarity. A string-based approach is compared to a more musically influenced geometric algorithm. Core differences in how both algorithms process the melodies are highlighted and illustrated with examples at the fine-tuning, verification and generalisation stages. The edit distance algorithm is among the most commonly used in determining melodic similarity and so the evaluation of its performance is noteworthy.

The verification of the results and evaluation of how the algorithms generalise to music beyond that of the testbed melodies is also an important extension to the evaluation of the algorithms.

## **1.6 Thesis Layout**

Chapter 2 of this thesis is introduced with a list of the various features present in a musical score. Music perception research, specifically that concerned with melodic memory and melody identification is then presented. These areas are explored as a potential source of ideas for the musical features that are most appropriate to use in the implementation of the algorithms, particularly the geometric algorithm, which was designed specifically for calculating melodic similarity. Finally, a survey of the features used in prominent melodic similarity research is included.

Chapter 3 details the testbed of melodies selected for the optimisation process and discusses a listening experiment conducted to gather similarity judgements for these melodies.

Chapter 4 presents the geometric algorithm and the two string-matching algorithms investigated in this research. A number of variants of these three algorithms are constructed so that factors highlighted by the perceptual literature in Chapter 2 can be explored.

In Chapter 5 the internal weights and parameters of the algorithms are fine-tuned according to a selection of the similarity judgements detailed in Chapter 3. The adjustment of the weights in the case of the geometric algorithm is carried out using a hill-climbing technique. The results are compared and discussed. The fine-tuned values

are then verified using a set of additional melodies and similarity judgements from Chapter 3.

In Chapter 6, the algorithms with the fine-tuned weights are run on two stylistically different collections of melodies, for which some degree of similarity is already known. The ability of the algorithms to successfully generalise to melodies other than the initial testbed is examined.

Chapter 7 summarises the research carried out, identifies the findings and discusses the limitations of the both algorithmic approaches. An overview is given of possible extensions to these algorithms and further work that might be carried out.

## Chapter 2

### Score Features for use with Melodic Similarity Algorithms

One of the motivations of this research is to determine which of the musical features available in a score may be useful for melodic similarity algorithms. In this chapter an overview is given of the musical features present in a score. Music perception literature relevant to melodic similarity is then investigated as an aid to selecting the most appropriate score features for the task. This is followed by a survey of the features used in some of the most cited research on melodic similarity algorithms and an overview of the features selected for use in this work.

#### 2.1 Musical Features of a Score

Table 2.1 illustrates the main features or elements used in music notation.

Feature	Example/Comment
clef	treble, bass, soprano,
time signature	
key signature	(no of sharps, flats etc. not interpretation of key – could be minor) C (none), A (3 sharps)
instrumentation	violin, piano etc. (usually only relevant in polyphonic music)
note pitch	
accidentals	sharp, flat, natural
note duration	crotchet, quaver etc.
ties	
barlines	normal, repeat, final
rest	
articulation	staccato, tenuto, pizzicato,
dynamics	f, p, mf, cresc., desc., graphic hairpins
ornamentation	trills, turns, grace notes, glissando
phrasing	slurs (implying legato playing or phrasing)
tempo directions	rubato, accelerando,
extra pitch instructions	8ve down, up etc.
specific instrument instructions	mute, bowing directions, pedal
beaming	groups of 2 or 4 quavers etc
other	fermata,

**Table 2.1: The main musical features of a score.**

In addition to these explicitly notated elements, an implied regular pattern of strong and weak beats reinforces the meter in each bar of music when the score is played. This is

relevant to the score representation since it is implicit according to the time signature and grouping of music into bars, even if it is not explicitly notated. It should be noted that syncopation is not considered here as part of this work. Melodies involving such displacement from the main beats of the bar would require specialist treatment and is beyond the scope of this research.

## **2.2 Melodic Similarity in Music Perception Literature**

The choice of musical features is a key issue in developing perceptually accurate melodic similarity measures. In most instances, however, researchers do not give reasons for or discuss the implications of choices they have made in this regard. In the context of this research a typically successful algorithm produces results that reflect human concepts of similarity/dissimilarity between melodies. It seems a natural step, therefore, to use music perception literature as a guide to deciding which features are most relevant. The intention is not to construct a model of the human perception process but look to this research for pointers to the main features that people use in comparing melodies and assess their relevance to this score-driven approach to melodic similarity algorithms. A body of research exists that investigates the human capacity to remember melodies (often referred to as memory for melody or melodic memory) and how we identify melodies. Both of these areas have relevance for comparing melodies and judging the degree of similarity between melodies. Indeed, many of the listening experiments used in this type of research include tasks to judge the similarity between melodies in order to make assumptions about memory and identification capabilities.

There is a need to be cautious when interpreting the melodic memory research. The features that people remember may not be the features they use to compare and judge the similarity of melodies. However, as Schulkind et al. (2003, p.218) note, “a musical feature that is difficult to perceive will be unlikely to aid identification” and so this review of the literature indicates, at the very least, which features are least likely to be useful for comparison and similarity.

### **2.2.1 Overview of Melodic Memory**

Melodic memory research encompasses a broad range of topics that attempt to examine and explore how we store and retrieve melodies from our memory. Often familiar and unfamiliar melodies are examined separately as it is thought that the encoding and

retrieval process is different for each. Much of the melodic memory research focuses on pitch aspects of memory including intervals, contour and key distance. Some attention is paid to duration and rhythm. A further branch of related research examines the concept of musical accents, which are described by Parncutt (1997, p.16) as “any relatively salient event, or any event that attracts the attention of a listener more than surrounding events”.

The concepts of short-term memory and long-term memory are important in the context of melodic memory. Short-term memory has a small capacity (7 +/- 2 items) and the contents of this type of memory only last for up to 30 seconds (Dowling and Harwood 1986, p.139). Long-term memory has an unlimited capacity and can store memories for up to a lifetime. Research that examines new melodies usually deals with short-term memory, while those that use known melodies investigate long-term memory.

### **2.2.2 Pitch**

Much of the perceptual research involving pitch and melodic memory focuses on the use of pitch contour and pitch intervals. Known and unknown melodies are usually investigated separately as the internal representation of melody is thought to vary with familiarity and time. Dowling and Fujitani (1971), Dowling (1978), Edworthy (1985), and Dewitt and Crowder (1986) found that contour was used for unknown/new melodies. Dowling and Fujitani (1971) and Dowling (1978) also found that the mode/scale was also important and that contour was not used on its own. Deutsch (cited in Dowling 1978) found that melodies were stored as sequences of pitches for familiar melodies and both Attneave and Olsen (cited in Dowling 1978) and Dowling and Fujitani (1971) found that the exact interval sizes between pitches were remembered. So it seems that contour with some reference to mode or scale is used to store and/or retrieve new unfamiliar melodies from memory and more detailed information about exact intervals is used for known melodies. Many of the listening experiments used in this research involve playing subjects a reference melody and a comparison melody that has one or more aspects altered and asking them to compare the melodies. Thus, these experiments use short-term memory for unfamiliar melodies. Familiar melodies are stored in long-term memory. It could be inferred that since short-term memory has a low capacity, contour information is usually good enough for comparison and differentiation between melodies. Conversely long-term memory has such large

capacity that more detailed information is needed to retrieve data. A number of studies have examined the change in memory and recognition for melodies over time. Dowling et al. (2002) tested memory over short and long delays and found contour and key were remembered over short delays but that the features used became more precise over a period of up to 4 minutes. Similarly, Dewitt and Crowder (1986) found in an experiment that contour is used for short delays and intervals for longer delays of up to 20-30 seconds (the upper limits of short-term memory). However, a second experiment they conducted did not corroborate this. There is a suggestion that on initially hearing a melody people use a rough storage mechanism (contour and key/scale information) and that over time this information is processed and stored in a more detailed format (intervals).

Levitin has examined the concept of Absolute Pitch, which allows people to produce or recognise specific pitches without reference to a standard pitch (1994, 1999). In a listening experiment that asked subjects to perform (sing, hum, or whistle) any song they knew well from 58 CDs provided, 12% of people performed the song without any pitch errors in two trials and 40% performed without error on at least one trial (1994). Levitin does not suggest that Absolute Pitch alone is used, which would mean that a melody would only be recognisable in that key. This indicates that some degree of information relating to intervals must also be available alongside the Absolute Pitch information.

### **2.2.3 Transposition of Melodies**

A further area of research related to melody recognition looks at transposition and how it affects our perception of similarity between melodies. It is generally agreed that melodies played an octave higher or lower are still essentially the same melody, with Francès (1988) noting that “a melody and its transposition to the octave (or double octave) are indistinguishable”. Deutsch (1972) investigated the ability to recognise familiar tunes played in different octaves. She states that memory for pitch generalises over octaves so that we recognise a tune that is played an octave higher or lower. (When the familiar tune was played with each note in a randomised octave subjects were no longer able to recognise the tune.)

There are two schools of thought regarding melodic transposition to different keys. Key distance refers to the distance between the keys of two melodies. A melody in C major

is closer in key distance to G major than to E major or A flat major. Each flat or sharp added to the key signature increases the distance between keys. Transposed melodies that are nearer in key distance are considered more similar to those transposed to farther key distances by Cuddy, Cohen, Mewhort (1981), Cuddy, Cohen, Miller (1979), Takeuchi and Hulse (cited by Deutsch 1999).

However, Francès (1988), Hershman (1994), and van Egmond and Povel (1994, 1996) found that the similarity judgement between the original and transposed melody was influenced more by pitch distance. The pitch distance is calculated by counting the number of semi-tones each notes is shifted up or down when transposed. Thus, an original melody transposed from C major to G major is considered less similar than a melody transposed to D major or E flat majors.

Despite this difference in opinion regarding transposition distances and similarity, all do agree that melodies transposed to different keys are still highly recognisable as the same melody. White (1960, p.100) noted that “simple transposition has virtually no effect on the ease with which a melody is recognized”.

#### **2.2.4 Rhythm and memory for melody**

Research into melody recognition and the role of rhythm has found that although rhythm alone is not a good feature for recalling and recognising melodies, it is important when used in conjunction with pitch. In this context rhythm is referred to as “the serial pattern of variable note durations in a melody”, as in Schulkind (1999, p.896).

Hébert and Peretz (1997) note that variation in dynamics, timbre, and tempo probably facilitate recognition but that these are considered secondary features and not determining factors. They conducted two experiments to explore the role of melodic and temporal cues. In an experiment that replicates earlier work by White (1960), they played two different groups of subjects a number of known melodic excerpts that had been modified in one of two ways. The first group were played an isochronous (all notes having equal duration) version of the melodies, while the other group were played a version that featured the same rhythm as the original melodies but had just a single constant pitch for every note. The results of this experiment showed that about 50% of subjects could name the tunes using the pitch information alone but that only 6% of subjects recognised the melodies when the original rhythm was used but the pitch kept

constant for each note. This indicates that pitch alone is more useful than rhythm alone in identifying known melodies. In a second experiment, Hébert and Peretz (1997) interchanged the rhythm and melodies of these musical excerpts to create new melodies and asked subjects to identify the melodies. One group was instructed to base their judgement on the melodic pattern, ignoring the rhythm, while another group was told to pay attention to the rhythm and to ignore the pitches of the tunes. Similar results to the first experiment were obtained, with 53% of subjects correctly identifying the melodies based on pitch and only 8% of those who based their judgements on the rhythm recognising the tunes correctly. Hébert and Peretz also played subjects the unaltered standard tunes and reported high recognition rates (>92%) as a comparison with the altered melodies. They do note that the changes they made to rhythm also affected the temporal (long notes) and melodic accents (the presence of the tonic, median, and dominant) according to Boltz (cited in Hébert and Peretz 1997). Accents are mentioned in the next section so this will not be discussed in-depth here but when more melodic and temporal accents combine they draw attention to a particular point in the music and hence aid the recognition process. Hébert and Peretz note that by changing the rhythm of the melodies they interrupt this joint accent structure and thus affect the ability of the subjects to identify the tune. They conclude that melody information is at least required for recognising familiar/stored melodies but that the “optimal code seems to require a combination of both melodic and rhythmic information” (p.528).

Schulkind (1999) investigates the interdependence of meter, phrasing, rhythmic contour, and the ratio of successive durations. Subjects in an experiment were asked to identify well-known songs that were presented in the original rhythm and with three alterations: isochronous (all notes same duration), with the original rhythm shifted on by one note, and with the durations of notes from the original randomised. As with Hébert and Peretz, Schulkind found that the recognition of the unaltered melodies were higher than the altered versions with isochronous rhythm. The mean percentage of songs identified was 96.4% and 85.8% respectively, when subjects identified the songs from a list of target and foil songs, and a much larger difference of 82.5% and 55.5% when they were asked to identify the songs without the help of a list of titles. Indeed, all three alterations to rhythm resulted in worse performance. A further experiment sought to control the four features being investigated more closely. He concludes that the more of

these features that were preserved, the better the melody identification. He proposes that each of these four features acts interdependently.

Demorest and Kim (2002) also studied temporal and melodic cues for recognition of familiar melodies. Their work follows on from that of Hébert and Peretz (1997) and Schulkind (1999). Subjects were asked to identify the title of a number of well-known melodies and to rate how well they knew the song. Subjects were divided into three groups and played either a pitched version with isochronous rhythm, a rhythmic version without pitch, or an un-pitched version with metric context. The metric context was given by playing a two bar metronome introduction before the rhythm. In both of the rhythmic versions, an accent was given on the first beat of the bar to instil a sense of strong and weak beats. They reported better recognition for melodies than Hébert and Peretz for the rhythmic version of the melodies, which is attributed to the use of the accent on the first beat of the bar. The addition of the rhythmic context by giving the metronome did not further improve results. They conclude from the experiment that certain temporal accents such as metrical accents are effective cues for melodic memory. In all cases musically trained subjects performed much better than untrained subjects.

### **2.2.5 Tempo**

Levitin and Cook (1996) note from a review of research in tempo that “people easily recognize songs in which the relation between rhythmic elements (the rhythmic pattern) is held constant, but the overall timing or musical tempo has been changed.” (Monahan 1993 and Serafine 1979, cited in Levitin and Cook 1996, p.928). They used data from an earlier experiment for absolute pitch in which subjects picked a song they knew very well from a collection of 600 on 58 CDs. They could whistle, sing, or hum the song, could start anywhere in the song and typically sang a four-bar phrase. None of the subjects had heard their chosen song in the preceding 72 hours. In one trial, 72% of the subjects reproduced a tempo that was within 4% of the actual song tempo. This study shows that even non-musicians can reproduce tempo with great accuracy.

In a second experiment that replicated an earlier one by Halpern (1988 cited in Levitin and Cook 1996), subjects were asked to sing what the author called three familiar folk songs (Happy Birthday, We Wish you a Merry Christmas, and Row Row Row Your Boat). The variability in tempo was in the 10-20% range for these songs. These songs

are mostly heard in informal situations and do not have a standardised tempo whereas the songs from the previous CDs were songs that would have been heard many times in the same tempo.

### **2.2.6 Melody Identification and Music Features**

Schulkind, Posner, and Rubin carried out an experiment to identify the musical features that facilitate melody identification (2003). They first survey music perception research to determine which musical features to use in their experiment. Seven features are identified that should help in melody identification: intervals, durations, pitch height, tonal function, contour, metrical accent, rhythm, and phrase boundaries. Each category of feature was further subdivided and 25 features in total were investigated. A listening experiment was conducted in which subjects were asked to identify some well-known songs in as few notes as possible. Starting with two notes, one note was added at a time until they could identify the melody with certainty or reached the 20<sup>th</sup> note. Notes 3-12 were analysed according to the 25 features previously identified as useful and regression analysis was used to find the predictive variables responsible for the identification.

The most significant features for melody identification found by Schulkind et al. were duration (long notes), metrical accent, contour patterns of alternations and pairs, and phrase boundaries. More than one change in the contour direction of pitch is considered an alternation as based on research by Restle (cited in Schulkind et al. 2003). A pair is defined as “any repeated combination of two notes sharing an easily recognizable pitch relationship” (p.231). They state that this feature was selected due to the presence of such patterns in the melodies they used but this is an ambiguous definition. Identification was highest at phrase boundaries but they note that multiple accents often coincide at the beginning and end of phrases. This finding is discussed in a later section. Schulkind et al also noted that temporal features contributed more to melody identification than the melodic pitch features they investigated. These included intervals, contour, tonal function in key, and pitch height (distance from lowest or median note).

### **2.2.7 Accents**

Huron and Royal (1996) describe an accent as “an increased prominence, noticability, or salience ascribed to a given sound event” (p.489) and Jones (1987) refers to accents as “anything that is relatively attention-getting in a time pattern” (p.623). Musical

features that attract our attention are more likely to be used in comparing melodies and so research relating to musical accents is consulted here. Music perception researchers have identified a number of different accent types and research has been carried out into the role they play in helping people remember and perform melodies. These include accents based on the metrical stress of a melody, the duration of notes, changes in direction of the melodic line, the use of large intervals between adjoining notes, pauses and rests, among other musical features. The terminology used to discuss these accents varies from researcher to researcher and so each author's work is presented in turn before the common ideas are identified and discussed in the following section (2.2.8).

Drake, Dowling, and Palmer (1991) investigated metric, rhythmic grouping, and melodic accents. They consider metric accents to be an increase in the intensity of a note and that these accents create the system of strong and weak beats that make up the meter of a piece of music. They indicate that first note event in a bar is assigned a primary accent and usually the mid-point of a bar is given a secondary accent. According to their research a rhythmic grouping accent is created by duration, with a long note or a pause between notes signifying the end of a group. Finally, they report that a melodic accent is created by jumps or changes of direction in pitch. A jump is considered a leap of more than three semi-tones between notes and the accent on the change of direction is considered to be present on the extreme note, before the note that moves in the opposite direction. They theorised that if all three accents coincided they would reinforce each other and draw attention to that particular point. They ran experiments in which pianists were asked to play back simple tunes they heard. The performance of these tunes was better when the accents coincided. When the accents did not coincide the rhythmic playback was not affected but the pianists ability to play the correct melody did deteriorate. They also found that more accented notes were played correctly than non-accented notes.

Drake and Palmer (1993) followed this research with some further experiments to investigate whether performers emphasised these types of accents. They were asked to play simple musical sequences in which there was only one accent present and more complex sequences in which all three types of accents were included in a coinciding or conflicting manner. They alter the definition of the melodic accent created by a change

of pitch direction so that the accent may be on the extreme note or on the following note which starts the actual change of direction.

Huron and Royal (1996) focus on melodic accents, which they consider to be the most “contentious”. They identify seven melodic accents from music theory literature: treble, bass, registral extreme, interval size, interval descent, interval ascent and contour pivot accents. Treble and bass accents are formed by extremes of high or low notes respectively, while both extreme higher and lower pitches can be considered registral extreme accents. Jumps in pitch are considered interval size accents, and interval descent and interval ascent accent are jumps in pitch with the direction taken into account. Finally, contour pivot accent places an accent on a change of direction in pitch. Perceptual research on melodic accents is then examined and they find support for the contour pivot accent, interval ascent, and interval size accent. They feel that treble and bass accents are called into question by perceptual literature. The contour pivot accent was supported by Thomassen (1982), who developed a model that assigned accent values for notes based on a three-note contour that took the preceding and subsequent notes into account. They conducted a number of experiments, including one that took 100 Western folk melodies and calculated the correlation between each type of accent and metric position in the bar. They consider duration to be an agogic accent and include it here also. Their approach is to determine if these suggested accents coincide at particular locations in the music, which would suggest that the composer recognised them as accents and intended to use them in this way. They found Thomassen’s contour pivot model to be important. However, they found that this type of accent was only about 1/10<sup>th</sup> of the magnitude of agogic (durational) accents and conclude that “melodic accent(s) may be a weak factor in rhythmic perception”. They did find that melodic accent was significantly positively correlated with metre and is therefore used to “establish, maintain, or strengthen the perceived sense of meter”. They also conclude that the variance normally attributed to interval size ascent is actually accounted for by contour pivot accents.

Lerdahl and Jackendoff (1983) also discuss the role of accents in the context of their model of musical understanding. They identify three categories of accent: phenomenal, structural, or metrical. Here, a phenomenal accent is defined as an “event in the musical surface that give emphasis or stress to a moment in the musical flow” (p.17). This

includes attack points, local changes such as sforzando, sudden changes in dynamics or timbre, long notes, leaps to high or low notes, and harmonic changes. Structural accents are considered to be “an accent caused by the harmonic/melodic points of gravity in a phrase/section – especially the cadence”. Finally, metrical accents are defined as “any beat that is relatively strong in its metrical context”. These accents are quite different to the main accents discussed by others. Parncutt (2003) points out ambiguities in Lerdahl and Jackendoff’s descriptions and categories. He proposes alternative categories and terminology to these in the context of his own research.

Parncutt (1997, 2003) discusses the role of accents in creating expressive computer simulations of piano performance. In the earlier work he classifies musical accents as follows based on Lerdahl and Jackendoff’s (1993) accent types, although they do not directly compare:

	<b>Structural Accents</b>	<b>Expressive Accents</b>
<b>time</b>	grouping	onset time (agogic)
	metrical	duration (articulatory)
		amplitude envelope
<b>pitch</b>	melodic (contour)	intonation
	harmonic	
<b>loudness</b>	dynamic	stress
<b>timbre</b>	instrument/orchestration	coloration

**Table 2.2: Accent categories defined by Parncutt (1997).**

The accents are divided according to the four main perceptual attributes of sound: time, pitch, loudness, and timbre. The expressive accents are those used by musicians to emphasise the structural accents. Since these are performed accents, and not those indicated by the notated score, these are included here for completeness but not discussed further. Grouping accents are defined as existing on the start and end notes of phrases, sections, and pieces. These accents are created according to Parncutt by the occurrence of larger time or pitch intervals between notes than occurs in the surrounding context. This means that a pitch leap of a large interval occurring between small interval relations and a long note surrounded by relatively shorter notes functions as a grouping accent to define phrase and higher structures.

Lerdahl and Jackdoff’s metrical accents are described as occurring “within and between beats, and within and between measures” in a hierarchical way (Parncutt 1997, p.16). Parncutt recognises the work of Drake and Palmer (1993) and Huron and Royal (1996)

and includes turns and skips in the melodic accent category, considering turns (changes in direction) to be more important, and peaks to be more important than valleys in these changes of direction. Harmonic accents are caused by changes of harmony on the horizontal axis or harmonic dissonances on the vertical axis. In the former case the salience of the accent depends on the distance between the chord and its surrounding context. Dynamic accents come from explicit notations in the score, and timbral accents arise from changes of instrument or articulation.

In a follow-on paper (2003), Parncutt refines these categories that were largely based on Lerdahl and Jackendoff's theory. He now prefers the main categories to be called immanent and performed accents. Immanent accents are simply those notated in the score and performed accents are added to the score as it is played. Parncutt finds some ambiguities in the division of accent types. Phenomenal accents had been defined by Lerdahl and Jackendoff (1983) as any surface event that emphasises or stresses a point in the music. Parncutt now argues that this definition is unsuitable as it could describe any kind of accent. Their structural accents, he also argues, refer to only one kind of musical structure and leave out further accents that may contribute to structure. A further immanent accent is added here by Parncutt, to include accents that fall on the notes of linear progression according to linear analysis such as Schenkarian analysis and Lerdahl and Jackendoff's prolongation reduction (1983). This accent is called a reductional accent and is placed in the pitch accent category.

Jones (1987) defines three types of melodic accent and notes that such accents are likely to mark the start and end of important time spans in music sequences. Thus, along with causing emphasis in the melodic line, these accents impact the perception of time periods. According to Jones, contour-based melodic accents are caused by changes in contour direction and the accent can be on the beginning or end point of a rising or falling pitch line. Such high and low pitch points are considered melodic markers for duration spans. The second melodic accent defined by Jones is a pitch-interval melodic accent. This is equivalent to the previously mentioned leaps or jumps in pitch. She states that since Western music uses predominantly small pitch changes in melodies (one to two semi-tones), a listener's attention is drawn to a note when a leap of a large interval occurs between it and the preceding note. Jones gives an example of four or five semi-tones as a large interval. Points of emphasis that arise from tonal expectation are

also considered accents. Tonal melodic accents include resolved notes at the end of phrases. This concept may have some commonality with Lerdahl and Jackendoff's structural accents.

Along with melodic accents, Jones defines temporal accents as caused by notes that are relatively longer or shorter than those surrounding them and by silences. Notes that are relatively longer or shorter than those around it draw the listener's attention. Silences arising from rests in music are also considered temporal accents but this depends on the metric grouping, pause length, and tempo involved.

Jones investigates the way in which these melodic and temporal accents interact. The way in which the accents interact, the strength of accent, and the regular temporal pattern between occurrences are combined to form the joint accent structure, which marks a higher-order time structure. Accents that coincide are judged to provide stronger emphasis than single accents alone, and the aim of this particular research is to investigate how this occurs through the use of listening experiments. As previously mentioned Herbert and Peretz (1997) found that Jones's joint accent structure was important for identifying melodies and that when the pattern of accents was changed people had difficulty with this task.

Schulkind et al's (2003) experiment findings are also relevant in the context of accents. Some overlap occurs between the general musical features they find useful in identifying melodies and the current discussion on accents. Metrical accents are explicitly mentioned as a factor in melody identification. (Two different sets of metrical stress patterns were used with different values for notes on strong and weak beats of the bar). Long notes were also found to be a factor and as has been shown, are considered by some to be a type of accent. They attribute the high identification-rate they found at phrase boundaries to the fact that notes in such locations usually having more coinciding accents than those in the middle of phrases.

The work of Demorest and Kim (2002) was mentioned in section 2.2.4 on general music features. They found that the identification of melodies from rhythm alone improved if metrical accents were used.

### **2.2.8 Summary of Accents and Relevance to Melodic Similarity Algorithms**

The occurrence of accents causes points of salience in music that attract the listener's attention. Notes at these points of emphasis form key points in a melody that are

remembered more precisely and performed more correctly than those non-accented notes. Therefore, accented notes may be important in determining the similarity between melodies and are potentially useful in similarity algorithms. Although different researchers use different terminology and categorise accents into different types, there is some agreement on the actual musical features that act as accents in melodies. There does seem to be some consensus that changes in contour and leaps cause melodic accents. Huron and Royal (1996), Drake, Dowling and Palmer (1991), Drake and Palmer (1993), and Jones all agree on this. Lerdahl and Jackendoff (1983) consider only leaps as accents. Parncutt (1997, 2003) does consider changes in contour as a melodic accent but defines leaps in pitch as a grouping accent.

Drake, Dowling and Palmer (1991), Parncutt (1997, 2003), Demorest and Kim (2002) and Schulkind et al. (2003) use the term metric accent and consider it to be important. Although, defined slightly differently in some cases, all of these authors refer to the metric hierarchy of strong and weak beats in a bar that are implicit according to the given time signature.

A relative change in duration, in relation to context, is viewed by some researchers as a type of accent. In this way, duration is shown to be an important aspect of how we hear and consequently how we remember, recognise and compare melodies. This is corroborated by the research findings presented in section 2.2.4, in which the role of rhythm in melodic memory is discussed.

Jones (1987) categorises relatively long and short notes as temporal accents. Drake, Dowling, and Palmer (1993) consider duration as a type of rhythmic grouping accent with long notes or pauses marking the end of a group. Parncutt (2003) also considers longer notes as a form of grouping accents. Grouping accents could help determine the end of phrases and hence divide a movement or piece into manageable phrases for processing with a similarity algorithm. Although the function of long notes as accents may be somewhat ambiguous, they are considered to be accents by these researchers and so this reiterates the fact that the duration of notes may be an important factor in the salience of melodies.

A number of accents related to harmonic change and tonality are also mentioned in these studies. Lerdahl and Jackendoff (1983) theorise that “harmonic/melodic points of gravity” (p.18) form accents, especially when they occur at cadence points. Changes of

harmony form accents according to Parncutt (1997) and Jones (1987) mentions tonal accents that may be comparable to those of Lerdahl and Jackendoff. Compared to the clear definitions of metrical accents, for example, there is a level of ambiguity in these definitions. Such features would also require the use of key-finding algorithms and harmonic/tonality modelling systems, which are outside the scope of this research.

### 2.3 Features Used by Other Researchers

A survey of the features used in some of the most commonly cited research in the area of melodic similarity shows that a number of researchers use a pitch feature only in similarity/search algorithms (see Table 2.3).

<b>Feature</b>	<b>Research/Reference</b>
pitch only	Blackburn and DeRoure (1998)
	Downie and Nelson (2000)
	Uitdenbogerd and Zobel (1998, 1999)
	Dovey (2001)
	McNab, Smith, Bainbridge, and Witten (1997)
	Themefinder (CCARH)
	Prechelt and Typke – Musipedia (formerly Tuneserve) (cited in Typke et al. 2005a)
	Ghias, Logan, Chamberlin and Smith (1995)
pitch and duration	Crawford, Iliopoulos, and Raman (1998)
	Smith, McNab and Witten (1998)
	Hoos, Renz and Gorg (2001)
	Doraisamy and Rürger (2001)
	Byrd (2001)
	Cambouropoulos (2001)
	Melucci and Orio (2004)
	Lemström’s C-Brahms system (2003)
	Typke et al. (2004a)
	Clausen et al.’s PROMS system (2000)
pitch, duration and metrical stress	Typke et al. (2003)
	Ó Maidín (1998a)
	Selfridge-Field (2003)
pitch, duration and dynamics	Hofman-Engl (2001, 2002, 2003)
other features	Cambouropoulos (2001)

**Table 2.3: Features used for determining similarity in symbolic music.**

The melodic similarity research listed here uses various forms of symbolic representations. In some cases, such as Ghias et al. (1995), the original database and search melody are in audio form but pre-processing takes place to convert the audio to a symbolic representation of some kind. Although some of this work is not directly comparable with the research being carried out here on musical scores, it is useful to investigate which features are commonly used and which are rarely used in determining melodic similarity.

Pitch is represented in many formats in the work referenced here. These include representing the pitch by MIDI note number, octave and pitch class (e.g. C4), in terms of the interval from the tonic, in terms of the relative interval to the preceding note and as contour (both simple up/down and more complex contour representations). The broadest category of research surveyed here uses both pitch and duration information, while only a very small body of work uses some of the additional musical information that is available from a score representation. It should be noted that not all of the systems/algorithms mentioned above have actually been implemented. In some cases a theoretical melodic similarity algorithm is proposed and so results and analysis of the performance are not available.

### **2.3.1 Features Other than Pitch and Duration**

It can be clearly seen at a glance that the majority of algorithms explored use pitch or pitch with duration. Table 2.1 listed many more musical features that can be present in a score and so particular attention is paid here to the work that takes features additional to pitch and duration into account.

Typke et al. represent melodies as weighted point sets (2003, 2004a). Each note is represented as a point with a time and pitch co-ordinate and duration used as a weight. Further features are listed as possible weights but are not implemented by the author. These include “inter-onset intervals, metric stress, melodic contour, position within a measure, piece, or chord, accents, or a combination of these and possibly other features.” (Typke et al. 2004a, p.129).

In his geometric algorithm, Ó Mairín (1998a) uses the absolute pitch difference between notes weighted by the duration and metrical stress. Here, the metrical stress refers to the implicit stress that occurs at certain positions of the bar according to the time signature. In 4/4 time, for example, a strong emphasis is placed on the first beat of

every bar, with a secondary emphasis occurring on the third beat of the bar, exactly half way through. Processing of the score is achieved by traversing through time-windows of the score. The two melodies to be compared are aligned on a common time-axis and temporal distance between the notes at that point in the score is the width of the current time-window. The width of the window and the metrical stress of that window are then used as the weighting factor.

Selfridge-Field (2004) proposes a scoring system for melodies that incorporates metrical accents with pitch or harmony. In the pitch with accent system, a hierarchy of scores is awarded depending on the similarity of meter, the subunit of the meter, the percentage of pitches that match on primary beats (100%, >90%, and >80%), and the percentage of pitches that match on secondary beats (percentages are as before). The “harmonic with accent system” is similar with scores for chord that match on primary, secondary and tertiary beats.

Hofman-Engl (2001, 2002, 2003) represents cognitive features using concepts he calls melotons, chronoton, and dynamons, which are somewhat equivalent to pitch, duration, and dynamics. These are “psychological concept[s] whereby a listener to a sound directs her/his attention to the sound with the intention to decide whether the sound is high or low [short or long and loud or soft]” (Hofman-Engl 2001, p.144). This representation and the transformations he uses to measure similarity are not directly relevant to this research but it is notable that he does use some form of dynamics as an integral part of determining the melodic similarity.

Cambouropoulos (2001) uses a notion of similarity in his clustering algorithm for categorising melodies. His basic representation for melodies uses pitch and duration. Some coarse statistics are calculated for leaps, repetitiveness, and changes of direction. A count of these features is done in a similar way to Selfridge-Field. The number of leaps, repeated notes, and changes of direction are counted and <30%, <30% and <50% respectively are used as threshold values for these feature in contributing to the overall measure.

It can be seen that although the majority of research in this area of melodic similarity uses just pitch, or pitch and duration together to measure similarity, there is a small body of research that investigates or at least proposes the use of additional features.

## 2.4 Musical Features Chosen for Implementation

In section 2.2, research relating to melodic memory, recognition and identification was discussed as it is relevant to the development of perceptually accurate melodic similarity algorithms. The pitch component of melody is the feature most studied in this context. The various findings presented regarding the use of pitch intervals, contour, mode and absolute pitches illustrate aspects of how the perceptual system is thought to function. The most detailed specific data is used when recognising and remembering known melodies stored in long-term memory, while a rougher mechanism is used to remember new unknown melodies stored in short-term memory. In retrieving melodies from long-term memory exact intervals are thought to be used (Deutsch 1969; Attneave and Olsen 1971; Dowling and Fujitani 1971), allowing the melody to be recognised even when transposed. However, Levitin suggests that absolute pitch is used for at least some melodies that are very well known.

In designing algorithms to judge similarity between melodies it is appropriate to use the most detailed pitch information available. Since more precise methods are used for more detailed memory and retrieval tasks involving life-long and large capacity long-term memory, contour representations of pitch are not considered as useful in the context of this research as absolute pitch or exact intervals.

People perceive directly transposed melodies to be essentially the same melody and so it is important for algorithms to recognise transpositions. Intervals are used in many melodic similarity algorithms as this easily allows the algorithm to be transposition invariant. However, absolute pitch can also be used in such algorithms as long as there is some means of taking transposition into other keys into account.

Rhythm (the pattern of note durations) was also found to be an important feature in melodic memory and identification although it has not been found to be as important as pitch. It was not found to be particularly useful on its own but improves the ability of people to recognise melodies than when pitch alone is used.

Aside from tempo, there is little in the literature on melodic memory that investigates features other than pitch and duration. This suggests that other musical features available from the score are not as important and would not be as useful to incorporate into melodic similarity algorithms. This is mirrored in the survey of features used by

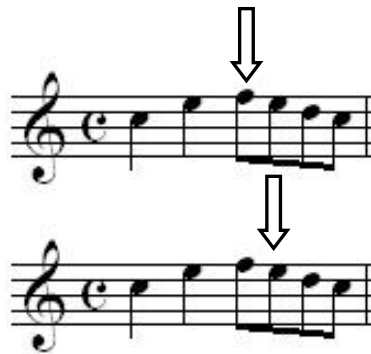
other researchers in various melodic similarity algorithms presented in section 2.3. Most of the researchers use pitch alone or pitch with duration but only a few suggest the use of other musical features. The lack of discussion of musical features other than pitch and duration in relation to melodic similarity algorithms and melodic memory research is explained in part by Parncutt (2003). He states that according to music notation and music theory, pitch and time play more important roles in creating musical structure than loudness and timbre. He goes on to give two explanations for this. The first is that the spectral frequencies and time intervals from which we derive pitch and temporal information are not adversely affected by the reflection of sound waves off walls and other objects in a space. The spectral amplitudes, on the other hand, from which loudness and timbre are derived, are affected by the reflections of the sound off surfaces. Parncutt suggests that because of this our ear has evolved to become less sensitive to amplitude. His second explanation is that the auditory environment we inhabit in everyday life is full of “equally spaced patterns of frequency ... and time” (p.165) such as the complex harmonics of speech and the regular rhythm of our heartbeat and footsteps. Because of this, he argues that pitch and time are the two musical features that are processed with relation to “defined reference frames (scales, metre)” (p.165) and notes that these are the only two features that are invariant to transposition.

The relative contributions of pitch, duration, and other features to the overall similarity measure is an important question and one that has not been explored in any great detail up to now. Some research has explored the interplay of pitch and rhythm and how one without the other affects memory and identification (Hébert and Peretz 1997; Schulkind 1999; Demorest and Kim 2002; Schulkind 2003) but there has not been comprehensive research into the proportional contributions of these features.

According to the music perception literature on accents the most commonly agreed upon and potentially useful accents for melodic similarity purposes were found to be pitch contour accents, pitch leaps/jumps and metrical accents.

The concept of the metrical accent is clearly defined and is well suited to algorithmic implementation. However, there are some ambiguities with the definition of the melodic accents listed above. A pitch contour accent occurs due to a change of direction in a melody. There is not clear agreement regarding which note of the melody the accent

actually occurs on (see Figure 2.1 below). In some cases it is believed to occur on the note before the actual change takes place (Jones 1987; Drake et. al 1991; Huron and Royal 1996), while other researchers regard the accent to be located on the first note in the opposite direction (Schulkind et al. 2003) and still more research indicates that the accent may occur on either of these notes (Drake and Palmer 1993). Thomassen (1982) regards a change in contour from rising to falling notes to be more important than a change of direction from falling to rising. Thomassen’s view is reiterated by Parncutt (1997) but the main body of research consulted did not make this differentiation.



**Figure 2.1: The possible locations of the pitch contour accent is shown by the arrow.**

Also, the definition of the accent itself is somewhat problematic when applied to real music. Figure 2.2, for example, shows an extract from one of the testbed melodies used



**Figure 2.2: The first two bars of Variation IX containing many contour changes.**

in this research (this testbed is discussed fully in Chapter 3). There are many changes of direction that could potentially signify a pitch contour accent. In the context of the definition of an accent itself as “an increased prominence, noticability, or salience ascribed to a given sound event” (Huron and Royal 1996, p.489) or as an attention-grabbing features (Jones 1983), it does not seem altogether credible to suggest that every change of contour present in such a melody catches the listeners attention and causes this note to particularly stand out from the surrounding notes.

There is also some ambiguity present in the definition of pitch leaps as accents. There is no clear idea of what constitutes a leap. Should intervals between notes larger than a certain value be included? What intervals leaps should be considered to form accents? Jones (1983) when discussing pitch leap accents, points out that Western music contains

a lot of step-wise melodic movement but intervals of a 3<sup>rd</sup> and a 5<sup>th</sup> are also commonly found since they may be based on the tonality of the music. Yet Jones suggests that an interval of 4 or 5 semi-tones (major 3<sup>rd</sup> or perfect 4<sup>th</sup>) forms such an accent. The remainder of the perceptual research surveyed here suggests that such a melodic accent is formed by large interval leaps between notes but does not give an indication of what should be considered “large”.

There is a general consensus that when multiple accents coincide they strengthen the accent and give greater emphasis to that point in the melodic surface. Drake et al. (1991) theorised that coinciding accents reinforced each other and they confirmed this through a playback experiment. Jones (1983) also believes that coinciding accents create stronger points of emphasis than single accents and that their interaction plays a role in the perception of hierarchical time structures. Hébert and Peretz (1997) found that joint accent structure contributed to melody recognition. Huron and Royal (1996), and Lerdahl and Jackendoff (1983) also discuss the interaction of accents with some similar concepts between the two. Huron and Royal found that the location of melodic accents was strongly correlated with metre and consider melodic accent a means of strengthening the perceived meter. Somewhat similarly, Lerdahl and Jackendoff consider their phenomenal accents (attacks, long notes, timbre and dynamic changes etc.) as “a perceptual input to metrical accent” (1983, p.17). Drake et al. (1991) experimentally found that the ability to play back simple tunes deteriorated when the accents did not coincide but that this did not affect rhythmic playback. Schulkind et al (2003) note that coinciding pitch and temporal accents occur at phrase boundaries. Thus, although there are many theories about the function of coinciding accents, there is agreement that jointly occurring accents create greater emphasis and salience. Hébert and Peretz (1997) found that Jones’s joint accent structure was important for identifying melodies and that when the pattern of accents was changed people had difficulty with this task.

Since there are such strong indications that different accent types coincide on the same points in a melody, and in particular because Huron and Royal (1996) and Lerdahl and Jackendoff (1983) indicate that metrical and melodic accents coincide, and also because of the ambiguity in the definitions of the contour change and pitch leap accent, it was decided to explicitly implement metrical accents only in the geometric algorithm. It is likely that important contour change accents and pitch leaps occur in the same place as

the metrical accents and so these may be seen as being implicitly rather than explicitly taken into account.

As a result of this survey of the music perception literature, the musical features chosen for implementation in Ó Maidín's geometric algorithm are pitch only, pitch with duration, pitch with metrical accent and pitch with both duration and metrical accent.

The relative contribution of pitch and duration when used together is not known and forms part of the investigation. Similarly, appropriate values are not known for the metrical accents and so these too are explored in Chapters 4 and 5.

A number of different versions of Smith et al.'s (1998) and Mongeau and Sankoff's (1990) algorithms are implemented so that the performance of pitch alone and pitch with duration can be assessed separately. These were the only musical features of the score taken into account by these authors. A number of extensions to the basic edit distance algorithm introduced by Mongeau and Sankoff are also implemented in various different versions of the algorithm so that each facet of their approach can be explored. Twenty-three different versions of the edit distance algorithms in total are implemented in this way.

## **Chapter 3**

### **Gathering Human Similarity Judgements for a Testbed of Melodies**

This chapter describes the testbed of melodies to be used in the fine-tuning of the algorithms. A listening experiment is developed to gather human judgements of similarity for the melodies in the testbed. The design considerations for the experiment are discussed and the musical testbed is presented. The results of the listening experiment are then analyzed for subject consistency and compared to similar work.

#### **3.1 Design Considerations for the Listening Experiment**

It is important for the success of this research that the human judgements of similarity are as accurate as possible. The performance of the algorithms will be compared to the human judgements and the algorithms modified accordingly. An essential part of this research relies on having human judgments of similarity that are as perceptually accurate as possible and that reflect an agreed notion of similarity. The task of gathering these judgements needs to be approached carefully. It is important to structure this experiment in a way that facilitates the listener giving as “true” a judgement as possible. The choice of musical material and procedures used in the experiment were intended to make the comparison task as easy as possible for the listener.

Similar research that involved gathering similarity judgements (and sometimes other related musical judgements) was examined. This included research that compared human similarity judgements with algorithmic measures which was directly related to our research, but also more general music perception experiments related to gathering judgements on melodies. A paper by Bonebright et al. (1998) titled “Data Collection and Analysis Techniques for Evaluating the Perceptual Qualities of Auditory Stimulus” was also useful as a general guide to designing the layout of the experiment. Some of the main points taken from this paper were:

- Subjects should be representative of the population that the experiment findings will apply to.

- The repetition of some of the experiment stimuli in random order can be used to measure the reliability of subjects.
- Pilot testing should be used “to validate experimental procedures, to help ensure that instructions are clear and specific, and to test the equipment and software” (p.506). The authors recommend using 3 to 5 pilot test subjects.
- Limiting the actual test time to about 30 minutes for fatigue and motivational reasons
- Incorporating practice trials (with similar but not identical material) so that the subject knows what is required and how to submit answers.

More specific points about the experiment procedures will be mentioned later.

### **3.2 Related Listening Experiments**

At the time of designing the experiment the closest related research was that of Müllensiefen and Frieler (2004a, 2004b, 2004c) and Eerola et al. (2001). The work of Typke (2004b, 2005b) in collecting similarity judgements for melodies from the RISM AI/II collection is also relevant but had not been published at the time this aspect of the work was carried out.

Müllensiefen and Frieler (2004a, 2004b, 2004c) gathered similarity ratings on melodies and processed these using a large number of similarity algorithms found in the literature on melodic similarity. These ranged from edit distance algorithms, to n-gram and correlation measurements. Algorithms were organised into interval, contour, rhythm, harmonic content and characteristic motif categories, and linear regression analysis was used on the best measure in each category to find the optimal measures in each category and the best overall measure. The internal components of individual algorithms are not adjusted. This approach is related to the work presented in this thesis but rather than run a large number of algorithms the focus here is on two particular algorithms with optimisation of the internal components of these algorithms. Eerola et al. (2001) compare measures of similarity based on statistical properties of folk melodies with human judgements. They recorded the frequency distribution of tones, intervals, durations, two-tone transitions, interval transitions and duration transitions for melodies, along with some variations of these measures with weighted durations and a further set

which incorporated some analysis of hierarchical levels of metre according to Lerdahl and Jackendoff's model. Along with the statistical measures mentioned, a set of measures based on a number of descriptive variables are also used for each melody. These variables include notions such as registral direction and closure from Narmour's model<sup>1</sup> and syncopation, rhythmic activity, tonal stability and other descriptors based on melodic perception research. The degree of similarity for each pair of melodies was calculated by using the city block distance between distributions (i.e. the sum of the absolute value of A-B, where A is the distribution of, for example, intervals in the first melody and B is the distribution in the comparison melody). In the case of the descriptive variables, the absolute difference in values between the descriptive variables was calculated for the degree of similarity. Similarity ratings were gathered from subjects and compared with the similarity measures using stepwise linear regression.

### **3.3 Design Considerations**

A number of design choices were made regarding the structure and content of the experiment and these are discussed in detail in the following sections. They include:

- the musical material to use - known or unknown melodies
- how the judgements should be gathered - rating, ranking or other task
- the number of pairs of melodies experiment subjects are asked to compare
- the duration/length of these melodies
- the duration of the experiment as a whole

#### **3.3.1 Music Material Used in Related Experiments**

In their experiments Müllensiefen and Frieler (2004a, 2004b, 2004c) used 8 bar segments from pop music songs and created 6 variants of each. These variants of the main melodies included "errors" based on rhythm, pitch, contour, phrase order and modulation errors that the authors state are documented in the literature on melodic memory. Initially the authors had intended to use melodies that were unknown to the subjects and participants were asked to record if they knew the melodies played. This was done in order to rule out effects due to previous knowledge. No differences were

---

<sup>1</sup> Coded according to Krumhansl (1995) but no further details are given.

found between the judgements of subjects who did and didn't previously know the melodies so both sets of subjects were included in the experiment.

Eerola et al. (2001) used 15 folk melodies with 3 melodies representing music from each of five different countries. They chose folk melodies because they are simple, real musical pieces and because they were assuming that taking melodies from these "different national styles would possess the natural variation of differences within melodies that is necessary for a similarity experiment" (p.279). They did ask subjects in a questionnaire to indicate their familiarity with the melodies and on analysis of the judgements given, they too found no familiarity effects.

Much of the literature on music perception and melodic perception in particular uses unknown melodies in experiments. Often this is to reduce any effects due to the subjects being familiar with the melody so as to control as many variables as possible in the experiment. There are examples however of experiments relating to recognition of melodies that do use well known melodies such as "Yankee Doodle", "Deck the Halls", "Bicycle Built for Two" (Dowling and Fujitani 1971), "Auld Lang Syne", "Oh Susanna" and "Twinkle, Twinkle" (White 1960).

Many past experiments related to the perception of melodies were carried out using melodies and short fragments specially written by the researchers for the experiment. Many experiments relating to melodic perception were run using very artificial sets of as little as 3 notes. Cuddy, Cohen and Miller (1979), Dowling (1978) and Dewitt and Crowder (1986), for example, used melodic segments of 3, 5 and 7 notes respectively. The aim of many of these experiments was to determine which features were used to identify and remember melodies. The ability of the subject to identify the melodic segment when one or more notes were changed in a particular dimension (contour, pitch etc.) was used to identify the particular features that made a set of notes or a melody identifiable or memorable.

As the aim of this research is to investigate algorithms that would be useful in comparing melodies in collections of real music stored on computer it was decided to fine-tune and evaluate the algorithms using real pieces of music rather than melodies composed by the author which may introduce unnecessary bias into the experiments.

### 3.3.2 The Choice of Testbed Melodies

It was decided to use a piece of music in Theme and Variation style for the melodies of the listening experiment and the later optimisation of the algorithms. Theme and Variations consists of a main musical theme which is repeated a number of times, with each repetition of the theme receiving a different musical treatment. This may include changes of key and time signature, embellishment or simplification of the melody, changes in accompaniment/harmony in the case of polyphonic music and rhythmic changes. Theme and Variations consist of musical material that is of varying degrees of similarity, where some variations may be very similar to the theme while others are less recognisable as being related to the theme. Such melodies provide a useful range of similarities for testing the algorithms. The use of Theme and Variation melodies also provides the subjects in the listening experiment with a context for giving their judgements. If the listener is asked to consider the similarity of a range of variations to one main reference melody then it is an easier task than asking them to choose a degree of similarity for many pairs of unrelated melodies. It was also apparent that short melodies or single melodic phrases would be most appropriate for this task. Long melodies make it difficult for people to remember and compare with each other.

A theme and set of variations on Twinkle, Twinkle, Little Star composed for recorder by Duschenes (1962) was chosen for the experiment. This piece was also used by Mongeau and Sankoff (1990) for evaluating the success of their edit distance algorithm. Initially this collection was considered because the Mongeau and Sankoff paper provided some discussion and comment on the musical material and research findings. On further examination this test collection demonstrated very good examples of the various kinds of problems that a melodic similarity algorithm is faced with. Different



Figure 3.1: The first two bars of the Theme and Variations IV and VII.

time signatures, different keys, augmentation of melodies (1 bar stretched to 2 bars), notes replaced by shorter repeated notes, triplets, elaborations of theme by stepwise motions and by leaps, notes occurring an octave higher and hiding of theme notes are all included in the 9 short variations (see Figure 3.1 for an example).

Also, although this set of Theme with nine variations consists mostly of 12 bars in  $\frac{4}{4}$  time, one variation is in  $\frac{3}{4}$  time and is 24 bars long and a further variation is in  $\frac{6}{8}$  time. This allows us to examine how successful the geometric and string-matching algorithms are at comparing melodies in different time signatures.

The two previously mentioned related similarity experiments found that using known or unknown melodies had no effect on the results. There are also many examples of using known melodies in melodic recognition experiments as reported in section 3.4.1 also so problems were not anticipated with the use of a known melody. Indeed the fact that the Theme melody is so well known means that the subjects should not have difficulty remembering the melody and therefore this makes the task easier for subjects and increases the likelihood of gathering consistent and accurate similarity judgments.

The test material also segments quite naturally into short melodies. The form of the piece is very obviously ABA (or ABA' in some cases where very slight changes appear in the reiteration of section A), with each section lasting for four bars. This meant that the Theme and Variations could be segmented into distinct four bar phrases for use in the experiment. Some examples of the four-bar phrases taken from section A of the piece are shown in Figure 3.2. The full set of melodies used are included in Appendix A.



Figure 3.2: The first four bars of the Theme, Variation III and Variation V

The first four bars and the middle four bars were used in two separate parts of the experiment. Although these melodies do not form a large test collection, as Pampalk et al. indicate “...even a tiny music collection can be used to identify weaknesses of a measure and compare measures to each other” (2003, p.207).

### **3.3.3 The Type of Similarity Judgements**

Another important design consideration for the experiment was the type of judgement subjects were asked to make. Some earlier melody perception research organised the melodies into pairs and asked the subjects to indicate if the melodies were the same, using the text descriptions “sure same”, “same”, “different”, “sure-different” (Dowling and Fujitani 1971; Dowling 1978). More commonly, similarity judgements are collected using rating scales. These scales usually have a number of points representing the level of similarity/dissimilarity and text descriptions on each end of the scale to indicate similarity or dissimilarity. In listening experiments run by Bartlett and Dowling (1988), Lamont and Dibben (2001), Müllensiefen and Frieler (2004a, 2004b, 2004c), Eerola et al. (2001), Hofman-Engl (2003), McAdams and Matzkin (2001) and Rosner and Meyer (1986) rating scales ranging from 6 to 11 points are used.

It is also possible to use a ranking method in the experiment. This would involve asking the subject to compare all melodies to each other and to list the melodies in order of similarity to the Theme melody. This method of gathering similarity judgements was rejected for a number of reasons. It discourages subjects from judging two melodies as being equally similar, while with the rating scales subjects can choose a particular similarity value a number of times if the melodies involved merit this. Since the melodies used in the testbed are variations on a theme, all of the melodies are somewhat related and potentially similar to the Theme melody. This makes the potential task of ensuring that all melodies are ranked correctly in order of similarity cumbersome for subjects, involving many repeated comparisons between melodies. Also, having similarity information in rank form means that a certain amount of detail is missing. The extent of the difference in similarity between each melody and the Theme is not known, only that certain melodies are less similarity than others. Melodies that vary greatly or very little in similarity may be reported as occupying subsequent ranks on the list.

### 3.3.4 The Rating Scale Used

Since most of the experiments reported in music perception and melodic similarity research used rating scales it was deemed the most appropriate approach. In deciding how many points of similarity should be included in the ratings scale similar experiments were examined. The number of points and text descriptions used for the opposite end of the rating scales in each of the research projects mentioned in section 3.3.3 is shown in Table 3.1.

While it is important to provide the subjects with a range of degrees of similarity to choose from, having a large number of values on the scale might also result in less consistent results as it might not be totally clear to all subjects what each value on the scale really represented in terms of similarity to the Theme. The initial design of the listening experiment used a 7-point scale that ranged from “very dissimilar” to “very similar”. The use of this scale and the text descriptions used at the end points of the scale were included in the pilot test, which is detailed in the next section.

<b>Authors</b>	<b>Number of points in scale</b>	<b>Left side of scale</b>	<b>Right side of scale</b>
Rossner & Meyer (1986)	9	very similar	not similar at all
McAdams and Matzkin (2001)	9/10	unknown	maximum similarity
Bartlett & Dowling (1988)	6	very different	very similar
Hofmann-Engl (2003)	9		
Lamont and Dibben (2001)	11	minimal similarity	maximal similarity
Eerola et al. (2001)	9	very similar	very dissimilar
Müllensiefen and Frieler (2004a, 2004b, 2004c)	7	unknown	maximal similarity

**Table 3.1: Examples of other rating scales for melodic similarity.**

### 3.3.5 Pilot Tests

Pilot tests were suggested by Bonebright et al. (1998) to examine the clarity of procedures and instructions before running the main experiment. Pairs of four-bar melodies were constructed consisting of the Theme melody and a variation melody, with the first melody of each pair always being the Theme. Subjects were asked to rate the similarity of each pair of melodies using the 7-point scale mentioned above. They were not told that the melodies were from a Theme and Variations piece or that all melodies would be somewhat similar to the first melody of each pair. It was found that

subjects found the use of the words dissimilar and similar on the same scale confusing as this meant there was uncertainty about which dimension they were listening for and rating on i.e. if they were listening and judging similarity or dissimilarity. Therefore, it was decided to express both of these text descriptions on the scale in terms of similarity. The text descriptions at each end of the scale were changed to read “hardly similar at all” and “very similar”. The number of points was not changed.

### 3.3.6 The Structure of the Listening Experiment

The segmentation of the Theme and Variations into two 4-bar melodic phrases easily allowed for the formation of pairs of melodies. This meant that there were nine pairs of melodies based on the first four bars of the Theme and Variations and a further nine melodies based on the second four bars, giving eighteen melodies in total.

Bonebright suggested randomly choosing a small number of stimuli to repeat and using the answers given to check subject reliability. In fact there are a number of ways to assess the consistency of subject ratings in such an experiment. Measures such as split halves, re-tests and parallel forms can be used to estimate the reliability of the results and experiment. Re-tests involve running the test again on the same set of subjects with an appropriate time interval between tests and correlating the two sets of scores. Such repetitions were included in Müllensiefen and Frieler’s (2004a, 2004b, 2004c) experiment, although they did a re-test proper a week after the initial experiment. The author did not want to burden volunteers by asking them to undertake the experiment twice so instead each of the nine pairs of melodies in our experiment were repeated in random order. The structure of the experiment is shown in Table 3.2 below. A short break was included between Part A and B of the experiment. The melodic segments used in Part A and B of the experiment are included in Appendix A.

Part A	Theme & Variations 1-9 Bars 1-4	sequential order
	Theme & Variations 1-9 Bars 1-4	random order
	Short Break	
Part B	Theme & Variations 1-9 Bars 5-9	sequential order
	Theme & Variations 1-9 Bars 5-9	random order

**Table 3.2: The melodies used in each part of the listening experiment.**

Having now extended the experiment to 36 pairs of melodies there was some concern that the task might have become too long and potentially tiresome for participants. Each melody is about eight seconds in duration. This results in Part A and Part B each lasting 12-15 minutes (given time for making the similarity judgements), and gives a total experiment duration of 25-30 minutes. This brings the experiment to the limit of Bonebright's recommended listening time of 30 minutes (1998).

In similar experiments where subjects were asked to make judgements about the similarity of melodies examples are found of using 48 pairs of seven note melodies (Bartlett and Dowling 1988), 36 pairs of 8-9 bar long melodies (Lamont and Dibben 2001), 105 pairs of 17 second (mean length) long melodies (Eerola et al. 2001), 42 pairs of 15-20 second long melodies (Müllensiefen and Frieler 2004a, 2004b, 2004c), and 66 pairs of 8-bar long melodies in an experiment lasting about 104 minutes in total (Rosner and Meyer 1986). It seems that the length of the melodies and the total listening time chosen in this research compares favourably with similar research. The use by Eerola et al. (2001) of 105 pairs of melodies that last on average 17 seconds seems very long in comparison to the 36 pairs of eight second long melodies used in this listening experiment.

### **3.3.7 Practice runs, demonstrations etc.**

The advice of Bonebright et al. (1998) was taken on allowing subjects to run through practice examples of the experiment. This was to ensure that they were familiar with the notion of different degrees of similarity and the scale being used, as well as the actual procedures involved. In total about ten minutes of an introduction to the experiment was given to each subject. In this introduction subjects were told that all the melodies they would hear would be related in some way and that the first melody of each pair they would hear would always be the same in each part of the experiment. A demonstration using seven variations constructed by the author on another well-known tune were played and discussed with the subject. The variations were based on the sort of musical modifications made to the "Twinkle, Twinkle" theme melody. A practice run of the experiment was developed using this melody and three variations. This practice experiment was used to familiarise subjects with the task and the rating scale being used. Subjects could ask questions and discuss the issue of similarity during this introduction period but no comments were made about which musical features subjects

should use to make their judgements. Subjects were encouraged to use the full range of the scale. These aspects of the introduction are included to ensure that each subject has a good understanding of the sort of melodic similarities they will encounter during the experiment and to ensure that they have thought about the sort of similarities that might be deserving of a rating of 1 or 7 at the extremes of the scale.

In related experiments, Eerola et al. (2001) gave the subjects instructions to read and used three practice pairs of melodies before the main experiment commenced. Müllensiefen and Frieler (2004a, 2004b, 2004c) asked subjects to imagine the scenario that the first melody of each pair was played by a teacher and the second melody by a student and that they were being asked to rate the performance of the student in terms of the “severeness” of any errors they may have played. This notion of errors in the melody would not be appropriate for our experiment where each comparison melody is a valid variant of the original and a range of different variation techniques are employed. It is important for this sort of experiment that each subject has a similar understanding of what the labelled endpoints of the rating scale represent so that if all participants consider something to be “fairly similar” they are using the same values or a common range of values. There may be subjective differences between people’s perception of the degree of similarity but if the data is to be useful there is a need to ensure that the recording of the ratings is as accurate as possible. This is why discussions around the idea of similarity are included in the introduction alongside the demonstration melodies. Subjects are told that they will not receive guidance regarding which musical features to use in making the similarity judgement and neither will they be asked to explain their choices. It is suggested to them that it may be useful to sing the theme melody in their heads while listening to the second melody of each pair.

### **3.3.8 Comments and Feedback**

Most experiments of this nature include some sort of question or comment sheet to be filled in by the researcher or subject regarding the musical background of the subject. This is done so that the results can be analysed for effects due to musical background or in order to remove or leave in the data from subjects that fit certain criteria. In this case subjects were given the opportunity to comment on the task involved, the procedures and any other issues related to the experiment by filling in a feedback sheet. The complete layout of the listening experiment is shown in Table 3.3.

<b>Section</b>			<b>Duration</b>
Introduction	Demonstration, questions, practice run		10 minutes
Part A	Theme & Variations 1-9 Bars 1-4	sequential order	12-15 minutes
	Theme & Variations 1-9 Bars 1-4	random order	
	Short Break		1 minute
Part B	Theme & Variations 1-9 Bars 5-9	sequential order	12-15 minutes
	Theme & Variations 1-9 Bars 5-9	random order	
Feedback	Musical background and comments		1-2 minutes

**Table 3.3: The layout of the listening experiment.**

### **3.4 The Listening Experiment – Implementation and Method**

#### **3.4.1 Subjects**

34 subjects participated in the listening experiment. Volunteer subjects were sought from university students enrolled in music-related postgraduate courses. A small number of postgraduate students who were musicians but not actually studying on music courses were also included. Six subjects had previously obtained an undergraduate or postgraduate music degree and were enrolled in a postgraduate music-related degree at the time of the experiment, 26 of the remaining subjects were current postgraduates in a music-related area. Seven subjects included in the experiment were musicians but did not have and were not studying for a music degree. All subjects, except one, report that they played an instrument. Subjects reported playing music in a variety of styles (i.e. all were not classically trained musicians), having had varying amounts of formal lessons ranging from 0 to 20 years.

#### **3.4.2 Generation of MIDI Files**

The experiment was run on computer using MIDI files of the melodies. The MIDI files were generated from the musical score using notation software and some relevant adjustments were made to ensure that the version of the melodies represented in the MIDI files reflected a performance of the score as closely as possible. In the particular notation software used, a default setting for playback and rendering to MIDI file stopped the sounding of notes 5% of their duration from the end of the note. The

purpose of this feature of the notation software is to allow the listener to hear a clear separation between the end of a note and the start of the next note. This may be impossible to hear if the note is held for its full length and if both notes are the same pitch. This ability to control the articulation of individual notes was further used in designing the experiment to enable slurred and tied notes and tenuto markings to be performed differently to notes that did not have such performance indications. Staccato notes were performed as half the written duration of the note. A change in the velocity of notes was also used to produce a slight metrical stress denoting the primary and secondary beat pattern of a bar. As previously mentioned, this feature is implicit in scores and would be automatically incorporated into the performance of a score by a musician. An increased by 10% on the first beat of 4/4, 3/4, 6/8 bars and by 5% on the 2<sup>nd</sup> main beat in appropriate time signatures was used. A MIDI piano instrument was used to play the melodies and all were played at the same tempo (110 BPM).

### **3.4.3 Procedure**

After the initial introduction stage and practice listening experiment, subjects progressed to the experiment proper. The experiment was divided into two Parts A and B, with a 1 minute break between each part. The main theme melody used in each part was played on its own before the melody pair to be rated. A .5 second break was included between the melodies in each pair and after the second melody was heard, the rating scale was displayed on the computer screen. Once the rating was made the experiment proceeded to the next pair of melodies. Subjects could not replay melodies. When the experiment ended subjects were invited to fill in a feedback sheet with a number of questions regarding musical background and space for comments on the experiment and procedures used.

## **3.5 Results**

### **3.5.1 Part A (bars 1-4 of the variations)**

The frequency distributions of the ratings are shown in graph form in Figure B.1 of Appendix B. In the case of the sequential playing of the melodies there is general agreement among subjects concerning the melodies that were considered very similar or just barely similar i.e. the extreme points of the scale used. The distributions of ratings were wider for melodies that were not so obviously at the extremes of the given scale.

This is reflected in the range of ratings that vary from 3 up to 6 with an average of 4.3 (see Table B.1 of the same Appendix). The standard deviation of ratings ranges from .74 to 1.53 and the average of the standard deviations was 1.12. There were no outliers present in the ratings given for any of the variations. Shapiro-Wilks normality tests (used when  $n < 50$ ) were carried out and the ratings for most of the variations are shown to be non-normal (see Table B.2). A significance or p-value for this test  $< .05$  indicates a non-normal distribution. Variation II appears to be almost normal in the distribution of ratings. The normal curve has been drawn over the frequency distributions shown in Figure B.1.

When these same melodies are repeated in random order the range decreases by one in two instances, bringing the average range down to 4.1 from 4.3 previously (see Table B.1 again). The average of the standard deviations is 1.10, as against 1.12 previously. An outlier appears in the ratings for Variation 5 (see Figure B.2 in Appendix B for distributions) and this time there are no melodies with normally distributed ratings (see Table B.3). In all but one case the mean rating for each melody increased on the second, random playing. Essentially this means that there was a tendency to give higher results the second time the melody was heard.

### **3.5.2 Part B (bars 5-9 of the variations)**

Part B of the experiment uses bars 5-9 of the Theme and Variations and the results are examined separately to those of Part A. The range of ratings for the sequential playing of the melodies is shown in Table B.4 of Appendix B and can be seen to vary from 2 to 6 with an average of 4.3. There are fewer melodies to Part A for which there is very strong agreement on the rating but the average of the standard deviations increases only slightly to 1.13. There are no outliers in the ratings. Variation IX is clearly normally distributed but all other Variations have a non-normal distribution of ratings. Figure B.3 shows the frequency distribution of ratings for each variation and Table B.5 contains the results of the Shapiro-Wilks normality tests.

The ratings for the randomly repeated melodies show that the range of ratings given increases (by 1 or 2 on the scale) in the case of four of the melodies and decreases in the case of three other melodies giving an average range of 4.4, as against 4.3 previously (see Table B.4). Similar to the Part A melodies, the mean rating increases in all but one case on this second randomly order playing of the melodies. The average of the standard

deviation of ratings increases to 1.18 (from 1.13 previously) and there are two outliers in the ratings for Variations V and I (see graphs in B. 4, Appendix B). The distribution of ratings is normal for Variation VII and almost normal for Variation II (see Figure B.4 again and Table B.6 for the Shapiro-Wilks test results).

This brief analysis of the results shows that there is an element of agreement between people on the suitable rating for each melody. The smaller range of ratings and frequency distributions of those melodies considered to be at the extremes of the similarity scale and the very low numbers of outliers show that there is a common notion of similarity emerging for some melodies. For other melodies that are somewhere in the middle of the similarity scale one might have expected to have ratings that were normally distributed but this has only happened in a small number of cases. The mean ratings given in Part A and B of the experiment for the second (random) playing of the melodies was higher than the first (sequential) and there are evident small differences in the range and standard deviation of the ratings. This shows that there are differences in the ratings given on the repeat playing of the melodies by some subjects. Further analysis is carried out to determine if particular subjects are more consistent in their ratings than others. Non-parametric tests were chosen in subsequent analysis due to the distribution of the data.

### **3.5.3 Consistency of Subjects**

The pairs of melodies were repeated in random order so that the ratings given by individual subjects could be checked for consistency using Spearman's non-parametric correlation. This is the approach used in Müllensiefen and Frieler (2004a, 2004b, 2004c), where subjects had to have a correlation not less than .5 (using Kendall's Tau).

One might expect some small differences between the ratings given for the first and second playing but if a subject was to give a rating of 1, for example, for the first time they heard a variation and 7 the next time they heard the same variation, the consistency and reliability of that subject would be questionable. The introduction and demonstration phase was designed to reduce the possibility of such occurrences and the feedback stage was designed to provide a record of subjects who had serious problems or lack of understanding of the task. Using scatter plots the linear relationship between the ratings each subjects gave for the original and repeat playing of each melody was examined. Generally a linear relationship between the results was observed, although

the linearity of the relationship was observed to be stronger for particular subjects. (A sample of six scatter plots from the first six subjects is included in Appendix B.) The correlation coefficient (Spearman's non-parametric coefficient) was calculated for each individual subject based on the ratings they gave for the first (sequential) and second (random) playings. The results are presented in Table B.7. In Part A 26 subjects showed significance at the .01 level (.798), another three at the .05 level (.750) and the remaining five were not considered significant (significance was lower than at the .05 level). Although Müllensiefen and Frieler (2004a, 2004b, 2004c) used a cut-off correlation value of .5, we considered this too generous and instead decided to use subjects who had a correlation value  $>.798$  which is the .01 level. This essentially means that there is only a 1% chance of this result having happened by chance. This chosen cut-off point was also corroborated by guideline correlation values found in both a statistics and research methods book (Witte and Witte 2004; Shaughnessy 2006). Both sources suggest that when using a repeat test mechanism that only data from subjects with correlation values  $>.8$  should be used.

The same cut-off mechanisms were used for the data in Part B and it was found that the individual subjects correlation between the ratings given for the first and second playings of the melodies were much reduced. The correlation data is included in Table B.8 in Appendix B. Here, only 16 subjects showed significant correlation at the .01 level, another 9 showed significance at the .05 level, leaving a further 9 subjects that did not show significant correlation. The musical experience and comments left by subjects were examined to see if the subjects in groups with high or low correlation values had common traits or problems with the experiment.

#### **3.5.4 The Reduced Data Set – Part A**

Having removed the ratings of those subjects not considered consistent the frequency distributions and basic statistical features of the ratings were once again examined to observe the effects this would have before further analysis of the overall reliability of the data was carried out.

The frequency distributions of the ratings for the 26 subjects now remaining in the dataset are shown in Figure B.6, with the basic analysis in Table B.9 and the results of the tests for normality in Table B.10. The range of ratings for the sequential playing of the melodies is the same when all the subject ratings were included, with an average

range of 4.3. In all but one case the mean has either increased or decreased from the original and the average of the standard deviations is 1.10, which is a slight reduction from the 1.12 value seen in the larger set. Interestingly, the melodies from Variations II, IV, and VI now clearly have normally distributed ratings. These are melodies with median ratings of 4.5, 3 and 3 respectively. The remaining melodies have median ratings of 7, 6, 5, 1, 1 and 2.

### **3.5.5 The Reduced Data Set – Part B**

The reduced set of ratings for the Part B melodies include the ratings of just 16 subjects, those showing significant correlation at the .01 level. Looking at the frequency distributions of the group of 16 subjects (see Figure B.6), it is clear to see that reducing the dataset in this way does give more consolidated results. Compared to the original set of 34 subjects' ratings the range of ratings has been reduced in many cases with an average range of 3.3 (down from 4.3) for the first playing of the melodies (see Table B.11). The mean ratings are slightly increased in some cases and slightly decreased in others and the average of the standard deviations is 1. The ratings for Variation V, VI and IX are normally distributed with Variation VII showing a borderline normal distribution (see Figure B.6 and Table B.12).

### **3.5.6 Consistency and Reliability of the Human Similarity Judgements Gathered**

The consistency of individual subjects was examined in the previous section and the ratings of subjects who proved to be least consistent were removed. In order to assess the overall reliability and level of agreement present in these human similarity judgements and to allow comparison with related work, further statistical tests were carried out.

The mean inter-subject correlation value has been used in similar experiments to provide a measure of overall agreement between subjects and although Cronbach's alpha is more often used to examine the reliability of the testing mechanism itself, it has also been used to measure reliability of subjects in Müllensiefen and Frieler (2004a, 2004b, 2004c) and Lamont and Dibben (2001). Eerola et al. (2001) use mean inter-subject correlation to assess the reliability of results from a similar listening experiment.

Correlation matrices were calculated for the ratings from Part A and B of the experiment so that the relationship between subjects' ratings could be examined. Spearman's rank correlation coefficient (non-parametric) was used. This inter-subject

correlation gives some indication of how much agreement there was between subjects and therefore how successful this experiment has been in collecting accurate human judgements of similarity. Table 3.4 below summarises some of the features of these matrices. The full tables (Tables B.11-B.12) are included in Appendix B. Although the minimum correlation between subjects is quite low in some cases, it is obvious from observing the overall mean and median correlation values that many of the inter-subject correlations were very high. The standard deviation gives a rough idea of how much these correlation values vary from the mean. The median values for the inter-subject correlation are significant at the .05 level ( $>.750$ ) and those from Part A are significant at the .01 level ( $>.798$ ). The median value is considered more reliable as it is not affected by low or high outlying values in the inter-subject correlation matrix. The percentage of subjects who had high inter-subject correlation is also given as that shows that there was considerable correlation between the ratings given by each subject, especially in the case of Part A of the experiment.

	<b>Minimum inter-subject corr.</b>	<b>Maximum inter-subject corr.</b>	<b>Mean inter-subject corr.</b>	<b>Median inter-subject corr.</b>	<b>Standard deviation</b>	<b>% corr &gt; .7</b>	<b>% corr &gt; .8</b>
Part A sequential	.25	.99	.78	.84	.16	73%	59%
Part B sequential	.21	.97	.71	.75	.19	55%	40%

**Table 3.4: Inter-subject correlation calculations for the reduced data sets.**

Eerola et al. (2001) used the mean inter-subject correlation as a measure of the reliability of their results and reported the following values:  $r = .41$ ,  $df = 103$ ,  $p < .001$ , where the  $r$  value represents the correlation coefficient. The mean inter-subject correlation in each section of this experiment is higher but the correlation value given in their research is statistically significant at a lower level of  $p$  due to the large number of melodies they used (105 melodies giving 103 degrees of freedom).

The correlation coefficient is not a definitive measure of similarity between the ratings given by subjects. It is quite reasonable to expect the similarity ratings from two different subjects to differ by one or two points of the scale for some melodies. Small differences in the similarity ratings such as these can result in different values for the correlation coefficient if the values are not all positive or negative but a mixture of both.

Tables 3.5 and 3.6 demonstrate an example of this issue. A set of possible ratings from two subjects are presented in which different rating values have been awarded for six of the seven melodies rated. In Table 3.5, the first example, the second subject gave lower ratings than the first subject by one or two points of the scale and the correlation between the two subjects is shown to be .945.

subject1	subject2	
1	1	same
4	2	lower
6	5	lower
2	1	lower
3	1	lower
4	3	lower
7	6	lower
	correlation	0.945708

**Table 3.5: An example of the inter-subject correlation using two subjects.**

In the next example in Table 3.6, both subjects give the same rating three times out of seven and when differences occur between subject ratings these are no more than 1 or 2 points on the scale. One might expect the correlation coefficient to be higher to reflect this increased level of agreement between the subjects.

subject1	subject2	
1	1	same
4	6	higher
6	6	same
2	1	lower
3	4	higher
4	4	same
7	6	lower
	correlation	0.880744

**Table 3.6: An example of the inter-subject correlation using two subjects.**

However, the correlation coefficient is lower in this instance as it is intended to measure the strength of the linear relationship between two sets of values. Even a small number of differences in the ratings can greatly affect the value of the correlation coefficient when the differences are both higher and lower than other subjects, rather than all higher or all lower. In this way correlation may not be the ideal way of examining the relationship between subjects' ratings.

Since Müllensiefen and Frieler (2004a, 2004b, 2004c) and Lamont and Dibben (2001) used Cronbach's alpha as a measure of the reliability this is also calculated here (see

Table 3.7) and allows the judgements gathered as part of this research to be compared with similar work for reliability. Usually Cronbach's alpha or Kuder-Richardson's Formula (20 or 21) are used to assess the internal consistency or reliability of a particular scale where subjects may be answering related questions.

	<b>Cronbach's alpha</b>
<b>Part A seq</b>	.987
<b>Part B seq</b>	.975

**Table 3.7: The Cronbach's alpha reliability values.**

Lamont and Dibben (2001) demonstrated a Cronbach's alpha value of .84 for their results and Müllensiefen and Frieler (2004a, 2004b, 2004c) presented values of .962 and .978 for two sets of data. In the context of this research, the data for Part A of the listening experiment appears to be quite reliable and while not so high for Part B it is on a par with the higher value found in Müllensiefen and Frieler and considerably higher than in Lamont and Dibben's work.

The results and analysis presented in section 3.5 show that omitting the ratings of subjects that did not show significant correlation between the ratings they gave for the sequential and random playing resulted in a reduction in the average of the standard deviations, an increase in the number of melodies with normally distributed ratings and a reduction in the range of ratings in many cases. The reduced data set was then tested for reliability using the inter-subject correlation and Cronbach's alpha. Very favourable results were found and shown to be comparable and higher in terms of reliability to other work that involved gathering similar judgements from people. Care was taken in the design of the experiment so that accurate human judgements could be collected and because they were intended for use in optimising a number of algorithms and this is reflected in the statistical analysis performed on the ratings.

### **3.5.7 Pooling the Ratings**

In order to pool the ratings together so that a single similarity rating/value could be brought forward for use in the optimisation of the algorithms, the median rating value was taken for each variation (see Table 3.8 below). The use of only the most consistent subjects' ratings did result in an increased number of normally distributed ratings for the mid-range values of the similarity scale. However, since the ratings for some of the variations still had significantly positively or negatively skewed distributions the

median was chosen as the measure of central tendency of the results. The median similarity ratings from the sequential playing of the melodies were carried forward for use in the remaining steps of this research. The ratings from the random repeat of the melodies were used in assessing the consistency of individual subjects only and are not used in subsequent aspects of this research.

<b>Median ratings for each variation</b>		
	<b>Part A</b>	<b>Part B</b>
<b>Variation 1</b>	7	7
<b>Variation 2</b>	4.5	5
<b>Variation 3</b>	6	6
<b>Variation 4</b>	3	3
<b>Variation 5</b>	5	4
<b>Variation 6</b>	3	3
<b>Variation 7</b>	1	3
<b>Variation 8</b>	1	1
<b>Variation 9</b>	2	4

**Table 3.8: The median ratings gathered in the listening experiment.**

## Chapter 4

### The Algorithms

In this chapter, the algorithms being evaluated are presented first in their original form and subsequently in the various versions implemented as part of this research. The contrasting way in which each type of algorithm processes melodies is discussed, with musical examples included to illustrate particular properties of each approach.

#### 4.1 Implementation Environment

The algorithms are implemented in C.P.N.View (Common Practice Notation View), an object-oriented software environment developed by Ó Maidín (1995, 1998b). In C.P.N.View score objects are created algorithmically or using convertors for music notation formats such as ALMA, \*kern and EsAC. The software is written in C++ and provides a large number of objects, classes, types and iterators that allow the user to retrieve information about the objects in a score. The most commonly used objects for processing scores along with examples of code written in C.P.N.View can be found in a 2001 paper (Ó Maidín and Cahill 2001).

#### 4.2 The Geometric Melodic Similarity Algorithm

The first algorithm under investigation in this thesis is an algorithm from Ó Maidín (1982; 1998a), shown below as equation 4.1.

$$difference = \sum_{k=1}^n | p_{1k} - p_{2k} | W_k \quad 4.1$$

where

k is the time window of the score

p<sub>1</sub>, p<sub>2</sub> are the pitch values of the first and second melodies in the current window

W<sub>k</sub> is the weight associated with that time window

totaldur is the total duration processed

This algorithm was chosen because it showed some success in measuring the similarity of folk melodies in Ó Maidín (1995). It also includes a mechanism for processing the relevant features identified from the music perception research in Chapter 2.

The author categorises this algorithm as a geometric algorithm since it calculates the distance between two melodies. One of the key aspects of this algorithm is the manner in which it processes the score as a series of time windows. Each window has an associated width, which corresponds to a time span. The time window to be processed at a particular point in the score is calculated by aligning both melodies on a common time axis. The left and right borders of each window correspond to onsets or offsets of notes or rests. In this way the width of each time window reflects the longest time for which there is uniform activity and there are no onsets and offsets within the window. An example of the time windows used to process two melodies is shown in Figure 4.1.



**Figure 4.1: The division of the score into time window.**

The algorithm calculates the absolute pitch difference of the notes contained in each window of the score and multiplies this value by a corresponding weight for that window. The results for all windows are then added together to provide an overall difference measure. The algorithm essentially uses the pitch difference between individual notes in a corresponding point in each melody to calculate the overall difference between the melodies. Zero implies that there is no difference between the melodies i.e. that they are exactly the same or that a note is broken down into a series of shorter repeated in one of the melodies. Small difference values indicate that the melodies are very similar and larger difference values indicate less similarity. In Ó Maidín's (1998) implementation the weight consists of the rational duration of the current time window combined with a metrical stress weight based on the position of the note within the bar. This metrical stress weight is in effect an implementation of the metrical accent discussed in Chapter 2 and is referred to as a metrical accent weight in

the remainder of this thesis. The values for the metrical accents were arbitrarily chosen by Ó Mairín and are based on the hierarchical stress pattern indicated by the time signatures in question. In order to assess the performance of each of these features individually and collectively, a number of versions of the geometric algorithm are implemented. The nature of this algorithm means that when a long note in one melody is compared to a series of short notes in the other melody, the longer note spans more than one time window, as can be seen in Figure 4.1 above. There are a number of ways in which the duration of the longer note may be used in each of the time windows it spans and a number of approaches are implemented and discussed in section 4.3.

#### 4.2.1 Eliminating the Effect of Melody Length

Since the geometric algorithm takes the sum of all the weighted pitch differences between notes, longer melodies will naturally result in a higher overall difference value. A small pitch difference between one or 2 notes when comparing 20-note melodies, for example, should result in a smaller difference value than when the same pitch difference occurs between melodies that are only five notes long. The algorithm was modified at a later stage by Ó Mairín so that the final difference value is divided by the total duration processed, as shown in equation 4.2 below. This length adjustment is included in all implementations of the algorithm presented here but is not explicitly stated.

$$difference = \frac{\sum_{k=1}^n |p_{1k} - p_{2k}|}{totaldur} \quad 4.2$$

#### 4.2.2 Transposition to different keys

Ó Mairín (1998a) outlines an approach to dealing with the problems raised by comparing melodies in different key signatures. As discussed in section 2.2.3 of Chapter 2, people regard melodies in different keys as being highly recognisable as the same melody and here melodies that differ by key only are regarded as being identical. Ó Mairín's approach is to transpose one of the melodies to all other possible keys. The similarity/dissimilarity value is then calculated for each of the transposed melodies and the key that results in the smallest overall difference value is identified as the correct transposition. A basic implementation is to subtract a constant value from the pitch difference as each window is processed ( $p_{1k} - p_{2k} - value$ ). However, a shortcut is taken in which the median value of the pitch differences with the associated weight value is

used to find the number of semi-tones a melody has been transposed by (Aitken 1939 cited in Ó Maidín 1998a).

$$difference = \sum_{k=1}^n | p_{1k} - p_{2k} - mptw_k | W_k \quad 4.3$$

where:

$mptw_k$  (median pitch according to total weight for window  $k$ ) is the median pitch value calculated using the sequence of pitch differences and their associated total weights for each window.

Again, although this feature of the algorithm is implemented as part of this research, it is not explicitly stated in the remainder of the thesis.

### 4.3 Implementing the Geometric Melodic Similarity Algorithm

Chapter 2 featured an exploration of the musical features that were likely to be most useful for melodic similarity algorithms and section 2.4 outlined those chosen for implementation here. Pitch only, pitch with duration, pitch with metrical accent and pitch with both duration and metrical accent are implemented in the geometric melodic similarity algorithm and run on the testbed melodies discussed in Chapter 3. The issues involved with implementing this algorithm are discussed in the following sections.

#### 4.3.1 Pitch difference only

$$difference = \sum_{k=1}^n | p_{1k} - p_{2k} | \quad 4.4$$

where:

$k$  is the window number

$n$  is the total number of windows to be processed

$p_1$  is the pitch (MIDI note number) of the note in melody 1

$p_2$  is the pitch (MIDI note number) of the note in melody 2

Pitch differences of more than an octave are treated as if they occurred within the same octave (see section 2.2.3). The modulus operator is used to calculate the pitch difference as follows but is omitted from the equation for succinctness:

$$| (P_{1k} - P_{2k}) | \% 12 \quad 4.5$$

### 4.3.2 Pitch difference weighted by duration

There are a number of different ways in which duration can be incorporated into this algorithm and four methods have been implemented here. The original Ó Maidín algorithm (1998a) used the duration (width) of the time window as the duration weight. The three additional methods used here use the duration of the notes involved but differ according to how this is implemented. The musical implications of each method are discussed in section 4.3.3.

#### Method 1

In this version of the algorithm, the duration weight is calculated directly from the rational duration of the window width as in Ó Maidín (1998a).

$$difference = \sum_{k=1}^n |p_{1k} - p_{2k}| dw1_k \quad 4.6$$

where:

$dw1_k$  is a weight based on window width (the rational duration in this case)

#### Method 2

In this version of the algorithm, the rational duration value of the notes from melodies 1 and 2 are added together to provide the duration weight. This weight is only used in the onset window of the notes. In this way the weight is only applied in the first window of long notes that span multiple windows.

$$difference = \sum_{k=1}^n |p_{1k} - p_{2k}| dw2_k \quad 4.7$$

where:

$dw2_k = (\text{duration from melody 1} + \text{duration from melody 2})$ , where the value of the duration is used at the note onset only

### Method 3

Here, the rational duration of the notes from both melodies are added together but the value of the duration is used in every window of the note. This means that for long notes, the full duration of the note is used in every window it spans.

$$difference = \sum_{k=1}^n |p_{1k} - p_{2k}| dw3_k \quad 4.8$$

where:

$dw3_k$  = (duration from melody 1 + duration from melody 2), where the value of the duration is used in every window of a note

### Method 4

Here, combined duration from both melodies is used at the onset of a note and the window width only (as in Method 1) is used in any subsequent windows that a note may span.

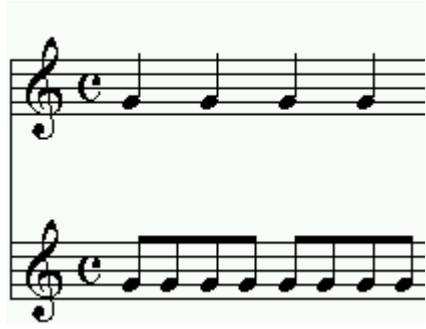
$$difference = \sum_{k=1}^n |p_{1k} - p_{2k}| dw4_k \quad 4.9$$

where:

$dw4_k$  = (duration from melody 1 + duration from melody 2), the value of the duration is used at the onset of a note and the window width is used in subsequent windows of a note

#### 4.3.3 Musical Implications of the Four Duration Methods

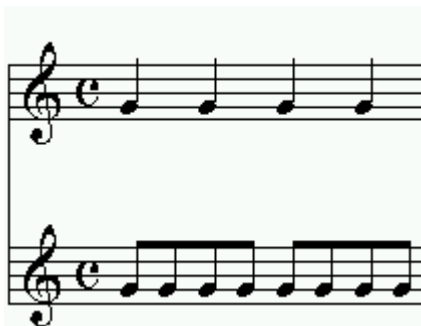
These different methods of incorporating duration are designed to give different results in cases in which a long note is present in one melody and is being compared to a series of shorter notes in the second melody. When the basic window weight is used as in Ó Maidín and Method 1 above, then there is no added emphasis on pitch differences that occur between two long notes or between a long note and a series of shorter notes. When a long note is compared to one or more shorter notes, it is essentially treated as if it was itself a series of shorter notes of the same pitch because of the use of the time windows. In the example shown in Figure 4.2 below, it is clear to see that there is no difference in the duration weight when quarter notes are compared to eighth notes, than would occur if eighth notes were compared to eighth notes.



Window	Duration weight (window width)
1	0.125
2	0.125
3	0.125
4	0.125
5	0.125
6	0.125
7	0.125
8	0.125

Figure 4.2: An example of the duration weights when Method 1 is used.

In Method 2, the duration value from both notes is added together to obtain the duration weight, but the weight is used at the note onset only and nothing is used in subsequent windows of a note. The duration of a long note (when compared with short notes) is used in the first window of the note only as shown in Figure 4.3. This means that pitch differences contribute a lot more to the overall similarity result when they occur at the beginning of a long note, since the pitch difference will be multiplied by this weight. It also means that pitch differences between two long notes will contribute more to the overall difference value than a pitch difference that occurs between a long and short note. An example is shown below in Figure 4.4.



Window	DurationWeight Melody 1	DurationWeight Melody2
1	0.25	0.125
2	0	0.125
3	0.25	0.125
4	0	0.125
5	0.25	0.125
6	0	0.125
7	0.25	0.125
8	0	0.125

Figure 4.3: An example of the duration weights when Method 2 is used.

Pitch differences at the start of a long note contribute more than on subsequent semi-quavers:

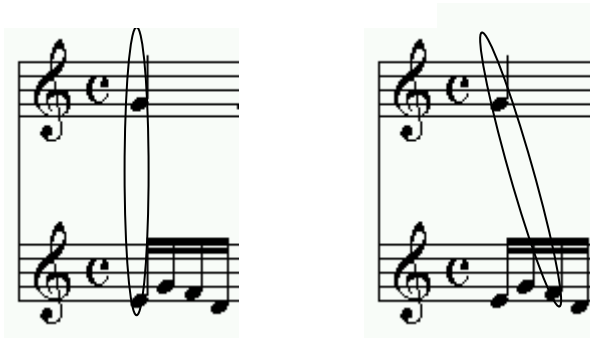
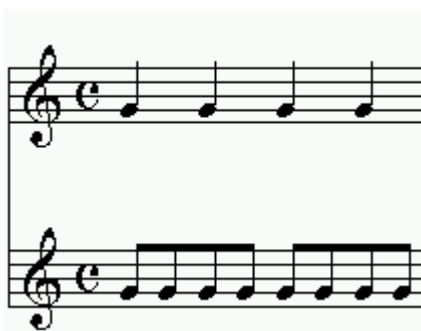


Figure 4.4: The implications of using Method 2 for calculating the duration weights.

In Method 3, where the summed durations are used in every window of a note, the duration weights are as illustrated in Figure 4.5.



Window	DurationWeight Melody 1	DurationWeight Melody2
1	0.25	0.125
2	0.25	0.125
3	0.25	0.125
4	0.25	0.125
5	0.25	0.125
6	0.25	0.125
7	0.25	0.125
8	0.25	0.125

Figure 4.5: An example of the duration weights when Method 3 is used.

The result is that pitch differences in any window occupied by a long note (compared with a series of short notes) will contribute more than the same pitch difference between short notes, as shown in Figure 4.6.

Pitch differences that occur in any window of a long note contribute more than the same differences between short notes

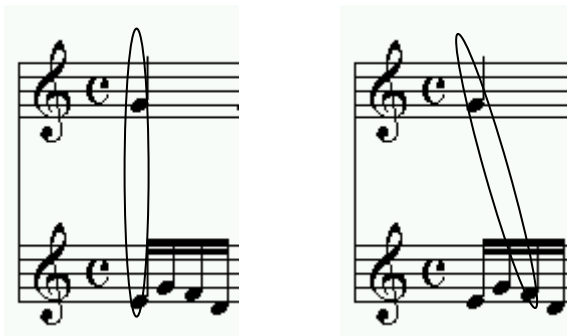
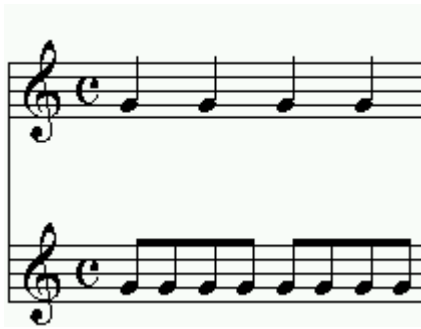


Figure 4.6: The implications of using Method 3 for calculating the duration weights.

In Method 4, the values for windows after the initial onset of long notes is higher than in Method 2, but still less than the actual duration value used in Method 3 (see Figure 4.7).



Window	DurationWeight Melody 1	DurationWeight Melody2
1	0.25	0.125
2	0.125	0.125
3	0.25	0.125
4	0.125	0.125
5	0.25	0.125
6	0.125	0.125
7	0.25	0.125
8	0.125	0.125

Figure 4.7: An example of the duration weights when Method 4 is used.

#### 4.3.4 Metrical accents

Metrical accents are implemented as weights as follows:

$$difference = \sum_{k=1}^n |p_{1k} - p_{2k}| (mw_{1k} \cdot mw_{2k}) \quad 4.10$$

where:

$mw_{1k}$  is the accent applied to melody 1 according to the position in the bar

$mw_{2k}$  is the accent applied to melody 2 according to the position in the bar

The values for the pitch and duration components of the algorithm are taken directly from the score and so there is no interpretation by the researcher or ambiguity involved regarding their values. The values for the metrical accents to be used as weights in the algorithm are not so clear however. Although there is a consensus in music perception research that there is a hierarchical pattern of stress/accent on notes according to their position in the bar, there is little or no discussion of the ratio relationship between such accents (see Chapter 2 for a discussion on metrical accents). There is general agreement that in a bar of 4/4 there is a primary accent on the 1<sup>st</sup> beat of the bar and a secondary accent half way through on the 3<sup>rd</sup> beat of the bar. There is one main accent in 2/4 and 3/4 bars, occurring on the 1<sup>st</sup> beat. Most research does not make a distinction in terms of hierarchical stress for the remaining notes in the bar. Ó Muidín does give the values used for two example melodies in 6/8 time, but notes that these values have been arbitrarily chosen. In this research, the values for the metrical accents are fine-tuned

using the testbed melodies and the human similarity judgements discussed in Chapter 3. This process is discussed in Chapter 5.

#### 4.3.5 Combining pitch, duration, and metrical accent

Each of the four methods of incorporating duration is also combined with the use of the metrical accent weight to provide further versions of the geometric algorithm, as shown in equations 4.11-4.14 below.

$$difference = \sum_{k=1}^n |p_{1k} - p_{2k}| dw1_k (mw_{1k} \cdot mw_{2k}) \quad 4.11$$

$$difference = \sum_{k=1}^n |p_{1k} - p_{2k}| dw2_k (mw_{1k} \cdot mw_{2k}) \quad 4.12$$

$$difference = \sum_{k=1}^n |p_{1k} - p_{2k}| dw3_k (mw_{1k} \cdot mw_{2k}) \quad 4.13$$

$$difference = \sum_{k=1}^n |p_{1k} - p_{2k}| dw4_k (mw_{1k} \cdot mw_{2k}) \quad 4.14$$

In total, ten variations of the geometric melodic similarity algorithm have been implemented.

#### 4.3.6 Rests

The geometric algorithm used here does not include a mechanism for dealing with rests in a score. It can take the duration of rests as well as notes into account when calculating time windows when traversing the score but when a note in one melody occurs in the same window as a rest on the other melody it is unclear how the rest should contribute to the similarity calculations. Only two rests occurred in the testbed melodies being used here. Both are eighth note rests and are located at the very end of the segments used. A long note of the same pitch was found at this point in the score in the Theme and most of the Variations and so it was decided to treat these short rests as if they were actually a continuation of the previous note, making a quarter note instead of an eighth note and an eighth note rest.

## 4.4 Edit Distance Algorithms – An Introduction

One of the most commonly used string-matching algorithms for measuring melodic similarity is the Edit Distance algorithm (Mongeau and Sankoff 1990; Orpen and Huron 1992; Smith, McNab and Witten 1998). This algorithm calculates the number and type of changes needed to transform one melody into another in order to measure the similarity of the two melodies. The more changes needed, the less similar the melodies are to each other. A number of edit operations are defined and a cost is assigned to each operation to arrive at the edit distance. Multiple sets of edit operations are investigated so that the combination of operations that provides the lowest edit distance is found. The melody to be altered is usually referred to as the source melody and the melody it is to be transformed into, as the target melody. The most common edit operations used are insertion of a note, deletion of a note, and replacement of one note with another. Mongeau and Sankoff (1990) define two further operations that will be discussed at a later stage and Orpen and Huron (1992) implements eight different types of edit operations in total, treating repeated and non-repeated notes differently.

A dynamic programming approach is used for implementations of the edit distance algorithm in a melodic similarity context and as such is usually implied and not explicitly stated by authors. This technique speeds up the process by which the edit distance and associated edit operations are found so that every possible set of values and results does not have to be calculated.

Two particular implementations of the edit distance algorithms are investigated here and the performance on the testbed melodies is compared to the results for the geometric algorithm, presented in the previous chapter. The first algorithm from Smith, McNab and Witten (1998), is chosen as an example of a basic version of the edit distance algorithm. The second algorithm, from Mongeau and Sankoff (1990), is chosen because it extends this basic algorithm in an effort to make it more suitable to the task of comparing melodies rather than words. Also, this algorithm is highly referenced in melodic similarity research, so it would be valuable to compare the success of this algorithm with both the basic edit distance version and the geometric algorithm. An overview of both algorithms is given in this chapter, along with details of the implementation. The cost values for the edit operations are found by fine-tuning the

algorithms for best performance using the testbed melodies and the human judgements of similarity detailed in Chapter 3.

#### 4.5 Smith McNab and Witten – A Basic Edit Distance Algorithm

Smith, McNab, and Witten (1998) define the edit distance dynamic programming algorithm as:

$$d_{ij} = \min [ d_{i-1,j} + w(a_i, \emptyset), d_{i-1,j-1} + w(a_i, b_j), d_{i,j-1} + w(\emptyset, b_j) ] \quad 4.15$$

where:

$$1 \leq i \leq \text{length of melody a}$$

$$1 \leq j \leq \text{length of melody b}$$

$$w(a_i, b_j) = \text{cost of replacing } a_i \text{ with } b_j$$

$$w(a_i, \emptyset) = \text{cost of inserting } a_i$$

$$w(\emptyset, b_j) = \text{cost of deleting } b_j$$

and initial conditions are:

$$d_{00} = 0$$

$$d_{i0} = d_{i-1,j} + w(a_i, \emptyset), i \geq 1$$

$$d_{0j} = d_{i,j-1} + w(\emptyset, b_j), j \geq 1$$

The implementation of the algorithm involves the use of a two-dimensional matrix. It should be noted that the authors label the horizontal axis of the matrix as  $i$  and the vertical axis as  $j$ . This is the opposite of the axis labels used in many other accounts of the edit distance algorithm, including that given by Mongeau and Sankoff, which follows in section 4.6. After the initial entry of the cost of insertion and deletion into the top row and left hand column of the matrix, the remainder of the values in the matrix are calculated a cell at a time as the minimum of:

- the cell above + the cost of deleting note  $a_i$
- the cell to the left + the cost of inserting note  $b_j$
- the cell diagonally above and to the left + the cost of a replacing note  $a_i$  with  $b_j$

The authors implement the algorithm using the following edit costs:

Insert cost      4

Delete cost     4

Replace cost    (absolute pitch difference) + weight \* (duration difference)

The duration is expressed in terms of 16<sup>th</sup> note units and the pitch as a MIDI note number.

#### 4.5.1 Filling the Matrix

The edit distance matrix is filled from the top left to bottom right corner. Each cell is filled with a value that is the minimum of the three costs stated above. The value in a particular cell represents the distance between the two melodies up to that point e.g the value in the last cell in the 3<sup>rd</sup> row gives the cost for converting the first three notes of one melody into the other melody, the value in the 5<sup>th</sup> cell of this row gives the cost for converting the first three notes of one melody into the first five notes of the other. The bottom right cell is the overall edit distance value between the two melodies. The best path/set of edits is found by tracing back from the bottom right to the top left cell in the least number of moves. More than one such path may be found.

#### 4.6 Mongeau and Sankoff – Extensions to the Basic Algorithm

The edit distance algorithm used by Mongeau and Sankoff (1990) is similar in structure to Smith et al.'s algorithm but rather than use a predefined value for the cost of insert and delete operations (4 in Smith et al.'s (1998) example above), the cost of these edit operations is dependent on the duration of the note in question. The cost of a replace operation is expressed as a combination of two weight values based on pitch and duration:

$$w(a_i, b_j) = w_{\text{interval}}(a_i, b_j) + k_1 w_{\text{length}}(a_i, b_j) \quad 4.16$$

$k_1$  controls the relative contribution of duration to pitch. The authors define insert and delete operations ( $w(\emptyset, b_j)$  and  $w(a_i, \emptyset)$ ) as  $k_1$  times the duration of the note to be inserted or deleted. Unlike the previous algorithm, here the pitch component of the replace cost is not simply calculated from the difference in pitches between the notes but instead is a weight based on the consonance of the interval between the two notes. This is an attempt to introduce music perception principles into what was originally a string algorithm for processing text.

The pitch of a note is converted to a format that represents the degree of the scale in relation to the tonic. The authors indicate that the weight associated with the interval between notes is independent of key and mode i.e. major and minor keys are regarded

as being equal. Essentially, the difference is calculated in terms of steps of the diatonic scale instead of simply using the chromatic pitch difference. If both notes are in a major or minor scale the following set of weights is used for the diatonic interval between the notes:

Difference in degree scale	Semi-tone diff major key	Semi-tone diff minor key	Interval	Associated Weight
0	0	0	8 <sup>ve</sup>	0
1	2	2	2 <sup>nd</sup>	.9
2	3	4	3 <sup>rd</sup>	.2
3	5	5	4 <sup>th</sup>	.5
4	7	7	5 <sup>th</sup>	.1
5	8	9	6 <sup>th</sup>	.35
6	11	11	7 <sup>th</sup>	.8

**Table 4.1: Mongeau and Sankoff's (1990) weights based on the consonance of the intervals.**

If either of the notes is not in the major or minor key the weight is calculated from the difference in semitones between the notes according to the values given in Table 4.2.

Difference in semi-tones between notes	Associated Weights
0	.6
1	2.6
2	2.3
3	1
4	1
5	1.6
6	1.8
7	.8
8	1.3
9	1.3
10	2.2
11	2.5

**Table 4.2: Mongeau and Sankoff's (1990) weights when either note is not in a major or minor key.**

If both notes do not belong to the major or minor scale, a semi-tone is subtracted from both notes and the difference in degrees of the scale calculated using these adjusted notes. If the notes are still not in the major or minor scale the second set of weights based on the semi-tone difference is used.

Two additional edit operations are also used in this implementation of the algorithm. Consolidation and fragmentation are terms used to describe the replacement of a number of notes of the same pitch with a single note and vice versa. Rather than treating such occurrences as needing separate edit operations for each note involved, these new

edit operations are intended to take the place of a replacement and number of insertions/deletions.



Fragmentation: a single semi-breve is replaced by four crotchets of the same pitch



Figure 4.8: An example of the fragmentation of a note.



Consolidation: four crotchets of the same pitch are replaced by a single semi-breve.

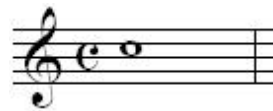


Figure 4.9: An example of the consolidation of a group of notes.

In this case, instead of the cells in the matrix being calculated as the minimum of three possible edit operations (insert, delete, and replace), the minimum of five numbers is used.

#### 4.17

$$d_{ij} = \min [$$

$d_{i-1, j} + w(a_i, \emptyset),$	deletion
$d_{i-1, j-1} + w(a_i, b_j),$	replacement
$d_{i, j-1} + w(\emptyset, b_j),$	insertion
$\{d_{i-1, j-k} + w(a_i, b_{j-k+1}, \dots, b_j)$	fragmentation
$2 \leq k \leq j\}$	
$\{d_{i-k, j-1} + w(a_{i-k+1}, \dots, a_i, b_j)$	consolidation
$2 \leq k \leq i\}$	

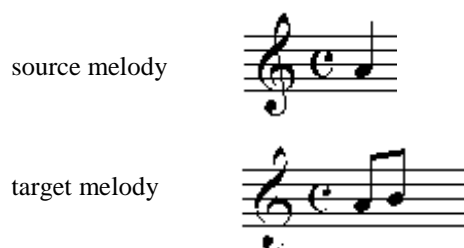
$$]$$

The authors discuss a method of cutting down some of the increased running time that these additions result in. They suggest that the fragmentation of a note into a number of notes of smaller duration can be stopped as soon as the total length of those fragmented notes is less than or equal to the duration of the note being fragmented.

## 4.7 Tracing the Edits Operations and Checking Note Alignments

The dynamic programming approach to the edit distance algorithm results in the lowest possible edit distance cost being entered in the bottom right hand cell of the matrix. The set of edit operations that produced this value can be determined by tracing back through the matrix from the bottom right hand cell to the top left hand cell, calculating at each cell which of the three edit operations was used to arrive at the value in that cell. The edit operations used is noted and a move is made to the appropriate cell; left for insertions, above for deletions, and above diagonal left for replacements. The process is repeated until all moves have been traced to arrive at cell the top left cell of the matrix.

However, both Smith et al. (1998) and Mongeau and Sankoff (1990) draw the reader's attention to the fact that the lowest edit distance alone may not be a reliable measure when using this algorithm in a musical context. Both sets of authors refer to the working out of the edits operations as "tracing" through the matrix. Particular combinations of edit operations imply an "alignment" between the notes of both melodies. By this they mean that the edits imply a matching up or lining up of related notes in both melodies. The example below in Figure 4.10 shows an excerpt from a potential set of melodies in which the cost of transforming a crotchet A into two quavers (A and B) is calculated.



**Figure 4.10: Excerpt from an example source and target melody.**

One set of edit operations might be to simply delete the A crotchet in the first melody and insert two quavers A and B. Another set of operations would be to replace the crotchet A with a quaver A and insert the quaver B. The latter combination of edit operations would imply an alignment of the note A in both melodies according to the above-mentioned authors. Smith et al. state that identifying the set of edit operations is "often useful....to make sure that the match "makes sense"... " (p.107). The values chosen for the edit operations and weights in these algorithms directly affect both the final edit distance and the associated set of edit operations. A series of edit operations or

alignments that do not make musical sense may indicate the use of unsuitable values for the edit operations.

Mongeau and Sankoff explore the issue of alignments in some depth as they study the effect that different values for the weight  $k_1$  has on the alignment between melodies. They “tweak” the value of this weighting factor so that it does not align lots of highly dissonant notes of the same length, link notes of the same pitches but different lengths, or compare too many notes across the bar line in different bars.

## **4.8 Implementing the Edit Distance Algorithms**

Smith et al. used a pre-defined value for the insert and delete costs in their implementation of the edit distance algorithm. In both examples discussed by the authors, these edit operations are assigned the exact same values. Smith et al. also use a weight value in calculating the replace costs so that the relative contribution of the duration difference and pitch difference can be controlled.

In an effort to find the most appropriate values for the two edit operations and the weight used in the replace cost, a number of different values are used for repeated processing of the testbed melodies. The results produced by the algorithms are then compared with the human similarity judgements and the best performing values are identified. While Smith et al. set the insert and delete costs at the same value, here a number of combinations of values are run including equal insert and delete costs and independently varied insert and delete costs. The duration weight used in calculating the replace cost is also set to zero in one instance so that the performance of the algorithm with pitch only can be assessed.

Mongeau and Sankoff (1990) use a weight factor  $k_1$  for the calculation of all edit distance costs (see 4.16). The cost for the insertion or deletion of a note is calculated by multiplying this weight by the duration of the note in question. The cost for a replace operation is calculated using the pitch consonance model discussed above and the weight multiplied by the difference in duration between the notes.

A number of different versions of the Mongeau and Sankoff algorithm are implemented here in order to judge the success of each of the components of their algorithm. Mongeau and Sankoff implemented a version of the edit distance that used pitch

weights based on the consonance of the intervals involved along with fragmentation and consolidation. They did not implement the algorithm without these features and so it is not evident from their work if these extensions of the algorithm improve or disimprove the performance. Here, their algorithm is implemented in a number of different ways so that the performance of each element can be judged. The first version of this algorithm implemented here uses the basic pitch difference between notes when calculating the replace cost, rather than the pitch consonance model given. This is similar to running Smith et al.'s algorithm but using the weighted note duration for the insert and delete costs rather than a predefined value. The pitch consonance model proposed by Mongeau and Sankoff is also implemented. In both cases the weighting factor  $k_1$  is varied independently for insert, delete and replace costs, which essentially means that three separate weights are used rather than just one and a variety of combinations of values are run to determine which algorithms and parameters result in the best performance. The fragmentation and consolidation costs have been implemented separately to the above algorithms so that the effect of this aspect of Mongeau and Sankoff's algorithm can be examined also. These additional edit operations are incorporated with the previously mentioned versions of the algorithms that use pitch difference and the pitch consonance values to calculate the replace cost. Table 4.3 below contains a list of all the versions of both edit distance algorithms that were implemented.

The algorithm description that is shown in bold is the actual version of the algorithm that was implemented by Mongeau and Sankoff (1990). They used weights based on pitch consonance rather than the actual pitch difference value, included fragmentation and consolidation operations, and used the same weight value for calculating insert, delete, and replace operations.

<b>Edit Distance Algorithm</b>	<b>Features, weights and costs implemented</b>
Smith	Insert cost $\neq$ Delete cost. Duration used in replace cost.
	Insert cost $\neq$ Delete cost. No duration used in the replace cost.
	Insert cost = Delete cost. Duration used in replace cost.
	Insert cost = Delete cost. No duration used in the replace cost.
M&S <sup>2</sup> pitch difference	Same insert, delete and replace weights. Duration used in replace cost.
	Same insert and delete weight. Duration used in replace cost.
	Same insert and delete weight. No duration used in replace cost.

<sup>2</sup> M&S is an abbreviation for the Mongeau and Sankoff (1990) algorithm.

Edit Distance Algorithm	Features, weights and costs implemented
	Different insert and delete weights. No duration used in replace cost.
	Different insert, delete and replace weights. Duration used in replace cost.
M&S pitch difference with fragmentation and consolidation	Same insert, delete and replace weights. Duration used in replace cost.
	Same insert and delete weight. Duration used in replace cost.
	Same insert and delete weight. No duration used in replace cost.
	Different insert and delete weights. No duration used in replace cost.
	Different insert, delete and replace weights. Duration used in replace cost.
M&S pitch consonance	Same insert, delete and replace weights. Duration used in replace cost.
	Same insert and delete weight. Duration used in replace cost.
	Same insert and delete weight. No duration used in replace cost.
	Different insert and delete weights. No duration used in replace cost.
	Different insert, delete and replace weights. Duration used in replace cost.
M&S pitch consonance with fragmentation and consolidation	<b>Same insert, delete and replace weights. Duration used in replace cost.</b>
	Same insert and delete weight. Duration used in replace cost.
	Same insert and delete weight. No duration used in replace cost.
	Different insert and delete weights. No duration used in replace cost.
	Different insert, delete and replace weights. Duration used in replace cost.

**Table 4.3: The versions of the edit distance algorithms implemented.**

#### 4.8.1 Expected Results

Due to the nature of the edit distance algorithms some specific sensitivities to the musical material of the testbed melodies was predicted. A cost is associated with inserting, deleting, and replacing notes (as well as the fragmentation and consolidation costs used by Mongeau and Sankoff). The more notes that are different between melodies, the higher the overall edit distance will become. The nature of the musical material used in this case suggests that insertions will be used far more than deletions to transform the Theme into each Variation in turn. Variations VIII and IX in particular are very elaborate versions of the Theme with long passages of semi-quaver movement as shown in Figure 4.11. Each quarter note in the Theme is embellished to become four



**Figure 4.11 The first two bars of the Theme and Variation VIII and Variation IX**

sixteenth notes, involving the insertion of at least three notes and the half note in bar 1 of the Theme requires the insertion of at least seven notes to turn it into the variation melody.

Variations that are considered perceptually similar to the Theme as evidenced by the human similarity judgements may involve a number of insert operations as the notes of the Theme melody are embellished and ornamented to form the variation melodies. It was expected that the algorithms would return low values for insert operations in such cases so that the overall edit distance would remain relatively low and provide a result that corresponded to the human similarity judgements.

It was noted that, since the edit distance algorithms favour insertion operations over deletion operations due to the nature of the musical material in the testbed, the weights and costs identified as providing the best results for the testbed melodies might not be transferable to other melodies. This issue is investigated further through the use of additional melodies in different styles and is reported on in Chapter 6.

## Chapter 5

### Fine-tuning the Algorithms and Results

A method is presented here for comparing the algorithmic results with the human similarity judgements using the testbed from Chapter 3. The internal weights and parameter values of the algorithms are then fine-tuned using these human observations. The results are verified by the use of further melodies extracted from the Theme and Variations and associated human similarity judgements.

In the case of the geometric algorithm Ó Maidín (1998a) uses an arbitrary set of values for the metrical accent weight. Smith et al. (1998) assign subjective values to the insert and delete costs and also subjectively chose to add half of the duration difference to the pitch difference in calculating the replace cost. The authors do suggest checking that the alignments indicated by the costs make musical sense and that “In setting the parameters, the researcher must consider what the musical questions is, and what feature(s) form the basis for similarity measurement.” (Smith 1998, p.108). In their own implementation of the edit distance algorithm, Mongeau and Sankoff (1990) fine-tune the weight that controls the relative contribution of duration to pitch according to their own subjective perception of the similarity of the melodies. Here, a more objective approach is taken with the reliable consistent judgments from the listening experiment used to provide objective similarity judgments for each melody and the various weights fine-tuned accordingly.

#### 5.1 Related Research

There are comparisons with research by Müllensiefen and Frieler (2004a, 2004b, 2004c) who gather similarity ratings on melodies and seek to find an optimal similarity measure based on these ratings from a set of 34 algorithms implemented. The algorithms include edit distance, n-grams, tonality measures and vector correlation measures. Although many versions of each algorithm are run the internal algorithm parameters are often not adjusted. In the case of the edit distance algorithms, for example, “raw” pitches, pitch weighted by duration, contour, intervals, and classes of rhythm (long, short etc.) are implemented but what is not clear is what particular cost and weight values were used

for the algorithm. If set costs are used for insert and delete operations what values were chosen? Alternatively, if these costs are calculated from the weighted durations as in Mongeau and Sankoff (1990), what weight values are used? There does not appear to be any optimisation of the internal algorithm weights.

The authors note that:

“As it is probable that human music experts make use of the information on several dimensions simultaneously, an optimal algorithmic model of the human ratings would encompass measures from several dimensions in a linear combination.” (Müllensiefen and Frieler 2004c, p.164)

Rather than incorporate the information from different dimensions in individual algorithms, they choose to organise the algorithms according to category, identify the best algorithm in each category and then to use linear regression to find the best overall combination. The categories used were interval, contour, rhythm, harmonic content and characteristic motif categories (mostly n-grams) and the Euclidean distance from the subjects’ ratings was used to find the best algorithm in each of these categories. In an experiment that featured a number of reference melodies with variants created with “errors” such as rhythmic or contour differences from the reference melody. The optimal combination of algorithms and weights identified by the linear regression analysis was:

$$3.355 \cdot \text{rawedw} + 2.852 \cdot \text{ngrcoord} \qquad \mathbf{5.1}$$

The rawedw is a “rhythmically weighted” edit distance algorithm on the actual pitches of the melodies (as opposed to intervals or contour). No details are given on how rhythm is included in the algorithm and what the relative contribution of pitch and rhythm is. The ngrcoord algorithm is an implementation of n-grams that counts the number of n-grams both melodies have in common. A further experiment was run in which the variant melodies were all transposed and these two algorithmic measures were reported as the optimal combination for similarity but with different weights returned from the linear regression analysis. The final experiment run by Müllensiefen and Frieler (2004a, 2004b, 2004c) included a number of variant melodies that were not derived from the reference melody and so would not be considered very similar by the subjects. This experiment returned a different optimal similarity measure that was the sum of the following weighted algorithms: n-grams using the Ukkonen measure, edit distance using classes of duration (long, short etc.) and a tonality correlation measure.

A contrasting approach is taken in this research. The most likely features to be of use were identified from related literature on music perception. Various different versions of each algorithm were then implemented to assess the interaction of these features within the algorithms themselves and internal weights tuned according to the subjects' similarity ratings so that the most successful realisation of the algorithm (within the framework used) was identified.

## 5.2 Assessing the Performance of the Algorithms Using the Human Similarity Judgements

The implemented algorithms are designed to return a value that represents the degree of difference between the two melodies being compared. Zero indicates that there is no difference between the two melodies, small positive values indicate very similar melodies, while larger values indicate a decreased level of similarity between the melodies.

The human similarity judgements were collected so that they could be used to evaluate the success of such melodic similarity algorithms in correctly identifying the level of similarity between two melodies. In comparing the human similarity judgements with the algorithm output values, there are a number of issues that need to be considered. (It is noted as a reminder here that the median rating for each variation is used as discussed in section 3.6.7).

1. The ratings have a definite range limit while the algorithm output does not:

	Minimum value	Maximum value
Ratings (human measure)	1	7
Algorithmic measures	0	unknown

**Table 5.1: The range of the human and algorithmic similarity measures.**

2. The limits/extremes of the both sets of results have the opposite meaning to each other:

	Least similar	Most similar
Ratings (human measure)	1	7
Algorithmic measures	some unknown value > 0	0

**Table 5.2: The meaning of the extreme values of the human and algorithmic similarity measures.**

The range of difference values output from the algorithms is context dependant, while the human similarity ratings are always in the designated range of 1 to 7. The maximum possible value and its meaning in terms of similarity will be different for each version of the various algorithms used. Any measure used to compare the two human ratings with the algorithmic measures of similarity will need to take both of these factors into account. It was decided to reverse one set of results so they were aligned in terms of similarity, with 0 meaning most similar in both cases. The data was then normalised so that both sets of results were readable on the same scale. A measure was then applied that facilitated the comparison of the performance of the algorithms over all of the variation melodies of the testbed.

### 5.2.1 Normalising the data

There are a number of options for normalising data in this way ranging from a) z-scores, b) dividing each value by the maximum value, and c) non-linear transformations of the data into a new range. In this case the min-max normalisation was chosen. This normalisation technique linearly maps values as follows:

$$y' = \left( \frac{y - \min}{\max - \min} \right) (\text{newmax} - \text{newmin}) + \text{newmin} \quad 5.2$$

The values are normalised so that they lie between 0 and 1 in both cases. Before normalisation of the pooled human similarity ratings, the values were reversed so that the most similar melody corresponded to the minimum value and the least similar melody corresponded to the maximum value, as is the case with the algorithmic similarity measure. The minimum and maximum human similarity values will always be the same for the set of Theme and Variation melodies but the algorithmic values of similarity will vary according to the algorithm used.

### 5.2.2 Comparing the Algorithm Output with the Human Similarity Judgements

The measure chosen for the actual comparison of the algorithmic output with the human judgements of similarity was the sum of the absolute differences between pairs. This measure is abbreviated to SumAbsDiff in the text and discussions that follow.

$$\sum |x_i - y_i|$$

5.4

where:

x is the normalised median human rating for the Theme and Variation i

y is the normalised algorithmic result for the Theme and Variation i

i = a pair of melodies consisting of the Theme and Variation i

1 < i <= number of variations

The differences between the human and algorithmic judgements of similarity are calculated for individual variation melodies in the testbed and added together to provide a single value that reflects how the algorithm performed in relation to the human similarity judgements.

An example of the process involved is shown below in Table 5.3 to illustrate the inversion and normalisation of the rating, a sample normalisation of the output from one algorithm, and the calculation of the SumAbsDiff measure. The results for six pairs of melodies are used in this example.

Melody Pair	Original geometric algorithm output	Original ratings	Reversed ratings	Normalised (reversed) ratings	Normalised algorithm output	Absolute Difference
1	2	7	0	0.00000	0.00000	0.00000
2	9.39496	4.5	2.5	0.41667	0.41370	0.00296
3	3.0625	6	1	0.16667	0.05944	0.10723
4	4.94499	5	2	0.33333	0.16475	0.16858
5	19.875	1	6	1.00000	1.00000	0.00000
6	17.87278	2	5	0.83333	0.88799	0.05465
					<b>SumAbsDiff</b>	0.33342

**Table 5.3: An example of the calculation of the SumAbsDiff value.**

This method of comparing the output of the algorithm with the human similarity judgements allows one to investigate which algorithms perform well and which don't across all of the testbed melodies.

Many of the geometric algorithms use a weight to represent metrical accent values (see section 4.3.4), but as was previously discussed, appropriate values are not known for these weights. Similarly, the edit distance algorithms have cost and weight parameters whose ideal values are not known. In the case of the geometric algorithm an optimising

approach was used to continually modify the values of the metrical accent weights until the best possible SumAbsDiff result is achieved. A hill-climbing algorithm was chosen to fine-tune these weights and the basic steps of this approach are discussed in the next section.

### **5.3 Using a Hill-Climbing Algorithm to Fine-tune the Metrical Accent Weights**

Hill-climbing algorithms are used to find the optimal point in a search space. The general steps in a basic hill-climbing algorithm are:

1. Start from a random point in the search space.
2. Consider the values of each of the neighbouring states.
3. Move to the neighbour that provides a higher/better result.
4. Repeat steps 2 and 3 until all neighbours give a lower result, signifying that the optimal point has been reached.
5. Return the current state as the best result.

In the context of this research, values are sought for the metrical accent weights that will result in the lowest SumAbsDiff value. The lower the SumAbsDiff value, the closer the algorithmic measure of similarity is to the human similarity judgements. In fine-tuning the metrical accent weights, the steepest descent version of the hill-climbing algorithm is most appropriate. A descending algorithm is used because the aim is to reach the lowest possible result for the SumAbsDiff value. The “steepest” aspect of this algorithm refers to the fact that all possible successors to the current state are examined and the successor closest to the best solution is used.

The suitability of such a hill-climbing algorithm for fine-tuning the metrical accent weights used in the geometric algorithm was investigated by calculating the SumAbsDiff for a range of experimental weight values. The value of the metrical accent weight on the first beat of the bar was varied from 1 to 9.99 in increments of .01 for each of the algorithms that feature this weight. The results are presented in a set of graphs in Appendix C. It can be clearly seen in all cases that the SumAbsDiff result decreases steadily towards a more favourable result as the metrical accent increases from 1 to

some point that ranges from slightly less than 2 to less than 4, depending on the particular algorithm in question. From that point the SumAbsDiff increases (worse results) as the metrical accent value is increased. All of the graphs have a single clear lowest SumAbsDiff result and do not feature multiple peaks or troughs. A hill-climbing algorithm is therefore appropriate for use with this data and the problem of getting stuck on local rather than globally optimal solutions is avoided. The metrical accent values that result in the best SumAbsDiff results also provide a rough guide to the range of values that might be most useful.

In terms of the melodic similarity algorithm the implementation was as follows:

1. Generate two random numbers for the metrical weights for the 1<sup>st</sup> and 3<sup>rd</sup> beats.
2. Generate a set of eight neighbouring states by increasing/decreasing one/both of the metrical weight values by a set value.
3. Move to the neighbouring state that gives the best SumAbsDiff result.
4. Repeat steps 2 and 3 until all neighbours give a worse result, signifying that the optimal weight values have been reached.

Two randomly generated numbers between 1 and 10 at the start of the fine-tuning process. In order to maintain the perceptual principle that the metrical accent on the 1<sup>st</sup> beat of the bar (the primary stress) is greater than that on the 3<sup>rd</sup> beat of the bar (the secondary stress), the condition is set that the first of these two random number is greater than the second number by at least 10%. Numbers of one and higher are generated because the implementation of the algorithm used here means that a value of one implies that no metrical stress/accent is present. An increment<sup>3</sup> of .01 is used to create the eight neighbouring states as follows:

- 1<sup>st</sup> number – increment, 2<sup>nd</sup> number
- 1<sup>st</sup> number – increment, 2<sup>nd</sup> number – increment
- 1<sup>st</sup> number, 2<sup>nd</sup> number – increment
- 1<sup>st</sup> number + increment, 2<sup>nd</sup> number – increment
- 1<sup>st</sup> number + increment, 2<sup>nd</sup> number

---

<sup>3</sup> The term increment can be used in this context to mean a positive or negative change in the value of a variable.

- 1<sup>st</sup> number + increment, 2<sup>nd</sup> number + increment
- 1<sup>st</sup> number, 2<sup>nd</sup> number + increment
- 1<sup>st</sup> number – increment, 2<sup>nd</sup> number + increment

In order to avoid finding local maxima, the process is repeated 30 times. The initial values for the metrical accent weights are different each time as they are randomly generated but each repetition would be expected to return the same or very similar values. The optimal weights are considered to be the values that gave the best (lowest SumAbsDiff) results over all 30 repetitions. The hill-climbing algorithm (with 30 repetitions) is run for each of the algorithms that include metrical accent weights (see sections 4.2.4 and 4.2.5).

## **5.4 Results for the Geometric Algorithm**

Although the majority of the variation melodies from the testbed are in the same time signature as the Theme (4/4), Variations IV and VI are in different time signatures. The division of the melodies into time windows and the use of metrical weights could be problematic when comparing melodies in different time signatures. Therefore it was decided to examine the performance of the algorithms on the six Variations in 4/4 time separately. Although Variation VII has a 4/4 time signature, it is made up entirely of triplets that divide the main beat into three rather than two, making it comparable to a different time signature. Therefore, Variation VII is also omitted from the initial group of melodies in a 4/4 time signature. Further discussion of Variation VII is included in section 5.4.7 and the issue of different time signatures is revisited later in this chapter in section 5.4.

The results for the first group of six melodies are shown below, in order of best performing algorithm. The algorithm, features of the melody used, and the SumAbsDiff result is given in each case. Where metrical accent weights are used, the optimal values for the primary and secondary accents of the bar are given.

Results for the six melodies in 4/4 sorted by SumAbsDiff. Each combination of features is run 30 times.	Features used	SumAbs Diff	Optimal metrical weight for 1st beat	Optimal metrical weight for 3rd beat
$\sum_{k=1}^n  p_{1k} - p_{2k}  dw1_k (mw_{1k} \cdot mw_{2k})$	Pitch Duration Metrical accents	0.17988	1.82	1.32
$\sum_{k=1}^n  p_{1k} - p_{2k}  dw4_k (mw_{1k} \cdot mw_{2k})$	Pitch Duration Metrical accents	0.24875	1.98	1
$\sum_{k=1}^n  p_{1k} - p_{2k}  dw2_k (mw_{1k} \cdot mw_{2k})$	Pitch Duration Metrical accents	0.31728	1.67	1.13
$\sum_{k=1}^n  p_{1k} - p_{2k}  (mw_{1k} \cdot mw_{2k})$	Pitch Metrical accents	0.4209	3.32	1
$\sum_{k=1}^n  p_{1k} - p_{2k}  dw3_k (mw_{1k} \cdot mw_{2k})$	Pitch Duration Metrical accents	0.457	3.2	1
$\sum_{k=1}^n  p_{1k} - p_{2k}  dw1_k$	Pitch Duration	0.46724	-	-
$\sum_{k=1}^n  p_{1k} - p_{2k}  dw4_k$	Pitch Duration	0.56308	-	-
$\sum_{k=1}^n  p_{1k} - p_{2k}  dw3_k$	Pitch Duration	0.60624	-	-
$\sum_{k=1}^n  p_{1k} - p_{2k} $	Pitch	0.62952	-	-
$\sum_{k=1}^n  p_{1k} - p_{2k}  dw2_k$	Pitch Duration	0.66346	-	-

**Table 5.4: Results for the six 4/4 variation melodies, sorted by the SumAbsDiff result.**

## **Discussion of Results**

These results clearly show that the three best performing algorithms use a combination of pitch, duration, and metrical accent. The use of a metrical accent weight improved the performance of all algorithms over those versions that used pitch or pitch with duration. In fact, the top five algorithms all incorporate a metrical accent and the five poorest performing algorithms, seen at the bottom of this table, used pitch only or pitch with duration. In all cases but one (duration Method 2), using duration along with pitch improved the performance over pitch alone. In Chapter 2, the findings of a number of experiments on the role of rhythm in melodic memory were discussed (see sections 2.4.4, 2.4.7 and 2.5) and it was suggested that that rhythm as well as pitch was important in allowing people to remember melodies and so might be useful for melodic similarity algorithms. Except for the implementation of duration Method 2, the results found here corroborate this theory. The assertion that metrical accents are points of added emphasis or salience in melodies is also validated (see sections 2.2.7 and 2.2.8). Since metrical accents are thought to coincide with melodic accents (see sections 2.2.8 and 2.4), these results might also be taken as indirect confirmation of the presence and usefulness of melodic accents in this context.

The best performing algorithm uses the window width as the duration weight (Method 1, see sections 4.2.2 and 4.2.3) and a metrical accent value of 1.82 on the first beat of the bar and 1.32 on the third beat of the bar for the secondary stress. The results for the remaining algorithms that incorporate a metrical accent weight give an optimised value of 1 or very close to 1 for the accent weight on the third beat of the bar. This essentially means that no extra emphasis is placed on the notes that occur on the third beat of the bar and that the best performance was reached when a metrical accent was placed on the first beat of the bar only. These algorithms return optimal ratios of roughly 2:1, 1.7:1, 3.3:1, and 3.2:1 depending on which duration method was also used.

The algorithm was repeated with the difference between the accent weights on the first and third beats of the bar ranging from 5% to 25% and no difference in results was observed.

### **5.4.1 Issues Encountered when Using All Nine Variation Melodies**

At the beginning of section 5.4, a number of reasons were mentioned for not running the melodic similarity algorithm on all nine melodies of the Theme and Variations

contained in the testbed. The ability of the algorithm to process melodies in different time and key signatures is discussed in the sections that follow, along with some other issues encountered in the three variations omitted in the first set of results.

### 5.4.2 Transpositions

The manner in which Ó Maidín's algorithm deals with transposed melodies was previously discussed in section 4.1.3. Seven of the nine variation melodies are in G major, the same key as the Theme. Variations V and VI are in G minor and so only two notes in the scale are affected – B and E become B flat and E flat. The algorithm is intended to identify melodies that have been transposed to different keys in which every note is shifted up or down a number of semitones and so this transposition to the relative minor key will not be identified. The issue of transposition to different keys is further explored in relation to this algorithm in Chapter 6.

### 5.4.3 Different Time Signatures

When both melodies being compared are in the same time signature it is quite easy to visualise the process by which this algorithm steps through the melodies comparing the pitch differences and weighting these by duration and metrical accent. An example of the Theme with Variation I is shown below in Figure 5.1 with vertical lines representing the time windows.



Figure 5.1: The time windows used to process the first bar of the Theme and Variation I.

It is not so obvious how one would compare melodies in different time signatures as shown in Figure 5.2 which contains excerpts from the Theme and Variations IV and VI. Three questions are raised by this example; how to handle compression/expansion in the time domain (Variation VI), how to use the time windows to compare melodies in different time signatures, and how to use metrical stress values appropriately when different time signatures are involved. Each of these issues is now discussed in turn.

The image shows three musical staves. The first staff, labeled 'Theme', is in 4/4 time and contains two measures of music with a key signature of one sharp (F#). The second staff, labeled 'Variation IV', is in 6/8 time and contains two measures of music with a key signature of one sharp. The third staff, labeled 'Variation VI', is in 3/4 time and contains two measures of music with a key signature of two flats (Bb, Eb). The first measure of Variation VI features a triplet of eighth notes, and the second measure features three triplet markings over eighth notes.

Figure 5.2: The first two bars of the Theme and Variations IV and VI.

#### 5.4.4 Compression/Expansion in the Time Domain

The issue of compression/expansion of the melody in the time domain is evident in the extract from Variation VI (Figure 5.2 above), but may also occur between melodies in the same time signature. In the sample melodies shown in Figure 5.3 below, for example, every quarter note in the first melody is expanded to form a half-note making the second melody twice as long as the first.

The image shows two musical staves, both in common time (C). Melody 1 consists of eight quarter notes. Melody 2 consists of eight half notes, where each half note in Melody 2 corresponds to a quarter note in Melody 1, illustrating a 1:2 time ratio.

Figure 5.3: Sample melodies showing expansion (or stretching) in the time domain.

Ó Mairín proposed further development of the algorithm to include the use of the time ratio between such melodies to adjust the calculation for the width of the time windows and for traversing or stepping through both melodies. This method is useful if there is a simple augmentation or diminution of a melody (compression/expansion in the time domain). In the above example, the time ratio of melody 1 to melody 2 is 1:2. The calculation for each window width to be processed (and for duration weights when used) involves the division of the duration of the current note in melody 2 by the time ratio of melody 1 to 2 (or vice-versa), as shown in equation 5.4 below.

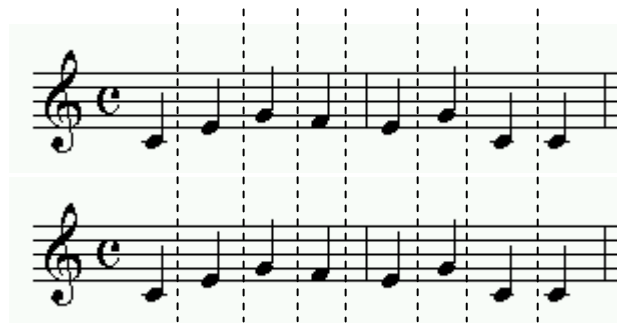
$$\text{Adjusted Melody 2 Duration} = \frac{\text{Melody2Duration}}{\text{Melody2 : Melody1}} \quad 5.4$$

An example based on the first notes of the melodies in Figure 5.3 is presented below:

Duration of note in melody 1	Duration of note in melody 2	Duration of melody2:melody1	Adjusted melody 2 duration
$\frac{1}{4}$	$\frac{1}{2}$	2:1	$\frac{\frac{1}{2}}{2} = \frac{1}{4}$

**Table 5.5: The duration/adjusted duration of the melodies from Figure 5.3.**

This adjustment of the duration allows such melodies to be treated as if they were both identical with windows of width  $\frac{1}{4}$  and duration weights of  $\frac{1}{4}$  for the notes in both melodies, as shown in Figure 5.4.



**Figure 5.4: The adjusted durations and time windows of Melody 2 from Figure 5.3.**

The use of the time ratio allows both to be aligned on a common time scale. Melodies that are compressed or stretched by a factor of 2 or 4 can be treated similarly.

#### 5.4.5 The Time Ratio Technique for Melodies in Different Time Signatures

The time ratio between melodies is also used to calculate the window widths of melodies that are in different time signatures and to subsequently step through each window to compare the melodies. In the example testbed melodies shown in Figure 5.5, the ratio is 6/8:4/4 or 6/8:1.



**Figure 5.5: Each bar of the Theme is matched against a bar of Variation IV.**

This ratio is used to calculate the width of the first time window to be processed, as shown in Table 5.6.

Window number	Duration of note in Theme	Duration of note in Variation	Duration of note in Variation/time ratio	Duration of time window to process (smallest of two windows)
1	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{\frac{1}{8}}{\frac{6}{8}}$	$\frac{1}{6}$

**Table 5.6: The calculation of the first time window adjusted by the time ratio.**

The details of the time windows for the first bar are given below in Table 5.7.

Window Width	Position in melody 1	Position in melody 2	Melody 1	Melody 2
$\frac{1}{6}$	$\frac{0}{1}$	$\frac{0}{1}$	start of note 1	start of note 1
$\frac{1}{12}$	$\frac{1}{6}$	$\frac{1}{8}$	middle of note 1	start of note 2
$\frac{1}{12}$	$\frac{1}{4}$	$\frac{3}{16}$	start of note 2	start of note 3
$\frac{1}{12}$	$\frac{1}{3}$	$\frac{1}{4}$	middle of note 2	start of note 4
$\frac{1}{12}$	$\frac{5}{12}$	$\frac{5}{16}$	middle of note 2	start of note 5
$\frac{1}{6}$	$\frac{1}{2}$	$\frac{3}{8}$	start of note 3/mid-point in bar	start of note 6/mid-point in bar
$\frac{1}{12}$	$\frac{2}{3}$	$\frac{1}{2}$	middle of note 3	start of note 5 (4/8 into bar)
$\frac{1}{12}$	$\frac{3}{4}$	$\frac{9}{16}$	start of note 4	middle of note 5S
$\frac{1}{6}$	$\frac{5}{6}$	$\frac{5}{8}$	middle of note 4	start of note 6

**Table 5.7: The calculation of the time windows of the first bar adjusted by the time ratio.**

The use of the time ratio in this way results in the two beat units of the 4/4 melody being matched against the three beat units of the 6/8 melody. The start of the bar (0/1) and the mid-point in each ( $\frac{2}{4}$  or  $\frac{1}{2}$  for the Theme and  $\frac{3}{8}$  for Variation IV) coincide and these notes are aligned for comparison. In order to compare the three quavers of the 6/8 variation with the two crotchets of the 4/4 Theme, the windows sometimes occur in the middle of notes so that the three against two time units can be compared. The first window width is  $\frac{1}{6}$ , for example. If one steps  $\frac{1}{6}$ <sup>th</sup> into the Theme melody this stops the score iterator  $\frac{1}{6}$ <sup>th</sup> of the way into the first quarter-note, then  $\frac{1}{12}$ <sup>th</sup> (the next window) and so on. The first beat and the mid-point of the bar (the third beat) always align/coincide. The method discussed here is intended for use when one bar of the

melody in a given time signature directly relates to another in a different time signature, as is the case in Figure 5.6.

In other cases the time relationship between melodies may be more complex than a straight bar-to-bar mapping. Variation VI, for example, is in  $\frac{3}{4}$  time but is also stretched in the time domain so that two bars of Variation VI matches up with one bar of the Theme as shown below in Figure 5.6. In the case of the Theme and Variations used for the testbed, the variations that are in different time signatures (Variation IV  $\frac{6}{8}$  and Variation VI  $\frac{3}{4}$ ) feature simple bar-to-bar or bar-to-two-bar ratios.

#### 5.4.6 Metrical Accent Weights in Melodies with Different Time Signatures

The values used for the metrical accent weights in the melodies in  $\frac{4}{4}$  have been adapted to suit other time signatures. In  $\frac{4}{4}$  melodies there is a primary accent on the first beat of the bar (0/4 rational position) and a secondary accent on the third beat of the bar ( $\frac{2}{4}$  rational position). In  $\frac{2}{4}$  and  $\frac{3}{4}$  time signatures there is one accent only on the first beat of the bar. In  $\frac{6}{8}$  time signatures there are two accents in the bar, the second being half-way through the bar at the  $\frac{3}{8}$  position. Table 5.8 below gives an example of how a set of sample weights (3, 1, 2, 1) assigned to a melody in  $\frac{4}{4}$  time signature can be applied to melodies in other time signatures.

The image shows two musical staves. The top staff is labeled 'Theme' and is in 4/4 time. It contains a melody of eight notes: G4, A4, B4, C5, B4, A4, G4, F4. The bottom staff is labeled 'Variation IV' and is in 6/8 time. It contains a melody of sixteen notes: G4, A4, B4, C5, B4, A4, G4, F4, G4, A4, B4, C5, B4, A4, G4, F4. Arrows point from the first and third notes of the Theme to the first and third notes of the first bar of Variation IV. Brackets with the number '3' are placed under the first three notes of the first bar and the first three notes of the second bar of Variation IV, indicating a triplet of eighth notes.

Figure 5.6: Each bar of the Theme is matched against two bars of Variation IV.

#### 5.4.7 The Variation Technique used in Variation VII

Variations IV, VI and VII were omitted from the first set of variations on which the algorithms were run and the results are presented in section 5.4. Variation IV and VI are in different time signatures to the theme and so the techniques presented above for handling these differences are evaluated separately. Although Variation VII does use the

same time signature as the Theme melody (4/4), it was omitted from the first set of six variations because it consists entirely of 1/8<sup>th</sup>-note triplets that divide the main beat into three instead of two.

Position in bar	Metrical accent weights
<b>4/4 time signature</b>	
0/4	3
1/4	1
2/4	2
3/4	1
<b>2/4 time signature</b>	
0/2	3
1/2	1
<b>3/4 time signature</b>	
0/3	3
1/3	1
2/3	1
<b>6/8 time signature</b>	
0/8	3
1/8	1
2/8	1
3/8	2
4/8	1
5/8	1

**Table 5.8: An example of metrical accent weights for a range of time signatures based on the 4/4 time signature weights.**

On closer inspection, however, an important stylistic difference between Variation VII and the remaining eight variations was discovered. In the case of all other variations, the whole Theme melody is altered or embellished, with a different musical style evident in each variation (see Appendix A). The first four bars of the Theme and Variation VII are shown below in Appendix A. Here, the composer has structured the variation in such a way that the first bar is repeated as a musical sequence in bars 2-4 and has not based the musical material around each bar of the Theme. It would not be appropriate to compare bars 1-4 of the Theme with bars 1-4 of Variation VII (Part A melodies from the listening experiment) as part of the fine-tuning and evaluation of the algorithms since bars 2-4 are not related in both melodies. All ten algorithms are run on a larger set of eight variations, which now includes Variations IV and VI. The techniques outlined for

comparing melodies in different time signatures are adopted and the metrical accent values are once again fine-tuned using the hill-climbing algorithm.

#### 5.4.8 Results when Variations IV and VI are included.

The techniques outlined in sections 5.3.3, 5.3.5 and 5.3.6 for comparing melodies in different time signatures are adopted here and the metrical accent values are once again fine-tuned using the hill-climbing algorithm. The results are presented in Table 5.9 before being compared to the results from section 5.4 for the initial set of six variations that were all in the same time signature.

Results for eight variations sorted by SumAbsDiff (Variation VII only omitted). Each combination of features is run 30 times.	Features used	SumAbs Diff	Optimal metrical weight for 1st beat	Optimal metrical weight for 3rd beat
$\sum_{k=1}^n  p_{1k} - p_{2k}  dw4_k (mw_{1k} \cdot mw_{2k})$	Pitch Duration Metrical accents	0.48603	1.98	1
$\sum_{k=1}^n  p_{1k} - p_{2k}  dw1_k (mw_{1k} \cdot mw_{2k})$	Pitch Duration Metrical accents	0.49485	1.82	1.32
$\sum_{k=1}^n  p_{1k} - p_{2k}  dw2_k (mw_{1k} \cdot mw_{2k})$	Pitch Duration Metrical accents	0.66835	1.65	1.14
$\sum_{k=1}^n  p_{1k} - p_{2k}  dw1_k$	Pitch Duration	0.86342	-	-
$\sum_{k=1}^n  p_{1k} - p_{2k}  dw3_k (mw_{1k} \cdot mw_{2k})$	Pitch Duration Metrical accents	0.8657	3.2	1
$\sum_{k=1}^n  p_{1k} - p_{2k}  dw4_k$	Pitch Duration	0.9281	-	-
$\sum_{k=1}^n  p_{1k} - p_{2k}  (mw_{1k} \cdot mw_{2k})$	Pitch Metrical accents	0.94742	3.32	1
$\sum_{k=1}^n  p_{1k} - p_{2k}  dw2_k$	Pitch Duration	1.1446	-	-

Results for eight variations sorted by SumAbsDiff (Variation VII only omitted). Each combination of features is run 30 times.	Features used	SumAbs Diff	Optimal metrical weight for 1st beat	Optimal metrical weight for 3rd beat
$\sum_{k=1}^n  p_{1k} - p_{2k}  dw3_k$	Pitch Duration	1.173	-	-
$\sum_{k=1}^n  p_{1k} - p_{2k} $	Pitch	1.3183	-	-

**Table 5.9: Results for the eight variation melodies sorted by the SumAbsDiff result.**

#### 5.4.9 Discussion of Results

The three best performing algorithms are the same as those found for the smaller set of variations in section 5.4 but the order within the top three is different. The optimal metrical accent weights are identical in the case of two of these three algorithms and very similar for the remaining algorithm. This indicates a number of things:

1. that the techniques proposed to allow melodies in different time signatures to be compared were successful.
2. that the optimal metrical accent weights found are appropriate and
3. that the best performing algorithms use a combination of pitch, duration (Methods 1, 4, or 2) and metrical accent weight

The use of any of the four methods of incorporating duration into the similarity calculation improved the performance over the use of pitch alone. The algorithm that used pitch and metrical accent only did not perform as well as previously. When pitch was used Duration Method 1 and no metrical accent, this algorithm performed higher than two algorithms that did use the metrical accent, which did not happen for the smaller set of variation melodies. Since there is not such a clear cut division between the performance of the algorithms that used/did not use metrical accent weights and because some of the less successful algorithms are in a different order in these results, it was decided to carry all of the algorithms forward to the evaluation stage of this research. Chapter 6 details further exploration of these algorithms with new melodies in a variety of musical styles. The fine-tuned metrical accent weights identified in this chapter are used in the evaluation stage discussed in Chapter 6. It is expected that the three top performing algorithms will use a combination of pitch, duration and metrical

accent as indicated above. The final assessment of which of these geometric algorithms are most useful in determining melodic similarity will then be carried out.

## **5.5 Fine-tuning the Edit Distance Algorithms**

It was decided not to use an optimising algorithm (such as hill-climbing) as had been used with the geometric algorithm, as discussed in Chapter 4. This was largely due to the fact that the sets of edits operations returned by the algorithms needed to be examined to ensure that they made musical sense (called the alignment by Smith et al. and Mongeau and Sankoff). At the very least, the edit operations used by the top performing algorithms should be studied.

The algorithms were run with a large range of values used for the edit costs and the weights, for example 1-20 in steps of .5. In each case, a range of values that gave good results was identified and the algorithms were run again, this time with a smaller range of appropriate numbers but greater resolution (two places of decimals).

Once an appropriate range of values for weights and parameters were identified the main steps in the evaluation process were:

1. run the algorithm multiple times with a range of weight values e.g. .01 to 2.5 in increments of .01
2. normalise the results for the theme and each variation (as in the previous chapter)
3. calculate the sum of the absolute differences between the normalised human ratings and the normalised algorithm results (as in the previous chapter)
4. identify the weight values and features that produce the best performance
5. check the trace/set of edit operations to make sure they make musical sense

### **5.5.1 Checking the Trace**

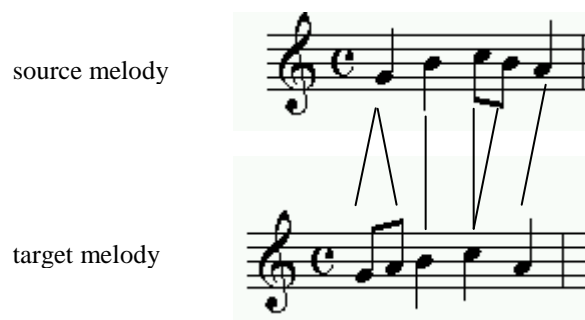
As mentioned in section 4.7, it is important to ensure that the set of edit operations produced by an algorithm are appropriate for the melodies involved. It would not be appropriate, for example, for an algorithm to indicate that all notes in the source melody should be deleted and replaced with all notes from the target melody. Choosing values for the edit operations can be seen as trying to achieve some balance between replace operations as opposed to a combination of insert and delete operations (or vice versa). If the combined cost of a deletion and insertion is less than the value for a replace, then the two individual operations will be chosen by the algorithm instead of the replace

operation. Since the Mongeau and Sankoff algorithm uses the note duration itself to calculate the cost of insert and delete operations, it is not altogether known before running the algorithm, what the relationship will be between the three basic operations (inset, delete and replace) without first carefully studying the musical material being processed.

It was decided in the context of this research to allow all traces that identified interspersed edit operations and to reject any that solely identified blocks of operations as shown in the examples below.

Good traces	R R R I R R R R I R R I I R I I R R R D R R R I D R I R R I R R R D R R
Bad traces	I I I I D D D D D R R R R R I I I I I

The first example of the bad trace means that in order to transform a source melody with five notes into a target melody of the same length, the first five notes from the target were inserted into the source melody and the original five notes of the source melody were then deleted. Figure 5.7 below shows two sample five-note melodies. The vertical lines illustrate the similar notes present in both melodies. If the bad trace above referred to these melodies it would mean that instead of the close relationship between the notes being recognised, the edit operations used would result in the insertion of all the notes of the target before the first note of the source melody, and then all the original notes of the source melody being deleted. Clearly, this is not desirable and therefore traces such as these are considered “bad” in the context of this research and this is not an appropriate set of edit operations.



**Figure 5.7: An example source and target melody with one possibility for the alignment of notes.**

Tracing back through the matrix from the bottom right to top left cell involves subtracting the cost of each edit operation from the current cell in order to determine which operation was used to fill that cell. A move to the relevant cell is made and the process is repeated until the beginning cell, the top left cell, is reached. Horizontal movements in the matrix represent insertions, vertical movements represent deletions and diagonal movements represent replacements of notes. A replacement of cost zero represents two notes that are regarded as the same notes in the context of the algorithm. This may depend on whether both pitch and duration of the notes are taken into consideration. Here, in tracing the edit operations an order of preference has been assigned so that if all three of the cells above, to the left, and above diagonal left are indicated as possible moves, the replace operation (signalled by a diagonal left move) is carried out.

## **5.6 Results for Smith et al.'s Edit Distance Algorithms**

The edit distance algorithms listed in Table 4.3 of section 4.8 are all used to calculate the similarity of the variations of the testbed melodies to the Theme and the results are compared to the human similarity judgements presented in Chapter 3. The algorithms are first run on the six variations that have the same time signature as the Theme. They are later run on the larger set of eight variation melodies and both sets of results are compared with the results from the geometric algorithms given in the previous section.

The best performing version of each of Smith et al.'s algorithms is shown below in Table 5.10, along with the weight and cost values that produce the result. A similar layout is used to present the results of all of the edit distance algorithms implemented. The result of the comparison with the human similarity judgements is reported using the SumAbsDiff value (as used with the geometric algorithm in the previous chapter). The trace column in each table reports that the trace of the edit operations has been examined for those values. G represents a good or appropriate trace and B represents a bad trace. A brief description of the algorithm is also given for clarity and is a shortened version of the descriptions given in the previous chapter.

Insert Cost	Delete Cost	Replace weight	SumAbsDiff	Trace	Description (I = insert cost, D = delete cost)
0.01	0.01	0	0.636	G	I = D. No duration used in replace cost. 50 values of I/D (.01 to .5) all give the exact same result.
0.5	0.5	0.01	0.643	G	I = D. Pitch and duration used in replace cost.
0.29	1.07	0	0.389	G	I ≠ D. No duration used in replace cost.
0.06	1.89	0.24	0.276	G	I ≠ D. Pitch and duration used in replace cost.

**Table 5.10: The results for the versions of Smith et al.'s (1998) edit distance algorithms.**

### Summary of results

The best performing of Smith et al.'s algorithms implemented has an insert cost of 0.06, a delete cost of 1.89 and a duration weight of 0.24. This indicates that both pitch and duration are taken into account when calculating the replace cost but that the duration difference between notes contributes significantly less than the pitch difference. When duration was used along with pitch to calculate the replace cost it gave better results than when pitch alone was used, although where insert and delete were assigned the same values, this difference was very small. Assigning the same value to the insert and delete edit operations gave a worse result than when different results were used for each.

## 5.7 Results of Mongeau and Sankoff's Edit Distance Algorithms

The values reported for the results of the Mongeau and Sankoff algorithms have a slightly different meaning to those reported in the previous Smith algorithm. In Mongeau and Sankoff's algorithm (see equations 4.16 and 4.17, section 4.6), the insert and delete costs are calculated by multiplying the note duration by a weight value. Then, as in the case of Smith et al.'s algorithm, a further weight is used to control the relative contribution of duration to pitch when calculating the replace cost. It should be noted that Mongeau and Sankoff use the same weight value in all three places and that although this is replicated here, separate weight values are also implemented. The actual version of the algorithm that was implemented by Mongeau and Sankoff (1990) included a weight based on the pitch consonance of the interval between the notes and used fragmentation and consolidation as well as the basic insert and delete operations. While the results for this version of the algorithm is given in section 5.7.6, the following

section details the results and optimal values found for all of the versions of Mongeau and Sankoff's edit distance algorithm listed in section 4.8.

### 5.7.1 Mongeau and Sankoff – pitch difference without fragmentation and consolidation

Insert Weight	Delete Weight	Replace weight	SumAbsDiff	Trace	Description (I = insert weight, D = delete weight, R = replace weight)
0.01	0.01	0.01	0.417	G	I = D = R.
0.5	0.5	0	0.412	G	I = D. No duration used in replace cost.
0.23	0.23	.08	0.292	G	I = D. Duration used in replace cost.
0.32	0.15	0	0.296	G	I ≠ D. No duration used in replace cost.
0.01	0.46	0.29	0.279	G	I ≠ D ≠ R. Duration used in replace cost. All 3 weights are independent.

**Table 5.11: The results for the pitch difference versions of Mongeau and Sankoff's (1990) edit distance algorithms.**

#### Summary of results

The algorithm performs best when duration is used in the replace cost and when the three weights are assigned separate values (last row of Table 5.11 above). The weight that controls the contribution of the duration difference in calculating the replace cost (.29) is similar to the best results for Smith et al.'s algorithm (.24). When each weight is assigned the same value (as Mongeau and Sankoff did) this results in the worst performing algorithm (see the top row of results in Table 5.11). It is noted that since Mongeau and Sankoff used fragmentation/consolidation and the pitch consonance weights in their implementation these results are not directly comparable. In both cases where duration is used to calculate the replace cost (I = D and I ≠ D), it improves on those results that use pitch only in the replace cost, although the difference in the SumAbsDiff result is very small in the latter case.

### 5.7.2 Mongeau and Sankoff – pitch consonance without fragmentation and consolidation.

These implementations of the algorithm use Mongeau and Sankoff's pitch consonance weights rather than a simple pitch difference as the pitch component of the replace cost. Tables 4.1 and 4.2 in section 4.6 give details of the weight values based on the intervals between the two notes being compared.

<b>Insert Weight</b>	<b>Delete Weight</b>	<b>Replace weight</b>	<b>SumAbsDiff</b>	<b>Trace</b>	<b>Description (I = insert weight, D = delete weight, R = replace weight)</b>
0.01	0.01	0.01	0.724	G	I = D = R.
0.04	0.04	0	0.737	G	I = D. No duration used in replace cost.
0.02	0.02	0.01	0.655	G	I = D. Duration used in replace cost.
0.02	0.01	0	0.627	G	I ≠ D. No duration used in replace cost.
0.01	0.02	0.01	0.621	G	I ≠ D ≠ R. Duration used in replace cost. All 3 weights are independent.

**Table 5.12: The results for the pitch consonance versions of Mongeau and Sankoff's (1990) edit distance algorithms.**

### **Summary of results**

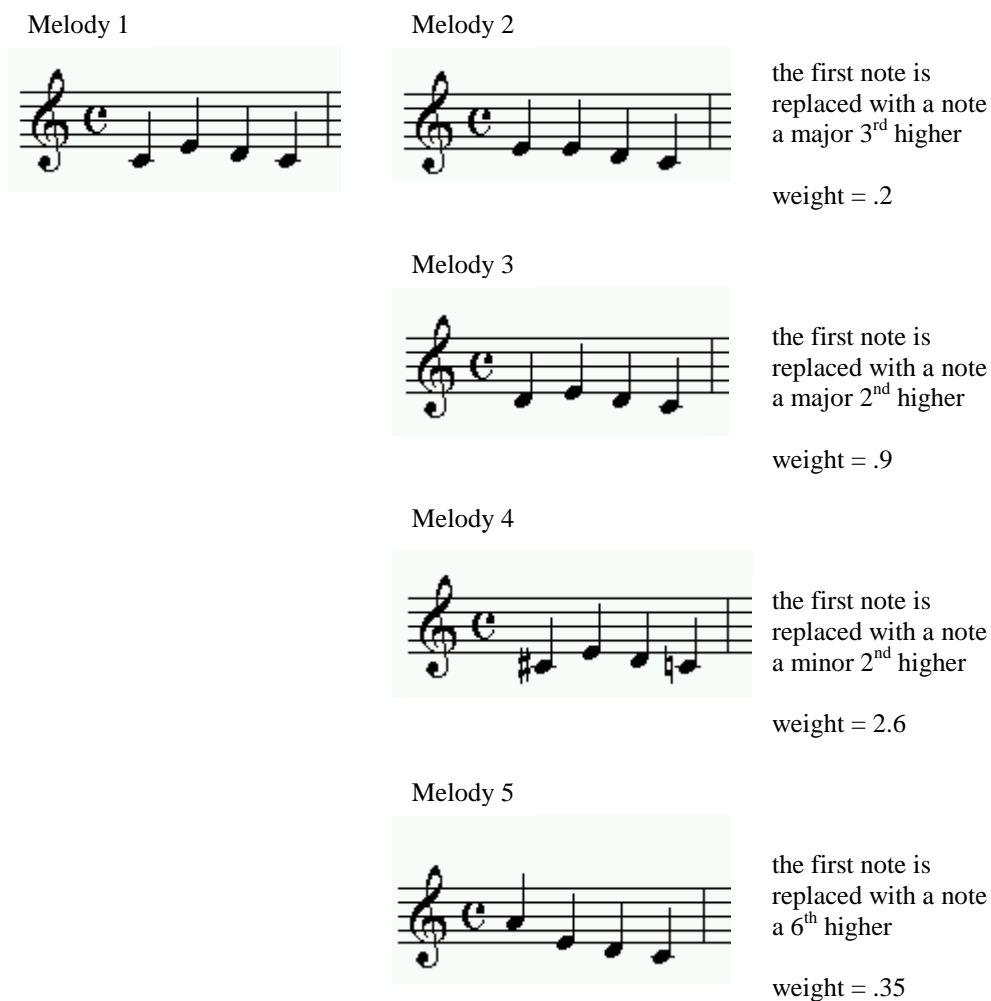
These algorithms all perform worse than the previous implementations of Mongeau and Sankoff's algorithms that used the basic pitch difference between notes to calculate the replace cost. When the insert, delete and replace weights are assigned the same value this results in the worst performing algorithm. This is significant as this version of the algorithm is very similar to Mongeau and Sankoff's implementation (without the use of fragmentation and consolidation) and is discussed in more detail below. The insert, delete and replace weights are all very small values, with a highest reported weight value of .04. Replacing the pitch difference with these pitch consonance weights results in a decrease in the replace cost. The lower replace costs have a knock-on effect on the values of the insert and delete weights as for insertions and deletions to be chosen over replacements by the algorithms, their combined values would have to be less than the cost of replacing the note.

### **5.7.3 Discussion of the Pitch Consonance Results.**

Mongeau and Sankoff's algorithm uses a table of pre-defined weights based on the interval between the note to be replaced and the note that replaces it. On first consideration, this may seem like an appropriate idea as it uses weights that reflect the consonance/dissonance of these intervals according to general music perception principles (see tables 12 and 13 in section 4.6). Intervals of a fifth, for example, are considered more consonant than a second, and are therefore assigned a smaller weight value. The authors do not refer to any source for the degree of consonance and

associated weight values. The assumption their approach makes is that replacing a note with one that has a consonant interval relationship with it makes the source melody more similar than if it was replaced with a note that formed a dissonant interval. Figure 5.8 below includes a sample source melody (Melody 1) and four potential target melodies (2-5). Each of the target melodies differs from the source melody by one note only. The pitch of the first note in the target melodies is different but the duration is the same so that it is clear to see the effect of the difference in pitch. The weight assigned to the interval between the first note of the source and target melody is given in each case. The question considered here is whether the weight values used by Mongeau and Snakoff (1990) reflect the perceptual difference between these source and target melodies.

If one examines the melodies that have diatonic pitch changes (within the key of the source and target melody), Melody 3 has by far the highest pitch weight (.9) and this would result in this particular melody being identified as the least similar of the three diatonic melodies. There may be some merit in the idea of considering the replacement of a note with a note a 3<sup>rd</sup> or a 5<sup>th</sup> higher as being quite similar and a similar harmonic scoring system was proposed by Selfridge-Field (2004). However, the system of weights used by Mongeau and Sankoff are questionable in the results they provide. The difference in pitch of a perfect 6<sup>th</sup> between the first notes of melodies 1 and 5 would result in the Melody 5 being considered far more similar to Melody 1 than Melody 3. This is despite the fact that the note in this melody is closer in pitch to the note in the target melody and is just a diatonic step from the next note, whereas the first note of Melody 5 is further away in pitch and involves a leap of a fourth to the following note.



**Figure 5.8: Five sample melodies, illustrating Mongeau and Sankoff's pitch consonance weights.**

The manner in which Mongeau and Sankoff divide their system of pitch consonance weights into two categories, depending on whether the notes to be compared are in the same mode, is an interesting idea and one that may prove useful. Another related approach would be to use the actual pitch difference (as used by Smith et al.) between the replaced and replacing note if both notes are in the same mode but to use a separate set of weights for accidental notes. In this way Melody 2, 3 (and possibly 5) would be considered more similar to Melody 1 than Melody 4. The presence of accidental/chromatic notes would be weighted higher than diatonic notes of the scale with the implication that they are considered perceptually less similar. Such an approach would involve the investigation and use of key-finding algorithms, which is outside of the scope of this project.

#### 5.7.4 Mongeau and Sankoff - pitch difference with fragmentation and consolidation.

Insert Weight	Delete Weight	Replace weight	SumAbsDiff	Description (I = insert weight, D = delete weight, R = replace weight)
0.01	0.01	0.01	0.340	I = D = R.
0.52	0.52	0	0.515	I = D. No duration used in replace cost.
0.23	0.23	0.08	0.515	I = D. Duration used in replace cost.
0.24	0.01	0	0.303	I ≠ D. No duration used in replace cost.
0.22	0.02	0.01	0.304	I ≠ D ≠ R. Duration used in replace cost. All 3 weights are independent.

**Table 5.13: The results for the pitch difference with fragmentation/consolidation versions of Mongeau and Sankoff's (1990) edit distance algorithms.**

#### Summary of results

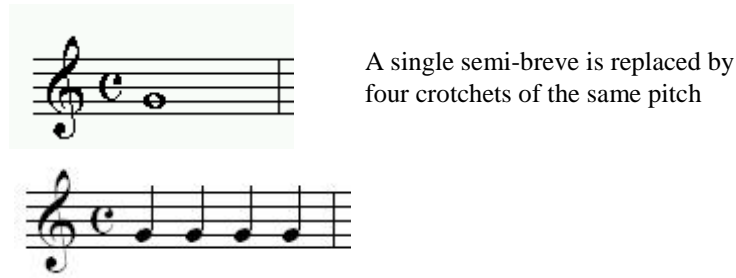
The optimal weight values identified for the first three algorithms listed in Table 5.13 are the same or very similar to those reported for the pitch difference versions of the algorithm that did not use fragmentation and consolidation. The remaining two algorithms identify different values for the weights used. When all three weights are assigned the same value the result is better than when fragmentation and consolidation was not used but no improvement is noted for other versions of the algorithm. The best performing of all the pitch difference versions of the algorithm implemented was the version that used pitch difference but not fragmentation and consolidation and that assigned different values to the weights for each of the edit operation costs.

#### 5.7.5 Discussion of the fragmentation and consolidation results

Mongeau and Sankoff's use of fragmentation and consolidation is an interesting extension to the basic edit distance algorithm. A number of questions and issues arise from the expected and actual results. The first issue worth noting is that it is expected that fragmentation only and not consolidation will be used by the algorithms because of the nature of the variation melodies used in the testbed. There are instances of long notes being embellished to form rhythmic repetitions of shorter notes of the same pitch but consolidations do not occur in this set of melodies and would generally not be expected to occur in music in Theme and Variation style.

Although the example used by Mongeau and Sankoff to explain the concept of fragmentation involved replacing a long note with multiple notes of the same pitch (see

Figure 5.9 below), the way in which they implement this feature actually allows for a broader definition of fragmentation. Any repeated notes of the same pitch in one



**Figure 5.9: Mongeau and Sankoff’s example of fragmentation**

melody may be compared with a longer note of a different pitch in the other melody when filling the matrix cells with values. Among the fragmentations we would expect to be identified by the algorithm include the following excerpts, which are fragmentations of a quarter-note (Variations III and IX) of pitch G (see Figure 5.10 below). The fragmentations are marked with brackets.



**Figure 5.10: Examples of the fragmentation of a quarter-note from the Theme.**

However, the implementation of the algorithm may also identify the following as a fragmentation (see Figure 5.11), although it is not likely that a person would identify it as such.



**Figure 5.11: An example of a questionable fragmentation identified by the algorithm.**

A similar instance of fragmentation occurs in Variation VIII, involving notes an octave apart (see Figure 5.12). This should not unduly affect the result since the combinations of values that give the closest results to those of the human similarity judgements will be identified and any fragmentations such as these that don’t make musical sense should be discarded in favour of edit operations that produce an appropriate overall result.



**Figure 5.12: Example of fragmentation from bar 4 of Variation VIII.**

A further question arises as to why the SumAbsDiff results for the first three versions of this algorithm reported here are different despite the fact that the same or very similar optimal weight values are identified. The instance of fragmentation mentioned above in Variation III does change the results slightly for that algorithm but a fragmentation in Variation IX also results in some differences. The instance of fragmentation in this variation isn't very musically significant and does not change the edit distance of that variation too much. The change to the overall results is significant though because of the normalisation process used to compare the algorithm and human similarity judgements on the same 0 to 1 scale. A min-max normalisation is used (see section 5.2) and since Variation IX tends to give the highest edit distance value (i.e. the max), this affects all the normalised values by a small amount.

The effects of fragmentation are expected to be more pronounced when the algorithm is run on the larger set of eight variations since both Variation IV and VI omitted here feature rhythmic passage of repeated short notes based on longer notes in the Theme melody.

#### **5.7.6 Mongeau and Sankoff - pitch consonance with fragmentation and consolidation.**

The results shown in the first two rows (identical results) of Table 5.14 are for the actual algorithm implemented by Mongeau and Sankoff, in which all three weights were assigned the same value, the pitch consonance weight system is used and fragmentation and consolidation is also incorporated (see section 4.6 for Mongeau and Sankoff's (1990) algorithm).

Insert Weight	Delete Weight	Replace weight	SumAbsDiff	Description (I = insert weight, D = delete weight, R = replace weight)
0.01	0.01	0.01	0.647	$I = D = R^4$
0.02	0.02	0.02	0.647	
0.04	0.04	0	0.686	I = D. No duration used in replace cost
0.02	0.02	0.01	0.644	I = D. Duration used in replace cost
0.03	0.01	0	0.636	$I \neq D$ . No duration used in replace cost
0.04	0.01	0.01	0.641	$I \neq D \neq R$ . Duration used in replace cost. All 3 weights are independent

**Table 5.14: The results for the pitch consonance with fragmentation/consolidation versions of Mongeau and Sankoff's (1990) edit distance algorithms.**

### Summary of results

Since the results for the pitch consonance weights without the use of fragmentation/consolidation were not as good as when the actual pitch difference value was used, these results were not expected to be very encouraging either. Some improvement on the version of the algorithm that uses pitch consonance without fragmentation/consolidation is reported in the case of the first three versions of the algorithm listed here, while the last two versions of the algorithm shown in Table 5.14 produce slightly worse results. The optimal weights reported are the same or very similar in most cases with the biggest difference being an increase in the insert weight in the last algorithm from .01 (for no fragmentation/consolidation) to .04 here. The best results for all of the pitch consonance algorithms is the version that uses different values for the weights in calculating the three edit operation costs and does not use fragmentation and consolidation. However, the SumAbsDiff results for all of the pitch consonance algorithms (including the algorithms that use fragmentation and consolidation) are worse than those found when using pitch difference for the replace cost (with or without fragmentation).

---

<sup>4</sup> Two different sets of weights produced the same SumAbsDiff result and so both are included here.

## 5.8 Combined Results for the Geometric and Edit Distance Algorithms.

The results for Smith et al.'s and Mongeau and Sankoff's edit distance algorithm for the set of six variations have been discussed and a summary of the results provided with further discussion of issues encountered where appropriate. An overall comparison of all algorithms (including the versions of the geometric algorithm) is presented in Table 5.15 below.

Six variations - I, II, III, V, VIII, IX Smith/M&S = edit distance algorithms Equations = Ó Maidín's geometric algorithms	Features used	SumAbs Diff
$\sum_{k=1}^n  p_{1k} - p_{2k}  dw1_k(mw_{1k} \cdot mw_{2k})$	Pitch Duration Metrical accents: 1.82 1.32	0.1799
$\sum_{k=1}^n  p_{1k} - p_{2k}  dw4_k(mw_{1k} \cdot mw_{2k})$	Pitch Duration Metrical accents: 1.98 1	0.2487
Smith	I 0.06 D 1.89 Dur .24	0.276
M&S pitch diff	I .01 D .46 Dur .29	0.279
M&S pitch diff. I = D.	I .23 D .23 Dur .08	0.292
M&S pitch diff. No duration in replace cost	I .32 D .15 Dur 0	0.296
M&S pitch diff, fragmentation/consolidation. No duration used in replace cost	I 0.24 D 0.01 Dur 0	.303
M&S pitch diff, fragmentation/consolidation	I 0.22 D 0.02 Dur 0.1	.304
M&S pitch diff, fragmentation/consolidation. I=D	I 0.23 D 0.23 Dur 0.08	.310
$\sum_{k=1}^n  p_{1k} - p_{2k}  dw2_k(mw_{1k} \cdot mw_{2k})$	Pitch Duration Metrical accents: 1.67 1.13	0.3173
M&S pitch diff, fragmentation/consolidation. All values equal.	I 0.01 D 0.01 Dur 0.01	0.340

Six variations - I, II, III, V, VIII, IX Smith/M&S = edit distance algorithms Equations = Ó Maidín's geometric algorithms	Features used	SumAbs Diff
Smith. No duration in replace cost	I .29 D 1.07 Dur 0	0.389
M&S pitch diff. I = D. No duration in replace cost	I .5 D .5 Dur 0	0.412
M&S pitch diff. All values equal.	I .01 D .01 Dur .01	0.417
$\sum_{k=1}^n  p_{1k} - p_{2k}  (mw_{1k} \cdot mw_{2k})$	Pitch Metrical accents: 3.32 1	0.4209
$\sum_{k=1}^n  p_{1k} - p_{2k}  dw_{3k} (mw_{1k} \cdot mw_{2k})$	Pitch Duration Metrical accents: 3.2 1	0.457
$\sum_{k=1}^n  p_{1k} - p_{2k}  dw_{1k}$	Pitch Duration	0.4672
M&S pitch diff, fragmentation/consolidation, I=D. No duration in replace cost	I 0.5 D 0.5 Dur 0	.515
$\sum_{k=1}^n  p_{1k} - p_{2k}  dw_{4k}$	Pitch Duration	0.5631
$\sum_{k=1}^n  p_{1k} - p_{2k}  dw_{3k}$	Pitch Duration	0.6062
M&S pitch consonance.	I .01 D .02 Dur .01	0.6218
M&S pitch consonance No duration in replace cost	I .02 D .01 Dur 0	0.6272
$\sum_{k=1}^n  p_{1k} - p_{2k} $	Pitch	0.6295
Smith. I = D. No duration in replace cost	I .01 D .01 Dur 0	0.636
M&S pitch consonance, fragmentation/consolidation. No duration in replace cost	I 0.03 D 0.01 Dur 0	0.636
M&S pitch consonance, fragmentation/consolidation. All 3 weights varied independently.	I 0.03 D 0.01 Dur 0	0.636

<b>Six variations - I, II, III, V, VIII, IX Smith/M&amp;S = edit distance algorithms Equations = Ó Maidín's geometric algorithms</b>	<b>Features used</b>	<b>SumAbs Diff</b>
Smith. I = D	I .5 D .5 Dur .01	0.643
M&S pitch consonance, fragmentation/consolidation. I = D. Variation varied independently	I 0.02 D 0.02 Dur 0.01	0.6437
<b>M&amp;S pitch consonance, fragmentation/consolidation. All values equal</b>	<b>I .01 D .01 Dur .01</b>	<b>0.6474</b>
M&S pitch consonance. I = D.	I .02 D .02 Dur .01	0.6552
$\sum_{k=1}^n  p_{1k} - p_{2k}  dw_{2k}$	Pitch Duration	0.6635
M&S pitch consonance, fragmentation/consolidation. I=D. No duration in replace cost	I 0.04 D 0.04 Dur 0	.686
M&S pitch consonance. All values equal	I .01 D .01 Dur .01	0.7244
M&S pitch consonance. I = D. No duration in replace cost	I .04 D .04 Dur 0	0.7372

**Table 5.15: The combined results for the geometric and edit distance algorithms for the six Part A 4/4 variation melodies, sorted by the SumAbsDiff result.**

### **Summary of overall results**

Two versions of the geometric algorithm that use pitch, duration and metrical accent weights perform better than any of the other algorithms. The third best performing algorithm is Smith et al.'s (1998) edit distance algorithm in which both costs are independent of each other and duration is used. The best performing version of Mongeau and Sankoff's (1990) algorithm uses pitch difference instead of their pitch consonance weights, incorporates duration and assigns a separate weight to all three edit operation costs. In the case of all of the algorithms, the inclusion of duration resulted in a better performance when duration was not used. This supports the findings of the perceptual research presented in Chapter 2 which suggested that duration along with pitch would be useful for similarity and that metrical accents may also be relevant.

In Smith et al.'s implementation the insert and delete costs were assigned the same values and the contribution of the duration difference in calculating the replace cost was half that of the pitch difference. Here, the algorithm performed better when the duration

difference was weighted by .24 instead of .5. As previously mentioned, the best performing version of Mongeau and Sankoff's (1990) algorithm used three independent weight values rather than using the same weight value (.348) for calculating all edit operations as in their implementation. The value of the weight applied to the duration difference between the notes is similar to the weight found for the previously mentioned Smith et al. (1998) algorithm at .29. The remaining algorithmic values are not directly comparable since the values represent the actual costs in the case of the Smith et al. algorithms but weights applied to the note durations in the case of Mongeau and Sankoff (1990) as discussed in sections 4.5 and 4.6. All of the versions of Mongeau and Sankoff's algorithms that used the pitch consonance weights instead of the basic pitch difference gave results that appear in the lower half of the table.

### 5.9 Results when Variations IV and VI are Included

The combined results for both the edit distance and geometric algorithms for this set of eight variations are shown below in Table 5.16. The algorithm highlighted in bold in the table is the actual one implemented by Mongeau and Sankoff. The traces of all edit distance algorithms except those that use fragmentation/consolidation have been checked.

<b>Eight Variations (Variation VII omitted) Smith/M&amp;S = edit distance algorithms Equations = Ó Maidín's geometric algorithms</b>	<b>Features used</b>	<b>SumAbs Diff</b>
M&S pitch diff, fragmentation/consolidation	I .01 D .43 Dur .29	0.3357
Smith	I .06 D 1.89 Dur .24	0.45588
M&S, pitch diff	I .01 D .21 Dur .22	0.484931
$\sum_{k=1}^n  p_{1k} - p_{2k}  dw4_k(mw_{1k} \cdot mw_{2k})$	Pitch Duration Metrical accents: 1.98 1	0.48603
$\sum_{k=1}^n  p_{1k} - p_{2k}  dw1_k(mw_{1k} \cdot mw_{2k})$	Pitch Duration Metrical accents: 1.82 1.32	0.49485
$\sum_{k=1}^n  p_{1k} - p_{2k}  dw2_k(mw_{1k} \cdot mw_{2k})$	Pitch Duration Metrical accents:	0.66835

<b>Eight Variations (Variation VII omitted) Smith/M&amp;S = edit distance algorithms Equations = Ó Maidín's geometric algorithms</b>	<b>Features used</b>	<b>SumAbs Diff</b>
	1.65 1.14	
M&S, pitch diff, fragmentation/consolidation, I=D	I .03 D .03 Dur .05	0.686242
M&S, pitch diff, fragmentation/consolidation, all values equal	I .01 D .01 Dur .01	0.720092
<b>M&amp;S, pitch consonance, fragmentation/consolidation, all values equal</b>	I .01 D .01 Dur .01	0.745733
M&S, pitch consonance, fragmentation/consolidation, I=D, no duration	I .28 D .28 Dur 1	0.751663
M&S, pitch consonance, fragmentation/consolidation	I .01 D .03 Dur .08	0.762281
M&S, pitch consonance	I .01 D .04 Dur .04	0.813973
Smith, no duration	I .29 D 1.71 Dur 0	0.8438
$\sum_{k=1}^n  p_{1k} - p_{2k}  dw1_k (mw_{1k} \cdot mw_{2k})$	Pitch Duration	0.86342
$\sum_{k=1}^n  p_{1k} - p_{2k}  dw3_k (mw_{1k} \cdot mw_{2k})$	Pitch Duration Metrical accents: 3.2 1	0.8657
M&S, pitch diff, I=D	I .05 D .05 Dur .04	0.92205
$\sum_{k=1}^n  p_{1k} - p_{2k}  dw4_k$	Pitch Duration	0.9281
M&S, pitch diff, all values equal	I .01 D .01 Dur .01	0.936441
$difference = \sum_{k=1}^n  p_{1k} - p_{2k}  (mw_{1k} \cdot mw_{2k})$	Pitch Metrical accents: 3.32 1	0.94742
M&S, pitch consonance, I=D	I .24 D .24 Dur .39	1.01762
M&S pitch consonance, all values equal	I .33 D .33 Dur .33	1.07858

<b>Eight Variations (Variation VII omitted) Smith/M&amp;S = edit distance algorithms Equations = Ó Maidín's geometric algorithms</b>	<b>Features used</b>	<b>SumAbs Diff</b>
Smith, I=D (bad trace on top results – look at later)	I .31 D .31 Dur .2	1.08119
M&S, pitch diff, fragmentation/consolidation, no duration	I .27 D .24 Dur 0	1.11504
Smith, I=D, no duration	I .01 D .01 Dur 0	1.13636
M&S, pitch diff, no duration	I .34 D .4 Dur 0	1.14077
$\sum_{k=1}^n  p_{1k} - p_{2k}  dw2_k$	Pitch Duration	1.1446
$\sum_{k=1}^n  p_{1k} - p_{2k}  dw3_k$	Pitch Duration	1.173
M&S, pitch diff, I=D, no duration	I 0.34 D 0.34 Dur 0	1.18607
M&S, pitch consonance, fragmentation/consolidation, no duration	I .15 D .01 Dur 0	1.22009
M&S, pitch consonance, fragmentation/consolidation, I=D, no duration	I .15 D .15 Dur 0	1.22009
$\sum_{k=1}^n  p_{1k} - p_{2k} $	Pitch	1.3183
M&S, pitch consonance, no duration	I .02 D .01 Dur 0	1.32143
M&S, pitch consonance, I=D, no duration	I .01 D .01 Dur 0	1.35
M&S, pitch diff, fragmentation/consolidation, I=D,no duration	I .26 D .26 Dur 0	1.7769

**Table 5.16: The combined results for the geometric and edit distance algorithms for the eight Part A variation melodies, sorted by the SumAbsDiff result.**

### Summary of the Results

Many of the Mongeau and Sankoff (1990) algorithms that used fragmentation/consolidation and duration performed higher here than for the smaller set

of six variations. The best performing algorithm uses pitch difference and duration to calculate the replace cost and incorporates fragmentation and consolidation. As noted in section 5.7.5, a number of rhythmic fragmentations are found in Variations IV and VI. When these variations melodies are included the performance of many of the algorithms that use fragmentation and consolidation improve, suggesting that this component of Mongeau and Sankoff's algorithm is useful when notes are rhythmically broken down or merged in this way. However, these algorithms did not perform as well in the previous results for the set of six variations where there was little evidence of fragmentation and consolidation and so such algorithms may not be appropriate for general application unless these sort of rhythmic features are likely to be present.

Many versions of the Mongeau and Sankoff algorithms that use pitch difference (rather than the weights based on pitch consonance) in the replace cost performed reasonably well in the previous set of results but show a clearly noticeable decline in performance in this case. It was noted previously that the Mongeau and Sankoff edit distance algorithms that use weights based on pitch consonance rather than simply pitch difference in the replace cost all appeared in the bottom half of the results presented in section 5.7. Here, it can be seen that while these algorithms are not top of the results list, they have clearly performed better on this larger set of eight variations. Variation IV contains a number of arpeggio figures centred on the notes of the theme, with a number of intervals of a third a sixth. Variation VI contains a number of pitches an interval of a third away from the notes of the theme. In both cases, the algorithms that use weights in the replace cost based the consonance of the intervals react favourable to the addition of these new melodies and they perform better than the algorithms that use pitch difference.

As discussed in section 5.4, in the case of the geometric algorithm, the top three algorithms and weight values were the same (or very similar) before and after Variations IV and VI in 6/8 and 3/4 time respectively were included. (Variation VI also involved a stretching of the length of the melody by a factor of two.) This suggests that the weights arrived at by the fine-tuning process do indeed represent some sort of optimal weights for these algorithms and that they may generalise well to other music. Most of the edit distance algorithms, on the other hand, return cost and weight values that are different to those identified when the algorithms were used with the smaller set

of melodies that were all in the same time signature. One of the exceptions is the version of Smith et al.'s (1998) algorithm that uses independent insert and delete costs and duration as well as pitch to calculate the replace cost. This algorithm is one of the top two performing edit distance algorithms in each case and the exact same values are identified when the fine-tuning is carried out using the melodies that are all in the same time signature and when the two variations in different time signatures are introduced. It is noted that the top three edit distance algorithms share a similar duration weight, with two versions of Mongeau and Sankoff's (1990) algorithm returning a similar weight as the previously mentioned version of Smith et al.'s algorithm. This suggests that a value in this range (.22 to .29) may be an optimal value.

The previous set of results for the smaller set of variations showed that two versions of the geometric algorithm performed better than all of the edit distance algorithms. The inclusion of Variations IV and VI resulted in two versions of the edit distance algorithms clearly performing better than all of the geometric algorithms (the opposite result to the previous one). Although three edit distance algorithms are actually listed above the geometric algorithms in Table 5.16, the third algorithm (Mongeau and Sankoff, pitch difference) and the top two geometric algorithms all have very similar SumAbsDiff values, ranging from .4849 to .4948, and hence could be considered to be equal in terms of performance.

In summary, the geometric algorithms showed consistency in the weight values identified and in the order of best performing algorithms, while there were a number of differences in the performance of the edit distance algorithms when two additional melodies were included in the testbed. The versions of the algorithms implemented by Smith et al. (1998) and Mongeau and Sankoff (1990) did not perform very well. The performance of algorithms that used Mongeau and Sankoff's fragmentation/consolidation technique and their pitch consonance weights did improve when consonant intervals and examples of fragmentation were present in the melodies but they showed relatively poor results generally. The optimal weight values for the duration difference in the top performing edit distance algorithms were very similar for this set of eight variations and the smaller set of six variations previously discussed indicating some consistency in the results for this algorithmic approach. The results are

explored further by running the algorithms and tuned weights on a further set of related melodies before being applied to a broader range of melodies in Chapter 6.

## 5.10 Verification of the Algorithmic Results Using the Part B Melodies from the Testbed

The listening experiment that was run to collect the human measures of similarity featured two distinct parts. Part A was concerned with the first four bars of each melody in the Duschenes Theme and Variations (1962), while Part B involved the second four bars of each melody (see section 3.4.8 of Chapter 3 for details). The similarity ratings from Part A were used to find the best weights for each algorithm, as detailed in chapter 5 but this is the first time the similarity ratings from Part B have been used in this research. The set of melodies that constituted Part B of the listening experiment are used here and are shown in Appendix A.

### 5.10.1 Assessing the Performance of the Algorithms

The median similarity ratings given by subjects in the listening experiment for these melodies is listed in Table 3.11 in section 3.5.7 of Chapter 3. The order of similar melodies returned by both the algorithm and human similarity judgements are used to evaluate the success of the algorithms and fine-tuned internal weights from the previous two chapters. The difference between the ranks is calculated as shown below and the sum of the difference in ranks taken as an overall indication of how close the results from the various algorithms are to the human similarity judgements. An example is shown in Table 5.17. (Melodies that had the same median rating were assigned the same rank order and so Variations V and IX are both ranked 4<sup>th</sup> in the example shown below.)

Comparison of ranks			
Variation	Ranked algorithm result	Ranked human judgement	Absolute difference between ranks
Var. I	1	1	0
Var. II	2	3	1
Var. III	3	2	1
Var. V	5	4	1
Var. VIII	6	6	0
Var. IX	4	4	0
<b>Sum of the difference between ranks</b>			<b>3</b>

Table 5.17: The sum of the difference in ranks.

However, it is noted that the use of ranks in this way somewhat reduces the sensitivity of the algorithmic measure used. The same level of similarity is not necessarily present between the melodies ranked first and second as those ranked second and third. It is possible that some sequentially ranked melodies will have very close results to each other, while others may have a much bigger difference between their results. The normalised algorithmic results for the best performing geometric algorithm are shown below in Table 5.18.

Variation	Actual algorithm results (normalised)	Ranked algorithm result
Var. I	0	1
Var. II	0.06247	3
Var. III	0.01775	2
Var. V	0.2586	5
Var. VIII	1	6
Var. IX	0.2307	4

**Table 5.18: An example of the actual algorithm results before the ranks are calculated.**

Here, it is clear to see that Variation I is considered the most similar melody to the Theme and Variation IX the most dissimilar. The difference between sequentially ranked variations ranges from .01 to .74. It is clear that some of the information related to the degree of similarity of the melodies is lost when the data is converted to ranked data. Therefore, the comparison with the ranked human judgments through the calculation of the sum of the difference between both sets of ranked data should be seen as an approximate indicator of the performance of the algorithms and may need some interpretation.

There was a marked difference in the performance of the geometric and edit distance algorithms when assessed in this way and so the results are presented and discussed separately here.

### **5.10.2 The Geometric Algorithms**

In the case of eight out of the ten versions of the geometric algorithms run, the only difference between the ranked order of the algorithm measures and the human similarity judgements is that the algorithms consider Variation IX to be slightly less similar (or more similar in some cases) to the query than Variation V, and so they are awarded sequential rather than equal ranks, as shown in Table 5.19 below.

Comparison of ranks			
Variation	Ranked algorithm result	Ranked human judgement	Absolute difference between ranks
Var. I	1	1	0
Var. II	3	3	0
Var. III	2	2	0
Var. V	5	4	1
Var. VIII	6	6	0
Var. IX	4	4	0
Sum of the difference between ranks			1

**Table 5.19: The comparison of the ranks between the human and algorithmic similarity measures.**

The results for all of the geometric algorithms are included in the table below and are presented in the order of the best performing algorithms identified for the Part A melodies in section 5.4.

Results for Part B melodies. Six variations - I, II, III, V, VIII, IX	Features used	Sum of the difference in ranks between the algorithm and human judgements
$\sum_{k=1}^n  p_{1k} - p_{2k}  dw1_k (mw_{1k} \cdot mw_{2k})$	Pitch Duration Metrical accents: 1.82 1.32	1
$\sum_{k=1}^n  p_{1k} - p_{2k}  dw4_k (mw_{1k} \cdot mw_{2k})$	Pitch Duration Metrical accents: 1.98 1	1
$\sum_{k=1}^n  p_{1k} - p_{2k}  dw2_k (mw_{1k} \cdot mw_{2k})$	Pitch Duration Metrical accents: 1.67 1.13	1
$\sum_{k=1}^n  p_{1k} - p_{2k}  (mw_{1k} \cdot mw_{2k})$	Pitch Metrical accents 3.32 1	1

Results for Part B melodies. Six variations - I, II, III, V, VIII, IX	Features used	Sum of the difference in ranks between the algorithm and human judgements
$\sum_{k=1}^n  p_{1k} - p_{2k}  dw3_k (mw_{1k} \cdot mw_{2k})$	Pitch Duration Metrical accents: 3.2 1	1
$\sum_{k=1}^n  p_{1k} - p_{2k}  dw1_k$	Pitch Duration	1
$\sum_{k=1}^n  p_{1k} - p_{2k}  dw4_k$	Pitch Duration	1
$\sum_{k=1}^n  p_{1k} - p_{2k}  dw3_k$	Pitch Duration	3
$\sum_{k=1}^n  p_{1k} - p_{2k} $	Pitch	3
$\sum_{k=1}^n  p_{1k} - p_{2k}  dw2_k$	Pitch Duration	1

**Table 5.20: The results of the geometric algorithms for the six 4/4 Part B variations melodies.**

The top seven algorithms listed here give a result of 1 here for the sum of the differences in the ranks, which is the best result found. Those algorithms that use pitch, duration and metrical accent weights are among the best performing algorithms, supporting previous evidence and results. Surprisingly, the worst performing algorithm from the fine-tuning process (section 5.4) pitch with duration method 2, performed quite well here with a difference of just 1. Finally, two further algorithms that use pitch only and pitch and duration respectively and that performed badly in the fine-tuning stage give the highest difference with the human judgements.

Although there was very little difference in the weights returned in section 5.4 when all eight melodies were fine-tuned, these values were run here for the six Part B melodies

and the results (sum of the difference in ranks) were exactly the same as those shown above.

### 5.10.3 The Edit Distance Algorithms

All of the edit distance algorithms perform worse than the geometric algorithms with the smallest sum of the difference between the ranks for human similarity judgements and the algorithmic measure being reported as 2. Smith et al.'s (1998) edit distance algorithms in particular perform badly here. The traces/alignments were checked for all of the algorithms except those that used fragmentation/consolidation and all had good traces (see section 4.8.1). This indicates that the weights used were in an appropriate range for the musical material. The results are shown in Table 5.21 below in order of the best performing algorithms from section 5.7. The actual algorithm run by Mongeau and Sankoff is highlighted in bold in the table of results.

<b>Results for Part B melodies. Six variations - I, II, III, V, VIII, IX</b>	<b>Features used</b>	<b>Sum of the difference in ranks between the algorithm and human judgements</b>
Smith	I 0.06 D 1.89 Dur .24	3
M&S pitch diff	I .01 D .46 Dur .29	2
M&S pitch diff. I = D.	I .23 D .23 Dur .08	2
M&S pitch diff. No duration in replace cost	I .32 D .15 Dur 0	2
M&S pitch diff, fragmentation/consolidation. No duration used in replace cost	I 0.24 D 0.01 Dur 0	2
M&S pitch diff, fragmentation/consolidation	I 0.22 D 0.02 Dur 0.1	2
M&S pitch diff, fragmentation/consolidation. I=D	I 0.23 D 0.23 Dur 0.08	3
M&S pitch diff, fragmentation/consolidation. All values equal.	I 0.01 D 0.01 Dur 0.01	2

<b>Results for Part B melodies. Six variations - I, II, III, V, VIII, IX</b>	<b>Features used</b>	<b>Sum of the difference in ranks between the algorithm and human judgements</b>
Smith. No duration in replace cost	I .29 D 1.07 Dur 0	3
M&S pitch diff. I = D. No duration in replace cost	I .5 D .5 Dur 0	2
M&S pitch diff. All values equal.	I .01 D .01 Dur .01	2
M&S pitch diff, fragmentation/consolidation, I=D. No duration in replace cost	I 0.5 D 0.5 Dur 0	2
M&S pitch consonance.	I .01 D .02 Dur .01	2
Smith. I = D. No duration in replace cost	I .01 D .01 Dur 0	4
M&S pitch consonance, fragmentation/consolidation. No duration in replace cost	I 0.03 D 0.01 Dur 0	2
Smith. I = D	I .5 D .5 Dur .01	5
M&S pitch consonance, fragmentation/consolidation. I = D. Variation varied independently	I 0.02 D 0.02 Dur 0.01	2
<b>M&amp;S pitch consonance, fragmentation/consolidation. All values equal</b>	<b>I .01 D .01 Dur .01</b>	<b>2</b>
M&S pitch consonance, fragmentation/consolidation. I=D. No duration in replace cost	I 0.04 D 0.04 Dur 0	2
M&S pitch consonance. I = D.	I .02 D .02 Dur .01	2
M&S pitch consonance. All values equal	I .01 D .01 Dur .01	2
M&S pitch consonance. I = D. No duration in replace cost	I .04 D .04 Dur 0	2

**Table 5.21: The results of the edit distance algorithms for the six 4/4 Part B variations melodies.**

It can be clearly seen that the edit distance algorithms produce more differences with the order of similar melodies identified by the human similarity judgements than the

geometric algorithms. As with the geometric algorithms, the edit distance algorithms do not award the same similarity value to Variation V and IX and this accounts for one of the differences in the ranked order. It is noted that many of the edit distance algorithms consider Variations II and III to be equal in similarity, as the results from one of the Mongeau and Sankoff algorithms show in Table 5.22 below.

<b>M&amp;S: pitch difference used in replace calculation</b>	
Weights	insert 0.01
	delete 0.46
	duration 0.29
<b>Variation</b>	<b>Normalised Results (0-1)</b>
Var. I	0
Var. II	0.1719
Var. III	0.1719
Var. V	0.4785
Var. VIII	1
Var. IX	0.8582

**Table 5.22: An example of the normalised results from the edit distance algorithms.**

This results in a further difference between the rank orders of both, since the human similarity judgements do not reflect this level of similarity.

<b>Comparison of ranks</b>			
<b>Variation</b>	<b>Ranked algorithm result</b>	<b>Ranked human judgement</b>	<b>Absolute difference between ranks</b>
Var. I	1	1	0
Var. II	2	3	1
Var. III	2	2	0
Var. V	4	4	0
Var. VIII	6	6	0
Var. IX	5	4	1
<b>Sum of the difference between ranks</b>			<b>2</b>

**Table 5.23: The calculation of the sum of the difference in ranks for the edit distance algorithm results.**

According to the human judgements and the geometric algorithms Variation III is more similar to the query than Variation II but the edit distance algorithms do not identify the same similarity relationship. It should be noted that many of the worst performing algorithms from section 5.7, located at the bottom of the table and showing a difference of 2, perform just as well as the top algorithms when run on these Part B melodies.

### 5.10.4 The Edit Distance Algorithms with the Fine-tuned Values from the set of Eight Variations

Internal parameter values for the geometric algorithms were shown to be almost identical for the smaller set of 6 variations (all in the same time signature) and the set of eight variations that included Variations IV and VI. Since there were many differences in the cost and weight values of the edit distance algorithms and in the order of best performing algorithms, the best values for the set of eight variations are evaluated here also (see section 5.8 for a discussion of these results).

<b>Eight Variations (Variation VII omitted)</b>	<b>Features used</b>	<b>Sum of difference in ranks</b>
M&S pitch diff, fragmentation/consolidation	I .01 D .43 Dur .29	2
Smith	I .06 D 1.89 Dur .24	3
M&S, pitch diff	I .01 D .21 Dur .22	2
M&S, pitch diff, fragmentation/consolidation, I=D	I .03 D .03 Dur .05	3
M&S, pitch diff, fragmentation/consolidation, all values equal	I .01 D .01 Dur .01	2
<b>M&amp;S, pitch consonance, fragmentation/consolidation, all values equal</b>	I .01 D .01 Dur .01	2
M&S, pitch consonance, fragmentation/consolidation, I=D, no duration	I .15 D .01 Dur 0	2
M&S, pitch consonance, fragmentation/consolidation	I .01 D .03 Dur .08	3
M&S, pitch consonance	I .01 D .04 Dur .04	2
Smith, no duration	I .29 D 1.71 Dur 0	3
M&S, pitch diff, I=D	I .05 D .05 Dur .04	2
M&S, pitch diff, all values equal	I .01 D .01 Dur .01	4

<b>Eight Variations (Variation VII omitted)</b>	<b>Features used</b>	<b>Sum of difference in ranks</b>
M&S, pitch consonance, I=D	I .24 D .24 Dur .39	4
M&S pitch consonance, all values equal	I .33 D .33 Dur .33	2
Smith, I=D (bad trace on top results – look at later)	I .31 D .31 Dur .2	5
M&S, pitch diff, fragmentation/consolidation, no duration	I .27 D .24 Dur 0	2
Smith, I=D, no duration	I .01 D .01 Dur 0	4
M&S, pitch diff, no duration	I .34 D .4 Dur 0	2
M&S, pitch diff, I=D, no duration	I 0.34 D 0.34 Dur 0	2
M&S, pitch consonance, fragmentation/consolidation, no duration	I .15 D .01 Dur 0	2
M&S, pitch consonance, fragmentation/consolidation, I=D, no duration	I .15 D .15 Dur 0	2
M&S, pitch consonance, no duration	I .02 D .01 Dur 0	2
M&S, pitch consonance, I=D, no duration	I .01 D .01 Dur 0	2
M&S, pitch diff, fragmentation/consolidation, I=D, no duration	I .26 D .26 Dur 0	2

**Table 5.24: The combined geometric and edit distance results for the eight Part B variation melodies.**

No overall improvement is produced by this combination of algorithm weights and costs. There are a higher number of sum of difference of 3 and 4 from those given in Table 5.23

### **5.10.5 Conclusion of Verification of Results using the Part B Melodies**

Most of the geometric algorithms performed well on these melodies using the fine-tuned weights identified in section 5.4 Those algorithms that originally performed well are among the algorithms that give almost the same order of similarity for the melodies

as the human subjects gave. The main difference that arises between both ordered lists of melodies occurs where the median rating for two different melodies are the same (i.e. they are judged to be the same by the listening experiment subjects) but the algorithms consider one to be just slightly more similar than the other. Two of the algorithms that do not incorporate metrical accents perform much worse than the remaining eight algorithms. The geometric algorithms that use pitch, duration and metrical accent weight are therefore considered to have been quite successful in identifying the similarity of the Part B melodies. This supports the perceptually relevant features identified in chapter 2. The fine-tuning of the weights using the testbed melodies and the human judgements of similarity is also considered to have been successful in light of these results.

The edit distance algorithms all perform worse than the geometric algorithms for these melodies. The same median rating for two different melodies is not reflected in the results from these algorithms (as was noted for the geometric algorithms) and many of the edit distance algorithms identify two melodies as being exactly the same when the subjects of the listening experiment consider one more similar to the Theme. The results for the edit distance algorithms are not as encouraging as for the geometric algorithms.

Although this section has explored the use of the algorithms and adjusted weights and parameters on melodies not used in the original testbed, in order to assess the usefulness of the algorithms in a broader context further melodies in contrasting styles are explored in the following chapter.

## Chapter 6

### Exploring the Ability of the Algorithms to Generalise to other Music

In Chapter 5, the best performing algorithms were identified using the testbed of melodies and human similarity judgements featured in Chapter 3. The testbed and human judgements were also used to fine-tune the values of various weights and internal parameters.

The same top three geometric algorithms were identified when the two melodies in different time signatures to the Theme were included and the fine-tuning of the internal weights gave almost identical values. These results were further verified by using the melodies from Part B of the listening experiment. It was therefore suggested that the geometric algorithms might also prove successful in identifying similarity on a wider range of music and not just the original testbed.

In the case of the edit distance algorithms there was a difference in the performance and the internal algorithm parameter values when the two additional variation melodies (in different time signature to the Theme and other variations) were included in the testbed set. The verification of the results using the melodies from Part B of the listening experiment produced results that were worse than the geometric algorithms with some algorithms ranking as many as five out of six melodies in a different order to that given by the human subjects. This suggests that at least some of the edit distance algorithms implemented may not generalise well to musical material other than the testbed they were fine-tuned on.

In order to explore the use of the results in a wider context, experiments involving new musical material are carried out and presented in this chapter. Although the weights, costs and internal parameter values based on the original six variation melodies (with a common time signature) are concentrated on for both the edit distance and geometric algorithms in the text of this chapter, the best performing values from running all eight variation melodies are also implemented and discussed at the end of each section.

## **6.1 The Additional Collections of Melodies**

Two different sets of real (i.e. not specially constructed here but taken from existing musical pieces) melodies have been selected for evaluating the fine-tuned algorithms. Each evaluation set features stylistically different music from different eras and musical traditions and has been chosen to test the ability of the fine-tuned algorithms to generalise to a broader range of music. Information is available regarding the similarity of the melodies in each collection but takes a different form in each case and so the evaluation method is also different for each evaluation set. The first set of melodies formed part of a melodic similarity competition at an international conference and partially ordered lists of similar melodies are available for these melodies. The performance of the algorithms on three “query” melodies from this collection is explored in sections 6.3-6.6. The second set of melodies are taken from a collection of Irish folk music for which expert judgements have been made regarding the similarity of melodies in this and other related collections.

While the testbed melodies were purposefully composed to have a similarity relationship (Theme and Variation form), these new melodies explore different kinds of similarities that exist in real music. These include coincidentally related melodies, borrowed melodies and musical melodies that may have altered slightly over their lifetime (the later are present in both collections presented here).

## **6.2 Exploring the Fine-tuned Algorithms using MIREX ground truth Data**

### **6.2.1 MIREX 2005 ground truth Data**

MIREX (Music Information Retrieval Evaluation eXchange) is a contest that compares and evaluates algorithms and systems designed for the many different types of media involved in Music Information Retrieval. It was first instigated in 2005 as part of ISMIR, the International Conference on Music Information Retrieval, and has been run alongside this conference for the past three years (MIREX 2005; Downie 2006). A set of training data based on the RISM A/II collection was made available for use by the Symbolic Melodic Similarity contestants. This RISM A/II collection is an annotated catalogue of music manuscripts from 1600-1800, containing more than 500,000 records.

Up to 100 fields of information are available for each manuscript, including a musical incipit, usually ranging from two to six bars long. The training data used was a “ground truth” for the RISM incipits, collected by Rainier Typke and colleagues (2004b, 2005b; Hoed and Nooijer 2004). The term ground truth is used by the author to indicate that the similarity judgments gathered reflect some kind of agreed notion of similarity; that people with musical experience would generally agree on the degree of similarity of the melodies. The ground truth/training data is used here to evaluate the algorithms under investigation. Some detail follows on the construction of the training data set as this has relevance for the results and evaluation of the algorithms being investigated here.

Typke et al. (2004b, 2005b; Hoed and Nooijer 2004) chose 11 melodies to be compared to the RISM incipit collection. The terms “query” and “candidate” melodies are used by Typke et al. to indicate a search melody and potential matches for it from the collection of RISM incipit melodies. The 11 query melodies were used to filter a set of potential candidate melodies for each query until there are 200 out of the 500,000 or so RISM incipits. The filtering process used feature extractors that describe particular information related to the melody in a text file. Note that some of these extractors use just a single value to describe a melody. Each feature is put in a column in a relational database and then the melodies are filtered based on arbitrary combinations of these features and arbitrary thresholds.

The feature extractors used were:

- pitch range filter – the interval between the highest and lowest note
- duration ratio filter – the ratio of the shortest to longest note
- maximum interval filter – the largest interval between subsequent notes
- Melodyhound filter – the edit distance between the gross contour (up, down or repeat) of two melodies
- Melodyhound rhythm filter – edit distance based on rhythm. No details are given on how the rhythm is represented and the edit distance calculated.
- Interval histogram – a frequency count of intervals between notes
- Interval strings – the diatonic and chromatic intervals between notes
- Motive repetitions - sequences of intervals that were repeated

Further manual filtering of these 200 melodies is carried out to decide if they are good candidates for comparison with the query melodies. The set of candidates for each query melody in the experiment is constructed so that there are 20% good candidates and 80% bad candidates. Fifty candidate melodies are used for each query melody. Some sets also include the query melody in the candidate set and that should be ranked highest. Typke et al.'s PTD & EMD algorithms (2003, 2004a) were run in an attempt to ensure that the filtering procedure had not omitted potentially good candidates.

Subjects were presented with the query and candidate melodies in score format in a browser window and asked to rank the candidates by similarity to the query. A sound icon could be used to play a MIDI version of the melody but subjects were asked to concentrate mainly on the stave when making their decisions. Subjects were instructed to regard transpositions of queries as being identical and were told that melodies that were notated slightly differently but in a way that wouldn't affect the sound of the melody (e.g. different clefs) should also be considered identical. If two incipits covered different amounts of musical material subjects were told to just consider the common material. It wasn't necessary to finish ranking all the candidates for all query melodies, just to do as many as they could. The order of queries presented was randomised, as were the candidates for each query. Subjects were asked to "rank all candidates that resemble the query by their melodic similarity to the query" (Typke et al. 2004b, p.4) with no further instructions given as to what was meant by melodic similarity.

### **6.2.2 The Edit Distance Algorithms and Transposed Melodies**

The melodies used up to this point in this thesis have all been taken from Duschenes Theme and Variations (1962) and although there were some minor key melodies, all had the same tonic and did not pose any processing problems for the edit distance algorithms. (A discussion is included in section 4.2.2 of Chapter 4 on the way in which Ó Maidín's algorithm handles transposed melodies). Smith et al. (1998) use the pitch difference between notes in their implementation of the edit distance algorithm but they do not discuss how melodies in different keys should be treated and transpositions taken into account. Mongeau and Sankoff do take transpositions into account by calculating the difference in degrees of the scale between notes and using a weight associated with the interval. The MIREX melodies and the Irish folk music collection that are used to further evaluate the success of the optimised algorithms include melodies in different

keys. It would be redundant to use Smith's algorithm and the versions of Mongeau and Sankoff's algorithm that use pitch difference only without making some modifications to take transpositions into account. The approach taken here is based on that of Mongeau and Sankoff. The tonic is identified for each melody based on the key signature. The pitch of each note is then calculated in terms of the number of semi-tones from the tonic and the absolute pitch difference is then calculated using these values. (a %12 calculation is carried out on the pitch difference value to allow for octave differences). The use of the key signature alone to identify the tonic is a rough method and would not correctly recognise the tonic notes in the case of melodies in a minor key. A more comprehensive approach to automatically finding the correct tonic note would require the use of key-finding algorithms (such as those by Chew (2001) and Huron and Parncutt (1993)), which is outside the scope of this research. The method adopted here should greatly improve the performance of the edit distance algorithms than if it had not been used.

### **6.2.3 The Construction of the ground truth**

The candidate melodies were ordered by their median and then mean rank. The Wilcoxon rank sum test was applied to the ranked candidate and every incipit that was ranked higher. The Wilcoxon ranked sum test uses the ranks of the medians to determine if there is a statistical significant difference between the distributions of two different groups of results/ranks. It tests the medians to see if they are equal (the Null hypothesis) against the alternative, that one population is greater or less than the other. The resulting p-value is used to determine if there is a statistically significant difference between a particular candidate and all other higher ranked ones. The Wilcoxon value is calculated on the ranks of a particular candidate melody against the ranks of each candidate with a higher median rank. This means that each candidate has multiple listed p-values, each reporting the difference between it and a higher candidate. Low p-values mean that there is a significant difference and that the authors are reasonably sure that higher ranked candidates would be ranked higher by the population as well as by this particular sample of subjects. A threshold value of .25 was set for p. In an email communication with the author (September 14<sup>th</sup>, 2005) Typke reported that this threshold value was chosen because it "seemed to give reasonable results" and that a more strict (but common) .05 value would not have resulted in many groups i.e. you would be unsure of the ranking of many more candidates. Currently when a p-value for

a candidate is  $<.25$  a heavy line is drawn above that candidate melody. (The earlier material from the Orpheus project (Hoed and Nooijer 2004) appears to have used a lower p value to judge the statistical significance.) This means that when the authors are not sure that there is a significant difference between the ranks assigned to two candidate melodies, these melodies form a group with no line between them. There may be many such candidate melodies in a group and what is known is that all melodies above those in the group should indeed be ranked higher and all melodies below the group should be ranked lower but the ordering within the group itself is not known. The last group in each set of candidate melodies for a query is usually a large group, consisting of melodies that are not considered to be very similar to the query melody. Thus, the ground truth is a set of partially ordered lists, in which the order of melodies within groups is not known.

#### **6.2.4 Assessing the Performance of the Algorithms using the MIREX Data**

Three query melodies from the MIREX 2005 training set for Symbolic Melodic Similarity are used to test the algorithms and the optimal weights and cost values associated with each, as detailed in Chapters 4 and 5. These melodies were chosen from the initial set of eleven query melodies used in the MIREX competition. Among the melodies not chosen was a melody purely based on leaps of a third, suggesting a more harmonic than melodic nature (240.001.397) and a very short segment with similar chord-based leaps (600.066.687). Both melodies are in fact incipits from multi-instrumental pieces and are inner voice parts (the RISM catalogue number indicates which stave the incipit comes from). These melodies are rejected here, as they are primarily harmonic in nature and not the sort of melodies the algorithms are specifically intended to process. The versions of the geometric algorithm that are being evaluated here require the melodies being compared to be of the same total duration/length or that their total durations form a simple ratio such as 2:1 or 1:2 to each other. Many of the query melodies had few candidate melodies that satisfied this criterion e.g. had many candidate melodies with radically different rhythms and time signatures. The three query melodies chosen do include candidate melodies in different time and key signatures and are representative of the kind of melodic segments this algorithm was designed for use with. Subjects who ranked the melodies from the MIREX ground truth experiments were asked to take only the common sections of melodies into account. A number of the candidate melody incipits are longer than the query melodies and so the

extra notes/bars have been omitted here. Melodies that have been shortened in this way are noted by the word cut after the melody number. The numbering scheme used here to refer to the melodies is taken directly from the RISM catalogue.

### 6.3 Query Melody 000.111.706

The first melody used from the MIREX 2005 is RISM 000.111.706 (The White Cockade). Five candidate melodies from the MIREX ground truth are used and are shown in Figure 6.1. A further melody is listed in the ground truth but it is omitted here because it does not have a time signature and includes a bar of notes whose combined duration is 5/8 in what otherwise appears to be a 2/4 melody.

#### Query Melody 000.111.706



#### Candidate Melodies for 000.111.706

000.116.073 (cut)



000.113.932 (cut)



000.127.493 (cut)



000.132.448 (cut)



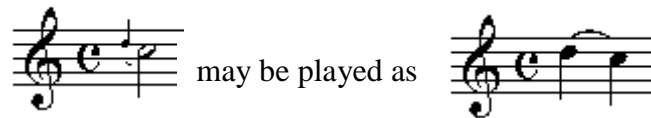
190.018.612



Figure 6.1: Query melody 000.111.706 and five candidate melodies.

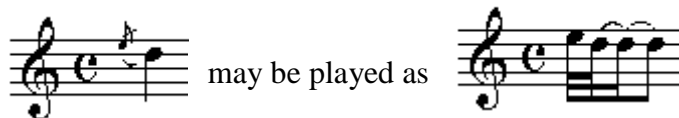
The ground truth identifies melody 000.116.073 to be the most similar melody to the query melody but the order of the remaining four melodies is not known.

Melody 000.113.932 includes a grace note as do some of the other MIREX melodies discussed later in this chapter. Both appoggiaturas and acciaccaturas are present in the melodies used. An Appoggiatura usually occurs without the oblique stroke and usually takes half the time of the note it prefixes as shown below.



**Figure 6.2: An example of an Appoggiatura.**

An Acciaccatura on the other hand is a shorter, less melodically significant variant of the appoggiatura above. The exact interpretation depends on the tempo of the piece and the question of whether the note should be played on or before the beat is a performance/stylistic issue. An example of an Acciaccatura and a possible execution is shown below:



**Figure 6.3: An example of an Acciaccatura.**

As the use of the words “may” and “usually” suggest, there is some ambiguity in how the grace notes should be executed and what the fully written out notation would look like. Since the durations of such grace notes are open to interpretation during performance the melodies are run without the inclusion of grace notes. It is noted that most of the participants in the MIREX 2005 Symbolic Melodic Similarity contest processed MIDI versions of the melodies rather than score and so would not have taken grace notes into account either.

### **6.3.1 Results**

The results for the geometric and edit distance algorithms are presented together in Table 6.1 in order of best performing algorithm as shown in Table 5.15 of Chapter 5. The costs and weights for the algorithms are those adjusted using the six melodies from the testbed variations. As mentioned in the previous section, the ground truth for this

particular query melody only reveals the known order for one melody, the most similar to the query melody. The candidate melody identified as most similar to the query melody by each of the algorithms is shown in this table. The algorithms are presented in order of the best performing algorithms from section 5.15 with the costs and weights that resulted in the best performance from the fine-tuning process in Chapter 5.

<b>Costs/Weights optimised on six variations of the testbed (I, II, III, V, VIII, IX) Smith/M&amp;S = edit distance algorithms Equations = Ó Maidín's geometric algorithms</b>	<b>Features used</b>	<b>Most similar melody</b>
$\sum_{k=1}^n  p_{1k} - p_{2k}  dw1_k (mw_{1k} \cdot mw_{2k})$	Pitch Duration Metrical accents: 1.82 1.32	000.113.932
$\sum_{k=1}^n  p_{1k} - p_{2k}  dw4_k (mw_{1k} \cdot mw_{2k})$	Pitch Duration Metrical accents: 1.98 1	000.113.932
Smith	I 0.06 D 1.89 Dur .24	000.116.073
M&S pitch diff	I .01 D .46 Dur .29	000.116.073
M&S pitch diff. I = D.	I .23 D .23 Dur .08	000.116.073
M&S pitch diff. No duration in replace cost	I .32 D .15 Dur 0	000.116.073
M&S pitch diff, fragmentation/consolidation. No duration used in replace cost	I 0.24 D 0.01 Dur 0	000.116.073
M&S pitch diff, fragmentation/consolidation. I=D	I 0.23 D 0.23 Dur 0.08	000.116.073
$\sum_{k=1}^n  p_{1k} - p_{2k}  dw2_k (mw_{1k} \cdot mw_{2k})$	Pitch Duration Metrical accents: 1.67 1.13	000.113.932
M&S pitch diff, fragmentation/consolidation. All values equal.	I 0.01 D 0.01 Dur 0.01	000.116.073
Smith. No duration in replace cost	I .29 D 1.07 Dur 0	000.116.073
M&S pitch diff. I = D. No duration in replace cost	I .5 D .5 Dur 0	000.116.073

<b>Costs/Weights optimised on six variations of the testbed (I, II, III, V, VIII, IX) Smith/M&amp;S = edit distance algorithms Equations = Ó Maidín's geometric algorithms</b>	<b>Features used</b>	<b>Most similar melody</b>
M&S pitch diff. All values equal.	I .01 D .01 Dur .01	000.116.073
$\sum_{k=1}^n  p_{1k} - p_{2k}  (mw_{1k} \cdot mw_{2k})$	Pitch Metrical accents 3.32 1	000.113.932
$\sum_{k=1}^n  p_{1k} - p_{2k}  dw_{3k} (mw_{1k} \cdot mw_{2k})$	Pitch Duration Metrical accents: 3.2 1	000.113.932
$\sum_{k=1}^n  p_{1k} - p_{2k}  dw_{1k}$	Pitch Duration	000.113.932
M&S pitch diff, fragmentation/consolidation, I=D. No duration in replace cost	I 0.5 D 0.5 Dur 0	000.116.073
$\sum_{k=1}^n  p_{1k} - p_{2k}  dw_{4k}$	Pitch Duration	000.113.932
$\sum_{k=1}^n  p_{1k} - p_{2k}  dw_{3k}$	Pitch Duration	000.113.932
M&S pitch consonance.	I .01 D .02 Dur .01	000.116.073
M&S pitch consonance No duration in replace cost	I .02 D .01 Dur 0	000.116.073
$\sum_{k=1}^n  p_{1k} - p_{2k} $	Pitch	000.113.932
Smith. I = D. No duration in replace cost	I .01 D .01 Dur 0	000.116.073
M&S pitch consonance, fragmentation/consolidation. No duration in replace cost	I 0.03 D 0.01 Dur 0	000.116.073
M&S pitch consonance, fragmentation/consolidation. All 3 weights varied independently.	I 0.03 D 0.01 Dur 0	000.116.073
Smith. I = D	I .5 D .5 Dur .01	000.116.073
M&S pitch consonance, fragmentation/consolidation. I = D. Variation varied independently	I 0.02 D 0.02 Dur 0.01	000.116.073

<b>Costs/Weights optimised on six variations of the testbed (I, II, III, V, VIII, IX) Smith/M&amp;S = edit distance algorithms Equations = Ó Maidín's geometric algorithms</b>	<b>Features used</b>	<b>Most similar melody</b>
M&S pitch consonance. I = D.	I .02 D .02 Dur .01	000.116.073
$\sum_{k=1}^n  p_{1k} - p_{2k}  dw_{2k}$	Pitch Duration	000.113.932
M&S pitch consonance, fragmentation/consolidation. I=D. No duration in replace cost	I 0.04 D 0.04 Dur 0	000.116.073
M&S pitch consonance. All values equal	I .01 D .01 Dur .01	000.116.073
M&S pitch consonance. I = D. No duration in replace cost	I .04 D .04 Dur 0	000.116.073

**Table 6.1: The combined results of the geometric and edit distance algorithms for RISM melody 000.111.706.**

Contrasting results are obtained for the versions of the geometric algorithms implemented and the edit distance algorithms. All of the geometric algorithms identify 000.113.932 as the most similar melody and all of the edit distance algorithms consider 000.116.073 to be most similar. Although the MIREX ground truth does give 000.116.073 as the most similar melody it is useful to consider the difference between these two melodies and the main approach of each of the two categories of algorithm under investigation.

### 6.3.2 The Geometric Algorithms

The geometric algorithm essentially adds up the weighted pitch difference between successive notes of each melody. The weights are derived from a combination of duration and metrical stress. The melodies are processed in time windows whose duration is equal to the longest time for which there is uniform activity (without onsets or offsets of notes or rest within the window) at that particular point in the score (see section 4.2 of Chapter 4). In the case of the 000.116.073 melody there are only four pitch differences between the notes of the melodies when they are processed in this way. These pitch differences are shown in Figure 6.4 below.

Query melody 000.111.706



000.116.073  
as processed by the geometric algorithms

**Figure 6.4: The pitch differences between the query melody and candidate melody 000.116.073.**

Three of the four pitch differences occur on very short notes (16<sup>th</sup>-notes) and no pitch differences occur on the strong beat of the bar. All pitch differences are small – one or two semi-tones. One pitch difference occurs between a slightly longer note than the others (8<sup>th</sup>-notes) and this is highlighted below. This was the melody identified as being most similar by the edit distance algorithms and the MIREX ground truth. All of the geometric algorithms however considered 000.113.932, shown below in Figure 6.5, to be the most similar melody to the query.



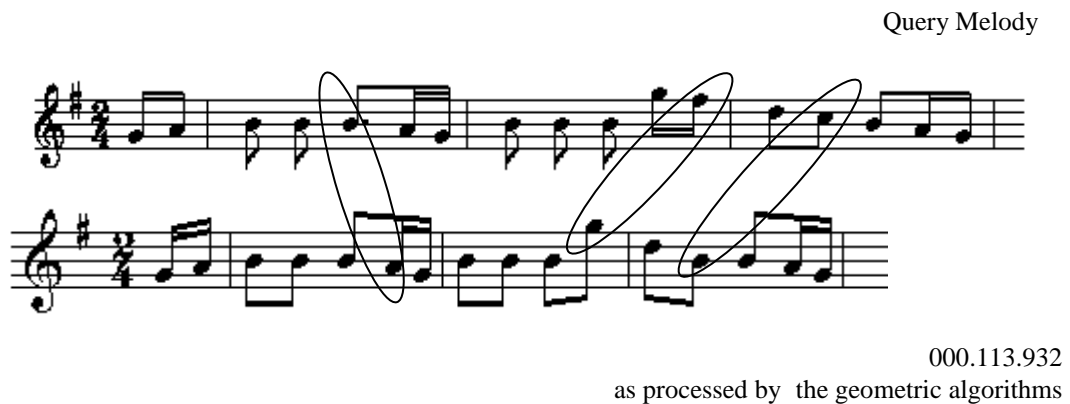
**Figure 6.5: Candidate melody 000.113.932.**

It is clear that although the melody is notated in cut common time (essentially 2/2), the ratio between the durations of notes is the same as that of the previous melody. The geometric algorithm takes the ratio between durations/lengths into account and so treats each notes as if it were half the duration when comparing it to the query melody (see sections 4.6.2-4.6.4 in Chapter 4). Additionally, the algorithm identifies that this melody has been transposed down a fifth and takes the difference of seven semi-tones into account when calculating the pitch difference (see section 5.4.2 for a reminder of how this algorithm handles transposition to different keys). Essentially the algorithm treats this melody as if it were notated as shown below in Figure 6.6.



**Figure 6.6: Candidate melody 000.113.932 as processed by the geometric algorithm**

This transposed, time-adjusted version of the candidate melody is compared to the query melody and the pitch differences identified when the time windows of the two melodies are processed are shown below in Figure 6.7.



**Figure 6.7: The pitch differences between the query melody and candidate melody 000.113.932.**

There are only three pitch differences between these two melodies as marked above. Where there were two pitch differences in the last bar between melody 000.116.073 and the query melody, here there is only one. This results in this melody being judged as more similar to the query than 000.116.073 by the geometric algorithms.

### 6.3.3 The Edit Distance Algorithms

When comparing the query melody with 000.116.073, all of the edit distance algorithms report the same edit distance operations needed to turn this melody into the query operation (see Figures 6.8 and 6.9).

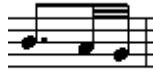


**Figure 6.8: The query melody aligned with candidate melody 000.116.073.**

Most of the edit operations required in this case involve replacing notes with notes of the same pitch but different durations. These replace operations therefore do not incorporate a pitch difference value (or pitch consonance weight) but a weighted duration difference value only. There is only one insert and one delete operation for notes with a 16th-note and 8<sup>th</sup>-note duration. The cost of these edit operations is simply a weighted duration difference.



No edits – the same pitches and durations used in each melody



Replace each of these notes with a note of the same pitch but different duration to form



No edits – the same pitches and durations used in each melody



Replace 1/16<sup>th</sup>-note G with 1/8th-note G  
Delete semiquaver F



No edit needed for the 1/8th-note D  
Insert 1/8<sup>th</sup> note B to form



Replace the 1/8<sup>th</sup> note C with a 1/16<sup>th</sup> note of the same pitch  
Replace the 1/8<sup>th</sup> note B in the segment shown on the left to a 16<sup>th</sup>-note of the same pitch  
No edits are required for the last two notes to form



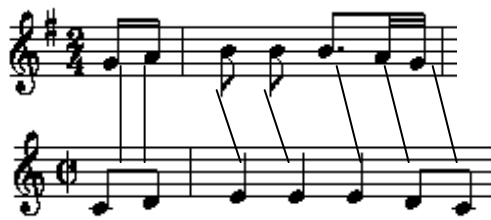
**Figure 6.9: The edit operations reported for candidate melody 000.116.073.**

There is much more deviation in the sets of edit operations reported by the edit distance algorithms for the 000.113.932 melody (shown below in Figure 6.10). Since



**Figure 6.10: Candidate melody 00.113.932.**

these algorithms process pitches in terms of the distance (in semi-tones) from the tonic, the difference in pitch is calculated as zero in most cases, since the notes in both melodies feature the same degrees of the scale. The main difference is therefore a durational difference with notes being aligned as follows as a result of the chosen edit operations, as shown in Figure 6.11 below. The difference in durations is apparent for all notes of these melodies and so it follows that using the edit distance algorithms, melody 000.113.932, which is stretched in time by a factor of two, is identified as being less similar than melody 000.116.073.



**Figure 6.11: The comparison of notes between the query melody and candidate melody 000.113.032.**

The best performing version of the Mongeau and Sankoff algorithm (as identified in Chapter 5) that uses pitch difference (rather than their full pitch consonance model with fragmentation and consolidation) has a higher duration weight than many similar algorithms. This algorithm actually judges the 000.113.932 candidate melody as being the least similar to the query melody due to this higher duration weighting.

### 6.3.4 Which Melody is More Similar to the Query Melody?

The question of which of these melodies should be considered more similar to the query relates to the issue of invariance in the time and pitch domains. While there is general acceptance that melodies that differ in key alone should be regarded as the same melody (see section in Chapter 2 for a discussion), the issue of invariance in the time domain has not had as much discussion and is explored here. Consider the melodies shown in Figure 6.12, all of which contain notes of the same pitch and relative duration.

Melody A



Melody B - same time signature, twice as long



Melody C – same time signature, half as long



Melody D – different time signature, half as long



**Figure 6.12: Four melodies featuring the same pitches and the same relative durations.**

Clearly, when altered by the simple ratios 1:2 or 2:1 the similarity is apparent but the similarity is not as evident when changes of time signature or rhythm are brought about by use of other ratios. The example in Figure 6.13 below shows a possible translation of this melody into  $\frac{3}{4}$  time. This changes the inherent beat structure of the melody and since different notes are now stressed, it becomes less recognisable as the same melody.



**Figure 6.13: Melody A from Figure 42 in  $\frac{3}{4}$  time.**

Pickens points out the fact that “most music IR researchers favour relative measures because a change in tempo or transposition across keys does not significantly alter the music information expressed” (2001, p.2). While this statement is not totally accepted because of the issues of differing beat structures that arises when the time signature is changed, it does show that the scaling of melodies in the time and frequency domain is not regarded as considerably changing the melody.

Levitin and Cook (1996), in referencing Monahan (1993) and Serafine (1979), also reflect this belief, indicating “the identity and recognisability of a song is maintained through transposition of pitch and changes in tempo” (p.928). Indeed, Typke et al. in a paper that evaluates their own EMD (Earth Mover’s Distance) algorithm (2006) say that “neither tempo changes nor transposition or the position of a melody within a piece fundamentally change the character of a melody” (p.2). Weyde (2001) adds the caveat that the amount by which the tempo is changed affects whether the melody will be considered to be the same or similar. Although tempo invariance is mentioned in some perceptual papers, there is no clear indication of what degree of tempo change is tolerated in this context.

Many Query-by-Humming systems allow for tempo and transposition invariance since they must account for users singing slower/faster in tempo and high/lower in pitch than the melodies stored in the system. Lemström et al. (2006) employ a “brute force tempo scaling version” (p.1) of his algorithm that is very similar to the method used here in the geometric algorithms to incorporate pitch invariance. The original algorithm is run “multiple times with the pattern scaled in time by predefined constant factors and retrieves the best match across the runs” (p.2). Lemström et al. use the ratios .5, .667, .8, 1, 1.25, 1.5, and 2. Not all of the ratios used by Lemström et al. are likely to be

encountered when dealing with scores. Scaling the length by  $\frac{1}{2}$  and 2 are much more likely ratios for scores, although other ratio values may be involved when comparing melodies in different time signature, for example 1.3 when the melodies are in 4/4 and 3/4 time.

The Edit Distance algorithms identify melody 000.116.073 as being the most similar to the query melody used here and this is also the melody regarded as being most similar by the MIREX ground truth. The geometric algorithms, on the other hand, identify melody 000.113.932 as being more similar to the query. Since melody 000.113.932 has less pitch differences with the query melody than melody 000.116.073 when transposition in the time and pitch domains are taken into account, there are grounds for suggesting that this melody is indeed more similar than the latter. The question remains then as to why the MIREX ground truth regards melody 000.116.073 as being more similar to the query than melody 000.113.932. In gathering the ground truth similarity judgements subjects were asked to rank melodies by primarily using the visual representation of the score and this may have had an influence on the judgements given. The high degree of similarity between the query melody and melody 000.113.932 may not have been noticed by some subjects because of it is in a different key and each notes of the melody is twice as long as the query melody. Differences in the notation of melodies that involve extra notes/bars, different keys, different duration/time signatures may make the accurate assessment of similarity between such melodies more difficult and lead to different judgements than if the melodies were auditioned aurally. The construction of the ground truth data is discussed in more detail in Chapter 7.

## **6.4 Query Melody 800.000.193**

The second melody from the MIREX 2005 ground truth used for evaluation is RISM melody 800.000.193 (Roslin Castle). In the case of the previous MIREX query melody, the correct order of similarity of only one candidate melody, the most similar melody to the query, was known. Here, the order of similarity of six melodies is known:

- 000.109.446
- 000.111.779/000.112.692
- 000.132.330
- 000.112.625
- 700.008.178

These melodies are shown below in Figure 6.14.

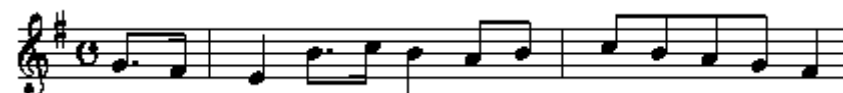
It is not known from the ground truth which of the two melodies 000.111.779 and 000.112.692 is more similar to the query melody but it is known that both melodies are less similar than 000.109.446 and more similar than the remaining melodies on the list. This ranked order of 000.111.779 and 000.112.692 are interchangeable in the results. All of the edit distance and geometric algorithms are run on these melodies and the order they return is compared to the MIREX ground truth in the table below, in order of the best performing algorithms from Chapter 5.

### Query Melody

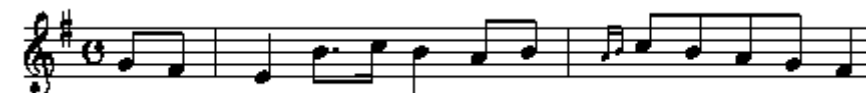


### Candidate Melodies

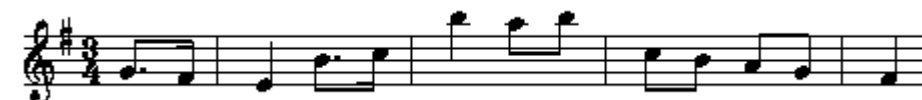
000.109.446 (cut)



000.111.779 (cut)



000.112.692



000.132.330 (cut)



000.112.625



700.008.178



Figure 6.14: Query melody 800.000.193 and six candidate melodies.

### 6.4.1 Results

Costs/Weights optimised on six variations of the testbed (I, II, III, V, VIII, IX) Smith/M&S = edit distance algorithms Equations = Ó Maidín's geometric algorithms	Features used	Comparison of order of similarity with the MIREX ground truth
$\sum_{k=1}^n  p_{1k} - p_{2k}  dw1_k (mw_{1k} \cdot mw_{2k})$	Pitch Duration Metrical accents: 1.82 1.32	same
$\sum_{k=1}^n  p_{1k} - p_{2k}  dw4_k (mw_{1k} \cdot mw_{2k})$	Pitch Duration Metrical accents: 1.98 1	same
Smith	I 0.06 D 1.89 Dur .24	Order different – 700.008.178 is most similar
M&S pitch diff	I .01 D .46 Dur .29	Order different – 700.008.178 is most similar
M&S pitch diff. I = D.	I .23 D .23 Dur .08	Order different – 700.008.178 is most similar
M&S pitch diff. No duration in replace cost	I .32 D .15 Dur 0	Order different – 700.008.178 is most similar
$\sum_{k=1}^n  p_{1k} - p_{2k}  dw2_k (mw_{1k} \cdot mw_{2k})$	Pitch Duration Metrical accents: 1.67 1.13	same
Smith. No duration in replace cost	I .29 D 1.07 Dur 0	Order different – 000.132.330 is most similar
M&S pitch diff. I = D. No duration in replace cost	I .5 D .5 Dur 0	Order different – 700.008.178 is most similar
M&S pitch diff. All values equal.	I .01 D .01 Dur .01	Order different – 700.008.178 is most similar
$\sum_{k=1}^n  p_{1k} - p_{2k}  (mw_{1k} \cdot mw_{2k})$	Pitch Metrical accents: 3.32 1	Similar – order of 000.112.625 and 000.132.330 swapped around
$\sum_{k=1}^n  p_{1k} - p_{2k}  dw3_k (mw_{1k} \cdot mw_{2k})$	Pitch Duration Metrical accents: 3.2 1	Similar – order of 000.112.625 and 000.132.330 swapped around
$\sum_{k=1}^n  p_{1k} - p_{2k}  dw1_k$	Pitch Duration	same
$\sum_{k=1}^n  p_{1k} - p_{2k}  dw4_k$	Pitch Duration	Similar – order of 000.112.692 and 000.111.779 swapped around

Costs/Weights optimised on six variations of the testbed (I, II, III, V, VIII, IX) Smith/M&S = edit distance algorithms Equations = Ó Maidín's geometric algorithms	Features used	Comparison of order of similarity with the MIREX ground truth
$\sum_{k=1}^n  p_{1k} - p_{2k}  dw3_k$	Pitch Duration	same
M&S pitch consonance.	I .01 D .02 Dur .01	Order different – 700.008.178 is most similar
M&S pitch consonance No duration in replace cost	I .02 D .01 Dur 0	Order different – 700.008.178 is most similar
$\sum_{k=1}^n  p_{1k} - p_{2k} $	Pitch	same
Smith. I = D. No duration in replace cost	I .01 D .01 Dur 0	Order different – 000.132.330 is most similar
M&S pitch consonance, fragmentation/consolidation. No duration in replace cost	I 0.03 D 0.01 Dur 0	700.008.178 is most similar. Order of 1 <sup>st</sup> and last melody swapped around
Smith. I = D	I .5 D .5 Dur .01	Order different – 700.008.178 is most similar
M&S pitch consonance, fragmentation/consolidation. I = D. Variation varied independently	I 0.02 D 0.02 Dur 0.01	Order different – 700.008.178 is most similar
M&S pitch consonance, fragmentation/consolidation. All values equal	I .01 D .01 Dur .01	Order different – 700.008.178 is most similar
$\sum_{k=1}^n  p_{1k} - p_{2k}  dw2_k$	Pitch Duration	same
M&S pitch consonance, fragmentation/consolidation. I=D. No duration in replace cost	I 0.04 D 0.04 Dur 0	Order different – 700.008.178 is most similar
M&S pitch consonance. All values equal	I .01 D .01 Dur .01	Order different – 700.008.178 is most similar
M&S pitch consonance. I = D. No duration in replace cost	I .04 D .04 Dur 0	Order different – 700.008.178 is most similar

**Table 6.2: The combined results of the geometric and edit distance algorithms for RISM melody 800.000.193.**

### **6.4.2 The Geometric Algorithms**

The geometric algorithms all identify the same melody as being most similar to the query as the MIREX ground truth. Seven out of the ten of these algorithms actually give the exact same order as the ground truth. Two of the remaining three algorithms give a similar order, with the order of the fourth and fifth melodies swapped around. The remaining geometric algorithm gives a different ordering for melodies two and three on the list. The top three performing algorithms identified by the Theme and Variations testbed in Chapter 5 are among those that match the order of similar melodies given by the MIREX ground truth.

### **6.4.3 The Edit Distance Algorithms**

None of the Edit Distance algorithms run match the order of the MIREX ground truth. In fact many of these algorithms report 700.008.178 as the most similar melody to the query. This particular candidate melody is actually identified as the least similar of the melodies by the MIREX ground truth. Almost all of the Edit Distance algorithms also resulted in “bad traces”, in which whole groups of notes were simply deleted and new notes then inserted (see section 4.7 in Chapter 4 for a discussion of “good” and “bad” traces). It is obvious even from a visual glance at the candidate melodies that many are quite similar to the query melody. In this case we would expect there to be lots of replace operations, as there are many pitches in common albeit in different keys. Also, we would expect there to be many more replace operations than insert and delete as we would expect the combination of these two operations to be more costly than replacing notes that are close in pitch and duration. Since the Mongeau and Sankoff (1990) basic pitch difference and the pitch consonance algorithms resulted in bad traces the versions of the algorithms that use fragmentation/consolidation were not run.

The reason the Edit Distance algorithms perform so badly compared to the geometric algorithms is due to the method used to represent pitch. Smith et al.’s (1998) algorithm was adapted to process transposed melodies by converting the MIDI note numbers to a value that represented the distance in semi-tones from the tonic. Using the difference between notes expressed in terms of the number of semi-tones from the tonic is a precursor to the pitch consonance weights implemented by Mongeau and Sankoff. The pitches in the melodies shown in Figure 6.15 would all be considered equal when

converted to the semi-tone from tonic format. (Octave equivalence is used so that notes an octave apart are considered the same pitch.)

Melody 1



Tonic: C (60)  
 MIDI note numbers:  
 72 71 69 67 65 67 69 65 67  
 Pitch of notes in terms of STs from tonic:  
 0 11 9 7 5 7 9 5 7

Melody 2



Tonic: G (67)  
 MIDI note numbers:  
 79 78 76 74 72 74 76 72 74  
 Pitch of notes in terms of STs from tonic:  
 0 11 9 7 5 7 9 5 7

Melody 3



Tonic: E<sup>b</sup> (63)  
 MIDI note numbers:  
 75 74 72 70 68 70 72 68 70  
 Pitch of notes in terms of STs from tonic:  
 0 11 9 7 5 7 9 5 7

**Figure 6.15: Sample melodies in different keys with common note pitches in terms of distance from the tonic.**

The query melody, 800.000.193, is shown below in Figure 6.16:



**Figure 6.16: The query melody.**

The melody is in the key of C with the first note an interval of a fifth (seven semi-tones) above the tonic. The MIDI note-number and pitches converted to semi-tones from the tonic are:

Tonic: C (60)  
 MIDI note number: 67 66 64 71 72 71 69 71 72 71 69 67 66  
 STs from tonic: 7 6 4 11 0 11 9 11 0 11 9 7 6

Most of the candidates melodies however, start on the tonic and feature similar pitches to 000.109.446, shown below in Figure 6.17 and described in terms of semi-tones from the tonic as:

Tonic: G (67)

MIDI note number: 67 66 64 71 72 71 69 71 72 71 69 67 66

ST from tonic: 0 11 9 4 5 4 2 4 5 4 2 0 11



Figure 6.17: Candidate melody 000.109.445.

When comparing the candidate melodies to the query melody using the Edit Distance algorithms, the fact that most of the pitches are the same is not identified. The replace cost is calculated by combining the pitch difference with a weighted duration difference. Since the pitch difference is quite high (as high as 7 in the above example (11-4)), we might expect the replace cost to be higher than a combined insert and delete cost and so making it “cheaper” to simply delete all notes in one melody and replace them with the other. Obviously, this would not be a solution that would make musical sense, since the pitches of the notes are the same or very similar in most cases. The geometric algorithms did not have this problem as they calculate the number of steps up or down the pitches of the original melody have been shifted by and recognise the similarity in the pitches.

The geometric algorithm is considered very successful here as it returns the same order of similarity for the melodies as the MIREX ground truth. The edit distance algorithms perform poorly due to their inability to identify the similarity of the pitches in melodies with different key signatures such as these. The issues of pitch representation and transpositions for edit distance algorithms are discussed further in Chapter 7.

## 6.5 Query Melody 230.005.489

The third and final melody chosen from the MIREX 2005 competition is RISM melody 230.005.489 (Liebster Jesu Wir Sind Hier, an arrangement of a Bach Chorale). A larger number of candidate melodies are included in this evaluation, with seventeen melodies in total, for which the order of similarity of six is known. It would be expected that the algorithms would be able to at least recognise the similarity of the six melodies and place the remaining eleven melodies at the bottom of the ordered list of similar melodies. According to Typke et al.’s ground truth the order of the most similar melodies are:

- 702.002.005
- 702.002.217
- 230.004.687/451.509.620
- 700.001.741/451.504.065

The query and candidate melodies are presented below in Figure 6.18.

### Query Melody



### Candidate Melodies

702.002.005



702.002.217 (cut)



230.004.687



451.509.620 (cut)



700.001.741



451.504.065 (cut)



451.512.833



230.005.490



400.215.226 (cut)



400.249.480



450.008.956 (cut)



450.032.447



000.051.782



150.200.732 (cut)



451.500.863



453.001.086



000.117.453 (cut)



Figure 6.18: Query melody 230.005.489 and seventeen candidate melodies.

### **6.5.1 Results**

The results for each type of algorithm show quite a number of differences from the ground truth order of melodies. None of the algorithms match the order of similar melodies provided by the ground truth. The ground truth distinctly lists melody 702.002.005 as being more similar to the query melody than 702.002.217 but all of the geometric and edit distance algorithms consider 702.002.005 and 702.002.217 melodies to be identical to each other and to the query melody (difference of 0). None of the algorithms consider 702.002.217 to be less similar to the query than the former melody.

### **6.5.2 Differences between the Algorithms and the ground truth**

The MIREX ground truth clearly indicates that melody 702.002.005 is more similar to the query melody than 702.002.217. The 702.002.217 incipit had three further bars of music but since subjects were asked to judge the segment of the melodies that was common to all, these bars were omitted in this experiment (see section 6.2.1). The only difference between these three melodies is the presence or absence of a fermata on the last note and the inclusion of trills on two notes in melody 702.002.217. It is possible that the inclusion of the extra bars of melody 702.002.217 in the visual and MIDI representation in the listening test run by Typke et al. (2004b, 2005b; Hoed and Nooijer 2004) and not the presence of two trills (which were not in the query melody) affected the perceptual judgement of the test subjects and caused this melody to be judged as less similar to the query.

Since all of the geometric and Edit Distance algorithms identified all three melodies as being the same, it is again suggested that the ordering information supplied by the MIREX ground truth is not entirely reliable. Some further interesting issues are raised by the order of similarity provided by the algorithms and these are discussed in the following section.

### **6.5.3 Further Discussion of the Results**

The order of similarity of the candidate melodies to the query melody according to most of the geometric algorithms is:

- 702.002.005
- 702.002.217
- 230.004.687
- 451.504.065
- 451.512.833

- 400.215.226
- 400.249.480
- 450.008.956
- 453.001.086

One of the most noticeable differences between the Edit Distance and geometric algorithms is the result for candidate melodies 451.504.065 and 451.512.833. These melodies are ranked as the third and fourth most similar melodies by all of the geometric algorithms but appear much further down the list of similar melodies when the Edit Distance algorithms are used.


Although stretched in the time domain to twice the length of the query melody and transposed to different keys, both show a strong resemblance to the query melody. One might expect such melodies to be judged as being quite similar to the query melody (see section 6.3.4) by both the human test subjects and by the computer algorithms. The Edit Distance algorithms place both of these melodies quite far down the ordered list of similar melodies, as there is no mechanism for taking the doubling of the duration of each note into account. A significant difference in durations is detected instead and contributes to the overall distance measure.

The geometric algorithms' handling of candidate melody 451.504.065 also raises an interesting issue in relation to the order of most similar melodies in the MIREX ground truth. This melody is an exact transposition of melody 230.004.687, as can be seen in Figure 6.19. The ground truth does not reflect this level of similarity however, with melody 230.004.687 clearly ranked higher than melody 451.504.065. Again, this is attributed to the visual presentation of the melodies in gathering the ground truth, with the difference in key, time stretching and the extra bars in melody 451.504.065 affecting the similarity judgements given.


One might expect all of the geometric algorithms to judge these melodies as identical to each other, receiving the same difference score. However, since the latter melody is notated in common time and the former in cut-common time, there are slightly different metrical stress weights applied to each melody. In some cases this results in 230.004.687 being regarded as slightly more similar to the query melody than 451.504.065, but in many cases both melodies are identified as being identical, depending on the optimised weight values applied. It seemed most appropriate to assign

the metrical stress weights based on the time signature of the notated score rather than to alter it to fit the particular comparison (in this case common versus cut-common time). In this way, if the two melodies being compared had exactly the same notes and differed only in time signature, the algorithm would reflect this and judge them as being the same. It is only when a difference in pitch and/or note duration occurs that this difference in metrical stress due to the time signatures influences the final similarity/difference measure for the two melodies.


Melody 230.004.687 in G major



Melody 451.504.065 (cut) in Ab major



Melody 451.504.065 (cut) as processed by the geometric algorithms



**Figure 6.19: Candidate melodies 451.504.065 - a time-stretched transposed version of melody 230.004.687.**

It is also noted that candidate melody 700.001.741 appears quite high in the ground truth list (joint fourth) but in the bottom half of the list generated by the geometric algorithms and is not one of the commonly identified similar melodies given at the start of this section. In Figure 6.20, an x marks the notes of 700.001.741 that match the pitch of the query melody in the case of the geometric algorithms where transposition to different keys is taken into account. The geometric algorithms score this melody lower than many others because there are two differences in pitch that occur on the primary and secondary stressed beats of the second bar.

Query melody (A major)



Melody 700.001.741 (G major)



### **6.5.5 Summary of Findings**

The use of the MIREX melodies and the associated ground truth has demonstrated that the geometric algorithm (implemented in many versions here) is better at handling melodies stretched (and compressed) in the time domain. The edit distance algorithms implemented are based on a string-matching rather than a musical approach and they do not identify the relationship between such melodies. The geometric algorithms were also better at identifying transpositions within keys and to different keys, where the pitches are all shifted up or down by a common number of semi-tones.

In the case of the first MIREX melody (000.111.706), where the order of one melody only was known, the geometric algorithms all gave a different result to the ground truth and the edit distance algorithms but it is suggested that the ground truth ordering of melodies is not perceptually accurate in this regard. The order of six of the most similar candidate melodies to query melody 800.000.193 is known from the ground truth. Most of the geometric algorithms returned the exact same order and all identified the most similar melody correctly. None of the edit distance algorithms matched the order of similarity indicated by the ground truth and it was noted that all indicated a set of edit operations that resulted in “bad” traces. Again, for query melody 230.005.48, there is some disagreement with the ground truth because of time-stretching and key changes of the candidate melodies. Most of the geometric algorithms identify provide the same top nine similar candidate melodies to the query melody. Differences with the MIREX ground truth are highlighted and discussed.

Overall, the geometric algorithms are deemed to have performed quite successfully in identifying similarities between the MIREX query and candidate melodies and to have outperformed the edit distance algorithms. A level of care is needed in interpreting results derived from the MIREX ground truth data. These similarity judgements have been shown here to be somewhat questionable.

## **6.6 Exploring the performance of the Algorithms Using A Collection of Irish Folk Music**

### **6.6.1 Overview of Ceol Rince na hÉireann and Irish Folk Music**

The second collection of music and similarity judgements used in the evaluation stage of this research is taken from Ceol Rince na hÉireann Vol. 1 (1963) (translates as The Dance Music of Ireland), an important collection of Irish folk music collected by Brendan Breathnach. Breathnach collected the tunes primarily from performers, although some are notated from commercial recordings, and published five volumes of Ceol Rince na hÉireann. Each volume in the collection includes annotations by Breathnach that attempt to identify the similarity of the melodies to others found in the Ceol Rince na hÉireann collection as well as in other well-known Irish music collections.

A large part of the repertoire of Irish folk music exists in the form of dance music from which it derives its structure. These dance music pieces are referred to as “tunes” in the context of Irish folk music and this terminology is used here to refer to an entire dance music piece. The tunes in this collection are categorised according to the dance type they were composed for, which include double jigs, single jigs, slip jigs, hornpipes, reels, waltzes and mazurkas, barn dances, flings, slides, highlands and polkas. The Double jigs from Ceol Rince na hÉireann Volume 1 along with the expert similarity observations made by Breathnach are used to evaluate the optimised algorithms.

### **6.6.2 The Tunes Chosen from Ceol Rince na hÉireann, Vol 1.**

The Double jigs from this first volume of Ceol Rince na hÉireann were chosen because they are the first category of tunes included in the collection and because they are the most common type of dance music in the repertoire. Double jigs use a 6/8 time signature. As is the case with most Irish dance music the tune is usually in two eight-bar parts, both of which may be repeated. Occasionally, three- and four-part tunes can be found and more rarely tunes which consist of five to seven parts. The first five tunes from this collection of music are shown in Appendix D. An upbeat of an 8<sup>th</sup>- or 16<sup>th</sup> – note is often included in jig tunes with an initial upbeat before the first bar and at the end of bar 8, leading into bar 9 and the second part of the tune (see melodies D.1-D.5 in Appendix D for examples). Since both parts of these dance tunes are often repeated this sometimes results in the use of first and second endings. The first eight bars only of

each tune was used and the upbeats to bars 1 and 9 omitted since not all of the jigs had a structure of two eight-bar parts and not all included the upbeats.

There were 54 double jigs in total included in Ceol Rince na hÉireann Volume 1. Breathnach's annotations noting the similarity of the tunes include references to sixteen other collections by seven other music collectors. Not all of these collections of Irish music have been encoded for processing with computer. Along with the Ceol Rince na hÉireann Volume 1 collection itself, a large number of jigs from The Dance Music of Ireland by O'Neill (1907) are available in encoded format and so the twenty-six jigs that had similarity annotations relating to melodies in both of these collections were used. Each of these twenty-six double jigs was compared in turn to an evaluation set of 426 melodies comprised of:

- 54 double jigs from Ceol Rince na hÉireann Vol. I
- 7 double jigs from Ceol Rince na hÉireann Vol. II, III and V (encoded for the purposes of this evaluation)
- 365 double jigs from the O' Neill's Dance Music of Ireland (1907)

In the previous evaluation collections presented, there were a smaller number of candidate melodies for each query melody (continuing with the terminology from the MIREX ground truth here) with similarity judgements given in terms of ratings on a scale or ranked melodies by subjects. Here, a much larger number of candidate melodies are each compared to the query melody and the results then compared to the similarity observations from a single person who was an expert in this type of music. The similarity scores provided for all 426 candidate melodies are used to rank these melodies in terms of similarity to the query melodies (each of the 26 jigs from Ceol Rince na hÉireann Vol. I) for each algorithm run. The performance of the algorithms is then compared to the similarity observations made by Breathnach. The first eight bars of these 426 melodies are used, with upbeats omitted.

It is noted that Breathnach includes a number of ornamentation symbols in the notation of some of these tunes. The actual notes that would be played are not notated because musicians would interpret such ornamentations differently, depending on the capabilities of the instrument being used. Although, the ornamentations are recorded in the encoded format of this collection, they are not taken into account here and the ornamented note is treated as a single continuous note.

### **6.6.3 Breathnach's Annotations**

The observations made by Breathnach are printed at the back of each of the *Ceol Rince na hÉireann* volumes in the Irish language. An online English translation of his comments by de Grae is used here (2000). Irish folk music is rooted in an oral tradition, with musicians learning tunes mainly from listening to them being played, rather than from notated scores of the music. This has resulted in many very similar tunes with the same and different names, as musicians changed tunes slightly to fit their instruments, their own style, and often simply because they did not quite remember the exact version they had heard. As a result there are often many slightly different versions of a tune in existence. Breathnach often uses the Irish word “leagan” to refer to related tunes. The translator of Breathnach’s annotations remarks that

“The word “leagan” may be translated as “version” or “setting; both forms are used in this translation, the choice depending on what seemed appropriate in the context.” (de Grae 2000)

Here, de Grae’s (2000) translations of Breathnach’s comments are used as written in the English translation. The relationship between melodies is noted and the algorithmic measures of similarity examined for evidence of agreement with Breathnach’s translated comments.

### **6.6.4 The Algorithms Implemented**

Not all algorithms that were implemented in Chapters 4 and 5 are evaluated here. Each of the 8-bar sections from the 26 chosen melodies from *Ceol Rince na hÉireann* Volume 1 are compared to the evaluation set of 426 tunes (again the first eight bars only). This produces a large volume of similarity results for even one algorithm (11,076 comparisons) and so it was decided to carry out the evaluation on the best performing algorithm of each type implemented and discussed in Chapter 4. These were the best:

- Geometric algorithm
- Smith edit distance algorithm
- Version of Mongeau and Sankoff’s edit distance algorithm that uses pitch difference when calculating the replace cost.
- Version of Mongeau and Sankoff’s algorithm that uses weights based on pitch consonance when calculating the replace cost

The geometric algorithm that uses pitch only is also included for comparison with the best performing geometric algorithm and the edit distance algorithms. This was one of two poorest performing algorithms identified in Chapter 5 when run on the testbed melodies.

Table 6.3 below features a list of each of the 26 melodies from *Ceol Rince na hÉireann* Volume 1. The second column of the table gives the name and source of the related melody. The third column contains the similarity observations and the remaining five columns indicate where this melody was placed by each of the algorithms in the ranked list of similar melodies. For ease of reading the word “same” is used when subsequent algorithms give the same result as those of the initial geometric algorithm. Where the letter “T” is used in the observation column, this indicates a comment on the similarity of melodies that has been made by the translator and not Breathnach himself. For ease of processing each query melody was compared to all 426 melodies in the evaluation set each time. This results in the query melody being considered first in terms of similarity in each case and ranked highest. Therefore, the most similar melody that is not that melody itself is identified as being the next most similar and second in terms of ranked order.

### **6.6.5 Results**

The best geometric algorithm performs very well here with most of the similar melodies being found in second place. In many cases, the version of this melody that uses pitch only returns the same results, although there are occasional minor differences with the best performing geometric algorithm. Of the 33 similarity observations included for these melodies the best geometric algorithm placed 18 of them in second place and a further seven in third place.

The geometric algorithms performed poorly for melodies 26, 27, 28, and 38 and on closer examination it was noted that the first parts of these melodies were not very similar to those mentioned in Breathnach’s comments. In the case of these three melodies, it is actually the second or later parts of the melodies that are actually related and not the first eight bars that are compared by the algorithms. This accounts for the poor performance of the geometric and edit distance algorithms for melodies 27, 28 and 38.

While the edit distance algorithms also identify the melody referenced by Breathnach as being most similar to the query melody in a number of cases, it is evident that these algorithms perform very badly for some melodies, for example, in melody 43 where the geometric algorithm identifies Breathnach's melody as the most similar melody but the edit distance algorithms place the melodies as low as 376<sup>th</sup>. The following melodies all show poor results for the edit distance algorithms used:

- 10
- 20
- 22
- 23
- 26
- 38
- 41
- 43
- 46

In every case, the query melody from Ceol Rince na hÉireann was in a different key to the melody Breathnach identified as being similar. The similarity between the melodies was evident and many notes were of the exact same pitch but with a different notated key signature. However, the edit distance algorithms do not detect this similarity because the pitch is processed in terms of degrees or semi-tones from the tonic in order to take transpositions into account, as discussed in section 6.2.2. A number of other transposed versions of the melodies are found in the evaluation set but most are transpositions in which the pitches of the notes are transposed up or down a number of pitches according to the key change. In these cases the method used in to adapt the edit distance algorithms to deal with transpositions seems to work effectively and the edit distance algorithms perform almost as well as the geometric algorithms for the remainder of the melodies used here.

The edit distance algorithms showed poorer results than the geometric algorithms for melodies 3 and 24. The version of Smith et al.'s algorithm (1998) used here was one of the best performing algorithms for the initial testbed melodies but in the case of these two melodies, performs worse than the other algorithms.

Overall, the geometric algorithms performed very well and can successfully identify a similar melody to the query melody from a set of 425 possible melodies. When the issues regarding transposition are taken into account, the edit distance algorithms also

perform well in general, although they are not seen to be as successful as the geometric algorithms.

<b>CRE Vol. I</b>	<b>Related melody according to Breathnach</b>	<b>Observation</b>	<b>Best Geometric</b>	<b>Geometric with pitch only</b>	<b>Smith ED</b>	<b>M&amp;S pitch diff</b>	<b>M&amp;S pitch con</b>
<b>1</b>	DMOI 94	poor version	2nd	same	same	same	same
<b>2</b>	CREIII 17	same tune	2nd	same	same	same	same
	DMOI 288	1st parts are the same	3rd	4 <sup>th</sup>	same	4th	same
<b>3</b>	DMOI 251	another version	2nd	same	same	same	ame
	CREI 14	another version	3rd	same	14th	7th	4th
<b>4</b>	DMOI 132	same tune	2nd	same	same	same	same
	DMOI 3	poor version	6th	10th	8th	9th	9th
<b>10</b>	DMOI 72	another setting	2nd	same	same	same	same
	CREV 26	another setting	3rd	same	153rd	195th	198th
<b>12</b>	DMOI 12	of the same stock	3rd	2nd	same	same	same
<b>13</b>	DMOI 24	same tune	2nd	same	same	same	same
<b>15</b>	DMOI 92	1st parts are the same	2nd	same	same	same	same
<b>16</b>	CREII 23	another setting	2nd	3rd	same	same	same
<b>19</b>	DMOI 145	same tune	2nd	same	same	same	same
<b>20</b>	DMOI 158	another setting	13th	11th	120th	97th	130th
<b>22</b>	DMOI 197	would remind you of this jig	5th	9th	244th	199th	225th
<b>23</b>	DMOI 187	another setting	2nd	same	298th	240th	278th
<b>24</b>	DMOI 162	another version	2nd	3rd	9th	4th	3rd
<b>26</b>	DMOI 226	T – another version	25th	14th	425th	424th	425th
<b>27</b>	DMOI 34	another setting	200th	240th	175th	199th	193rd
	DMOI 199	another setting	169th	256th	184th	142nd	152nd
<b>28</b>	CREIII 53	T - same tune	2nd	same	same	same	same
	DMOI 67	same tune	2nd	4th	same	same	same
<b>29</b>	DMOI 99	1st parts are the same	2nd	same	same	same	same
<b>33</b>	DMOI 88	another version	3rd	same	same	same	same
<b>36</b>	DMOI 267	1st parts are the same	2nd	same	same	same	same
<b>38</b>	DMOI 106	another setting	2nd	3rd	49th	59th	78th
	DMOI 150	another setting	4th	same	225th	83rd	90th

<b>CRE Vol. I</b>	<b>Related melody according to Breathnach</b>	<b>Observation</b>	<b>Best Geometric</b>	<b>Geometric with pitch only</b>	<b>Smith ED</b>	<b>M&amp;S pitch diff</b>	<b>M&amp;S pitch con</b>
	DMOI 53	another setting	111th	144th	52nd	108th	98th
<b>41</b>	DMOI 21	very similar	2nd	same	43rd	44th	53rd
<b>43</b>	CREII 29	T – better known as this tune but Breathnach had already used the name for another	3rd	same	373rd	376th	364th
<b>46</b>	DMOI 25	another version	2nd	same	196th	99th	118th
<b>53</b>	DMOI 214	based on no. 53	3rd	2nd	2nd	2nd	2nd

**Table 6.3: The similar melodies from Ceol Rince na hÉireann identified by Breathnach and the algorithms.**

## Chapter 7

### Conclusion

A comparative evaluation of two contrasting approaches to melodic similarity algorithms for music scores has been presented. Potentially useful features were identified from perceptual research on melodic memory and recognition. A number of versions of the geometric and edit distance algorithms were implemented so that the benefit of using various musical features from the score could be investigated. The algorithm weights and parameters were objectively fine-tuned using consistent and reliable human judgments of similarity collected in a listening experiment. The results were verified using an additional set of melodies and similarity ratings. The results for each algorithm were discussed and the strengths and weaknesses of each approach noted. Two further collections of melodies were used to assess how successfully the fine-tuned algorithms would generalise to use with a broader range of music. Over the three stages of fine-tuning, verification and generalisation, the results indicate that the geometric algorithms approximated human perception of similarity more successfully than the edit distance algorithms implemented.

Among the contributions to research in the field of melodic similarity algorithms are the review of music perception literature on melodic memory, the incorporation of music perception principles into the algorithms, the collection of human judgements of similarity for the testbed melodies, the use of these judgements to objectively fine-tune aspects of the algorithms and the fine-tuned weights and values themselves. The comparison of the performance of two contrasting algorithmic approaches and the use of verification and generalisation stages to further evaluate the initial results and suitability of the algorithms in a wider context, can also be seen as contributions.

#### 7.1 Summary of Findings

An overview of the best performing algorithms from the fine-tuning stage and the relevant weights are presented in the following sections. There were some differences in the results for the smaller set of testbed melodies that were all in the same time signature and the larger set that introduced different time signatures so both sets of results are summarised here. The verification of these results are summarised in section

7.1.4. Section 7.1.5 includes a discussion of the generalisation stage of the research. The geometric algorithms that use pitch, duration method 1 or 4 (see section 4.3) and metrical accent weights are shown to be the best performing overall algorithms.

### 7.1.1 Fine-tuning the Geometric Algorithms

The three best performing geometric algorithms used a combination of pitch, duration and metrical accents. When the six variation melodies in the same time signature were compared to the Theme melody, duration methods 1, 4 and 2 (with pitch and metrical accent) were identified as producing the best results (see Table 7.1 below).

Algorithms – in order of best performance	Features used	Metrical accent weight on 1st beat	Metrical accent weight on 3rd beat
$\sum_{k=1}^n  p_{1k} - p_{2k}  dw1_k (mw_{1k} \cdot mw_{2k})$	Pitch Duration Metrical accents	1.82	1.32
$\sum_{k=1}^n  p_{1k} - p_{2k}  dw4_k (mw_{1k} \cdot mw_{2k})$	Pitch Duration Metrical accents	1.98	1
$\sum_{k=1}^n  p_{1k} - p_{2k}  dw2_k (mw_{1k} \cdot mw_{2k})$	Pitch Duration Metrical accents	1.67	1.13

**Table 7.1: The top three geometric algorithms (6 variation melodies).**

The top two of these geometric algorithms performed better than all the others, including the edit distance algorithms (see Table 7.2 below for the top five algorithms and Table 5.15 in section 5.7 for detailed results). When the additional variations in different time signatures to the Theme were included (Variations IV and VI), the results for the geometric algorithms were very similar (see Table 7.3 below and Table 5.16 in section 5.8). In fact the same or very similar metrical accent values were identified for all of the geometric algorithms for both the smaller set of six melodies and the larger set that included the melodies with different time signatures (see Tables 5.4 and 5.9). This suggests that these are indeed optimal values and that the fine-tuning process was a success for the geometric algorithms.

All but one of the versions of the geometric algorithm that used pitch and duration without metrical accent perform better than the algorithm that uses pitch alone to calculate the similarity. This is in agreement with the earlier suggestion that since rhythm (a sequence of note durations) and pitch is important in remembering melodies

(sections 2.2.4, 2.4.7 and 2.5), both of these features might be useful for calculating melodic similarity. Similarly, the role of metrical accents as points of emphasis in the melody has been shown to be useful in the context of melodic similarity, producing the best results when combined with both pitch and duration.

Algorithms – in order of best performance	Features used	Weights/costs
$\sum_{k=1}^n  p_{1k} - p_{2k}  dw4_k (mw_{1k} \cdot mw_{2k})$	Pitch Duration Metrical accents	Accent weights: 1.98, 1
$\sum_{k=1}^n  p_{1k} - p_{2k}  dw1_k (mw_{1k} \cdot mw_{2k})$	Pitch Duration Metrical accents (1.82, 1.32)	Accents weights: 1.82, 1.32
Edit distance – Smith	Pitch Duration	Insert cost: 0.06 Delete cost: 1.89 Duration weight for replace cost: 0.24
Edit distance – Mongeau and Sankoff (pitch difference)	Pitch Duration	Insert weight: 0.01 Delete weight: 0.46 Duration weight for replace cost: 0.29
Edit distance – Mongeau and Sankoff (pitch difference, insert weight = delete weight)	Pitch Duration	Insert weight: 0.23 Delete weight: 0.23 Duration weight for replace cost: 0.08

**Table 7.2: The overall top five algorithms (6 variation melodies).**

Algorithms – in order of best performance	Features used	Metrical accent weight on 1st beat	Metrical accent weight on 3rd beat
$\sum_{k=1}^n  p_{1k} - p_{2k}  dw4_k (mw_{1k} \cdot mw_{2k})$	Pitch Duration Metrical accents	1.98	1
$\sum_{k=1}^n  p_{1k} - p_{2k}  dw1_k (mw_{1k} \cdot mw_{2k})$	Pitch Duration Metrical accents	1.82	1.32
$\sum_{k=1}^n  p_{1k} - p_{2k}  dw2_k (mw_{1k} \cdot mw_{2k})$	Pitch Duration Metrical accents	1.67	1.13

**Table 7.3: The top three geometric algorithms (8 variation melodies).**

### 7.1.2 Fine-tuning the Edit Distance Algorithms

The best performing edit distance algorithm for the six variations that all have the same time signature was the version of Smith et al.'s (1998) algorithm in which the insert and delete costs were assigned different values and the duration weight for the replace cost was assigned the value .24 (see Table 7.4 below). The next best edit distance algorithm was the version of Mongeau and Sankoff's (1990) algorithm in which pitch difference is used and all three weights (insert, delete and replace) are assigned separate values. The weight applied to the duration differences in calculating the replace cost is similar to Smith et al.'s at .29.

<b>Algorithms – in order of best performance</b>	<b>Weights/costs</b>
Smith	Insert cost: 0.06 Delete cost: 1.89 Duration weight for replace cost: 0.24
M&S pitch difference	Insert weight: 0.01 Delete weight: 0.46 Duration weight for replace cost: 0.29
M&S pitch difference insert weight = delete weight	Insert cost: 0.23 Delete cost: 1.23 Duration weight for replace cost: 0.08

**Table 7.4: The top three edit distance algorithms (6 variation melodies).**

The versions of Mongeau and Sankoff's edit distance algorithm that used their weights based on pitch consonance (instead of a basic pitch difference calculation) did not perform well (see Table 5.15). In fact, the actual algorithm they implemented (pitch consonance weights and fragmentation/consolidation with all three weights for calculating the insert, delete and replace costs assigned the same value) performed badly when compared to most other edit distance and geometric algorithms.

The inclusion of the additional melodies with a different time signature to the Theme did change these results somewhat. The version of Mongeau and Sankoff's edit distance algorithm that used pitch difference, fragmentation/consolidation and three different weight values performed better than all other edit distance and geometric algorithms (see Table 7.5 below). This was explained in section 5.9 by the increased occurrence of fragmented notes in Variations IV and VI. The version of Smith et al.'s algorithm mentioned above also performs better than the geometric algorithms for this set of eight melodies.

<b>Algorithms – in order of best performance</b>	<b>Features used and</b>	<b>Weights/costs</b>
Edit distance – Mongeau and Sankoff (pitch difference, fragmentation/consolidation)	Pitch Duration	Insert weight: 0.01 Delete weight: 0.43 Duration weight for replace cost: 0.29
Edit distance – Smith	Pitch Duration	Insert cost: 0.06 Delete cost: 1.89 Duration weight for replace cost: 0.24
Edit distance – Mongeau and Sankoff (pitch difference)	Pitch Duration	Insert weight: 0.01 Delete weight: 0.21 Duration weight for replace cost: 0.22
$\sum_{k=1}^n  p_{1k} - p_{2k}  dw4_k (mw_{1k} \cdot mw_{2k})$	Pitch Duration Metrical accents	Accents weights: 1.98, 1
$\sum_{k=1}^n  p_{1k} - p_{2k}  dw1_k (mw_{1k} \cdot mw_{2k})$	Pitch Duration Metrical accents (1.82, 1.32)	Accents weights: 1.82, 1.32

**Table 7.5: The overall top five algorithms (8 variation melodies).**

The versions of Mongeau and Sankoff’s algorithm that use pitch consonance do perform better than before but this is explained by the presence of notes in the additional melodies (Variations IV and VI) that form consonant intervals with the notes of the Theme (see section 5.9). The top three edit distance algorithms have a duration weight in the range of .22 to .29. The previous results for the set of six variations in the same time signature were also within this range, suggesting that this is indeed an optimal range of weights for the duration difference in the replace cost for these edit distance algorithms. Most of the remaining edit distance algorithms identify different weight values when these two additional melodies (Variations IV and VI) were included, suggesting that these algorithms may not generalise well to other melodies.

### **7.1.3 Conclusions of the Fine-Tuning Stage**

In conclusion, the best performing geometric algorithms for the initial testbed of six variation melodies were the two algorithms that used pitch, duration methods 1 or 4 and the relevant fine-tuned metrical accent weight values. These algorithms were also shown to be the best overall algorithms, performing better than all of the implemented edit distance algorithms. The actual version of the edit distance algorithms implemented by Mongeau and Sankoff (1990) performed quite poorly here. The modified version of

the algorithm that used pitch difference, instead of their weights based on pitch consonance, performed better. It was also shown that the best performing edit distance algorithms used three separate cost/weight values for insert, delete and replace edit operations. The presence of fragmented notes improved the performance of the edit distance algorithms that incorporated fragmentation/consolidation, but the performance was not as successful when these musical features were not present in the melodies.

#### **7.1.4 Verifying the Results**

In the verification stage, using the melodies from Part B of the listening experiment, it was found that all but two of the geometric algorithms performed better than all of the edit distance algorithms when the difference between the algorithmic and human similarity measures were compared. The top three geometric algorithms that used pitch, duration and metrical accent (see Table 7.1) are among those that produce the best results here. This is true for both the set of six melodies that share the same time signature and when the additional two variation melodies (Variations IV and VI) were included. This again suggests that the geometric algorithms perform better than the edit distance algorithms and that the weights identified for the geometric algorithms in the fine-tuning stage are indeed optimal weights that are suitable for use with a wider range of melodies.

#### **7.1.5 The Ability of the Algorithms to Generalise to Other Melodies**

A selection of melodies from the MIREX 2005 training set and from Brendan Breathnach's *Ceol Rince na hÉireann Vol. 1* (1963) were used to explore the ability of the algorithms to generalise to other musical material. Human observations of similarity were available for both of these collections in the form of collective ranked judgements (the ground truth) for the former and the observations of a sole expert for the latter. The geometric algorithms performed better than the edit distance algorithms overall. In the case of one of the MIREX melodies the geometric algorithms returned the same ranked order of melodies as the ground truth, where none of the edit distance algorithms returned the same order or even identified the most similar melody correctly. The results of the geometric algorithm are different from those of the edit distance algorithms and the ground truth in the case of another query melody but it is argued that the ground truth is flawed in this case. The third set of MIREX melodies raise questions about the way each algorithm processes time-stretched and transposed melodies. The geometric algorithms are again considered to perform better than the edit distance algorithms.

In the case of the Irish folk tunes from *Ceol Rince na hÉireann Vol. 1* (1963), the performance of the edit distance algorithms was worse than the geometric algorithms. Some, though not all, of these poor results can be explained by the issue of transposition which is discussed in section 7.2.1. The best geometric algorithm identified in Chapter 5 is the best performing algorithm here also.

The best performing geometric algorithms from the fine-tuning stage are considered to generalize to other melodies more successfully than the edit distance algorithms. The success of the geometric algorithm over all three stages (fine-tuning, verification and generalization) indicates that these geometric algorithms are preferable to the edit distance algorithms for computationally determining the similarity between melodies.

### **7.1.6 Issues Encountered with the MIREX Ground Truth**

Although not under direct investigation in this thesis, a number of issues relating to the reliability of the MIREX ground truth were raised in the previous chapter. Discussions in section 6.3 and 6.5 raised questions about the order of melodies given by the MIREX ground truth. In some cases candidate melodies that had more pitch differences than others to the query melodies, when transposition and stretching in the time domain were taken into account, were judged to be more similar to the query melody. In the case of another query melody, two of the candidate melodies are ranked differently in terms of similarity when the only difference between the melodies is the presence of a trill on two notes in one of the melodies. The issue is partly explained by the nature of the test procedure used to gather the ranked similarity judgements. Test subjects were asked to rank the candidate melodies in order of similarity to the query melody by examining the music notation rather than by listening to the melodies. A MIDI version of the melodies was provided but subjects were asked to primarily use the notation to make their similarity judgement. Some of the candidate melodies in the MIREX ground truth are notated in different keys, some are stretched in the time domain, others are notated using clefs not commonly in use today, and others include whole bars of music that should not be taken into account according to the instructions for the experiment. The mezzo-soprano and soprano clefs (a C clef on the 2<sup>nd</sup> line from the bottom of the stave and bottom line respectively) would rarely be seen by modern musicians but frequently occur in these melodies due to the historical span of the RISM A/II material (1600-1800). It is suggested that these aspects may have combined to confuse the issue of melodic similarity for the test subjects. Comparing melodies in different keys using

different clefs may make it difficult, for example, to differentiate between direct transpositions and transpositions that include different pitches. The candidates were asked to rank around 50 query candidate melodies for each of the 11 query melodies provided. Although they were asked to be thorough rather than finish ranking all of the melodies, the sheer volume of melodies to be examined and ranked for one query alone considering the difficulties mentioned above means that the judgements gathered may not be entirely perceptually accurate.

The MIREX ground truth data is one of the only publicly available collections of real world melodies (i.e. not made up for the experiment) with human similarity judgements. The particular dataset used here formed part of the MIREX 2005 Training Set for the Symbolic Melodic Similarity contest. The partially ordered list of candidate melodies has been useful for evaluating the algorithms, identifying issues and advantages/disadvantages of the algorithms, the features used and the particular implementations of the algorithms used in this research. However, results derived from using this ground truth may require some interpretation on the part of the researcher.

## **7.2 Limitations of the Algorithms and Potential Solutions**

### **7.2.1 Alterations in the Time Domain and Transposition of Melodies**

Among the limitations of the edit distance algorithms identified were the issues of identifying similarity in time-stretched/compressed and transposed melodies. In the case of the geometric algorithm a parameter value was manually set to indicate how many bars of the comparison melody were mapped to each bar of the query melody (section 5.3.4) and the results were shown to be quite successful. A potential solution proposed by Lemström et al. (2006) was discussed in section 6.3.4. This involves scaling one of the melodies in time and repeating the edit distance calculations until the minimum result is found. This automatic method would also be useful for the geometric algorithm.

As previously discussed (section 6.3.4), this time-scaling approach is somewhat similar to that used by Ó Maidín (1998) to process transposed melodies. Ó Maidín originally proposed to transpose one of the melodies up and down a number of semi-tones and to calculate the minimum overall similarity result for each of the transposed melodies. A similar technique could be used with the edit distance algorithms also to ensure that the algorithms identify transposed as well as time-stretched and compressed melodies.

Gomez et al. (2007) and Müllensiefen and Frieler (2004c) transpose the query melody to a number of different keys for comparison with the target melody, so there is a precedent for this approach.

An alternative approach to calculating the results for multiple transposed versions of the algorithms is to adjust the way in which the pitches of the melodies are represented for processing. A number of versions of both Smith et al.'s (1998) and Mongeau and Sankoff's (1990) algorithms were implemented here. Smith et al.'s edit distance algorithm did not include any mechanism for dealing with melodies transposed to different keys (they used MIDI note numbers and calculated the pitch difference using these values). Mongeau and Sankoff calculated the distance in semi-tones between each note and the tonic and so the notes of each melody were processed according to where they fit on the major/minor scale and therefore direct transpositions of melodies were identified. The implementations of Smith et al.'s algorithms in this research and the versions of Mongeau and Sankoff's that used pitch difference rather than the pitch consonance weights represent the pitches of the melodies in terms of the distance in semi-tones from the tonic. Mongeau and Sankoff used only a small number of melodies and the authors manually identified the tonic so that the melody pitches could be represented relative to this pitch. Here, the tonic was identified from the key signature, which allowed the tonic to be identified computationally. However, this is a simplistic method and would not identify minor keys or keys not explicitly notated by the key signature. Realistically, a more sophisticated approach to determining the key is required. Key-finding algorithms such as those mentioned in section 6.2.2 should be useful in this respect.

A number of researchers solve the problem of identifying transposed melodies by using the intervals between successive notes to represent the pitch of melodies. The pitch of each note is represented in terms of the distance from the previous note, usually the number of semi-tones by which the notes is higher or lower. In this way two transposed melodies with different MIDI note numbers, such as those shown below, produce the same pitch interval representation.

### **Melody 1**

MIDI note number            60 65 64 60 62

Interval representation      +5 -1 -4 +2

### **Melody 2**

MIDI note number            64 69 68 64 66

Interval representation      +5 -1 -4 +2

Examples of the use of intervals to represent pitch edit distance algorithms include Uitdenbogerd and Zobel (1999), Lemström and Ukkonen (2000) and Müllensiefen and Frieler (2004a, 2004b, 2004c). One of the problems with this approach is that incorporating duration into the algorithm can be problematic since it is the pitch interval successive between notes of a melody that are represented rather than the pitch of each note in turn. The perception research presented in Chapter 2 suggests that duration improves melodic memory over pitch alone and the results presented in this research show that using duration as well as pitch improves melodic similarity algorithms. Therefore, this approach to dealing with transposed melodies was not adopted.

#### **7.2.2 Rests**

The question of how to handle and interpret rests has not been considered in detail in this research. Only two rests occurred in the Duschenes' (1962) melodies used here. In the case of the geometric algorithm, these variation melodies were altered slightly to remove the rest and the previous notes extended in length by the duration of the rest (see section 4.3.6). There was some precedent for this as the longer note without the rest was present in the Theme melody and most of the Variation melodies. The same method was adopted for the edit distance algorithms in order to make the results comparable. Similar approaches are also found in the work of Mazonni and Dannenberg (2001), Soulez et al. (2003) and Uitdenbogerd and Zobel (1999), where the pitch of the note before the rest was used in calculating the similarity. While this method may be acceptable for short rests where it might be proposed that a listener is still processing the pitch of the previous note, it is not altogether clear how longer notes should be processed. The geometric algorithm calculates the pitch difference between notes in each window of the melodies being processed. The pitch differences are later multiplied by a weight that represents duration and/or metrical stress. If one of the melodies in the window contains a rest instead of a note, it is not possible to calculate the pitch

difference. It may seem reasonable to regard the pitch difference as 0 or to skip this window and proceed to the next window (and note) but that could potentially result in a situation where the following melodies are regarded as being exactly the same.



**Figure 7.1: Two melodies that differ only by the presence of a rest in melody 2.**

Along with supplying a set of weights based on the consonance of the intervals in question, Mongeau and Sankoff (1990) designate a weight to be used when one of the notes involved in a replace or fragmentation/consolidation operation is a rest. The weight value is the same as that used for an interval of a fifth when both melodies are in the same scale. This is considered the most consonant interval and so is assigned the lowest weight value. Although this approach is not entirely suitable as it effectively inserts a note into the melody in place of a rest, there may be merit in the idea of including some sort of a low value that can be used to represent a rest. If, for example, a pitch difference of .5 was used then this would be smaller than any possible real difference in pitch between notes (using pitch difference in semi-tones the lowest possible value is 1), and yet the presence of some value would mark the presence of the rest and some small difference between the melodies in Figure 7.1 would be identified.

There is some evidence from perceptual research that relatively long rests function as phrase boundaries and can be used to help segment the melody into phrases (segmentation is discussed in section 7.3.3). Longer rests could be used for extracting phrases and not be seen as part of the melody to be processed by the melodic similarity algorithm itself. The question of what constitutes a long and short rest is relative to the duration of the surrounding notes and the context in which the note appears should be taken into account.

### **7.2.3 The Length of Melodies**

One difference in the way in which the edit distance and geometric algorithms process melodies relates to the length of the melodies being compared. The geometric algorithm requires both melodies to be the same total rational duration, although compression and stretching of one of the melodies in the time domain can also be taken into account (see

section 5.4.4). This is owing to the way in which the algorithm breaks the melody into time windows for processing (see section 4.2). This could be seen as a limitation of this algorithm.

The edit distance algorithms, on the other hand, can work with melodies of different lengths, inserting, deleting and replacing notes as appropriate.

## **7.3 Future Work**

### **7.3.1 Extensions to the Edit Distance Algorithms**

Robine et al. (2007) propose some interesting extensions to the edit distance algorithm. All notes that are two semi-tones or fewer from the surrounding notes receive reduced insert and delete costs and metrical accents are incorporated by penalising insertions and deletions of notes on strong beats and not on weak beats. It would certainly be interesting to further explore these ideas and to find appropriate insert and delete costs for jumps of large intervals, as against small stepwise movement, and for notes on strong versus weak beats.

These ideas are related to the music perception theory and findings reported in Chapter 2 of this thesis. Although many different versions of the two edit distance algorithms were implemented as part of this research, no actual extensions were made to the algorithms of the original authors. It may be useful to explore methods of incorporating metrical accents into the edit distance algorithms and the melodic accents identified in section 2.4 into both algorithms.

### **7.3.2 Implementation and Investigation of Melodic Accents**

The definition of these melodic accents was shown to be somewhat ambiguous (see section 2.4) and investigating their definition is outside of the scope of this research. However, as part of a broader research project it would be useful to use these algorithms with a testbed and human similarity judgements to investigate the importance of interval leaps of various sizes and the exact note that pitch contour accents occur on.

### **7.3.3 Melodic Search Systems and Segmentation**

An obvious next step for this research would be to take the algorithms identified here as being successful (the geometric algorithms that use pitch, duration and metrical stress) and explore their use in a full search/retrieval system that would facilitate searching for melodies similar to a query melody in a database of scores. One of the main issues

involved in a search of this means would be deciding what melodic segments to compare with the melodic similarity algorithm(s). One approach is to identify potential melodic phrases by pre-processing using segmentation techniques.

A common approach to segmentation algorithms is based on Gestalt principles of proximity and similarity in which “relatively large changes or distances in any musical parameter like pitch, dynamics, or melodic movement marks segment boundaries” (Weyde 2004, p.128). Cambouropoulos’ (2001) local boundary detection model (LBDM) is an example of this and is used as an “essential reference amongst segmentation algorithms, mostly due to its simplicity and generality” (Ferrand et al. 2003). The strength of potential segmentation boundaries is calculated based on strength and degree of change between neighbouring intervals, with pitch, duration and rests taken into account. Spevak et al. (2002) explore the ambiguity inherent in segmentation using this model and Melucci and Orio (2002) extend the LBDM algorithm in a comparison with manual segmentation. Such approaches are related to the perceptual literature on accents discussed in section 2.2.8, with long notes considered as marking the end of phrases by Drake et al. (1993) and Parncutt (2003).

Memory-based models of segmentation were investigated in Bod (2001). Three different models were applied to 1,000 songs from the Essen Folksong Collection (EsAC 2005). This collection includes phrase annotations and recall and precision measure were used to evaluate the success of the models implemented.

A further approach is proposed by Chen et al. (2004), who combine two methods to extract musical phrases from pieces of music. They consider notes preceding rests and sudden changes in the duration of notes to be “terminative” notes in a phrase. They also use Huron’s findings (1995 cited in Chen et al. 2004) that musical phrases tend to have an arch-shaped contour and are generally between six and ten notes long to determine the start and end points of phrases.

#### **7.3.4 The Role of Context in Similarity**

The role that melodic context plays in influencing similarity judgements are investigated by Eerola and Bregman (2007). Their research suggests that the most salient features of a melody can change from one melody to another and that an ideal model of similarity would therefore dynamically change the melodic features it uses to reflect this. The research presented in this thesis attempts to identify features and

weights that are appropriate for use with a range of melodies. An alternative approach, and a possible extension to the research based on the work of Eerola and Bregman (2007), would be to analyse the melodies for particular characteristics and to use the results to inform the feature set and weights used in the melodic similarity algorithm.

The potential solutions to current limitations of the algorithms presented in section 7.2 and the possible extensions discussed in section 7.3 are recommended starting points for the further development of the research presented in this thesis.

## Appendix A

### The Listening Experiment/Testbed Melodies

This Appendix includes the melodic segments extracts from Duschenes (1962) Variations on ‘Twinkle, Twinkle, Little Star’ used in the listening experiments detailed in Chapter 3. See sections 3.3.2 and 3.3.6 for detailed discussion on the use of these melodies.

#### Part A melodies – the first four bars from Duschenes’ ‘Twinkle, Twinkle, Little Star’ variations

Theme



Variation I



Variation II



Variation III



Variation IV



Variation V



Variation VI

Musical notation for Variation VI, consisting of two staves of music in 3/4 time with a key signature of two flats. The first staff contains a melodic line with several triplet markings. The second staff contains a bass line with a triplet marking.

Variation VII

Musical notation for Variation VII, consisting of a single staff of music in 4/4 time with a key signature of one sharp. The melody is characterized by frequent triplet markings.

Variation VIII

Musical notation for Variation VIII, consisting of two staves of music in 4/4 time with a key signature of one sharp. The music features a steady eighth-note accompaniment in the bass and a more active melodic line in the treble.

Variation IX

Musical notation for Variation IX, consisting of two staves of music in 4/4 time with a key signature of one sharp. The piece is characterized by a consistent eighth-note accompaniment throughout.

**Part B melodies – the second four bars from Duschenes’ ‘Twinkle, Twinkle, Little Star’ variations**

Theme



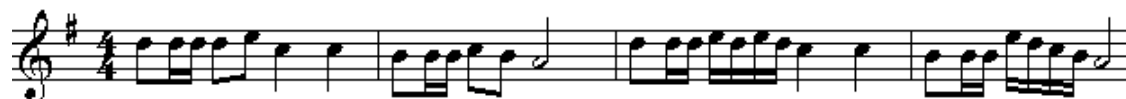
Variation I



Variation II



Variation III



Variation V



Variation VIII



Variation IX

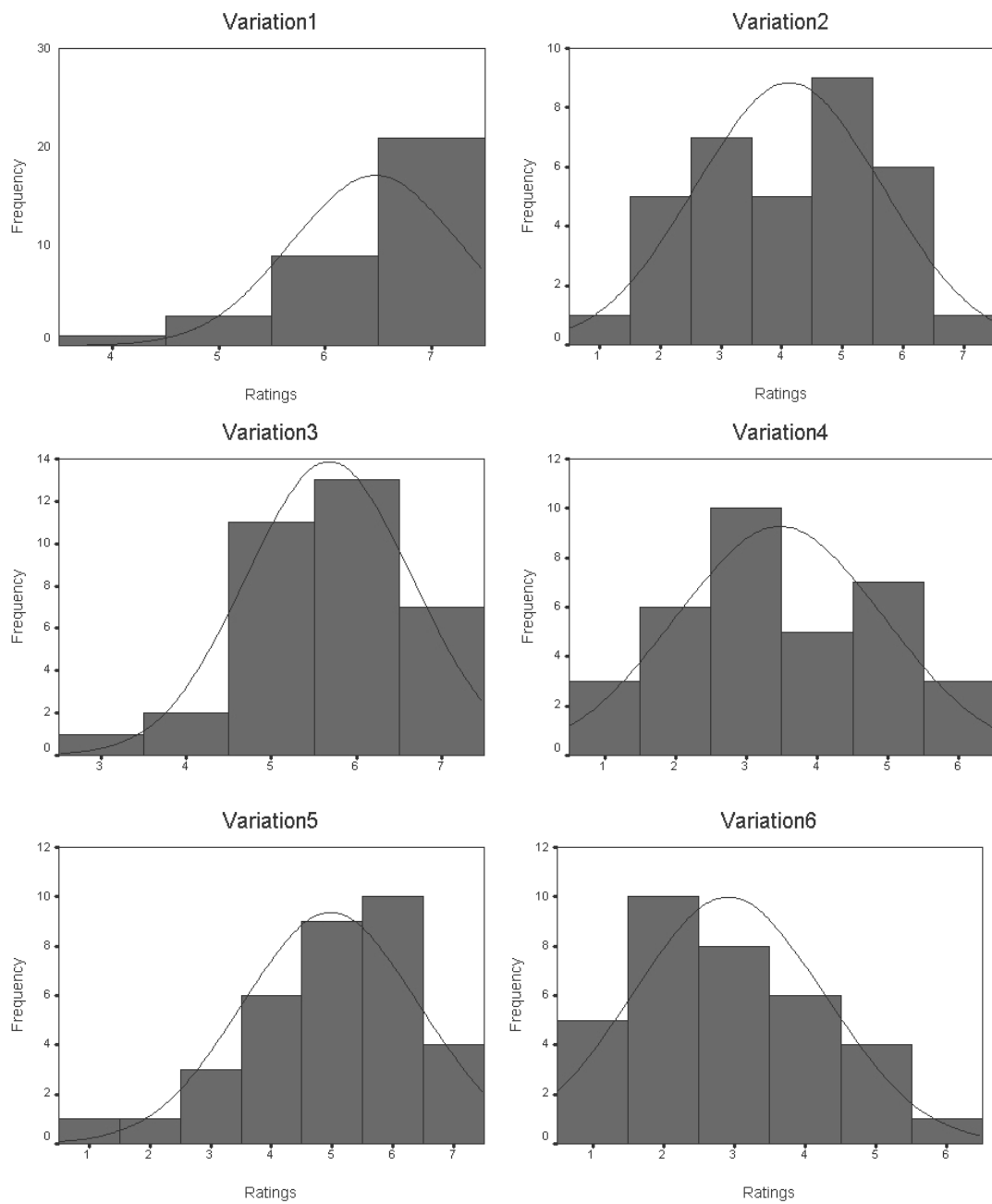


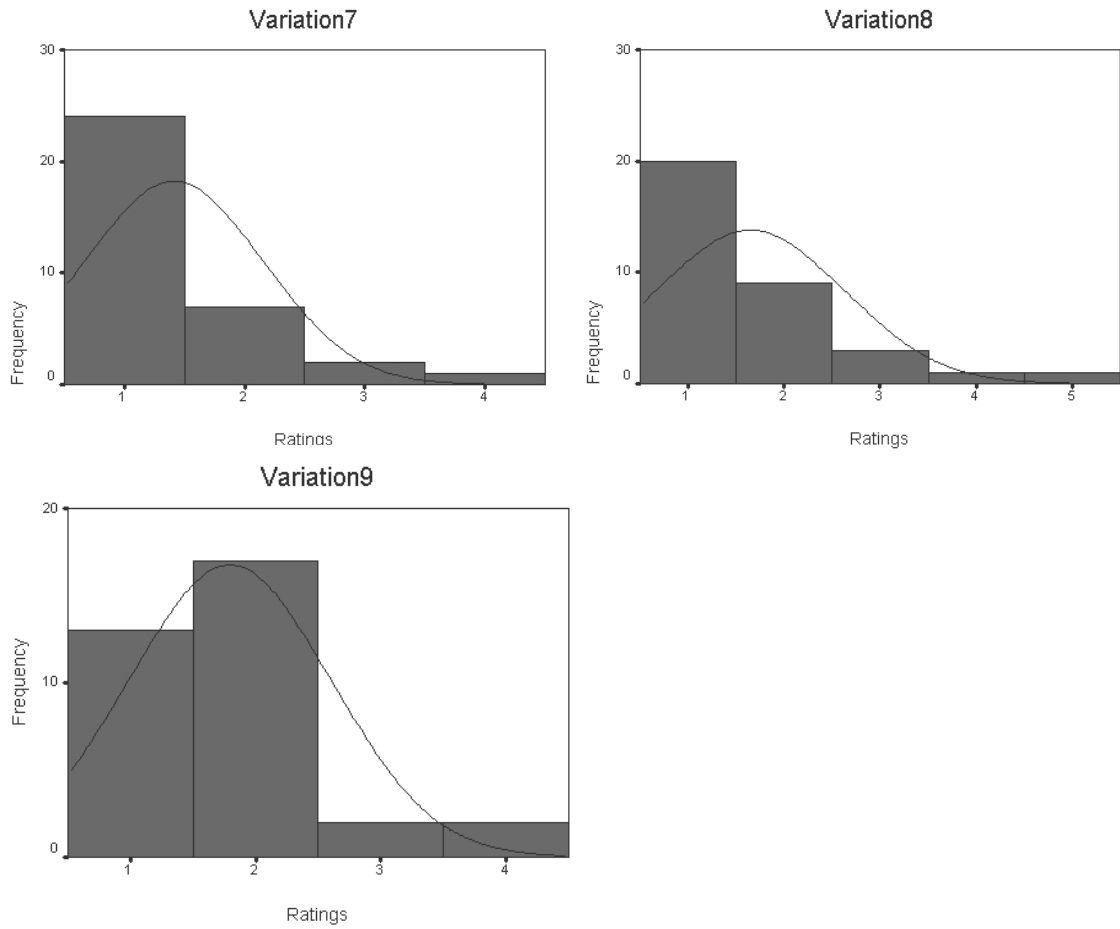
## Appendix B

### Results of the Listening Experiment

This Appendix contains graphs and tables of results from the listening experiment discussed in Chapter 3.

#### Analysis of the Ratings from Part A for all 34 subjects (see sections 3.6.1)





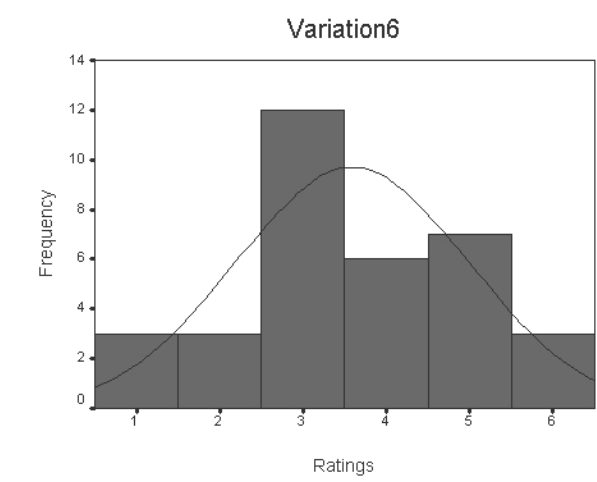
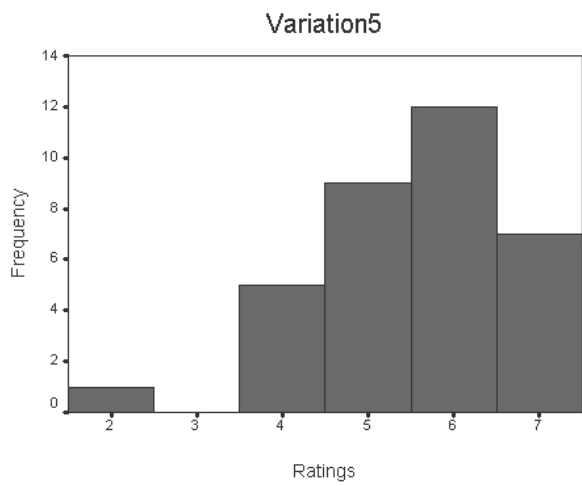
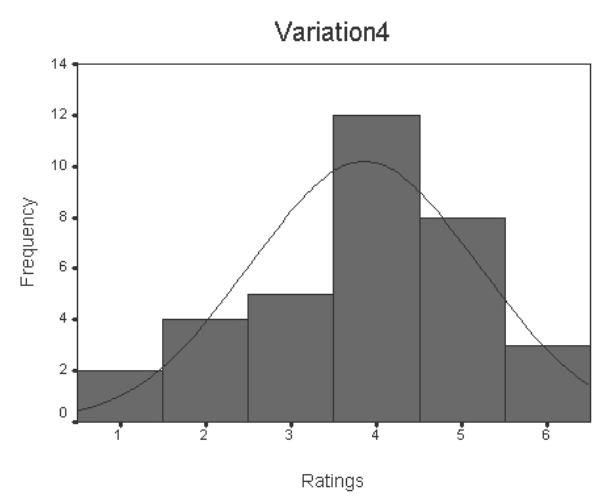
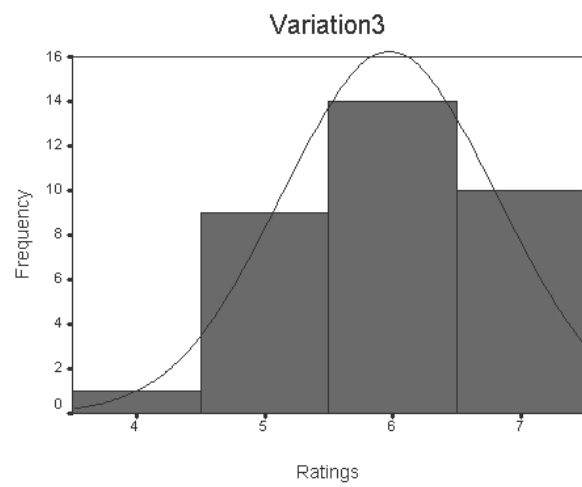
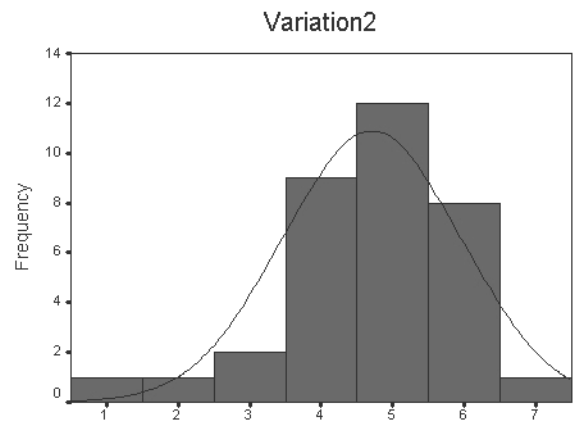
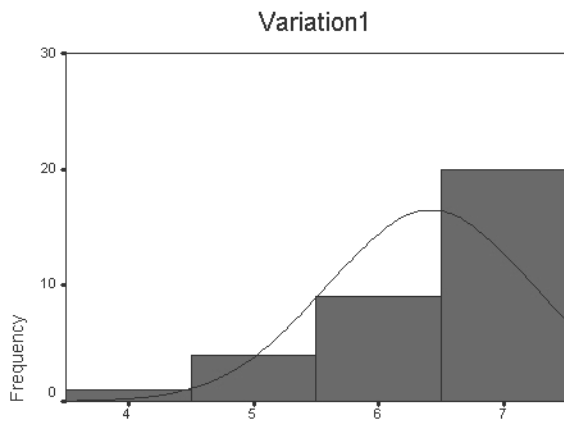
**Figure B.1: The frequency distributions of the ratings given by all 34 subjects for Part A (sequential) of the listening experiment.**

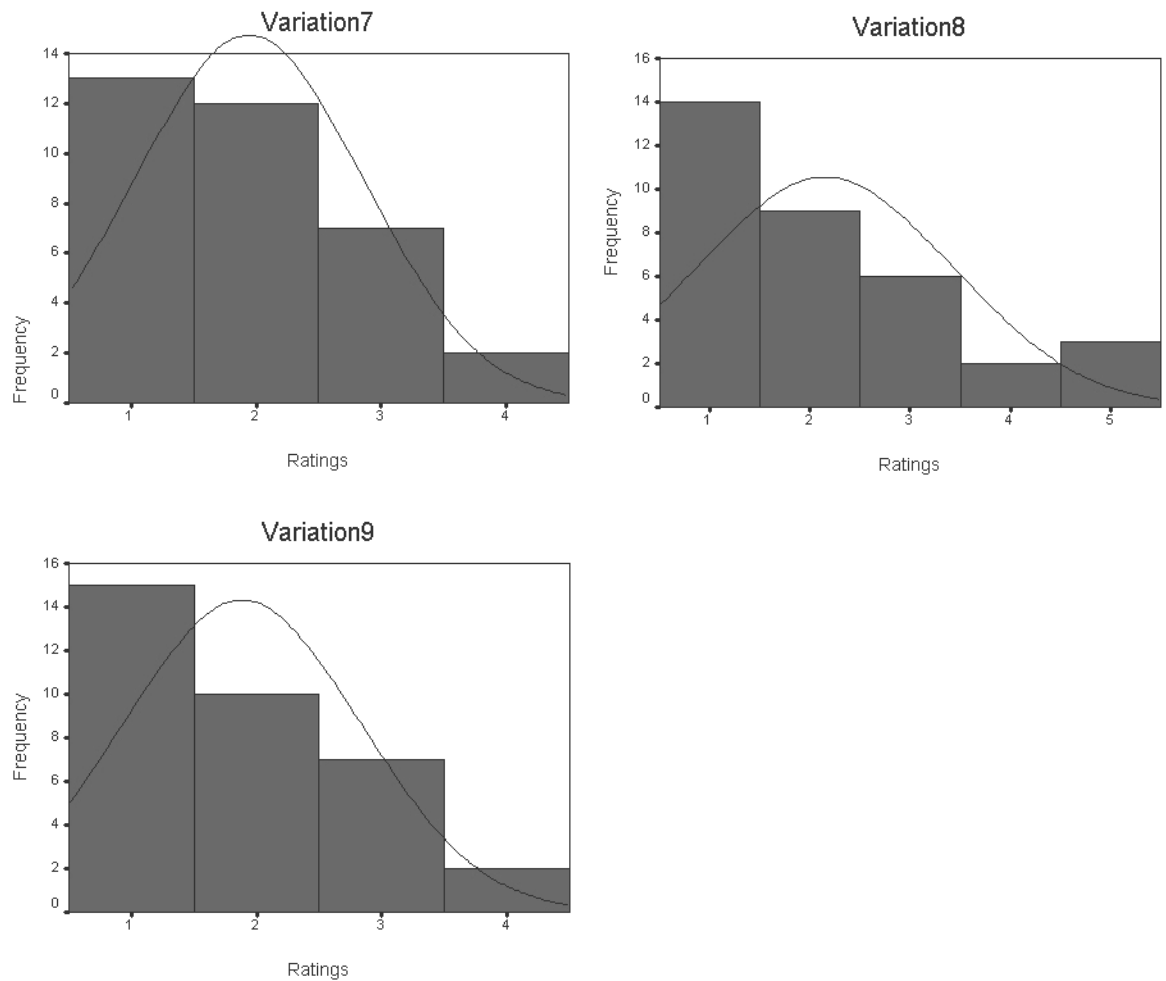
Part A ratings - sequential playing									
<b>Mode</b>	7	5	6	3	6	2	1	1	2
<b>Mean</b>	6.47	4.12	5.68	3.47	4.97	2.91	1.41	1.65	1.79
<b>Median</b>	7	4	6	3	5	3	1	1	2
<b>Range</b>	3	6	4	5	6	5	3	4	3
							<b>avg range</b>		4.3
<b>Std dev</b>	0.79	1.53	0.98	1.46	1.45	1.36	0.74	0.98	0.81
							<b>avg std dev</b>		1.12
Part A ratings - random playing									
<b>Mode</b>	7	5	6	4	6	3	1	1	1
<b>Mean</b>	6.41	4.71	5.97	3.85	5.53	3.59	1.94	2.15	1.88
<b>Median</b>	7	5	6	4	6	3	2	2	2
<b>Range</b>	3	6	3	5	5	5	3	4	3
							<b>avg range</b>		4.1
<b>Std dev</b>	0.82	1.24	0.83	1.33	1.16	1.40	0.92	1.28	0.95
							<b>avg std dev</b>		1.10
Difference between sequential and random ratings									
<b>Mode</b>	0	0	0	-1	0	-1	0	0	1
<b>Mean</b>	0.06	-0.59	-0.29	-0.38	-0.56	-0.68	-0.53	-0.50	-0.09
<b>Median</b>	0	-1	0	-1	-1	0	-1	-1	0
<b>Range</b>	0	0	1	0	1	0	0	0	0
<b>Std dev</b>	-0.03	0.29	0.14	0.13	0.28	-0.04	-0.18	-0.30	-0.14

Table B.1: Basic analysis of the ratings from all 34 subjects for Part A of the listening experiment.

Shapiro-Wilks test of normality			
	Statistic	df	Sig. (p)
<b>Variation 1</b>	.698	34	.000
<b>Variation 2</b>	.938	34	<b>.052</b>
<b>Variation 3</b>	.885	34	.002
<b>Variation 4</b>	.932	34	.035
<b>Variation 5</b>	.919	34	.015
<b>Variation 6</b>	.924	34	.021
<b>Variation 7</b>	.615	34	.000
<b>Variation 8</b>	.698	34	.000
<b>Variation 9</b>	.770	34	.000

Table B.2: The results for the Shapiro-Wilk test for normal distribution of the ratings from all 34 subjects for Part A (sequential).  $p < .05$  indicates a non-normal distribution.





**Figure B.2: The frequency distributions of the ratings given by all 34 subjects for Part A (random) of the listening experiment.**

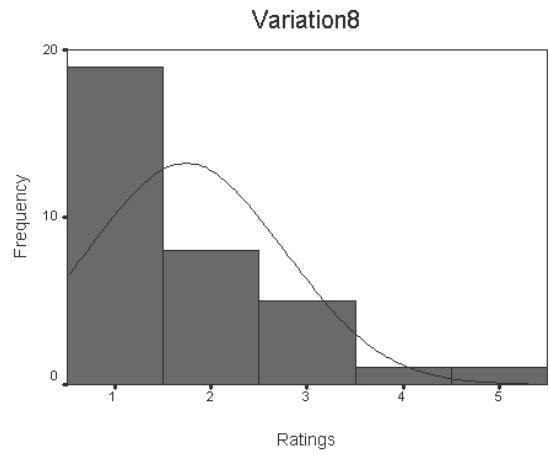
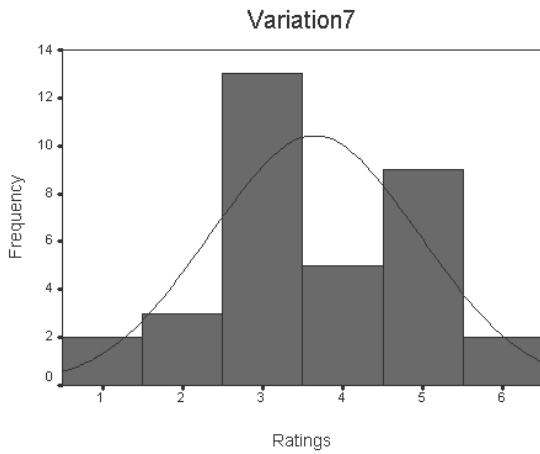
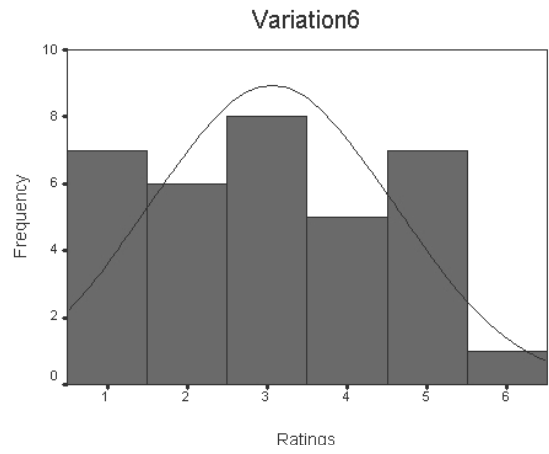
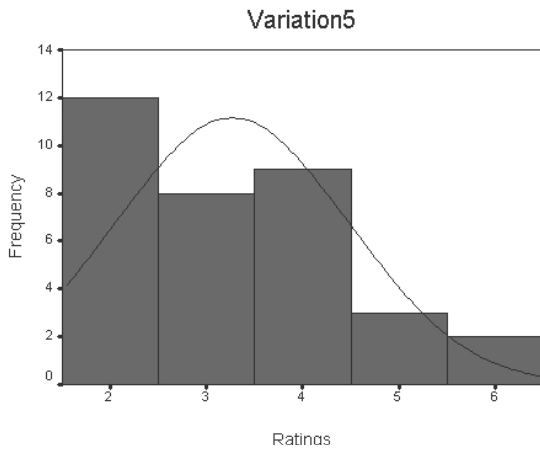
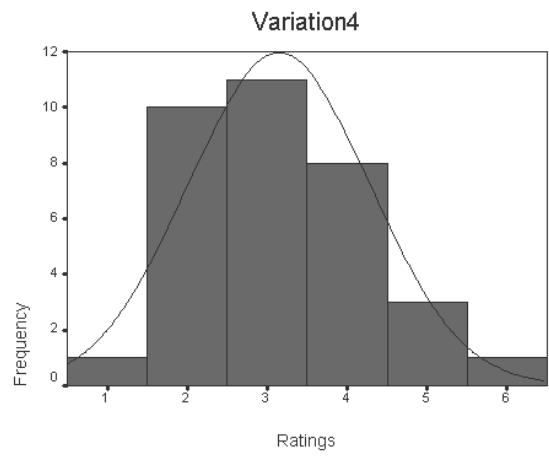
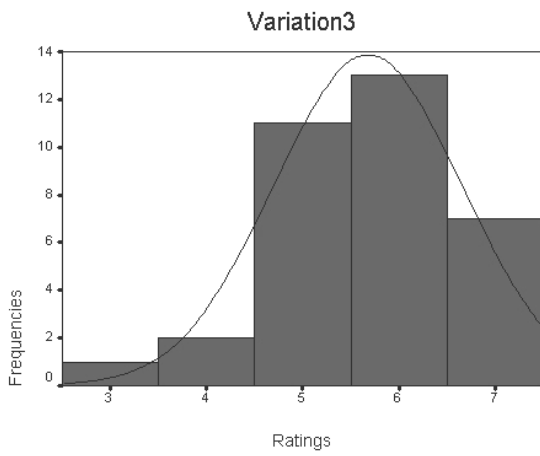
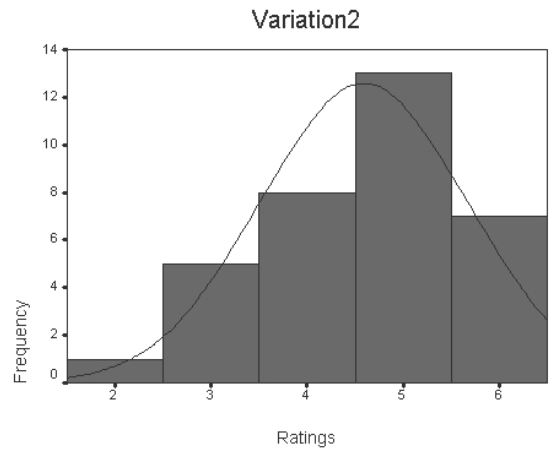
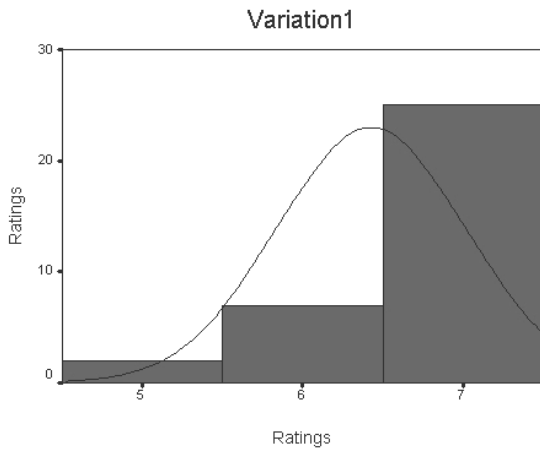
<b>Shapiro-Wilks test of normality</b>			
	<b>Statistic</b>	<b>df</b>	<b>Sig. (p)</b>
<b>Variation 1</b>	.725	34	.000
<b>Variation 2</b>	.901	34	.005
<b>Variation 3</b>	.851	34	.000
<b>Variation 4</b>	.924	34	.022
<b>Variation 5</b>	.887	34	.002
<b>Variation 6</b>	.929	34	.029
<b>Variation 7</b>	.837	34	.000
<b>Variation 8</b>	.816	34	.000
<b>Variation 9</b>	.815	34	.000

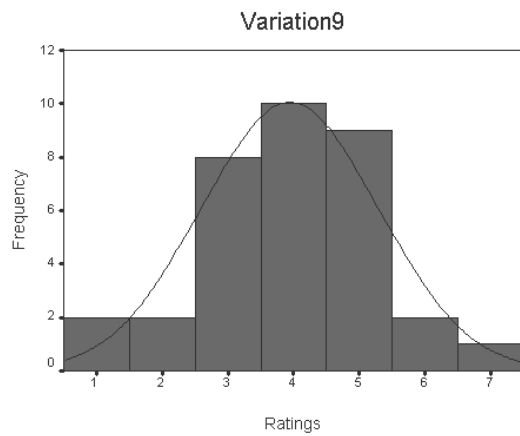
**Table B.3: The results for the Shapiro-Wilk test for normal distribution of the ratings from all 34 subjects for Part A (random).  $p < .05$  indicates a non-normal distribution.**

**Analysis of the Ratings from Part B for all 34 subjects (see sections 3.6.2)**

<b>Part B ratings – sequential playing</b>									
<b>Mode</b>	7	5	6	3	2	3	3	1	4
<b>Mean</b>	6.68	4.59	5.68	3.15	3.26	3.06	3.65	1.74	3.94
<b>Median</b>	7	5	6	3	3	3	3	1	4
<b>Range</b>	2	4	4	5	4	5	5	4	6
							avg range		4.3
<b>Std dev</b>	0.59	1.08	0.98	1.13	1.21	1.52	1.3	1.02	1.35
							avg std dev		1.13
<b>Part B ratings - random playing</b>									
<b>Mode</b>	7	4	6	5	5	4	4	1	5
<b>Mean</b>	6.5	4.71	5.94	3.85	4.29	3.59	3.71	2.21	4.53
<b>Median</b>	7	5	6	4	4.5	4	4	2	5
<b>Range</b>	4	5	3	4	5	6	5	4	4
							avg range		4.4
<b>Std dev</b>	0.83	1.24	0.85	1.26	1.19	1.4	1.31	1.34	1.16
							avg std dev		1.18
<b>Difference between sequential and random ratings</b>									
<b>Mode</b>	0	1	0	-2	-3	-1	-1	0	-1
<b>Mean</b>	0.18	-0.1	-0.3	-0.7	-1	-0.5	-0.1	-0.5	-0.6
<b>Median</b>	0	0	0	-1	-1.5	-1	-1	-1	-1
<b>Range</b>	-2	-1	1	1	-1	-1	0	0	2
<b>Std dev</b>	-0.2	-0.2	0.13	-0.1	0.02	0.12	-0	-0.3	0.19

**Table B.4: Basic analysis of the ratings from all 34 subjects for Part B of the listening experiment.**

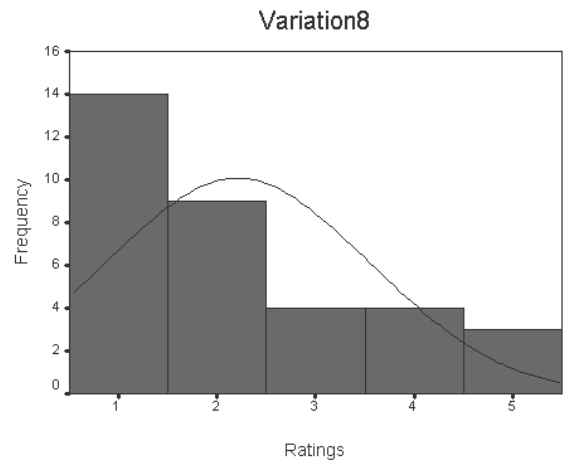
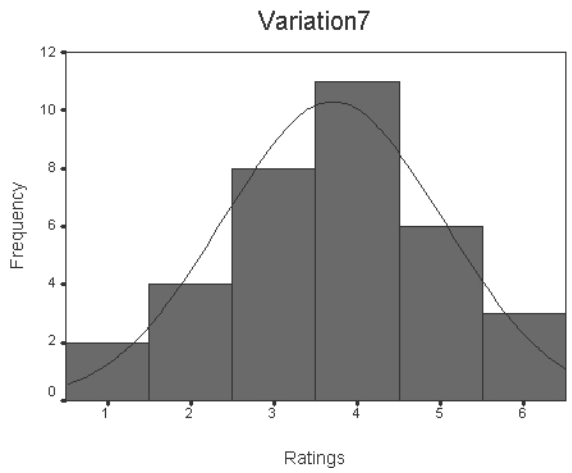
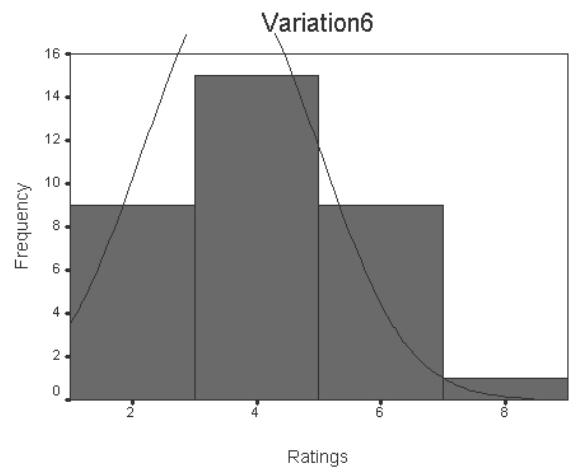
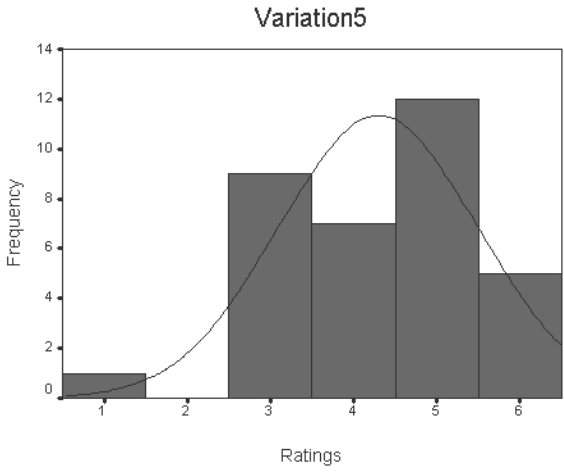
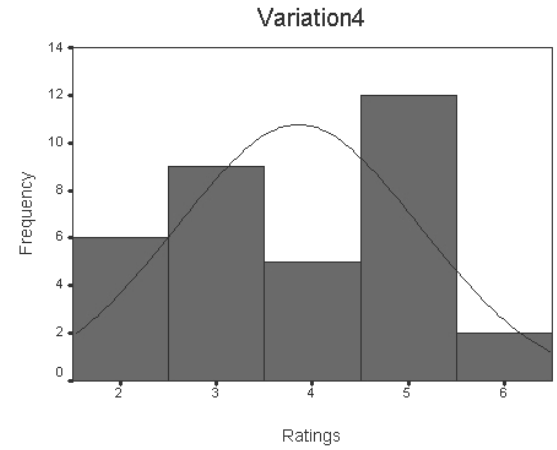
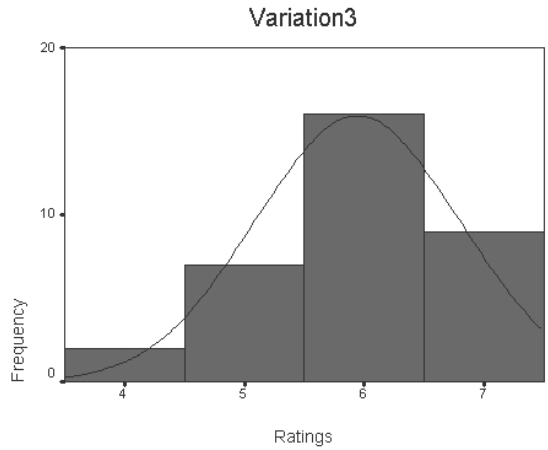
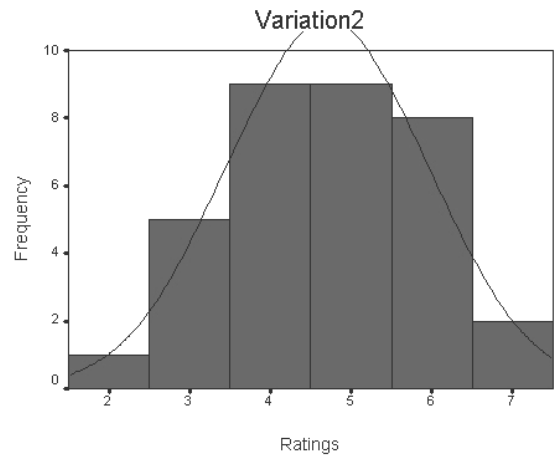
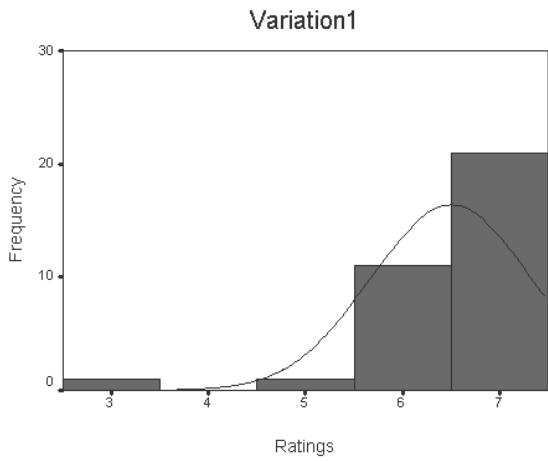


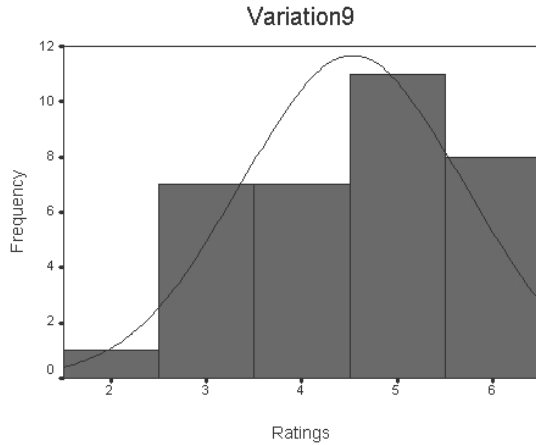


**Figure B.3:** The frequency distributions of the ratings given by all 34 subjects for Part B (sequential) of the listening experiment.

<b>Shapiro-Wilks test of normality</b>			
	<b>Statistic</b>	<b>df</b>	<b>Sig. (p)</b>
<b>Variation 1</b>	.592	34	.000
<b>Variation 2</b>	.895	34	.003
<b>Variation 3</b>	.885	34	.002
<b>Variation 4</b>	.915	34	.012
<b>Variation 5</b>	.863	34	.001
<b>Variation 6</b>	.910	34	.009
<b>Variation 7</b>	.916	34	.013
<b>Variation 8</b>	.739	34	.000
<b>Variation 9</b>	.945	34	<b>.087</b>

**Table B.5:** The results for the Shapiro-Wilk test for normal distribution of the ratings from all 34 subjects for Part B (sequential).  $p < .05$  indicates a non-normal distribution.





**Figure B.4:** The frequency distributions of the ratings given by all 34 subjects for Part B (random) of the listening experiment.

Shapiro-Wilks test of normality			
	Statistic	df	Sig. (p)
<b>Variation 1</b>	.615	34	.000
<b>Variation 2</b>	.938	34	<b>.054</b>
<b>Variation 3</b>	.853	34	.000
<b>Variation 4</b>	.880	34	.001
<b>Variation 5</b>	.894	34	.003
<b>Variation 6</b>	.927	34	.026
<b>Variation 7</b>	.941	34	<b>.066</b>
<b>Variation 8</b>	.816	34	.000
<b>Variation 9</b>	.891	34	.003

**Table B.6:** The results for the Shapiro-Wilk test for normal distribution of the ratings from all 34 subjects for Part B (random).  $p < .05$  indicates a non-normal distribution.

### Consistency of subjects (see section 3.6.3)

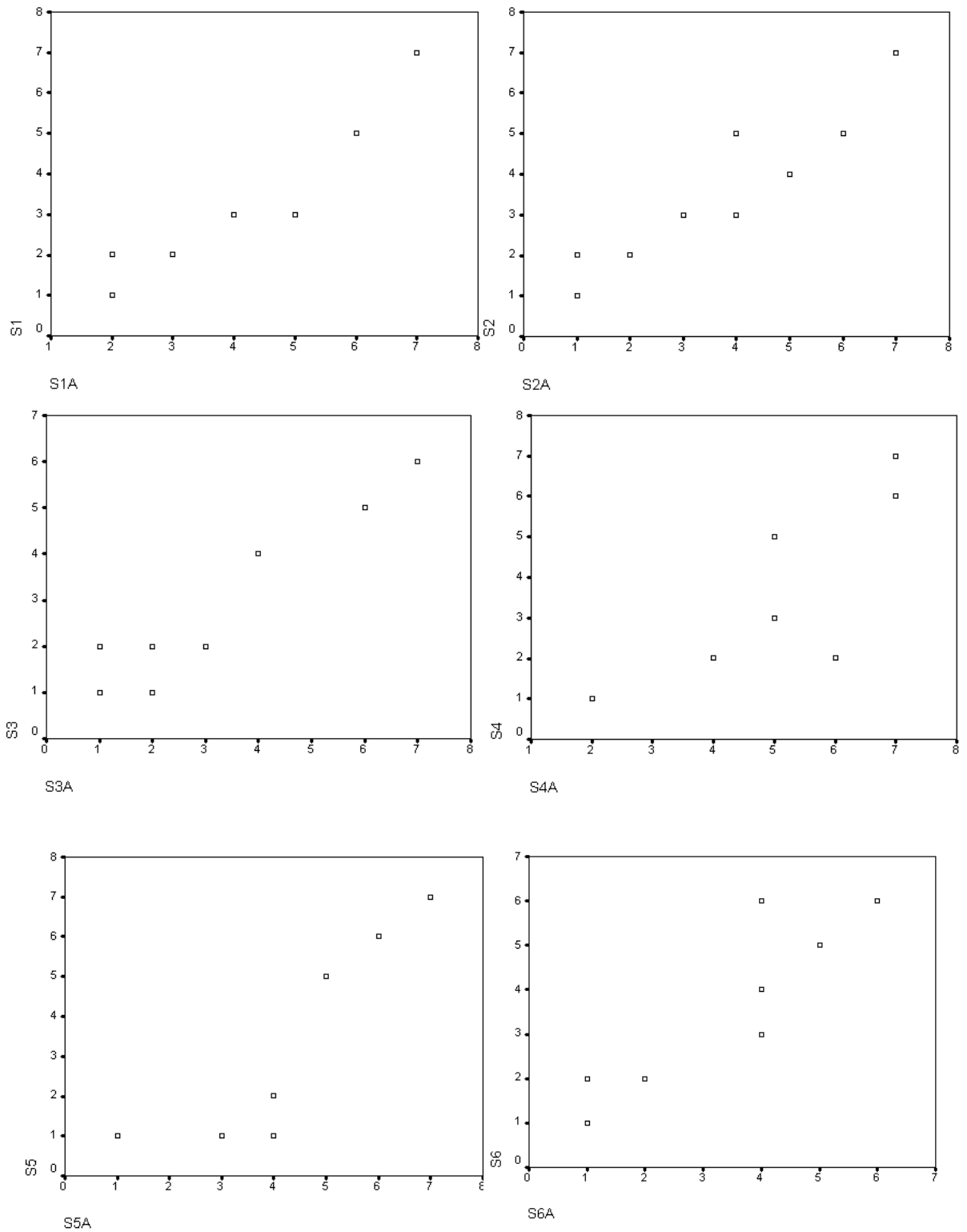


Figure B.5: A sample of scatterplots from the first six subjects' similarity judgements.

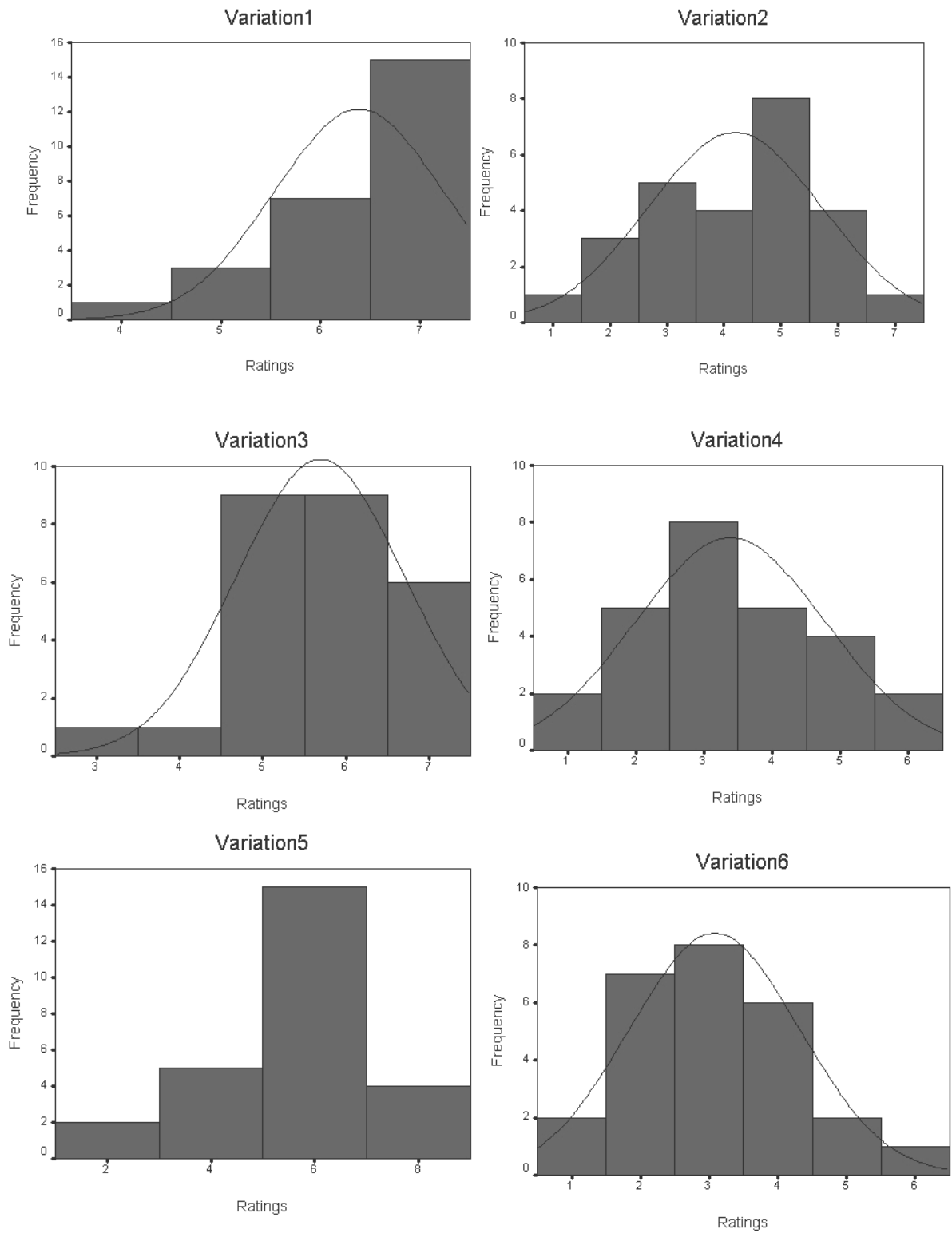
Subject	Correlation Coefficient (Spearman's rho)	Significance level
1	.97	.01
2	.932	.01
3	.901	.01
4	.849	.01
5	.905	.01
6	.9	.01
7	.987	.01
8	.94	.01
9	.491	<.05
10	.91	.01
11	.857	.01
12	.767	.05
13	.904	.01
14	.914	.01
15	.897	.01
16	.586	<.05
17	.292	<.05
18	.874	.01
19	.952	.01
20	.869	.01
21	.987	.01
22	.833	.01
23	.879	.01
24	.645	<.05
25	.668	.05
26	.805	.01
27	.921	.01
28	.628	<.05
29	.906	.01
30	.974	.01
31	.916	.01
32	.89	.01
33	.96	.01
34	.782	.05

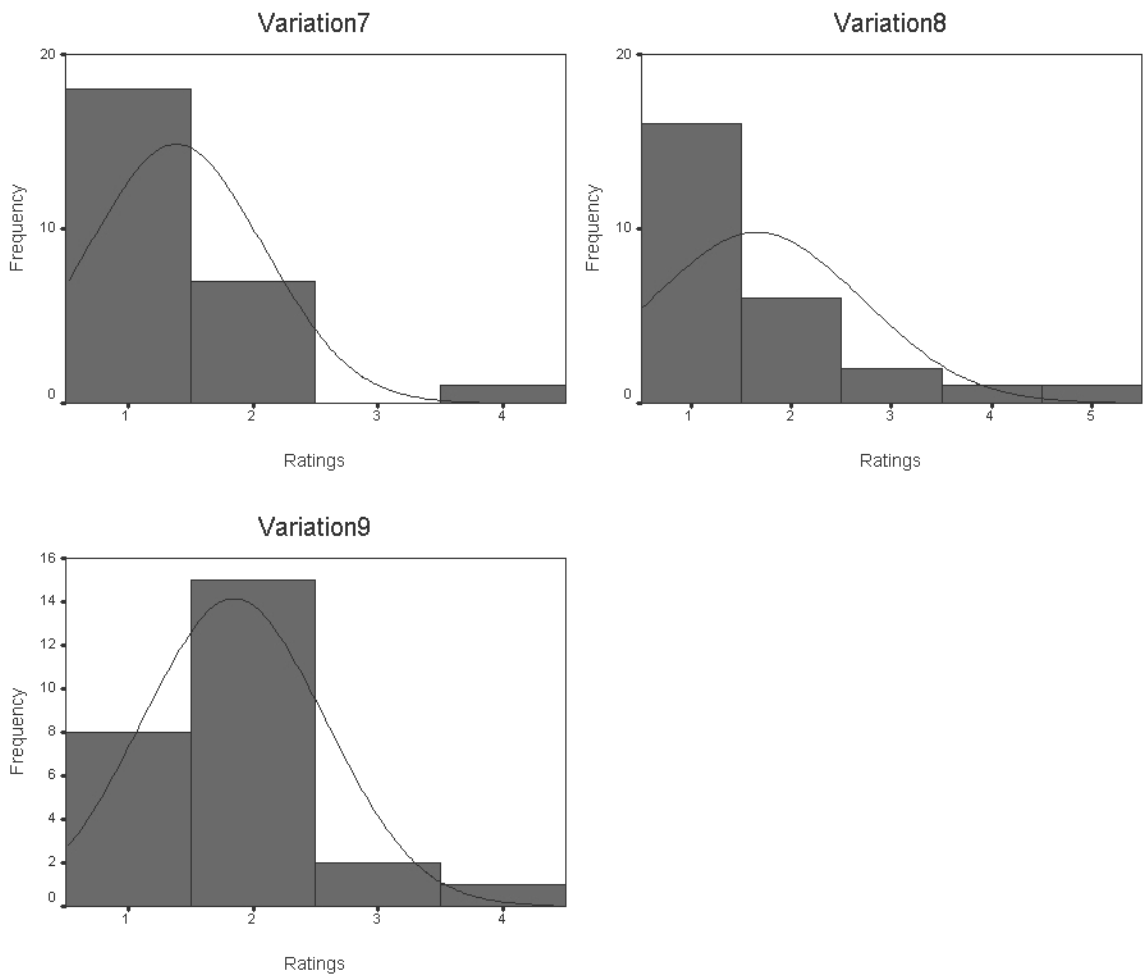
**Table B.7: Individual subject correlation between the ratings given for the 1st and 2nd playing of the melodies from Part A.**

Subject	Correlation Coefficient (Spearman's rho)	Significance level
1	.481	
2	.913	.01
3	.704	.05
4	.817	.01
5	.888	.01
6	.863	.01
7	.905	.01
8	.901	.01
9	.581	<.05
10	.849	.01
11	.839	.01
12	.911	.01
13	.690	.05
14	.949	.01
15	.839	.01
16	.848	.01
17	.346	<.05
18	.799	.01
19	.762	.05
20	.927	.01
21	.890	.01
22	.791	.05
23	.910	.01
24	.620	<.05
25	.441	<.05
26	.714	.05
27	.725	.05
28	.372	<.05
29	.772	.05
30	.320	<.05
31	.644	<.05
32	.723	.05
33	.623	<.05
34	.783	.05

**Table B.8: Individual subject correlation between the ratings given for the 1st and 2nd playing of the melodies from Part B.**

**The reduced data set for Part A - 26 subjects (see section 3.6.4)**





**Figure B.6: The frequency distributions of the ratings given by the reduced set of 26 subjects for Part A of the listening experiment.**

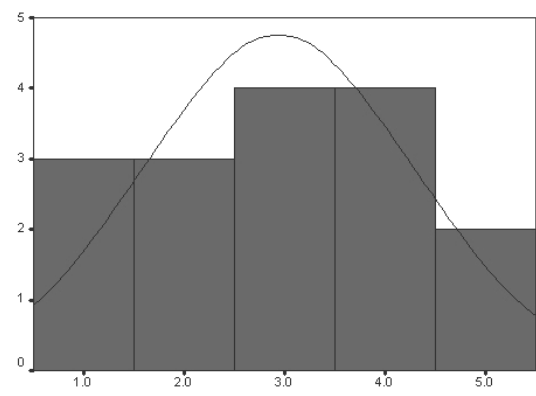
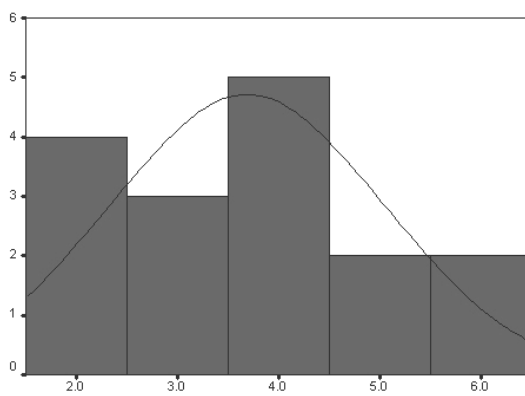
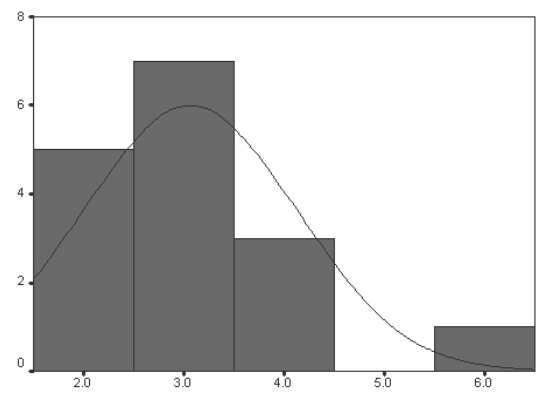
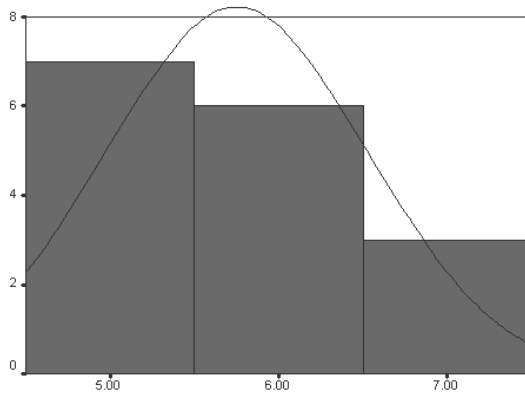
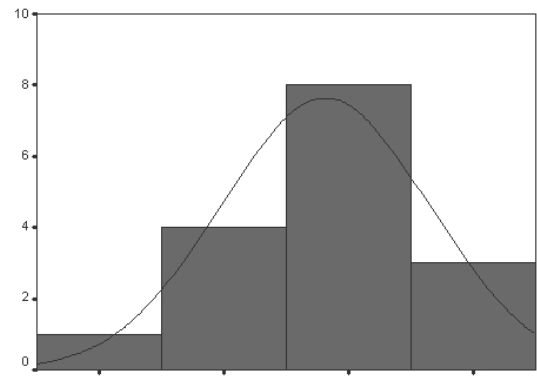
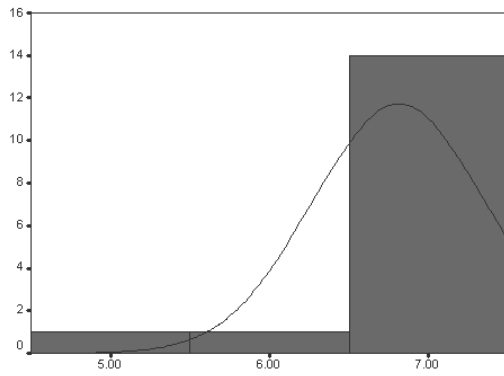
Part A ratings - sequential playing									
Mode	7	5	5	3	5	3	1	1	2
Mean	6.38	4.19	5.69	3.38	5.12	3.08	1.38	1.65	1.85
Median	7	4.5	6	3	5	3	1	1	2
Range	3	6	4	5	6	5	3	4	3
Std dev	0.85	1.52	1.01	1.39	1.45	1.23	0.70	1.06	0.73
Part A ratings - random playing									
Mode	7	5	6	4	6	3	1	1	1
Mean	6.62	4.73	6.15	4.12	5.62	3.692	1.73	1.96	1.88
Median	7	5	6	4	6	3	2	2	2
Range	3	6	2	4	5	5	2	4	3
Std dev	0.75	1.31	0.73	1.11	1.2	1.258	0.78	1.11	0.95
Difference between sequential and random ratings									
Mode	0	0	-1	-1	-1	0	0	0	1
Mean	-0.2	-0.5	-0.46	-0.73	-0.5	-0.62	-0.35	-0.3	-0
Median	0	-0.5	0	-1	-1	0	-1	-1	0
Range	0	0	2	1	1	0	1	0	0
Std dev	0.1	0.21	0.28	0.28	0.25	-0.03	-0.08	-0.1	-0.2

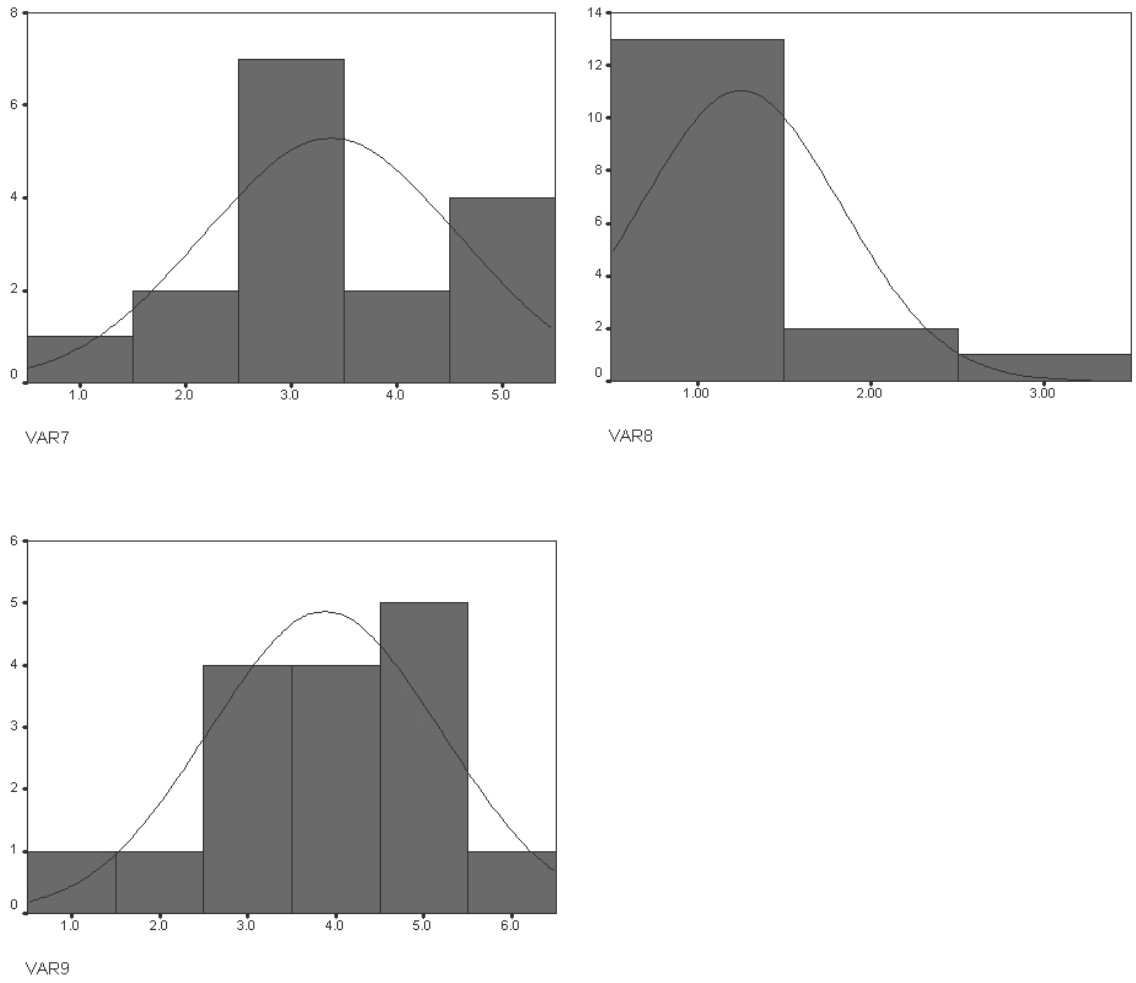
Table B.9: The basic analysis for the reduced set of 26 subjects - Part A (sequential).

Shapiro-Wilks test of normality			
	Statistic	df	Sig. (p)
Variation 1	.734	26	.000
Variation 2	.946	26	<b>.184</b>
Variation 3	.879	26	.005
Variation 4	.941	26	<b>.146</b>
Variation 5	.886	26	.008
Variation 6	.935	26	<b>.104</b>
Variation 7	.586	26	.000
Variation 8	.678	26	.000
Variation 9	.781	26	.000

Table B.10: The results for the Shapiro-Wilk test for normal distribution of the ratings from the reduced set of 26 subjects for Part A.  $p < .05$  indicates a non-normal distribution.

**The reduced data set for Part B - 16 subjects (see section 3.6.5)**





**Figure B.7: The frequency distributions of the ratings given by the reduced set of 16 subjects for Part A of the listening experiment.**

<b>Part B ratings – sequential playing</b>									
<b>Mode</b>	7	5	5	3	4	3	3	1	5
<b>Mean</b>	6.81	4.81	5.75	3.06	3.69	2.94	3.38	1.25	3.88
<b>Median</b>	7	5	6	3	4	3	3	1	4
<b>Range</b>	2	3	2	4	4	4	4	2	5
<b>Std dev</b>	0.54	0.83	0.77	1.06	1.35	1.34	1.20	0.58	1.31
<b>Part B ratings - random playing</b>									
<b>Mode</b>	7	5	6	5	5	5	4	1	5
<b>Mean</b>	6.94	5.06	6.19	3.63	4.81	3.44	3.44	1.38	4.31
<b>Median</b>	7	5	6	3.5	5	3.5	3.5	1	4.5
<b>Range</b>	1	4	2	4	3	4	4	2	4
<b>Std dev</b>	0.25	1.12	0.54	1.41	0.98	1.36	1.03	0.62	1.25
<b>Difference between sequential and random ratings</b>									
<b>Mode</b>	0	0	-1	-2	-1	-2	-1	0	0
<b>Mean</b>	-0.13	-0.25	-0.44	-0.56	-1.13	-0.50	-0.06	-0.13	-0.44
<b>Median</b>	0	0	0	-0.5	-1	-0.5	-0.5	0	-0.5
<b>Range</b>	1	-1	0	0	1	0	0	0	1
<b>Std dev</b>	0.29	-0.29	0.23	-0.35	0.37	-0.02	0.17	-0.04	0.06

Table B.11: The basic analysis for the reduced set of 16 subjects - Part B (sequential).

<b>Shapiro-Wilks test of normality</b>			
	<b>Statistic</b>	<b>df</b>	<b>Sig. (p)</b>
<b>Variation 1</b>	.405	16	.000
<b>Variation 2</b>	.872	16	.029
<b>Variation 3</b>	.793	16	.002
<b>Variation 4</b>	.816	16	.004
<b>Variation 5</b>	.904	16	<b>.094</b>
<b>Variation 6</b>	.918	16	<b>.158</b>
<b>Variation 7</b>	.889	16	<b>.055</b>
<b>Variation 8</b>	.507	16	.000
<b>Variation 9</b>	.932	16	<b>.259</b>

Table B.12: The results for the Shapiro-Wilk test for normal distribution of the ratings from the reduced set of 16 subjects for Part B.  $p < .05$  indicates a non-normal distribution.

### Consistency and Reliability of Ratings (see section 3.6.6)

Inter-subject correlation matrices for Part A and Part B. The results for the first sequential playing of the melodies only are given.

Inter-subject correlation using Spearman's correlation coefficient - Part A sequential 26 subjects																										
	S1A	S2A	S3A	S4A	S5A	S6A	S7A	S8A	S10A	S11A	S13A	S14A	S15A	S18A	S19A	S20A	S21A	S22A	S23A	S26A	S27A	S29A	S30A	S31A	S32A	S33A
S2A	0.96																									
S3A	0.87	0.82																								
S4A	0.80	0.85	0.55																							
S5A	0.89	0.92	0.66	0.77																						
S6A	0.93	0.96	0.72	0.94	0.92																					
S7A	0.89	0.88	0.95	0.59	0.74	0.76																				
S8A	0.93	0.96	0.75	0.93	0.92	0.99	0.77																			
S10A	0.95	0.89	0.90	0.66	0.77	0.82	0.93	0.80																		
S11A	0.92	0.96	0.75	0.88	0.91	0.94	0.83	0.96	0.79																	
S13A	0.57	0.68	0.41	0.44	0.60	0.54	0.57	0.49	0.61	0.52																
S14A	0.92	0.97	0.77	0.89	0.90	0.96	0.85	0.95	0.85	0.95	0.65															
S15A	0.89	0.94	0.86	0.76	0.86	0.89	0.93	0.90	0.86	0.90	0.60	0.95														
S18A	0.52	0.56	0.28	0.53	0.70	0.62	0.40	0.57	0.48	0.50	0.55	0.67	0.60													
S19A	0.88	0.91	0.90	0.68	0.81	0.82	0.95	0.86	0.84	0.92	0.46	0.87	0.94	0.35												
S20A	0.89	0.84	0.87	0.64	0.85	0.83	0.83	0.86	0.82	0.85	0.28	0.78	0.85	0.39	0.91											
S21A	0.91	0.94	0.93	0.69	0.81	0.84	0.96	0.87	0.88	0.90	0.55	0.88	0.94	0.35	0.98	0.89										
S22A	0.89	0.93	0.90	0.79	0.82	0.90	0.91	0.92	0.85	0.89	0.51	0.93	0.98	0.53	0.93	0.87	0.94									
S23A	0.92	0.97	0.76	0.79	0.95	0.93	0.80	0.93	0.84	0.90	0.70	0.91	0.89	0.58	0.84	0.84	0.89	0.88								
S26A	0.86	0.88	0.72	0.65	0.76	0.78	0.83	0.73	0.90	0.76	0.88	0.84	0.80	0.53	0.73	0.60	0.80	0.74	0.84							
S27A	0.96	0.94	0.79	0.78	0.85	0.89	0.89	0.86	0.94	0.90	0.69	0.93	0.88	0.56	0.85	0.77	0.86	0.84	0.87	0.93						
S29A	0.71	0.67	0.91	0.25	0.54	0.49	0.92	0.51	0.83	0.56	0.49	0.61	0.76	0.27	0.80	0.72	0.82	0.74	0.61	0.70	0.70					
S30A	0.97	0.97	0.83	0.75	0.92	0.91	0.88	0.91	0.91	0.92	0.68	0.91	0.89	0.49	0.89	0.87	0.93	0.87	0.97	0.89	0.93	0.71				
S31A	0.94	0.93	0.92	0.62	0.82	0.81	0.95	0.81	0.96	0.83	0.68	0.85	0.89	0.41	0.90	0.85	0.95	0.87	0.90	0.91	0.91	0.86	0.96			
S32A	0.80	0.84	0.64	0.82	0.80	0.85	0.75	0.88	0.67	0.96	0.32	0.85	0.82	0.37	0.88	0.78	0.82	0.80	0.75	0.60	0.81	0.46	0.79	0.69		
S33A	0.74	0.76	0.50	0.77	0.64	0.77	0.63	0.70	0.76	0.66	0.73	0.82	0.69	0.70	0.51	0.40	0.56	0.65	0.67	0.85	0.85	0.40	0.68	0.65	0.55	

Table B.13: The inter-subject correlation matrix for Part A (sequential).

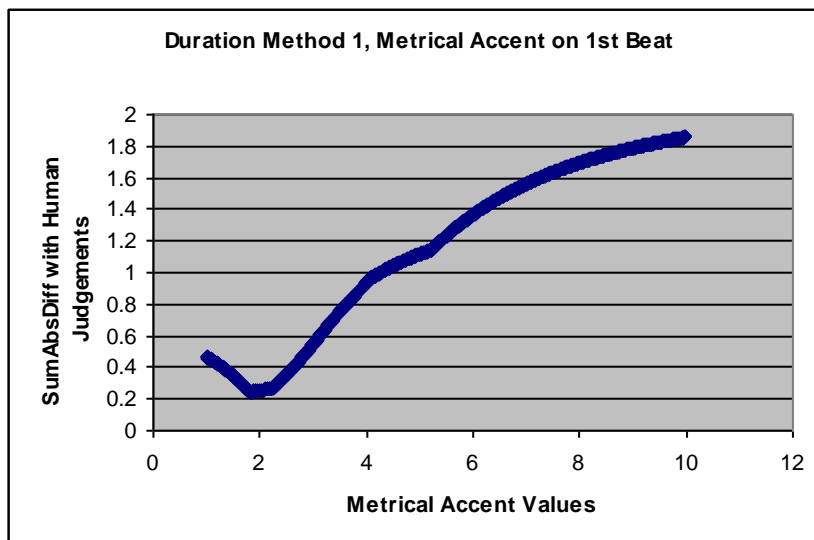
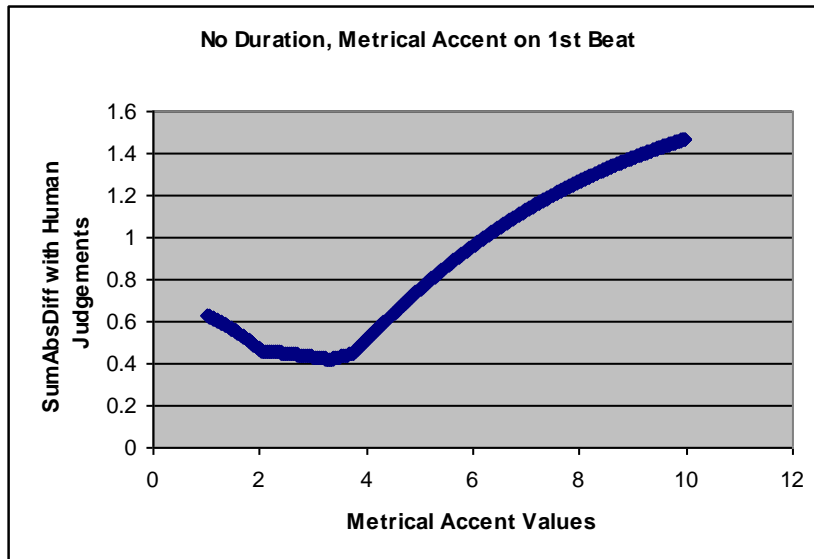
Inter-subject correlation using Spearman's correlation coefficient - Part B sequential 16 subjects																
	S2	S4	S5	S6	S7	S8	S10	S11	S12	S14	S15	S16	S18	S20	S21	S23
S2																
S4	0.43															
S5	0.89	0.52														
S6	0.40	0.92	0.52													
S7	0.32	0.86	0.31	0.89												
S8	0.31	0.66	0.49	0.87	0.62											
S10	0.76	0.66	0.90	0.64	0.36	0.58										
S11	0.85	0.75	0.91	0.67	0.52	0.54	0.90									
S12	0.65	0.83	0.78	0.80	0.61	0.69	0.87	0.94								
S14	0.34	0.72	0.56	0.89	0.61	0.97	0.69	0.62	0.78							
S15	0.59	0.92	0.69	0.88	0.80	0.71	0.73	0.88	0.93	0.75						
S16	0.59	0.85	0.65	0.86	0.80	0.65	0.69	0.70	0.67	0.66	0.81					
S18	0.85	0.61	0.84	0.50	0.39	0.34	0.84	0.94	0.86	0.46	0.71	0.51				
S20	0.52	0.92	0.61	0.93	0.82	0.80	0.71	0.83	0.93	0.84	0.97	0.76	0.69			
S21	0.83	0.78	0.88	0.75	0.60	0.65	0.89	0.96	0.91	0.68	0.91	0.80	0.83	0.86		
S23	0.88	0.42	0.87	0.35	0.21	0.31	0.76	0.90	0.78	0.39	0.64	0.36	0.93	0.59	0.80	

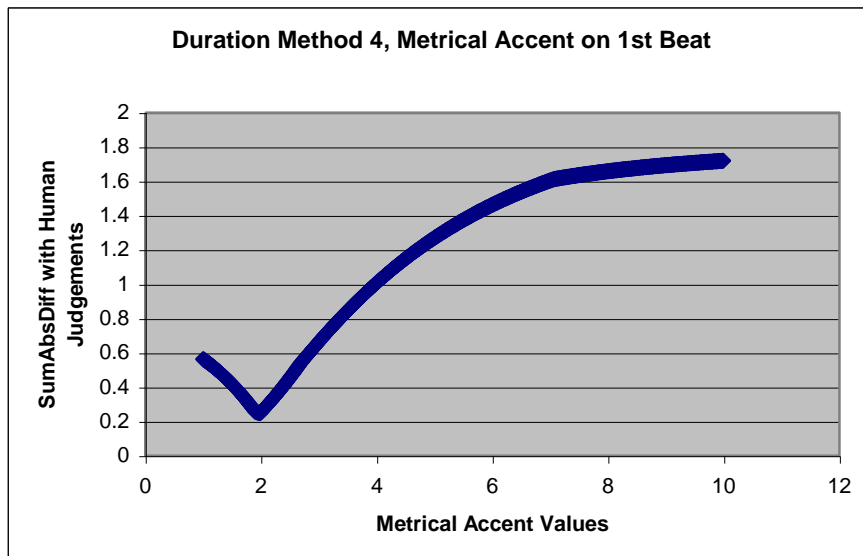
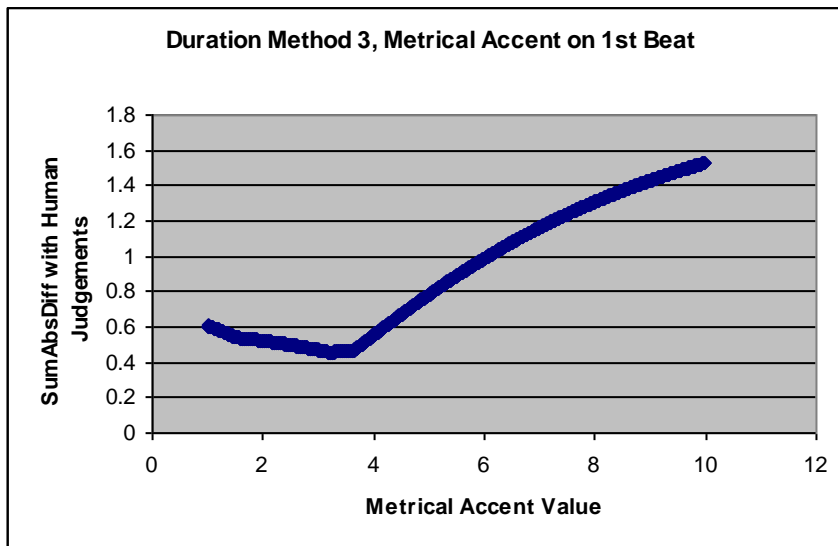
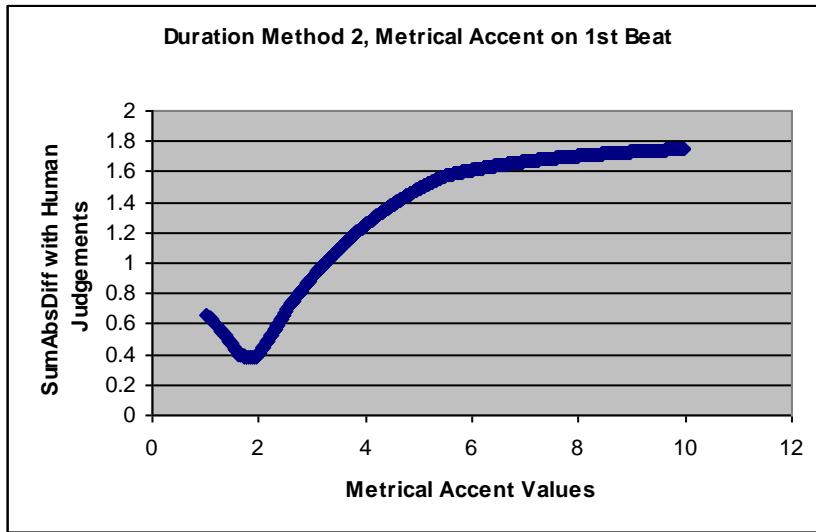
**Table B14: The inter-subject correlation matrix for Part B (sequential).**

## Appendix C

### Graphed Results for a Range of Metrical Accent Values

The graphs presented here illustrate an investigation of values for the metrical accent weight on the first beat of the bar in the geometric algorithm. See section 5.2 for details.





## Appendix D

### Sample Melodies from Ceol Rince na hÉireann

The first five tunes from Ceol Rince na hÉireann are shown here as a demonstration of the musical style of Irish folk music. The tunes from this collection of music are discussed in section 6.7.



Figure D.1: Cailleach an Túirne (The Maid/Hag of the Spinning Wheel)



Figure D.2: Pléaraca na Céise (The Humours of Kesh)

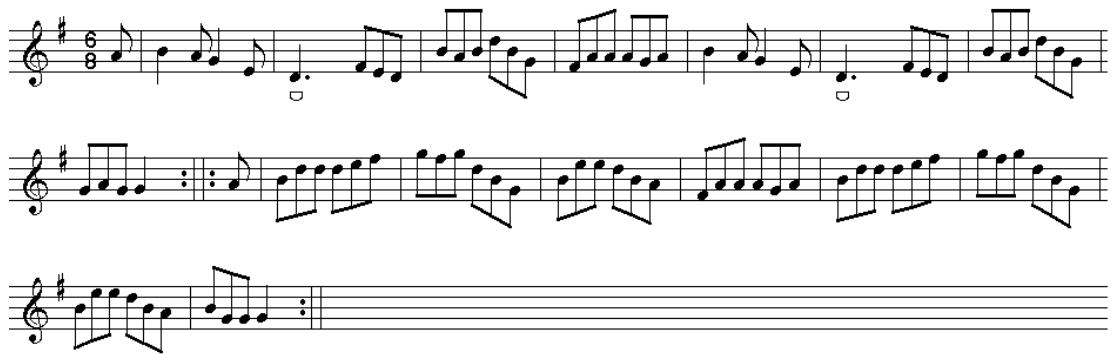


Figure D.3: Carraig an tSoip

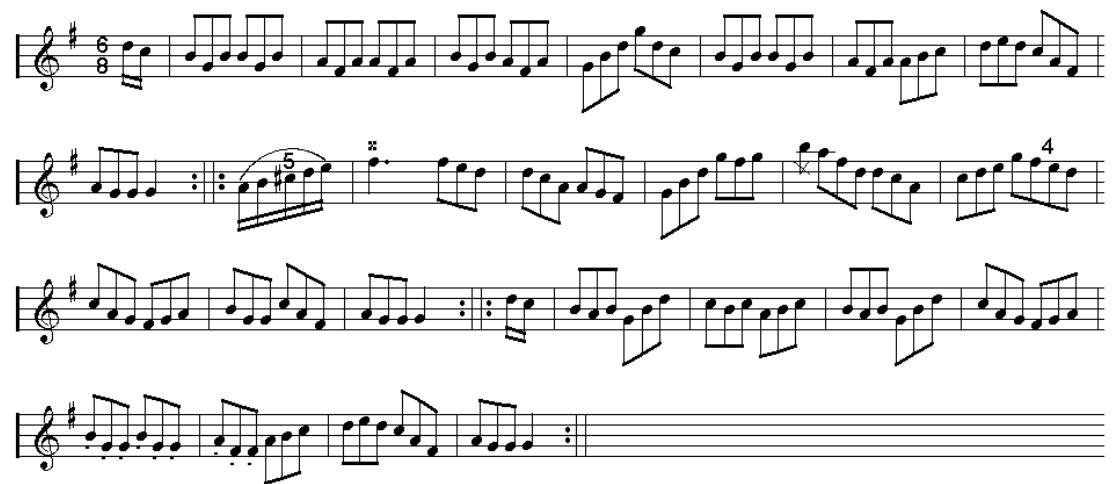


Figure D.4: Pingneacha Rua agus Prás (Coppers and Brass)

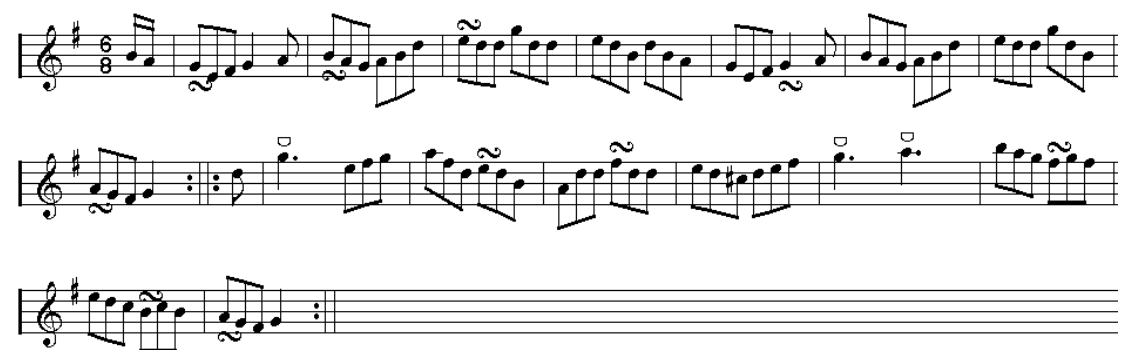


Figure D.5: Gleanntán na Samhaircíní (The Primrose Glen)

## Appendix E

### Publications

The following published conference papers are related to the work presented in this thesis:

Cahill, M. and Ó Maidín, D. (2007) ‘Melodic Similarity Algorithms – Computational Musicology for Music Analysis’, *The Fifth Annual Conference of the Society for Musicology in Ireland (SMI 2007)*, Dublin, 11-13 May.

Cahill, M. and Ó Maidín, D. (2006a) ‘Identifying Successful Melodic Similarity Algorithms for use in Music Retrieval’, *Proceedings of the Conference on Multidisciplinary Sciences and Technologies (INSCIT 2006)*, Merida, Spain, October 25-28, 355-356.

Cahill, M. and Ó Maidín, D. (2006b) ‘Assessing the Performance of Melodic Similarity Algorithms Using Human Judgments of Similarity’, in Lemström, K., Tindale, A. and Dannenberg, R. eds., *Proceedings of the 7th International Conference on Music Information Retrieval (ISMIR 2006)*, Victoria, BC, Canada, October 8-12, 355-356.

Cahill, M. and Ó Maidín, D. (2005) ‘Melodic similarity algorithms – using similarity ratings for development and early evaluation’, in Reiss, J. and Wiggins, G., eds., *Proceedings of the 6th International Conference on Music Information Retrieval (ISMIR 2005)*, Queen Mary, University of London & Goldsmiths College, September 11-15, London: University of London, 450-453.

## References

- Bartlett, J. C. and Dowling, W. J. (1988), 'Scale structure and similarity of melodies', *Music Perception* 5(3), 285–314.
- Blackburn, S., and DeRoure, D. (1998) 'A Tool for Content Based Navigation of Music', in *Proceedings of the 6th ACM International Conference on Multimedia '98*, September 12-16, Bristol, England, New York: ACM Press, 361-368.
- Bonebright, T., Miner, N., Goldsmith T. and Caudell, T. (1998) 'Data Collection and Analysis Techniques for Evaluating the Perceptual Qualities of Auditory Stimulus', in *Proceedings of the International Conference on Auditory Display (ICAD '98)*, Glasgow, United Kingdom, November 1-4, 505-516.
- Breathnach, B. (1963) *Ceol Rince na hEireann Vol. 1*, Dublin, Ireland: Department of Education.
- Byrd, D. (2001) 'Music-Notation Searching and Digital Libraries' in *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries (JCDL'01)*, June 24-28, 2001, Roanoke, Virginia, USA, New York: ACM Press, 239-246.
- Cahill, M. and Ó Maidín, D. (2005) 'Melodic similarity algorithms – using similarity ratings for development and early evaluation', in Reiss, J. and Wiggins, G., eds., *Proceedings of the 6th International Conference on Music Information Retrieval (ISMIR 2005)*, Queen Mary, University of London & Goldsmiths College, September 11-15, London: University of London, 450-453.
- Cahill, M. and Ó Maidín, D. (2006a) 'Assessing the Performance of Melodic Similarity Algorithms Using Human Judgments of Similarity', in Lemström, K., Tindale, A. and Dannenberg, R. eds., *Proceedings of the 7th International Conference on Music Information Retrieval (ISMIR 2006)*, Victoria, BC, Canada, October 8-12, 355-356.
- Cahill, M. and Ó Maidín, D. (2006b) 'Identifying Successful Melodic Similarity Algorithms for use in Music Retrieval', *Proceedings of the Conference on Multidisciplinary Sciences and Technologies (INSCIT 2006)*, Merida, Spain, October 25-28, 355-356.
- Cahill, M. and Ó Maidín, D. (2007) 'Melodic Similarity Algorithms – Computational Musicology for Music Analysis', *The Fifth Annual Conference of the Society for Musicology in Ireland (SMI 2007)*, Dublin, 11-13 May.
- Cambouropoulos, E. (2001) 'Melodic Cue Abstraction, Similarity, and Category Formation: A Formal Model', *Music Perception*, 18(3), 347-370.
- Chen, A.L.P. (2003) 'Music IR 201: Music Retrieval and Analysis', International Conference on Music Information Retrieval (ISMIR 2003) Tutorial, Baltimore, Maryland, USA, October 26-30, Baltimore, Maryland, available [http://make.cs.nthu.edu.tw/alp/ismir03\\_tutorial\\_new.ppt](http://make.cs.nthu.edu.tw/alp/ismir03_tutorial_new.ppt) [accessed 2 November 2007].

- Chen, H., Chih-Hsiang, L. and Chen, A.L.P. (2004) 'Music segmentation by rhythmic features and melodic shapes' *Proceedings of the 2004 IEEE International Conference on Multimedia and Expo (ICME 2004)*, 27-30 June, Taipei, Taiwan, 1643-1646.
- Chew, E. (2001) 'Modeling Tonality: Applications to Music Cognition', Proceedings of the 23rd Annual Meeting of the Cognitive Science Society, Edinburgh, Scotland, 1-4 Aug, NJ/London: Erlbaum, 206-211.
- Clausen, M., Engelbrecht, R., Meyer, D. and Schmitz, J. (2000) 'PROMS: A Web-based Tool for Searching in Polyphonic Music', in Byrd, D., *Proceedings of the 1<sup>st</sup> International Symposium on Music Information Retrieval (Music IR 2000/ISMIR 2000)*, Plymouth, Massachusetts, October 23-25.
- Conklin, D and Anagnostopoulou, C. (2006) 'Segmental Pattern Discovery in Music', *INFORMS Journal on Computing*, 18(3), 285-293.
- Crawford, T., Iliopoulos, C. and Raman, R. (1998) 'String-Matching Techniques for Musical Similarity and Melodic Recognition', *Computing in Musicology*, 11, 73-100.
- Cuddy, L.L, Cohen, A.J. and Miller, J. (1979) 'Melody Recognition: The Experimental Application of Musical Rules', *Canadian Journal of Psychology*, 33(3), 148-157.
- Cuddy, L.L., Cohen, A.J. and Mewhort, D.J.K. (1981) 'Perception of structure in short melodic sequences', *Journal of Experimental Psychology: Human Perception & Performance*, 7, 869-883.
- De Grae, P (2000) A Translation of the Irish text of Breathnach's Ceol Rince na hEireann Vol. 1, available <http://www.nigelgatherer.com/books/CRE/cre1.html> [15 August 2007].
- Demorest, S. and Kim, D. (2002) 'The Effect of Temporal Versus Melodic Cues on Recall of Familiar Songs', in Stevens, C., Burnham, D., McPherson, G., Schubert, E. and Renwick, J., eds., *Proceedings of the 7<sup>th</sup> International Conference on Music Perception and Conognition (ICMPC 2002)*, Sydney, Australia, July 17-21, Sydney: AMPS and Causal Productions (CD-ROM).
- Deutsch, D. (1972) 'Octave Generalization and Tune Recognition', *Perception and Psychophysics*, 11(6), 411-412.
- DeWitt, L. and Crowder, R. (1986) 'Recognition of Novel Melodies after Brief Delays', *Music Perception*, 3, 259-274.
- Doraisamy, S. and Ruger, S. (2001) 'An Approach Towards A Polyphonic Music Retrieval System', in Downie, J.S. and Bainbridge, D., eds., *Proceedings of the 2nd International Symposium on Music Information Retrieval (ISMIR 2001)*, Bloomington, Indiana, USA, October 15-17, 187-193.
- Dovey, M. (2001) 'A Technique for Regular Expression Style Searching in Polyphonic Music', in Downie, J.S. and Bainbridge, D., eds., *Proceedings of the*

- 2nd International Symposium on Music Information Retrieval (ISMIR 2001)*,  
Bloomington, Indiana, USA, October 15-17, 179-185.
- Dowling, J. and Fujitani, D. (1971) 'Contour, Interval, and Pitch Recognition in Memory for Melodies', *Journal of the Acoustical Society of America (JASA)*, 49(2), 524-531.
- Dowling, J. (1978) 'Scale and Contour: Two Components of a Theory of Memory for Melodies', *Psychological Review*, 85(4), 341-354.
- Dowling, J. and Harwood, D. (1986) *Music Cognition*, London: Academic Press.
- Dowling, J., Tillman, B. and Ayers, D. (2002) Memory and the Experience of Hearing Music. *Music Perception*, 19 (2), 249-276.
- Downie, J.S. (1999) 'Music Retrieval as Text Retrieval: Simple yet Effective', in *Proceedings of the ACM Special Interest Group on Information Retrieval Conference (SIGIR '99)*, Berkeley, California, United States, August 15 - 19, New York: ACM Press, 297-298.
- Downie, S. and Nelson, M. (2000) 'Evaluation of a Simple and Effective Music Information Retrieval Method', in Belkin, N., Ingwersen, P. and Leong, M., eds., *Proceedings of the ACM Special Interest Group on Information Retrieval Conference (SIGIR 2000)*, Athens, Greece, July 24-28, New York: ACM Press, 73-80.
- Downie, J. S. (2006) 'The Music Information Retrieval Evaluation eXchange (MIREX)', *D-Lib Magazine*, 12(12), December 2006, available <http://www.dlib.org/dlib/december06/downie/12downie.html> [accessed 5 December 2007].
- Drake, C., Dowling, W.J., Palmer, C. (1991) 'Accent Structures in the Reproduction of Simple Tunes by Children and Adult Pianists', *Music Perception*, 8(3), 315-334.
- Drake, C. and Palmer, C. (1993) 'Accent Structures in Music Performance', *Music Perception*, 10(3), 343-378.
- Deutsch, D. (1999) 'The Processing of Pitch Combinations' in Deutsch, D., ed., *The Psychology of Music*, 2<sup>nd</sup> Edition, San Diego: Academic Press, 349-411.
- Duschenes, M. (1962) 'Variations on Twinkle, Twinkle, Little Star', from *Method for the Recorder – Tunes and Exercises*, Ontario, Canada: Berandol Music Limited.
- Edworthy, J. (1985) 'Interval and Contour in Melodic Processing', *Music Perception*, 2, 375-388.
- Eerola, T., & Bregman, M. (2007) 'Melodic and contextual similarity of folk song phrases.' *Musicae Scientiae*, Discussion Forum 4A-2007, 211–233.
- Eerola, T., Järvinen, T., Louhivuori, J. and Toivainen, P. (2001) 'Statistical Features and Perceived Similarity of Folk Melodies', *Music Perception*, 18(3), 275-296.

- EsAC, Essen Associative Code and Folksong Database, available: <http://www.esac-data.org/> [accessed June 2005].
- Francès, R. (1988). *The Perception of Music*, translated by Dowling, W.J., Hillsdale, New Jersey: Erlbaum.
- Futrelle, J. and Downie, J.S. (2002) 'Interdisciplinary Communities and Research Issues in Music Information Retrieval', in Fingerhut, M., ed., *Proceedings of the Third International Conference on Music Information Retrieval (ISMIR 2002)*, Paris, France, October 13-17, Paris: IRCAM, 215-221.
- Ghias, A., Logan, J., Chamberlin, D., and Smith, B. (1995) 'Query By Humming: Music Information Retrieval in an Audio Database', in *Proceedings of the Third ACM International Conference on Multimedia '95*, November 5-9, San Francisco, California, New York: ACM Press, 231-236.
- Hébert, S. and Peretz, I. (1997) 'Recognition of music in long-term memory: Are melodic and temporal patterns equal partners?', *Memory & Cognition*, 25, 518–533.
- Hershman, D. (1994) 'Key Distance Effects in Ecological Contexts', in Deliège, I., ed., *Proceedings of the 3<sup>rd</sup> International Conference on Music Perception and Cognition (ICMPC 3)*, Liège, Belgium, Liège: ESCOM, 243-244.
- Hofman-Engl, L. (2001) 'Towards a Cognitive Model of Melodic Similarity', in Downie, J.S. and Bainbridge, D., eds., *Proceedings of the 2nd International Symposium on Music Information Retrieval (ISMIR 2001)*, Bloomington, Indiana, USA, October 15-17, 143-151.
- Hofman-Engl, L. (2002) 'Melodic Similarity – A Conceptual Framework', in Stevens, C., Burnham, D., McPherson, G., Schubert, E. and Renwick, J., eds., *Proceedings of the 7<sup>th</sup> International Conference on Music Perception and Cognition (ICMPC 2002)*, Sydney, Australia, July 17-21, Sydney: AMPS and Causal Productions (CD-ROM).
- Hofman-Engl, L. (2003) *Melodic Similarity and Transformations: A Theoretical and Empirical Approach*, unpublished thesis (Ph.D.), Keele University.
- Hoed, M. den and Nooijer, J. de (2004) *Orpheus - Music Retrieval. Obtaining ground truth*, University of Utrecht report, available: [http://www.cs.uu.nl/research/projects/i-cult/Publications/GroundTruth\\_Hoed.pdf](http://www.cs.uu.nl/research/projects/i-cult/Publications/GroundTruth_Hoed.pdf) [accessed July 2005].
- Hoos, H., Renz, K. and Görg, M. (2001) 'GUIDO/MIR – an Experimental Musical Information Retrieval System based on GUIDO Music Notation', in Downie, J.S. and Bainbridge, D., eds., *Proceedings of the 2nd International Symposium on Music Information Retrieval (ISMIR 2001)*, Bloomington, Indiana, USA, October 15-17, 41-50.
- Huron, D and Parncutt, R. (1993) 'An Improved Model of Tonality Perception Incorporating Pitch Saliency and Echoic Memory', *Psychomusicology*, 12(2), 154-171.

- Huron, D and Royal, M. (1996) 'What is Melodic Accent? Converging Evidence from Musical Practice', *Music Perception*, 13(4), 489-516.
- Jones, M. R (1987) 'Dynamic Pattern Structure in Music: Recent Theory and Research', *Perception & Psychophysics*, 41(6), 621-634.
- Kern Scores, A library of virtual musical scores in the Humdrum \*\*kern data format, available: <http://kern.ccarh.org/> [accessed 10/7/2006].
- Lamont, A. & Dibben, N. (2001) 'Motivic Structure and the Perception of Similarity', *Music Perception*, 18(3,) 245-274.
- Lartillot, O. (2007) 'Automated extraction of motivic patterns and application to the analysis of Debussy's *Syrinx*', *International Conference of the Society for Mathematics and Computation in Music*, May 18-20, Berlin.
- Lartillot, O. (2005) 'Efficient Extraction of Closed Motivic Patterns in Multi-Dimensional Symbolic Representations of Music', *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence (WI'05)*, Compiègne, Washington:IEEE Computer Society Press, 239-235.
- Lemström, K. (2000) *String Matching Techniques for Music Retrieval*, thesis (Ph.D.), Report A-2000-04, Department of Computer Science, University of Finland.
- Lemström, K. & Ukkonen, E. (2000) 'Including Interval Encoding into Edit Distance Based Music Comparison and Retrieval', *Proceedings of the AISB'00 Symposium on Creative & Cultural Aspects and Applications of AI & Cognitive Science*, April, Birmingham, 53-60.
- Lemström, K., Mäkinen, V., Pienimäki, A., Turkia, M. and Ukkonen, E. (2003) 'The C-BRAHMS Project', in Hoos, H. and Bainbridge, D., eds., *Proceeding of the Fourth International Conference on Music Information Retrieval (ISMIR 2003)*, Baltimore, Maryland, USA, October 26-30, Baltimore, Maryland: The Johns Hopkins University, 237-238.
- Lemström, K., Mikkilä, N., Mäkinen, V. and Ukkonen, E. (2006), 'Sweepline and Recursive Geometric Algorithms for Melodic Similarity', MIREX 2006 results, available: [http://www.music-ir.org/evaluation/MIREX/2006\\_abstracts/SMS\\_mikkila.pdf](http://www.music-ir.org/evaluation/MIREX/2006_abstracts/SMS_mikkila.pdf) [accessed 20 September 2007].
- Lerdahl, F. and Jackendoff, R. (1983) *A Generative Theory of Tonal Music*, Cambridge, Massachusetts: MIT Press.
- Levitin, D. (1994) 'Absolute Memory for Musical Pitch: Evidence from the Production of Learned Melodies', *Perception & Psychophysics*, 56(4), 414-423.
- Levitin, D. (1999), 'Absolute Pitch: Self-Reference and Human Memory', *International Journal of Computing and Anticipatory Systems*, 4, 255-266.
- Levitin, D. and Cook, P. (1996) 'Memory for Musical Tempo: Additional Evidence that Auditory Memory is Absolute', *Perception & Psychophysics*, 58, 927-935.

- Mazzoni, D. and Dannenberg, R. (2001) 'Melody Matching Directly From Audio', in Downie, J.S. and Bainbridge, D., eds., *Proceedings of the 2nd International Symposium on Music Information Retrieval (ISMIR 2001)*, Bloomington, Indiana, USA, October 15-17, 17-18.
- McAdams, S. and Matzkin, D. (2001) 'Similarity, Invariance, and Musical Variation', in Zatorre, R. and Peretz, I., eds., *The Biological Foundations of Music*, New York: Academy of Sciences.
- McNab, R., Smith, L., Bainbridge, D., & Witten, I. (1997) 'The New Zealand Digital Library MELody Index', in D-Lib magazine, 3(5), May 1997, available: <http://www.dlib.org/dlib/may97/meldex/05witten.html> [accessed 13 August 2006].
- Melucci, M., Orio, N. (2002) 'A Comparison of Manual and Automatic Melody Segmentation' in Fingerhut, M., ed., *Proceedings of the Third International Conference on Music Information Retrieval (ISMIR 2002)*, Paris, France, October 13-17, Paris: IRCAM, 7-14.
- Melucci, M., Orio, N. (2004) 'Combining Melody Processing and Information Retrieval Techniques: methodology, evaluation, and system implementation', *Journal of the American Society for Information Science and Technology (JASIST)*, 55(12), 1058-1066.
- MIREX 2005, The 1st Annual Music Information Retrieval Evaluation eXchange, run in conjunction with the 6th ISMIR Conference in London, United Kingdom, September 11-15, available: [http://www.music-ir.org/mirex/2005/index.php/Main\\_Page](http://www.music-ir.org/mirex/2005/index.php/Main_Page) [accessed 7 February 2006].
- Mongeau, M. and Sankoff, D. (1990) 'Comparison of Musical Sequences', *Computers and the Humanities*, 24, 161-175.
- Müllensiefen, D. and Frieler, K. (2004a) 'Measuring Melodic Similarity: Human vs. Algorithmic Judgements', in Parncutt, R., Kessler, A. and Zimmer, F., eds., *Proceedings of the First Conference on Interdisciplinary Musicology (CIM '04)*, Graz, Austria, April 15-18, Graz: Department of Musicology, University of Graz, available [http://www-gewi.uni-graz.at/staff/parncutt/cim04/CIM04\\_paper\\_pdf/Muellensiefen\\_Frieler\\_CIM04\\_proceedings.pdf](http://www-gewi.uni-graz.at/staff/parncutt/cim04/CIM04_paper_pdf/Muellensiefen_Frieler_CIM04_proceedings.pdf) [accessed 10 March 2006].
- Müllensiefen, D. and Frieler, K. (2004b) 'Optimizing Measures of Melodic Similarity for the Exploration of a Large Folk Song Database', in Lomeli Buyoli, C. and Loureiro, R., eds., *Proceedings of the 5th International Conference on Music Information Retrieval (ISMIR 2004)*, Barcelona, Spain, October 10-14, Barcelona: Universitat Pompeu Fabra, 274-280.
- MuseData, An Electronic Library of Classical Musical Scores, available: <http://www.musedata.org/>, [accessed 10/7/2006].
- Ó Maidín, D. (1995) *A Programmer's Environment for Music Analysis*, unpublished thesis (Ph.D), University College, Cork, Ireland.

- Ó Maidín, D. (1982) 'Computer Analysis of Irish and Scottish Jigs', in Baroni, M., Brunetti, R., Callegari, L. and Jacoboni C. eds., *Musical Grammars and Computer Analysis*, Firenze: Olschki, pp.329-336.
- Ó Maidín, D. (1998a) 'A Geometrical Algorithm for Melodic Difference', *Computing in Musicology 11*, 65-72.
- Ó Maidín, D. (1998b) *Common Practice Notation View User's Manual*, Technical Report UL-CSIS-98-02, University of Limerick.
- Ó Maidín, D. and Cahill, M. (2001) 'Score Processing for MIR', in Downie, J.S. and Bainbridge, D., eds., *Proceedings of the 2nd International Symposium on Music Information Retrieval (ISMIR 2001)*, Bloomington, Indiana, USA, October 15-17, 59-64.
- O'Neill, F. (1907) *The Dance Music of Ireland*, Chicago: Lyon & Healy.
- Orpen, K. and Huron, D. (1992) 'Measurement of Similarity in Music: A Quantitative Approach for Non-parametric Representations', *Computers in Music Research*, Vol. 4, (Fall 1992), 1-44.
- Pampalk, E., Dixon, S., and Widmer, G. (2003) 'Exploring Music Collections by Browsing Different Views', in *Proceedings of the 6th International Conference on Digital Audio Effects (DAFx-03)*, London, United Kingdom, September 8-11, 201-208, available <http://www.elec.qmul.ac.uk/dafx03/proceedings/pdfs/dafx02.pdf> [accessed 18 March 2004].
- Parncutt, R. (1997) 'Modeling piano performance: Physics and cognition of a virtual pianist', in *Proceedings of the International Computer Music Conference (ICMC 1997)*, Thessaloniki, Greece, September 25-30, 15-18.
- Parncutt, R. (2003) 'Accents and Expression in Piano Performance', in Niemöller, K.W., ed. , *Perspektiven und Methoden einer Systemischen Musikwissenschaft (Festschrift Fricke)*, Germany: Peter Lang, 163-185, available [http://www.gewi.uni-graz.at/staff/parncutt/publications/Pa03\\_AccentExpression.pdf](http://www.gewi.uni-graz.at/staff/parncutt/publications/Pa03_AccentExpression.pdf) [accessed 23 September 2006].
- Pickens, J. (2001) *A Survey of Feature Selection Techniques for Music Information Retrieval*, Center for Intelligent Information Retrieval (CIIR) Technical IR-238, Massachusetts: CIIR, University of Massachusetts, available <http://ciir-publications.cs.umass.edu/getpdf.php?id=238> [accessed 4 November 2007].
- Robine, M., Hanna, P., Ferraro, P. and Allali, J. (2007) 'Adaptation of String Matching Algorithms for Identification of Near-Duplicate Music Documents', in Stein, B., Koppel, M. and Stamatatos, E. eds., *Proceedings of SIGIR '07, Workshop on Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection (PAN)*, 27 July 2007, Amsterdam.
- Rosner, B. and Meyer, L. (1986) 'The Perceptual Roles of Melodic Process, Contour, and Form', *Music Perception*, 4(1), 1-40.

- Schulkind, M. D. (1999) 'Long term memory for temporal structure: Evidence from the identification of well-known and novel songs', *Memory & Cognition*, 27, 896-906.
- Schulkind, M., Posner, R. and Rubin, D. (2003) 'Music Features that Facilitate Melody Identification: How do you Know it's "your" Song when they Finally Play It?', *Music Perception*, 21(2), 217-249.
- Selfridge-Field, E. (2004) 'Towards a Measure of Cognitive Distance in Melodic Similarity', *Computing in Musicology* 13, 93-111.
- Shaughnessy, J., Zechmeister, E. and Zechmeister J. (2006) *Research Methods in Psychology*, 7th edition, London: McGraw.
- Smith, L., McNab, R. and Witten, I. (1998) 'Sequence-based Comparison: A Dynamic-Programming Approach', *Computing in Musicology*, 11, 101-118.
- Soulez, F., Rodet, X. and Schwarz, D. (2003) 'Improving Polyphonic and Poly-Instrumental Music to Score Alignment' in Proceedings of the *4th International Conference on Music Information Retrieval (ISMIR 2003)*, Baltimore, Maryland, USA, October 27-30, 143-148.
- Themefinder, A collaborative project between the Center for Computer Assisted Research in the Humanities (CCARH) at Stanford University and the Cognitive and Systematic Musicology Laboratory at the Ohio State University, available: <http://www.ccarh.org/themefinder/> and <http://www.themefinder.org/> [accessed 1 September 2006]].
- Thomassen, J. (1982) 'Melodic Accent: Experiments and a Tentative Model', *Journal of the Acoustical Society of America (JASA)*, 71(6), 1596-1605.
- Typke, R., Giannopoulos, P., Veltkamp, R., Wiering, F. and van Oostrum, R. (2003) 'Using Transportation Distances for Measuring Melodic Similarity', in Proceedings of the *4th International Conference on Music Information Retrieval (ISMIR 2003)*, Baltimore, Maryland, USA, October 27-30, 107-114.
- Typke, R., Veltkamp, R. and Wiering, F. (2004a) 'Searching Polyphonic Music Using Transportation Distances', in *Proceedings of ACM Multimedia 2004*, 128-135.
- Typke, R., den Hoed, M., de Nooijer, J., Wiering, F. and Veltkamp, R.C. (2004b) *A ground truth For Half A Million Musical Incipits*, Technical UU-CS-2004-060, Institute of Information and Computing Sciences, Utrecht University.
- Typke, R., Wiering, F. and Veltkamp, R. (2005a) 'A Survey of Music Information Retrieval Systems', in Reiss, J. and Wiggins, G., eds., *Proceedings of the 6th International Conference on Music Information Retrieval (ISMIR 2005)*, Queen Mary, University of London & Goldsmiths College, September 11-15, London: University of London, 153-160.
- Typke, R., den Hoed, M., de Nooijer, J., Wiering, F. and Veltkamp, R. (2005b) 'A ground truth For Half A Million Musical Incipits', in *Proceedings of the 5th*

- Dutch-Belgian Information Retrieval Workshop (DIR) 2005*, Utrecht, the Netherlands, 63-70.
- Typke, R., Wiering, F. and Veltkamp, R. (2006) ‘Evaluating The Earth Mover’s Distance For Measuring Symbolic Melodic Similarity’, MIREX 2006 results, available: <http://www.music-ir.org/evaluation/mirex-results/articles/similarity/typke.pdf> [accessed 4 November 2007].
- Uitdenbogerd, A., and Zobel, J. (1998) ‘Manipulation of Music for Melody Matching’, *Proceedings of the sixth ACM international conference on Multimedia 1998*, Bristol, United Kingdom, September 13 - 16, New York: ACM Press, 235-240.
- Uitdenbogerd, A., and Zobel, J. (1999) ‘Melodic Matching Techniques for Large Music Databases’, *Proceedings of the seventh ACM international conference on Multimedia 1999*, Orlando, Florida, United States, October 30 - November 05, New York: ACM Press, 57-66.
- van Egmond, E. and Povel, D. (1994) ‘Similarity Judgments on Transposed Melodies as a Function of Overlap and Key-distance’, in Deliège, I., ed., *Proceedings of the 3<sup>rd</sup> International Conference on Music Perception and Cognition (ICMPC 3)*, Liège, Belgium, Liège: ESCOM, 219-220.
- van Egmond, E. and Povel, D. (1996) ‘The Influence of Height and Key on the Perceptual Similarity of Transposed Melodies’, *Perception & Psychophysics*, 58(8), 1252-1259.
- Weyde, T. (2001) ‘Grouping, Similarity and the Recognition of Rhythmic Structure’, in *Proceedings of the International Computer Music Conference (ICMC) 2001*, Havana, Cuba, September 18-22, 475-478.
- White, B. W. (1960) ‘Recognition of Distorted Melodies’, *American Journal of Psychology*, 73(1), 100–107.
- Witte, R. and Witte, J. (2004) *Statistics*, 7<sup>th</sup> Edition, New York:Wiley.