

2016-6

## Proceedings of the 6th International Workshop on Folk Music Analysis, 15-17 June, 2016

Pierre Beauguitte

*Technological University Dublin, pierre.beauguitte@mydit.ie*

Bryan Duggan

*Technological University Dublin, bryan.duggan@tudublin.ie*

John D. Kelleher

*Technological University Dublin, john.d.kelleher@tudublin.ie*

Follow this and additional works at: <https://arrow.tudublin.ie/fema>

 Part of the [Musicology Commons](#)

---

### Recommended Citation

Beauguitte, P., Duggan, B., Kelleher, J. (eds.).(2016).*Proceedings of the 6th International Workshop on Folk Music Analysis*, Dublin, 15-17 June, 2016. ISBN: 978-1-900454-59-9

This Conference Paper is brought to you for free and open access by the 6th International Workshop on Folk Music Analysis, 15-17 June, 2016 at ARROW@TU Dublin. It has been accepted for inclusion in Papers by an authorized administrator of ARROW@TU Dublin. For more information, please contact [yvonne.desmond@tudublin.ie](mailto:yvonne.desmond@tudublin.ie), [arrow.admin@tudublin.ie](mailto:arrow.admin@tudublin.ie), [brian.widdis@tudublin.ie](mailto:brian.widdis@tudublin.ie).



This work is licensed under a [Creative Commons Attribution-NonCommercial-Share Alike 3.0 License](#)





6TH INTERNATIONAL WORKSHOP

# Folk Music Analysis

15 - 17 June 2016



DIT GRANGEGORMAN - DUBLIN

<http://fma-2016.sciencesconf.org> - hashtag #fmadit16



## Workshop organization

### Organizing committee

BEAUGUITTE Pierre (DIT, Dublin, Ireland)  
DUGGAN Bryan (DIT, Dublin, Ireland)  
KELLEHER John (DIT, Dublin, Ireland)

### Scientific committee

ADAM Olivier (UPMC, Paris, France)  
BEAUGUITTE Pierre (DIT, Dublin, Ireland)  
BENETOS Emmanouil (Queen Mary University, London, UK)  
BONINI BARALDI Filippo (CREM-LESC, Université Paris-Ouest Nanterre La Défense, France  
& Instituto de Etnomusicologia, Universidade Nova de Lisboa, Portugal)  
BURGOYNE John Ashley (University of Amsterdam, The Netherlands)  
CAMBOUROPOULOS Emilios (Aristotle University of Thessaloniki, Greece)  
CARROLL David (DIT, Dublin, Ireland)  
CAZAU Dorian (ENSTA Bretagne, Brest, France)  
CONKLIN Darrell (University of the Basque Country UPV/EHU, Donostia - San Sebastián,  
Spain)  
CULLEN Charlie (DIT, Dublin, Ireland)  
DELANY Sarah Jane (DIT, Dublin, Ireland)  
DOVAL Boris (LAM - d'Alembert, Paris, France)  
DUGGAN Bryan (DIT, Dublin, Ireland)  
FILLON Thomas (Parisson, Paris, France)  
GOMEZ Emilia (Universitat Pompeu Fabra, Barcelona, Spain)  
HOLZAPFEL Andre (Bogazici University, Istanbul, Turkey)  
KELLEHER John (DIT, Dublin, Ireland)  
MAROLT Matija (University of Ljubljana, Slovenia)  
O'SHEA Brendan (Dublin, Ireland)  
PICARD François (Paris-Sorbonne University, France)  
PIKRAKIS Aggelos (University of Piraeus, Greece)  
PINQUIER Julien (IRIT, Toulouse, France)  
SHIELDS Lisa (Dublin, Ireland)  
SU Norman Makoto (Indiana University Bloomington, USA)  
VAN KRANENBURG Peter (Meertens Institute, Amsterdam, The Netherlands)  
VOLK Anja (Utrecht University, The Netherlands)  
WALSHAW Chris (Old Royal Naval College, London, UK)  
WEYDE Tillman (City University London, UK)

### Invited speakers

VOLK Anja (Utrecht University, The Netherlands)  
BROWNE Peter (DIT, Dublin, Ireland)

## Folk Music Analysis 2016

The Folk Music Analysis workshop brings together computational music analysis and ethnomusicology. Both symbolic and audio representations of music are considered, with a broad range of scientific approaches being applied (signal processing, graph theory, deep learning). The workshop features a range of interesting talks from international researchers in areas such as Indian classical music, Iranian Singing, Ottoman-Turkish Makam music scores, Flamenco singing, Irish traditional music, Georgian traditional music and Dutch folk songs.

The 6th International Workshop on Folk Music Analysis, FMA 2016, is organised by a team of researchers from the School of Computing of the Dublin Institute of Technology (DIT) and hosted in the new DIT Grangegorman campus from 15th to 17th June 2016.

Some members of our committee have been involved in FMA from the very start whilst others are new to this community. Hence we have a continuity with previous workshops while also gaining new insights.

This year we established a collaboration between FMA and the AAWM (Analytical Approaches to World Music) journal. Our shared scientific interests led us to the following arrangement: a best paper will be elected and announced at the end of the FMA workshop, and its author(s) will be guaranteed to publish a revised and extended version of the paper in a forthcoming edition of the AAWM journal.



We want to thank the DIT School of Computing and School of Media for the financial support, as well as the Grangegorman campus who provided us with a great venue. Thank you also to the DIT Conservatory of Music and Drama, and to the Zurmukhti choir for their musical performances during the workshop. Thank you to all committee members and authors for the very interesting and diverse scientific program we had this year. Finally, thank you to everyone who participated in some way to this edition of FMA!

Pierre Beauguitte, Bryan Duggan and John Kelleher  
<http://fma-2016.sciencesconf.org>

Cover illustration by Olivier Chén  , 2016

# Contents

<b>Keynote: <i>Computational pattern search in folk music: challenges and opportunities</i></b>	<b>6</b>
<i>Anja Volk</i>	
<b>Keynote: <i>Tuning the radio</i></b>	<b>7</b>
<i>Peter Browne</i>	
<b>A revaluation of learning practices in Indian classical music using technological tools</b>	<b>8</b>
<i>Julien Debove, Dorian Cazau, Olivier Adam</i>	
<b>Detection of melodic patterns in automatic transcriptions of flamenco singing</b>	<b>14</b>
<i>Aggelos Pikrakis, Nadine Kroher, José Miguel Díaz-Báñez</i>	
<b>Publishing the James Goodman Irish music manuscript collection: how modern technology facilitated the editors' task</b>	<b>18</b>
<i>Lisa Shields</i>	
<b>Constructing proximity graphs to explore similarities in large-scale melodic datasets</b>	<b>22</b>
<i>Chris Walshaw</i>	
<b>Note, cut and strike detection for traditional Irish flute recordings</b>	<b>30</b>
<i>Islah Ali-MacLachlan, Maciej Tomczak, Jason Hockman, Carl Southall</i>	
<b>Formalising vocal production cross-culturally</b>	<b>36</b>
<i>Polina Proutskova, Christophe Rhodes, Tim Crawford, Geraint Wiggins</i>	
<b>A structure analysis method for Ottoman-Turkish Makam music scores</b>	<b>39</b>
<i>Sertan Şentürk, Xavier Serra</i>	
<b>After the <i>Harmonie Universelle</i> by Marin Mersenne (1636), what fingering for the chabrette in 2016?</b>	<b>47</b>
<i>Philippe Randonneix</i>	
<b>NeoMI: a new environment for the organization of musical instruments</b>	<b>50</b>
<i>Carolien Hulshof, Xavier Siebert, Hadrien Mélot</i>	
<b>Closed patterns in folk music and other genres</b>	<b>56</b>
<i>Iris Ren</i>	
<b>A graph-theoretical approach to the harmonic analysis of Georgian vocal polyphonic music</b>	<b>59</b>
<i>Frank Scherbaum, Simha Arom, Frank Kane</i>	
<b>Towards flamenco style recognition: the challenge of modelling the aficionado</b>	<b>61</b>
<i>Nadine Kroher, José-Miguel Díaz-Báñez</i>	

<b>A search through time: connecting live playing to archive recordings of traditional music</b>	<b>64</b>
<i>Bryan Duggan, Jianghan Xu, Lise Denbrok, Breandan Knowlton</i>	
<b>Human pattern recognition in data sonification</b>	<b>67</b>
<i>Charlie Cullen</i>	
<b>A pattern mining approach to study a collection of Dutch folk songs</b>	<b>71</b>
<i>Peter van Kranenburg, Darrell Conklin</i>	
<b>The Georgian traditional music system</b>	<b>74</b>
<i>Malkhaz Erkvandze</i>	
<b>On the benefit of larynx-microphone field recordings for the documentation and analysis of polyphonic vocal music</b>	<b>80</b>
<i>Frank Scherbaum</i>	
<b>Automatic alignment of long syllables in a cappella Beijing opera</b>	<b>88</b>
<i>Georgi Dzhambazov, Yile Yang, Rafael Caro Repetto, Xavier Serra</i>	
<b>Analysis of Tahreer in traditional Iranian singing</b>	<b>92</b>
<i>Parham Bahadoran</i>	
<b>Segmentation of folk songs with a probabilistic model</b>	<b>96</b>
<i>Ciril Bohak, Matija Marolt</i>	

**Keynote: Anja Volk, *Computational pattern search in folk music: challenges and opportunities***

**Abstract**

In this talk I address current challenges and opportunities of computational analysis of folk music, by taking the specific angle on how automatic pattern search enables us to scrutinize what it is that we really know about a specific folk music style, if we consider ourselves to be musical experts. I elaborate my hypothesis that musical knowledge is often implicit, while computation enables us to make part of this knowledge explicit and evaluate it on a data set. Specifically, I address the questions as to when we perceive two folk melodies to be variants of each other, and how to unravel style characteristics. With examples from my research on patterns in Dutch folk songs, Irish folk songs and Rags, I demonstrate what both experts and non-experts gain from developing computational methods for analysing folk music.

## Keynote: Peter Browne, *Tuning the radio*

### Abstract

Irish state radio broadcasting began with the creation of the station 2RN and its first evening's programming broadcast from Dublin on Jan 1st 1926 - 90 years ago this year.

From a simple 3 hour programme on that first night in 1926, comprised entirely of music (to all intents and purposes a broadcast concert), there has been an ongoing relationship between traditional Irish music/song and radio broadcasting; each could undoubtedly have had its own amply fulfilled existence without the other and each has experienced ever-present change and development over the years in their own separate spheres of activity. Yet there are many connections and influences and the purpose of the paper is to trace and attempt to enumerate and assess at least some of these.

There were determining factors on both sides of the relationship: on the radio side, technical issues such as transmission, staffing, audience reach, sound quality and also questions of awareness and judgement by the 2RN (later Raidió Éireann and RTÉ Radio) authorities of what might constitute good taste or competence in the playing and appreciation of traditional music and indeed this at times extended as far as knowing what would or would not be included in a definition of traditional music and song. On the traditional music side, there was a lack of engagement for various reasons, geographical, cultural and even class-based and the absence of a widespread language of criticism. Traditional music had (and continues to have) different forms such as solo instrumental expression, singing in both Irish and English, ensemble and orchestral playing and debates about authenticity have never been far away.

Among the historical points of interest considered here are: the early years of 2RN with only live broadcasting, auditions, the coming of the céilí band, orchestration, outside broadcasts and the M.R.U. (Mobile Recording Unit), Seán Ó Riada and Ceoltóirí Chualann, the pre- and post-television era, The Long Note, The Brendan Voyage, the change from RTÉ Radio as a sole player to a proliferation of radio stations and finally the present day and the era of literally instant worldwide digital access.

The presentation will include references from published sources, unpublished written archive material, transmitted radio programmes and other illustrative music audio as well as some personal communications and unedited interviews.



# A REVALUATION OF LEARNING PRACTICES IN INDIAN CLASSICAL MUSIC USING TECHNOLOGICAL TOOLS

**Julien Debove**

EHESS Paris – CAMS  
(UMR CNRS 8557),  
Paris, France  
juliendebove@hotmail.fr

**Dorian Cazau**

ENSTA Bretagne - Lab-STICC  
(UMR CNRS 6285),  
Brest, France

**Olivier Adam**

UPMC Univ. Paris 06,  
CNRS UMR 7190, Insti-  
tut Jean Le Rond  
d'Alembert, F-75005,  
Paris, France

## ABSTRACT

Each *khyāl* performance of Indian classical music is unique and unreproducible because it is mainly based on improvisation. As for most orally transmitted musical repertoires, learning practices are essential as they guarantee that the musical codes are properly reproduced from one generation to another. In Indian classical music, practice, tightly imbricated in the pupil – teacher relation, favors clearly the imitation. Students tend to reproduce more or less successfully their master's style. That's why in order to be creative, it is necessary that each musician develops his own skills of understanding, experimentation and invention.

Today, technological tools have considerably transformed our way of learning. From now on, it is possible to have access to considerable data for the understanding of traditional music, and to listen, record and analyse them via numerous audio softwares. Indeed, works by visualization allows reporting know-how common to all these musics (fingering, musical process, improvisation, patterns...). Through various softwares and practice examples from Rajam's dynasty (hindustani violinist players), hindustani violin lessons and *rāg* performances, we will present a "toolbox" useful for all musicians and musicologist to improve their self-study.

If the pedagogy and teaching can give us comprehension keys, the apprenticeship, such as it is practiced in North India and in the long master to pupil's tradition, favors clearly the imitation at the expense of the assimilation. The pupil learns above all by imitation and by impregnation, without taking the time to understand or to write. He learns to know a number of ingredients, but does not inevitably learn how to use it. In this way, the pupil tends inexorably to reproduce with varying degrees of acuteness the master's style. His space of creativity is extremely reduced even non-existent. The musician will feel difficulties finding his own style. For that purpose, it is necessary to him to be able to stand back, to be able to experiment, invent and understand.

The technological tools really transformed our way of learning in our daily practice. So the analysis via a number of IT data and software allows to understand and to learn musical processes, specific ornamentations, rarely taught. In addition, it is possible to question the relationship between what is taught by the master and what is produced on stage. Through the comparison of different performances, different performers and different learning lessons, one can clearly dissociates the stored material from the improvised material, i.e. the fixed components from the modular elements. This current work

aims to study this question, focusing on different *rāg* according to the vocal tradition of *khyāl* within the Rajam's Dynasty, violinist descendants.

In this communication, we investigate the possibility of using modern computer-based technologies as a teaching assistance system for Indian classical music. Due to its improvisation nature, a comparative approach is necessary to analyse it. For example, by comparing recordings between Hindustani violin lessons at the Hubli-Gurukul (India, August 2010-2012) and Hindustani *rāg* performances, it is possible to show up the way(s) Rajam Dynasty musicians transform the structural and structuring<sup>1</sup> elements of a *rāg*. At a larger scale of analysis, by multiplying the interpreters on a same *rāg*, we could quantitatively compare their different improvisation strategies, and better understanding the fundamental elements of a *rāg* that need to be properly taught to every musician.

## 1. LINKS BETWEEN PERFORMANCES AND APPRENTICESHIP

We notice that it exists a correlation between Rajam's lessons (Gangubai Gurukul, Hubli, India, August 2010-2012) and musical performances. Indeed, we can observe that a number of formulas, that Julien Debove learned in Gangubai Gurukul, are repeated in the musical performances. It means that during the performance, the musicians dig into his memory bank and add it formulas transposed from another *rāg* or improvised formulas.

As we can observe on the figure 1, the cycle is structured in the following way :

- establishment of a formula -red oval-
- suite of variations -oval yellow if it is played by one musician and orange rectangle if it is played by two different musicians-
- resolution on C medium -fuchsia rectangle- and chorus<sup>2</sup> -light pink rectangle-.

The chorus serves as a link from one cycle to another.

<sup>1</sup> The structuring elements are useful elements for the continuity of the structure. These elements are generally signals allowing the passage from a subsection to another or from one part to another. The structural elements are the elements forming part of the structure.

<sup>2</sup> Sometimes, the chorus can be substitute by a melodic phrase repeated three times (*tihāī*).

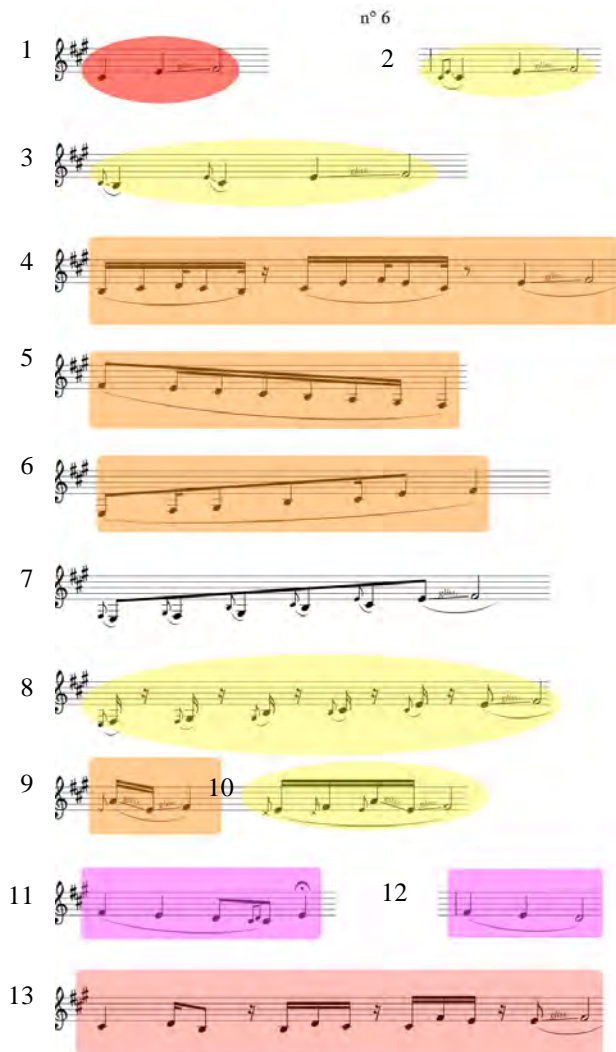


Figure 1. Scores of the 6<sup>th</sup> rhythmic cycle of *vilambit ektāl* (slow 12 beat cycle), *rāg* Yaman, N. Rajam lessons, August 2013 realized with ianalyse.

	Introductive formula		Variation play
	Conclusive formula		- by one musician
	Chorus		- by two musicians
Musicians	N. Rajam, 2012	R. Shankar, 04/06/13	J. Debove 08/16/13
Formulas & variations	1,5,6,9	1,2,4,7,11,13	1,3,4,5,6,9,10,11,12,13

## 2. TYPE OF IMPROVISATION AND SOUND REPRESENTATIONS

### 2.1 Thematic variations

If learning topics are strictly taught from master to student, during the performance, some rhythmic and timespan variations can be made. Thus, we perceive a slight difference in the notion of composition, that can be

used to talk about music of oral transmission or western music. The difference lies mainly in the medium used:

- external memory for music written
- internal memory for music of oral transmission.

Indeed, when we use the term composition to describe melodies laid down by oral transmission from individuals to individuals, internal memory to internal memory, whatever audio message's quality of memorization or assimilation, we can not imagine that the audio message will be transmitted from one generation to another without a slight modification, even if the references to earlier records can avoid excessive transformation.

So we can perceive within the various examples below duration or rhythmic's variations within theme's modules (transmitted orally in a strictest way).

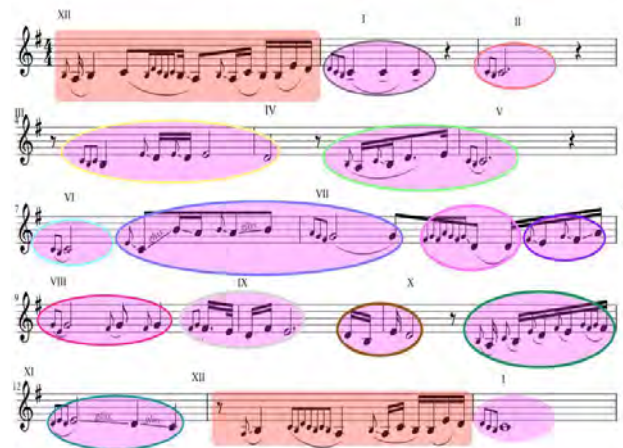


Figure 2. *Vilambit Ektāl* (slow 12 beat cycle), theme, *rāg* Yaman, N. Rajam, 2012, Hubli


















Figure 3. *Vilambit Ektāl* (slow 12 beat cycle), theme, *rāg* Yaman, Ragini Shankar, 04/06/13, Lille

Thematic modules	
Chorus	

## 2.2 Structural improvisation

As underlined by Nettl (1974), we consider a musical repertoire, composed or improvised, as the realization of a system. One of the approaches to describe such a system is to divide it into theoretical component units. These units are, so to speak, blocks of construction accumulated by tradition and of which the musicians (within the tradition) make use, by choosing, combining, recombining and rearranging them. These blocks of construction are, even in a single directory, of various types.

This type of improvisation is seen in action in North Indian classical music. So, in the *drut tintāl* (fast cycle of 16 beat), whatever performances, we find in each performance the same ingredients placed in a certain order and transposed according to the *rāg*. We notice also some fundamental processes of development defined by Widess (2006) as melodic expansion & rhythmic intensification.

Progressiv ascent		Specific bowing	
Trembling ascent		Changing octave	
Specific descent		Simple and complex <i>Tān</i>	
High <i>Alāp</i>		High rhythmic variation	
Theme		Musical phrases repeated three times ( <i>tihātī</i> )	
First part of the theme		Low rhythmic variation	
Thematic variations		Inflexion	
Suite of <i>tān</i> <sup>3</sup>			

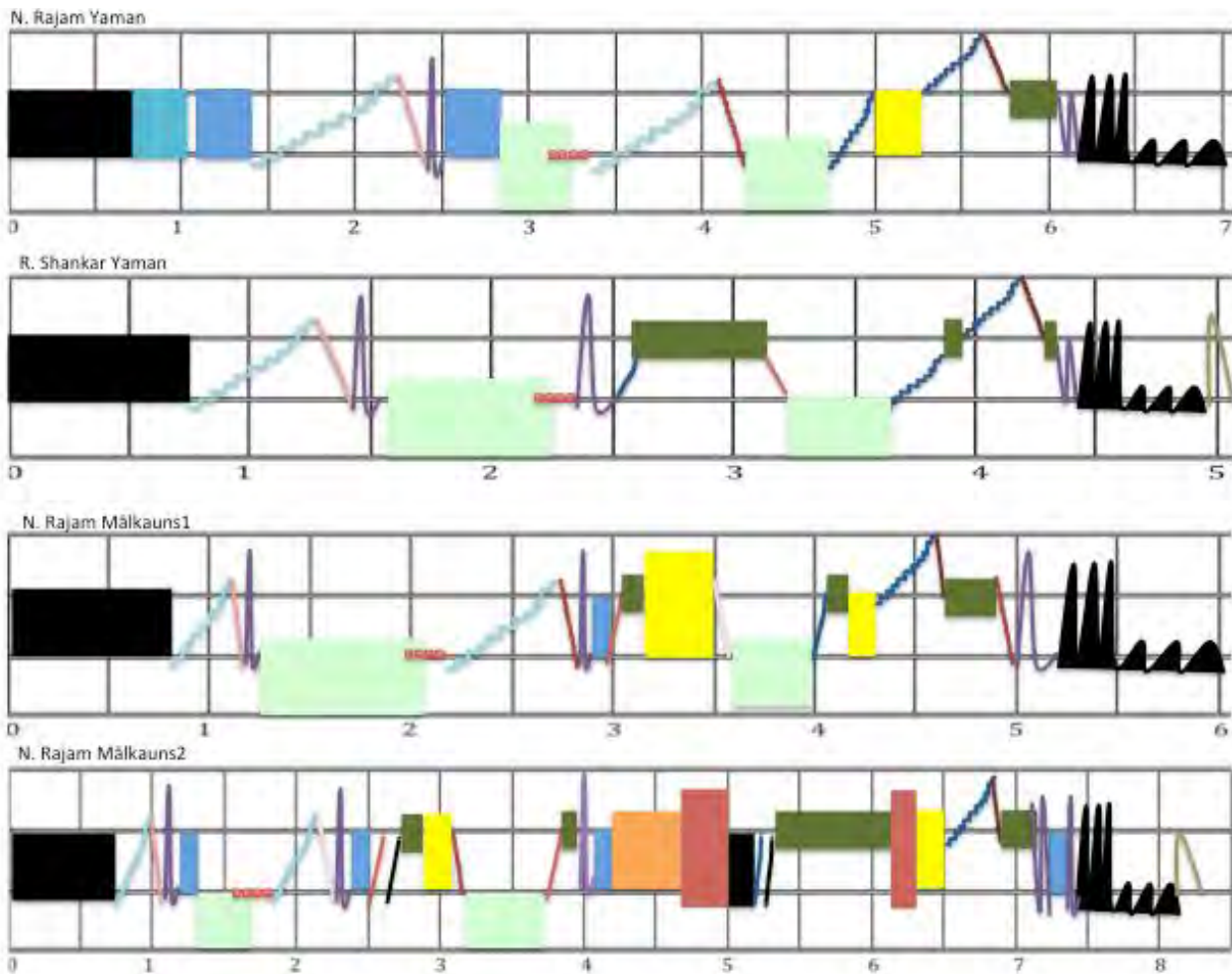


Figure 4. Structural evolution of *drut tintāl* (fast 16 beat cycle), *rāg* Mālkauns & Yaman (with Excel).

## 2.3 Melodic variations

Combined together, Sonic visualizer<sup>4</sup> and Acousmographie<sup>5</sup> can build genuine listening guides, associating music

<sup>3</sup> Fast improvised melodic lines

<sup>4</sup> Sonic Visualiser, developed by Queen Mary University (London) is an application for viewing and analysing the contents of music audio files.

<sup>5</sup> Acousmographie, developed by G.R.M (Paris) is a software of listening and visual representation of the music. He allows location, annotation and thorough description of any music or any sound document.



playback, sound visualization and precise and fine analysis of various extract. They allow to perceive melodic inflections difficult to hear, to highlight the notes and ornamentation and understand the overall shape of the different musical passages.

The only drawback is that the implementation of these tools requires a lot of time and work because all additions (image, text, scope, notes, ornaments) are done manually.

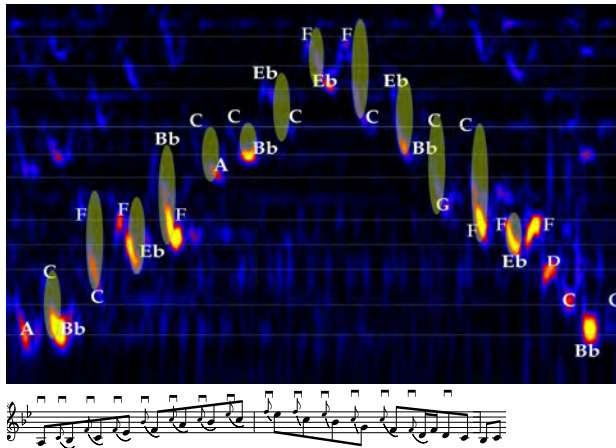


Figure 5. N. Rajam, fast melodic line, *rāg* Bāgeśrī Kānadā, 1991

Here, we perceive thanks to the visualization and annotation, the particular fingering of a fast melodic line feature. If they are played frequently with ornamentations from the bottom upward, in this case, it is inverted, what gives a particular effect.

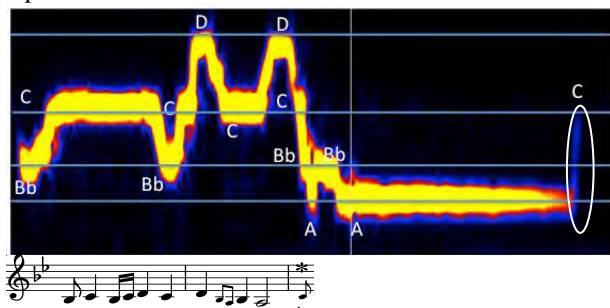


Figure 6. H. Chaurasia, improvised introduction (*ālāp*), *rāg* Bāgeśrī, 1994

In this example, it is possible to perceive an inflection of an almost inaudible sound upward played at the beginning or at the end of musical sentences

### 3. COMPREHENSION & THEORIZATION

#### 3.1 Monika<sup>6</sup> & Carnet de Notes

(<http://carnetdenotes.paris-sorbonne.fr/>)

So, it is possible to perceive some inflections of an almost inaudible sound upward played at the beginning or at the end of musical sentences

<sup>6</sup> Monika is a software of description of monodies drafted in VBA for Excel. Developed by N. Méus, professor at the University Paris Sorbonne

For the needs of intelligent practice respectful of traditions, it is advisable to set out fixed recordings<sup>7</sup> and diverse analyses to allow, by the empirical practice, access to many data. From file XML, the Monika software allows to do numerous statistics that are very useful for musical analysis (as for example, upper or lower melodic peaks, number of occurrences of each note, intervals most used, internal finales<sup>8</sup> etc.).

Here are some of these diagrams drafted thanks to Monika software:

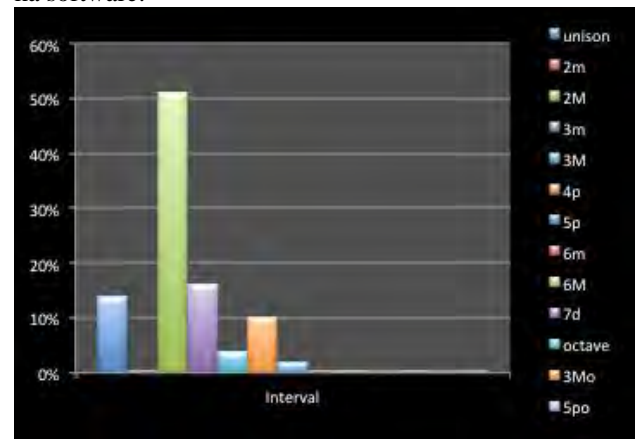


Figure 7. H. Chaurasia, *rāg* Mālkauns, interval, *Live in Stuttgart*, 1988

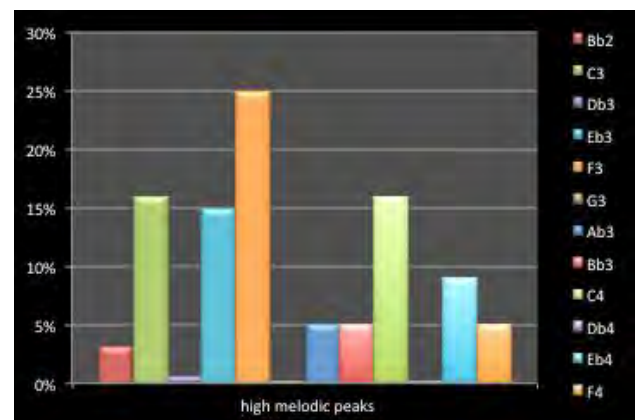


Figure 8. H. Chaurasia, *rāg* Mālkauns, high melodic peaks, *Live in Stuttgart*, 1988

The first diagram allows to show clearly the linearity of the musical way because the intervals superior to the fifth are almost non-existent. Furthermore, the musical way is mainly made by joint intervals. In a practical way, the analysis of the superior melodic peaks is very important for understanding the hierarchical organization of notes. So, F, although it is not strongly present in the *rāg* Mālkauns, is an important melodic pole because notes go there. All these data are in free access on the site “Carnet de notes”.

<sup>7</sup> These recordings are obviously a version among many others and don't represent models. It is thus that by the depth and comparative analysis we can have a minimum of objectivity.

<sup>8</sup> Monika software considers internal finales in its widest release. These are the notes preceding silence.



### 3.2 Musical strategy on rajam's style in *ālāp*

Melodic analysis of various *ālāp*<sup>9</sup> via a synoptic view can distinguish different melodic phases and better understand the way of improvising (see figure 9). By listing all internal finales in order chronological the first time when they appear (defined here as notes finishing a melodic sentence of a duration upper at three seconds) as well as the set of the internal finales on C, it is possible to distinguish four essentials phases. (symbolized by the grey sinusoidal curve).

These phases correspond to the various possibilities which offer themselves to the musician. None of these phases is compulsory, but the phases 1 and 2, 3 and 4 are consecutive (Gorakh Kalyān, Māru Bihāg, Mālkauns, Jog 1 & Bāgeśrī).

We also notice that the musicians can play solely phases 3 and 4 (Yaman & Jayjayvantī) or 1 and 2 (Jog 2 et Bāgeśrī Kānadā) or none of these phases when the *ālāp* is very short (Desī) by stressing only the tonic (C).

### 3.3 OpenMusic<sup>10</sup>

New technological tools like OpenMusic allow us to create musical processes modeling. So I have create a fast melodic line of synthesis. (see figure 10)

It consists of groups of 4 ratings whose first note is accentuated (represented by the sub-patch "segmentation 4") groups of 3 notes that the first note is accentuated (represented by the sub-patch "segmentation 3") , groups of 2 ratings whose 1st note is accentuated (represented by the sub-patch "segmentation 2").

These modules are played consecutively. Notes within these modules are played randomly, but linearly. *Tān* begins with an onset formula and ends with fixed phrases repeted three times (*tihār*) on the right side of the diagram. The scale of the *tān* is fixed on the left side of the diagram. Each time the object is revalued, music notation and music changes.

What is very interesting is that it corresponds perfectly to a *tān* as could play musicians in the performance.

## 4. PERSPECTIVES TOWARDS AUTOMATIC MUSIC TRANSCRIPTION

In this paper, we presented the use of audio software for the analysis of improvisation styles in Indian classical music. These software could be more efficient by implementing new methods from automatic music transcription (AMT). However, despite a large enthusiasm for AMT challenges, and several audio-to-MIDI converters available commercially, perfect polyphonic AMT systems are out of reach of today's algorithms. This lecture will be started from our previous works (Cazau et al., 2013 ; Cazau et al., 2015) to present

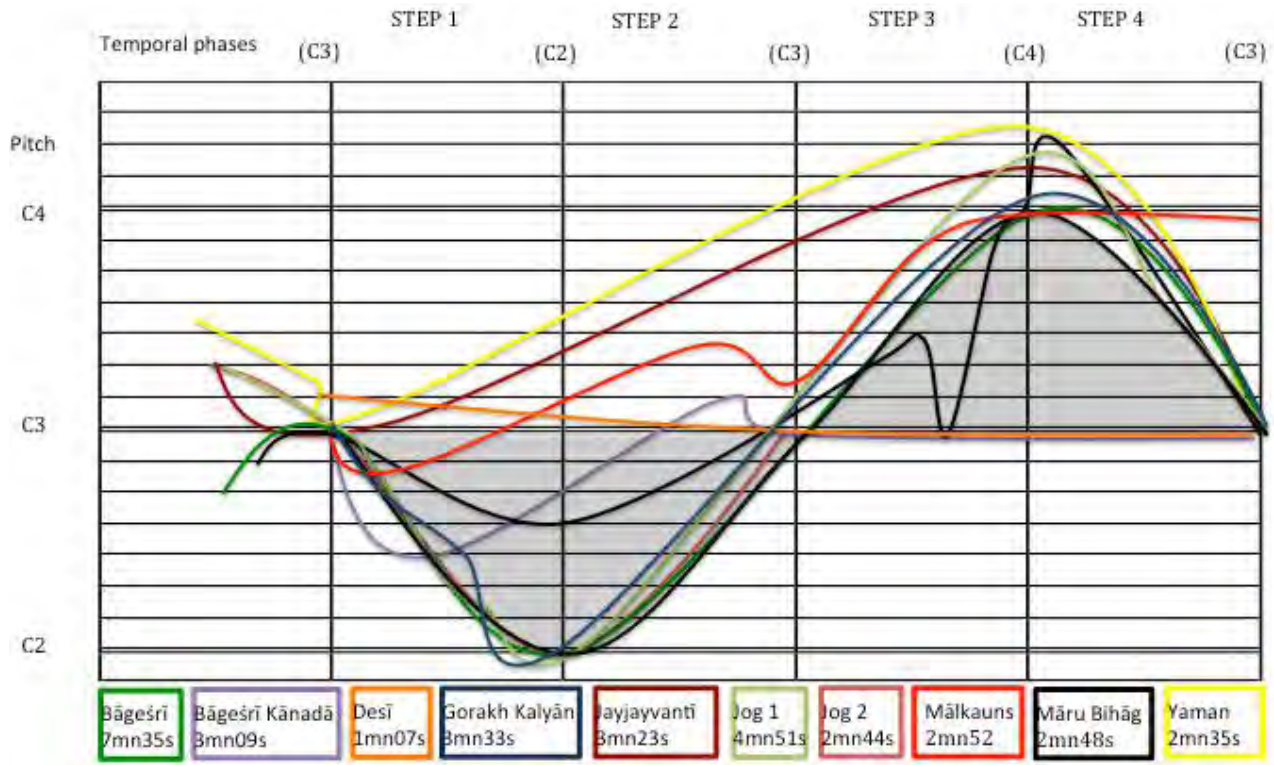
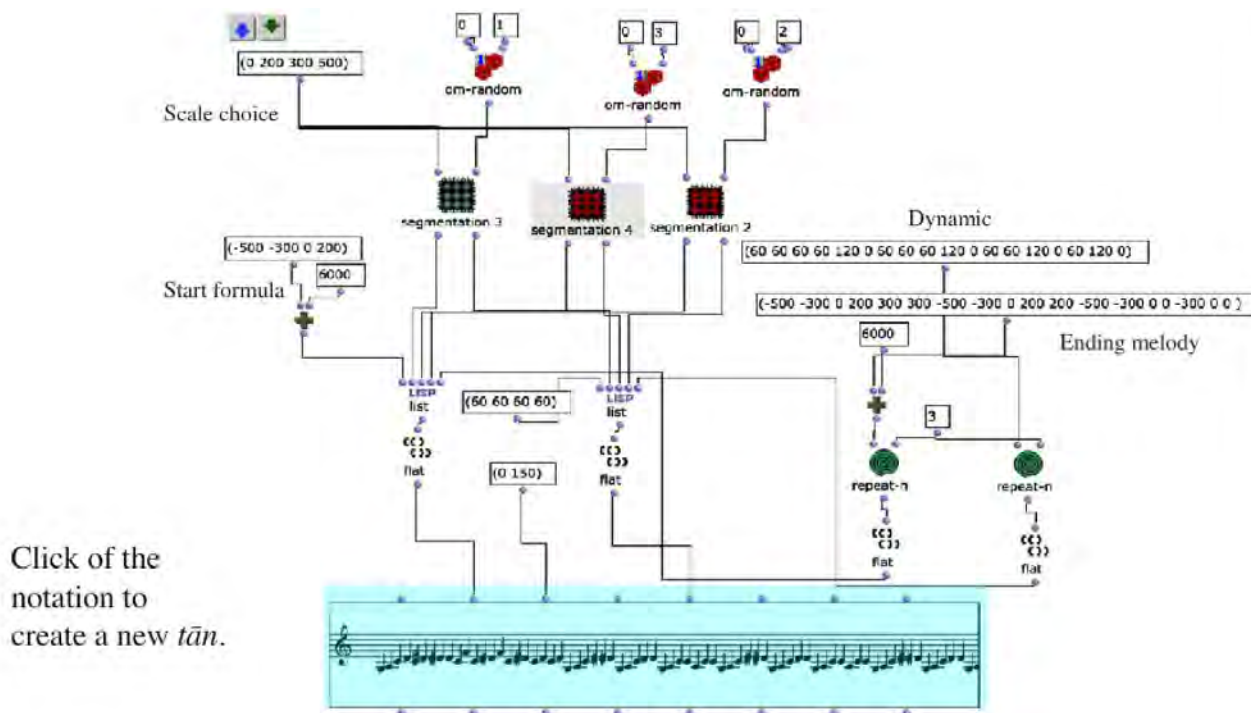
a new multichannel capturing sensory systems of traditional acoustic plucked string instruments, including the following traditional African zithers: the marovany zither (Madagascar), the Mvet lute (Cameroon), the N'Goni harp (Mali). These systems use multiple string-dependent sensors to retrieve discriminatingly some physical features of their vibrations. For the AMT task, such a system has an obvious advantage in this application, as it allows breaking down a polyphonic musical signal into the sum of monophonic signals respective to each string. The development of this technology has already allowed the constitution of a new sound dataset dedicated to AMT evaluation for plucked-string instrument repertoires, used in Cazau et al. (2015), and including audio recordings, MIDI-like transcripts and sound samples over the instrument pitch ranges. This technology is very convenient to develop extensive sound corpus for repertoires without written supports, including orally transmitted repertoires, as well as improvisation.

## 5. REFERENCES

- Cazau, D., Chemillier, M. and Adam, O. (2013). Information retrieval of marovany zither music with an original optical-based system. *Proceedings of DAFx*, Maynooth, Ireland, 1-6
- Cazau, D., Revillon, G., Krywyk, J. and Adam, O. (2015). An investigation of prior knowledge in Automatic Music Transcription systems. *The Journal of the Acoustical Society of America*, 138, 2561-2573
- Debove, J. (2015). *Approche de la musique modale et transmission orale de la musique hindoustanie*, thesis, supervised by M. Chemillier, School of Advanced Studies in Social Sciences, Paris
- Nettl B., (1974). Thoughts on Improvisation: A Comparative Approach, *The Musical Quarterly*, 60(1), 14.
- Nooshin, L., Widdess, R. (2006). Improvisation in Iranian and Indian music. *Journal of the Indian Musicological Society*, 112-113.

<sup>9</sup> Slow introduction where we discover the *rāg* and all its features. *Ālāp* also refers by extension slow improvised melodic phrases.

<sup>10</sup> OpenMusic (OM), developed by IRCAM Music representation research group (Paris) is a visual programming language based on Lisp. Visual programs are created by assembling and connecting icons representing functions and data structures.

Figure 9. Melodic ways of *ālāp*, N. Rajam & R. ShankarFigure 10. Creation of a synthesis « *tān* »

# DETECTION OF MELODIC PATTERNS IN AUTOMATIC TRANSCRIPTIONS OF FLAMENCO SINGING

Aggelos Pikrakis

University of Piraeus, Greece  
pikrakis@unipi.gr

Nadine Kroher, José-Miguel Díaz-Báñez

University of Seville, Spain  
nkroher@us.es, dbanez@us.es

## ABSTRACT

The spontaneous expressive interpretation of melodic templates is a fundamental concept in flamenco music. Consequently, the automatic detection of such patterns in music collections sets the basis for a number of challenging analysis and retrieval tasks. We present a novel algorithm for the automatic detection of manually defined melodies within a corpus of automatic transcriptions of flamenco recordings. We evaluate the performance on the example of five characteristic patterns from the *fandango de Valverde* style and demonstrate that the algorithm is capable of retrieving ornamented instances of query patterns. Furthermore, we discuss limitations, possible extensions and applications of the proposed system.

## 1. INTRODUCTION

Flamenco is a rich music tradition from the southern Spanish province of Andalucía. Having evolved from a singing tradition, the vocal melody remains the main musical element, accompanied by the guitar, rhythmical hand-clapping and dance. Gómez et al. (2016) mention, among others, the frequent appearance of glides and protamenti, sudden dynamic changes in volume and a small pitch range of less than an octave, as key characteristics of the flamenco singing voice. For a more detailed description of the genre, we refer to Gómez et al. (2014) and Gómez et al. (2016).

Flamenco singing is largely improvisational, in particular with respect to melody: during a performance, a melodic skeleton or a set of prototypical patterns are subject to spontaneous ornamentation and variation. Consequently, the automatic detection of modified instances of a given melodic sequence is a crucial step to a number of music information retrieval tasks. For example, most characteristic melodies are uniquely bound to a particular singing style. Consequently, detected melodic patterns provide important indications towards the style of an unknown recording. Furthermore, flamenco recordings often contain various songs and the location of pattern occurrences can assist the structural segmentation of a song. Moreover, the occurrence of common melodic patterns across tracks is crucial to characterising similarity among melodies which exhibit structural differences (Volk & van Kranenburg, 2012).

Given the absence of musical scores, related approaches in the context of flamenco (Pikrakis et al., 2012) but also in other non-Western oral music traditions (Gulati et al., 2014) have focused on the retrieval of melodic patterns from the fundamental frequency (f0) contour. The high degree of detail of this representation does not only increase computational complexity but is also prone to errors

arising from micro-tonal ornamentations. In this study, we present a novel approach which operates on symbolic representations obtained from an automatic transcription system (Kroher & Gómez, 2016).

We provide a detailed technical description of the method in Section 2. The experimental setup is described in Section 3 and results are given in Section 4. We conclude the paper in Section 5.

## 2. METHODOLOGY

The core of our method is a modification of the well known Needleman-Wunsch (NW) algorithm (Needleman & Wunsch, 1970) from the area of bioinformatics. The NW algorithm was proposed as a global alignment method of molecular sequences. The term global alignment refers to the fact that when two sequences of discrete symbols are being matched, the objective is to align them from the beginning to the end, without omitting parts around the endpoints. During the alignment procedure, gaps are allowed to be formed. In the original NW formulation gaps are not penalized. Given two sequences of discrete symbols, the original NW algorithm can be formulated as a dynamic programming method that creates a dot matrix and finds the best path of dots on it, i.e., a path of dots (nodes) of increasing index that accumulates the largest score (number of dots). The dot matrix (also known as similarity grid) is formed by placing one pattern on the x-axis and the other one on the y-axis. An element of the dot grid is set equal to “1” if the symbols corresponding to its coordinates coincide.

The problem that we are dealing with in this paper cannot be treated as a global alignment task because our goal is to detect occurrences of a pattern in a significantly longer stream of notes. We are therefore proposing a modification of the NW algorithm, that preserves its fundamental characteristics and adds the capability to retrieve a ranked list of subsequences from an automatic transcription. Each retrieved result aligns, in some optimal sense, with the given prototype pattern. The novelty of our approach lies in the fact that it introduces a systematic way to: **(a)** extract iteratively occurrences of the reference pattern, ranked with respect to similarity score, **(b)** embed endpoint constraints in the NW method, **(c)** ensure invariance to key changes because the alignment takes place on the sequences of intervals derived from the pitch sequences that are being matched, and, **(d)** formulate transition costs between nodes of the

similarity grid as a function of intervallic differences. At a first stage, the proposed method operates on pitch sequences only, ignoring note durations. At a second stage, the results are refined by removing alignments that correspond to excessive local time-stretching. In the rest of this paper, we will use the abbreviation *mNW* for the proposed method.

In order to describe *mNW*, let  $A = \{a_i; i = 1, 2, \dots, I\}$  and  $P = \{p_j; j = 1, 2, \dots, J\}$  be the pitch sequences of the automatic transcription and the search pattern, respectively, where the  $a_i$ 's and  $p_j$ 's are pitch values in some symbolic (MIDI-like) format. We therefore ignore note durations at this stage. Sequence  $P$  is manually defined and reflects our musicological knowledge of the pattern to be detected. For example, pattern "A" of our experimental setup (Section 3) is represented by the following sequence of MIDI values:

$$\{64, 67, 65, 64, 67, 65, 65, 64, 62, 60, 58, 57\}$$

We now define that,

$$\delta_P(j_2, j_1) = p_{j_2} - p_{j_1}, j_2 > j_1,$$

is the music interval formed between the  $j_1$ -th and  $j_2$ -th note (pitch value) of the prototype pattern, which are not necessarily adjacent, and, similarly

$$\delta_A(i_2, i_1) = a_{i_2} - a_{i_1}, i_2 > i_1,$$

is the music interval formed between the  $i_1$ -th and  $i_2$ -th note (pitch value) of the automatically generated transcription. Therefore, the proposed *mNW* algorithm seeks a subsequence (chain) of  $a_i$ 's, of increasing index (not necessarily adjacent), such that the resulting sequence of intervals matches in some optimal scoring sense, a sequence of intervals formed by a subsequence of  $p_j$ 's of increasing index (also not necessarily adjacent).

To solve this problem from a dynamic programming perspective,  $A$  is placed on the vertical axis and  $P$  on the horizontal one, forming a scoring grid,  $S$ . Let

$$(i, j), i = 1, 2, \dots, I, j = 1, 2, \dots, J$$

be a node on this grid, which aligns the  $i$ -th note of  $A$  with the  $j$ -th note of  $P$ , and let  $S(i, j)$  be the respective accumulated alignment score. The grid is initialized by setting the elements of the last row and column of the grid equal to zero, i.e.,  $S(I, j) = 0, j = 1, 2, \dots, J$  and  $S(i, J) = 0, i = 1, 2, \dots, I$ .

We then proceed row-wise, decreasing the row index and examining the nodes of each row at decreasing column index, which stands for a standard zig-zag scanning procedure. The accumulated score,  $S(i, j)$ , at node  $(i, j)$ , where  $i < I$  and  $j < J$  is computed as follows:

$$h = \max\{S(i+1, k) + \gamma(\delta_A(i+1, i), \delta_P(k, j));$$

$$k = j+1, \dots, j+G_h\}, \quad (1)$$

$$v = \max\{S(m, j+1) + \gamma(\delta_A(m, i), \delta_P(j+1, j));$$

$$m = i+1, \dots, i+G_v\}, \quad (2)$$

$$S(i, j) = \max\{h, v\}, \quad (3)$$

where parameters  $G_h$  and  $G_v$  are positive integers that define the search radius for successors on the horizontal and vertical axis, respectively, and function  $\gamma(\cdot)$  is defined as:

$$\gamma(x, y) = \begin{cases} 1, & \text{if } x = y, \\ -1, & \text{if } |x - y| = 1, \\ -\infty, & \text{if } |x - y| > 1, \end{cases}$$

The first two equations impose that the best successor of node  $(i, j)$  resides either on the next row (the  $(i+1)$ -th row) or on the next column (the  $(j+1)$ -th column). Parameters  $G_h$  and  $G_v$  control the horizontal and vertical gap length, respectively. In other words, they control how many pitch values can be skipped horizontally or vertically when searching for the best successor of the node. Function  $\gamma$  rewards equal intervals with a score equal to +1, penalizes with -1 any pair of intervals that differ by one semitone and forbids intervallic differences larger than a semitone to take place, hence the  $-\infty$  penalty. After a node has been processed, the coordinates,  $(i_B, j_B)$ , of its best successor, are stored in a separate matrix,  $\Psi$ , where  $\Psi = \{\psi(i, j) = (i_B, j_B); i = 1, \dots, I, j = 1, \dots, J\}$ .

After the whole grid has been scanned, the highest accumulated score on the first  $E_1$  columns is selected and forward tracking on matrix  $\Psi$  reveals the best alignment path. However, this path will be rejected if it does not end in one of the last  $E_2$  columns of the grid. Therefore, parameters  $E_1$  and  $E_2$  stand for the endpoint constraints of the alignment procedure, i.e., we permit that at most  $E_1 - 1$  and  $E_2 - 1$  notes are omitted from the left and right endpoints of the prototype pattern, respectively. If a path is rejected, we repeat from the second highest score until a valid path is detected or until all nodes of the first  $E_1$  columns have been processed as candidate starting points of the best path. Obviously, if we want the algorithm to return two pattern occurrences, the procedure will be repeated until a second path is revealed, and, of course, this can be readily extended to address any number of desired occurrences.

Table 1 presents the best alignment result between pattern A of the experimental setup and a Valverde transcription. In this example, two notes are skipped from the automatically generated transcription (5th and 10th note from the first column) and this is shown with one inserted gap (symbol "-") per deleted note in the second column, in the respective rows. It is also worth observing that the matched subsequences are performed in different keys.

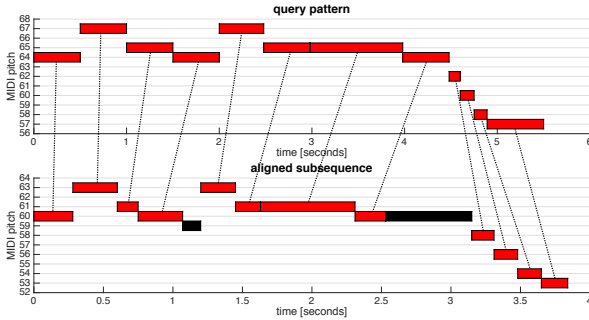
The example is further illustrated in Figure 1, where the dotted lines connect aligned notes and the two black notes are the ones that have been skipped on the automatic transcription sequence.

After the first processing stage has been completed, the obtained results are subsequently filtered at a second stage. More specifically, in order to restrain note duration variability, we compute the sequence of inter-onset differences of the notes of a formed path on both axes and discard any path for which at least two ratios of aligned inter-onset durations exceed a predefined stretching threshold (equal to 3 or 1/3 in our study). This is equivalent to imposing, at a



**Table 1:** Best alignment result of pattern A against an automatically generated Valverde transcription: symbol “-” marks a skipped note (gap insertion).

transcription		query pattern (A)	
pitch	duration	pitch	duration
60	0.28	64	0.50
63	0.32	67	0.50
61	0.15	65	0.50
60	0.32	64	0.50
59	0.13	-	-
63	0.25	67	0.48
61	0.18	65	0.50
61	0.68	65	1.00
60	0.22	64	0.50
60	0.62	-	-
58	0.16	62	0.12
56	0.17	60	0.15
54	0.17	58	0.14
53	0.19	57	0.61



**Figure 1:** Illustration of the alignment shown in Table 1.

post-preprocessing stage, a local time-warping threshold.

### 3. EXPERIMENTAL SETUP

We demonstrate the performance of the proposed algorithm in a query-by-example task. We aim at detecting occurrences of manually annotated MIDI sequences in a corpus of automatic transcriptions of polyphonic flamenco recordings. In this study, we focus on *fandangos de Valverde* (FV), a singing style belonging to the family of the *fandangos* (Kroher et al., 2016).

Like most *fandangos*, the *fandangos de Valverde* are bimodal in a structural sense (Fernández-Marín, 2011): solo guitar sections are set in *flamenco mode*, a scale with the diatonic structure of the Phrygian scale but with the dominant and sub-dominant located on the second and third scale degree, respectively (Figure 2). Singing voice sections are set in major mode and modulate only in the last verse back to *flamenco mode*.

Having evolved from Spanish folk tunes, songs belonging to this style are based on a particular melodic skeleton which, during interpretation, is subject to melodic and rhythmic modifications in terms of an expressive perfor-



**Figure 2:** The flamenco mode: The tonic is located on the first, the dominant on the second and the sub-dominant on the third scale degree.

mance. The skeleton is composed of five distinct patterns (Figure 3) which occur in the form A-B-A'-C-A-D (where A' refers to a variant of A).

In this study, we use as query patterns manual transcriptions of the five phrases constituting the *fandango de Valverde* skeleton (Figure 3) and aim to retrieve their ornamented and modified occurrences in automatic transcriptions of real performances. To this end, we gathered a collection of 20 *fandangos de Valverde* taken from commercial recordings. The *cante100* dataset (Kroher et al., 2016) was added as noise to the collection: The contained 100 accompanied flamenco recordings cover a variety of singing styles and serve as a representative sample of flamenco music. None of the tracks contained in the *cante100* dataset belong to the *fandangos de Valverde* style. For each of the 120 tracks of the resulting collection we generated an automatic note-level transcription of the vocal melody using the algorithm described by Kroher & Gómez (2016).

The retrieved results are evaluated by means of the precision of the top 5 (P@5) and top 10 (P@10) ranking. A query result is considered relevant if its origin is a *fandangos de Valverde* recording and the detected melodic sequence corresponds to the query phrase.



**Figure 3:** MIDI representations of the query patterns.

### 4. RESULTS

Table 2 gives the quantitative evaluation of all five query patterns and the top 5 results for pattern A are shown in Figure 4. It can be seen that the percentage of relevant melodic sequences in the top ranked results is significantly higher for patterns A, A' and B compared to patterns C and D. In particular, for patterns A' and B, all of the 5 highest ranked results are relevant with respect to the query, while for pattern D only one relevant result is retrieved.

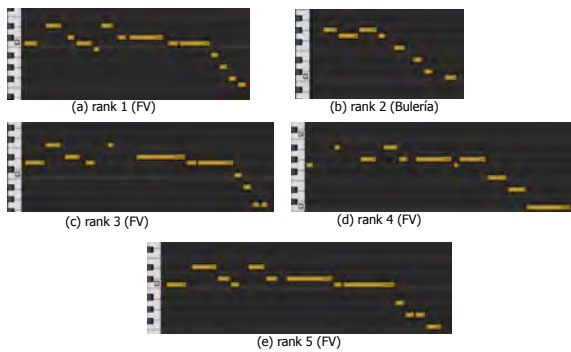
A reasonable explanation for this behaviour is related to the amount of variation a pattern is subjected to during

performance: Pattern D, referred to as *caída* in flamenco terminology, constitutes the end phrase and, at the same time, the musical "highlight" of the interpretation. During this phrase, the melody modulates from major mode to flamenco mode and resolves in the Andalusian cadence. Consequently, singers tend to apply more expressive resources, which result in a higher performance variance. Within a lesser extent, the same applies to pattern C, where a high degree of ornamentation, in particular prolongation through a sequence of grace notes, tends to appear during the last two bars. Four examples of manual MIDI transcriptions of *caídas* are shown in Figure 5 in order to highlight observed performance variation, free of possible transcription errors. Furthermore, automatic transcriptions are particularly prone to errors in the end of the singing voice section, since the guitar accompaniment tends to significantly increase in volume. As a result, notes belonging to the singing voice melody might be missed and guitar notes might be transcribed instead.

Nevertheless, it can be seen from Figure 4 that the algorithm is capable of detecting ornamented and modified occurrences of a query pattern. It is also interesting to note that the obtained results contain a similar melodic sequence that was found in a recording of a different style (Figure 4 (b)), a *Bulería*. Despite this result being rated as not relevant in this task, it nevertheless demonstrates the potential of this tool for uncovering hidden structures and similarities in the context of large mining studies.

**Table 2:** P@5 and P@10 measures among queries.

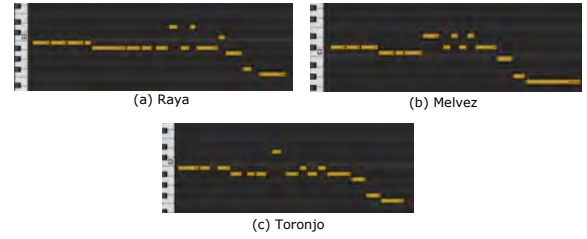
query	P@5	P@10
A	80%	60%
A'	100%	70%
B	100%	70%
C	40%	40%
D	20%	10%



**Figure 4:** MIDI representations of the top 5 results for query pattern A.

## 5. CONCLUSIONS

We presented an algorithm for melodic pattern retrieval based on automatic transcriptions and demonstrated ex-



**Figure 5:** Manual transcriptions of pattern D for three singers: (a) A. Raya, (b) M. Vélez and (c) P. Toronjo.

amples of the capabilities and limitations of the system. Future applications are expected to include the incorporation of the algorithm in a framework for unsupervised pattern detection, the retrieval of typical ornamentations from music recordings and the detection of short melodic guitar fragments (*falsestas*) in the melody of the singing voice.

## 6. REFERENCES

- Fernández-Marín, L. (2011). La bimodalidad en las formas del fandango y en los cantes de levante: origen y evolución. *La Madrugá*, 5(1), 37–53.
- Gómez, F., Díaz-Báñez, J. M., Gómez, E., & Mora, J. (2014). Flamenco music and its computational study. In *BRIDGES: Mathematical Connections in Art, Music, and Science*.
- Gómez, F., Mora, J., Gómez, E., & Díaz-Báñez, J. M. (2016). Melodic contour and mid-level global features applied to the analysis of flamenco cantes. *Journal of New Music Research*, In Press.
- Gulati, S., Serrá, J., Ishwar, V., & Serra, X. (2014). Melodic pattern extraction in large audio collections of indian art music. In *International Conference on Signal Image Technology and Internet Based Systems - Multimedia Information Retrieval and Applications.*, (pp. 264–271).
- Kroher, N., Díaz-Báñez, J. M., Mora, J., & Gómez, E. (2016). Corpus cofla: A research corpus for the computational study of flamenco music (in press). *ACM Journal on Computing and Cultural Heritage*.
- Kroher, N. & Gómez, E. (2016). Automatic transcription of flamenco singing from polyphonic music recordings. *IEEE-Transactions on Audio, Speech and Language Processing*, 24(5), 901–913.
- Needleman, S. B. & Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology*, 48(3), 443–453.
- Pikrakis, A., Gómez, F., Oramas, S., Díaz-Báñez, J. M., Mora, J., Escobar-Borrego, F., Gómez, E., & Salamon, J. (2012). Tracking melodic patterns in flamenco singing by analyzing polyphonic music recordings. In *13th International Society for Music Information Retrieval Conference (ISMIR)*.
- Volk, A. & van Kranenburg, P. (2012). Melodic similarity among folk songs: An annotation study on similarity-based categorization in music. *Musicae Scientiae*.

# PUBLISHING THE JAMES GOODMAN IRISH MUSIC MANUSCRIPT COLLECTION: HOW MODERN TECHNOLOGY FACILITATED THE EDITORS' TASK

**Lisa Shields**

Irish Traditional Music Archive  
Lisashields1425@gmail.com

## ABSTRACT

The paper gives a description of an important mid-nineteenth-century manuscript Irish music collection. It outlines the history of the edition and the work involved. The use of modern technology in the editorial process is considered. Undoubtedly these technological advances have been very helpful. However, they have also enlarged the scope of the project, creating new kinds of work which are seen as adding value to the product.

## 1. INTRODUCTION

This paper gives an overview of the James Goodman collection of Irish traditional music and the work undertaken by its editors in bringing about its publication, in print and on line, by the Irish Traditional Music Archive. The main focus of this paper is to consider to what extent the advent of new technology aided and expanded the task of the editors, and made the fruits of their work more accessible to the public.

## 2. CANON JAMES GOODMAN (1828–1896)

James Goodman was a native of Dingle, Co. Kerry, who spoke Irish from childhood. He was also a Church of Ireland clergyman who, towards the end of his life, was Professor of Irish at Trinity College Dublin (TCD). He was a proficient performer on the uilleann pipes.

Goodman collected a large number of local traditional piping tunes and songs in Irish. In the 1860s he compiled a very large manuscript collection of these tunes and song airs, much of it directly from live performances.

## 3. THE MANUSCRIPTS

The manuscripts (containing 2,300 tunes in four volumes) were donated after his death to the TCD Library, where they lay unpublished until 1998. As well as the tunes Goodman collected directly 'from Munster pipers' the manuscript also contains copies from borrowed manuscripts and pieces deriving directly or indirectly from printed sources.

## 4. THE WORK OF THE EDITORS

An edition was envisaged which would comprise those tunes that Goodman took from oral tradition or from lost manuscripts, reset in staff notation, with errors corrected and tunes evidently taken from printed sources

eliminated. The edition was also to provide indexes, background material and information about the individual tunes—including those not selected.

## 5. HISTORY OF THE EDITION

The piper and music scholar Breandán Breathnach carried out some preparatory work for the edition, but this was interrupted by his death in 1985. After his death the song collector and music scholar Hugh Shields took over the editing. He compiled a database of the whole collection, with eight fields containing information about the sources and structure of each tune, notes and the numerical codes devised by Breathnach (Breathnach 1982).<sup>1</sup> This database (written specially in Fortran) served as a tool for tune analysis, generating indexes and eliminating duplicates by means of Breathnach codes. It also formed the basis for the eventual online annotated index. It had a Boolean search facility which retrieved items using search terms from a combination of fields.

The Irish Traditional Music Archive (ITMA) published the first volume of the two-volume edition in 1998, edited by Hugh Shields (Shields 1998). This consisted of a selection of 516 tunes, all of which were in Goodman's own settings from local musicians, identified by Goodman himself with the letter 'K'. (His main informant seems to have been the piper Thomas Kennedy.) The volume had a biographical introduction, description of editorial procedures, musical scores and an index of the tunes.

Hugh Shields had done much preliminary work towards a second volume, and after his death in 2008 his wife Lisa undertook the editing of the second volume (Shields & Shields 2013). This was to include a further selection of over 500 tunes, an introduction about the sources, description of the manuscripts, bibliography, title index of the tunes in the edition and a further title index of all the 2,300 tunes in the manuscripts.

### 5.1 Electronic Supplements

A free supplemental index, based on the database, is available from ITMA on line as searchable PDF and

---

<sup>1</sup> These codes provide a method of identifying tunes by assigning the numbers 1–7 to represent degrees of the musical scale to the stressed notes in the first couple of bars. The tonic note, usually the final, is given the number 1. Upper and lower octaves are indicated by a rule above or below the number. Anacruses and ornamentation are ignored. See Figure 1.

HTML downloads.<sup>2</sup> This gives information about the structure and provenance of each of the tunes in the whole manuscript collection, briefly indicating parallels in printed sources and modern practice of session musicians. To have printed this in the edition would have added another 70 or so pages to the book, making it unwieldy and expensive. Also there would have been no way to enter corrections or add further information to the entries.

Since the edition included a printed list of the sources referenced by abbreviations in the online index it was found necessary to put on line a free electronic version of this bibliography. This was enhanced by including some 150 live links leading directly to those publications which are publicly available in electronic format—a laborious task, but worth the extra trouble. Permanence of these online supplements is ensured by their being hosted on a stable archival website.

## 6. HOW MODERN TECHNOLOGY HELPED THE EDITORS

A cooperative publishing venture requires constant communication between the various people involved. Obviously email and the possibility of sending large files electronically via Dropbox has made the proofing and correction of text and music scores much more efficient.

### 6.1 Resources used for Recognizing Tunes from Print in order to Eliminate them

One major problem in preparing the second volume of the edition was how to identify items taken from print (including print-derived tunes copied from manuscripts Goodman had on loan). Goodman acknowledged a good few of his printed sources but on the whole he is reticent about the provenance of his tunes.

Formerly, in order to identify tunes suspected as being from print the editors would have had to travel to libraries near and far to consult rare early volumes. Between the publication of the first and second ITMA volumes (a fifteen-year gap) facsimiles of a multitude of these early music collections have become available on and freely downloadable on the Internet. Goodman has many tunes of Scottish origin, and the collections of the National Library of Scotland<sup>3</sup> have been particularly useful. The interactive online Irish collections hosted by ITMA<sup>4</sup> and Na Piobairi Uilleann (Irish pipers' organization)<sup>5</sup> are also very valuable.

#### 6.1.1 Advantages and Limitations of Breathnach Codes

The starting point in identifying tunes from print is normally the title. However, titles are notoriously variable and the same tune may have several titles. Various aids were available to the editors. Initially the editors mostly used the Breathnach codes mentioned above. These were used in Breathnach's own card index of tunes (now housed in ITMA), also by James Gore's *Scottish Fiddle*

*Index* (Gore et al. 1994) and by the very large online databases of EAMES (Colonial Music Institute 2002). The codes had the great merit in being simple and also independent of written pitch, but they had serious limitations. They took into account merely the first few bars of the first part of the tune. Problems arose too when it was not easy to discern the correct tonic for the tune, or when reels and hornpipes had been written in 2:4 rather than 4:4 or common time. Nevertheless the codes were considered useful enough to be included in the online index to the edition, with alternatives suggested in doubtful cases.

The editors found them of practical use in many ways. Being armed with a print-out of doubtful tune titles with their codes, they were able easily to recognize similar tunes in library collections being consulted. They were also able quickly to answer queries as to whether a particular melody (which might be nameless) existed in the Goodman collection. Breathnach's system predates the digital era but, if the computer could be taught to recognize the tonic reliably, the codes would (if stripped of the indications of octave position) lend themselves readily to computerized retrieval techniques.

#### 6.1.2 Electronic Tune-finding and Tune-recognizing Methods

More sophisticated web-based tune-finding strategies have recently been developed, mainly based on large internet collections of traditional dance music in ABC format<sup>6</sup> (such as the mainly Irish ones by Norbek (1996–2016) and *The Session*<sup>7</sup>). Bryan Duggan's query-by-playing *Tunepal* program<sup>8</sup> proved extremely useful in providing clues as to alternative titles and related tunes. It can recognize similarities from a short sample played instrumentally into the program from any part of the tune, with the limitation (at present) that it is dependent on the written pitch of the ABC files. An innovative newcomer not available to the editors at the time is Chris Walshaw's *TuneGraph* (like Breathnach's system this is pitch-independent, and relies mainly on the stressed notes). It is integrated into his *TuneSearch*<sup>9</sup> site (Walshaw 2015).

There are some other very good tune-finding sites based on ABC notation, a notably informative one being Andrew Kuntz's *Fiddler's Companion* (Kuntz 2003–2012). A reliable and well-organized site, but confined mainly to the era of recorded sound, is Alan Ng's *Irish Traditional Music Tune Index* (Ng 2000–2015). It is unusual in presenting results not as ABC or midi files, but as short samples of actual recorded music. It is likely that great advances will be continue to be made in the area of tune-similarity recognition (perhaps by applying some of the machine-learning techniques employed by scientists in recognizing patterns in gene sequences).

<sup>2</sup> <http://www.itma.ie/digitallibrary/print-collection/tunes-of-the-munster-pipers-vol-2>

<sup>3</sup> *Music at the NLS*. <https://archive.org/details/nlsmusic>

<sup>4</sup> <http://www.itma.ie/digitallibrary/interactivescores-all>

<sup>5</sup> *Irish Music Collections Online*. <http://pipers.ie/imco/>

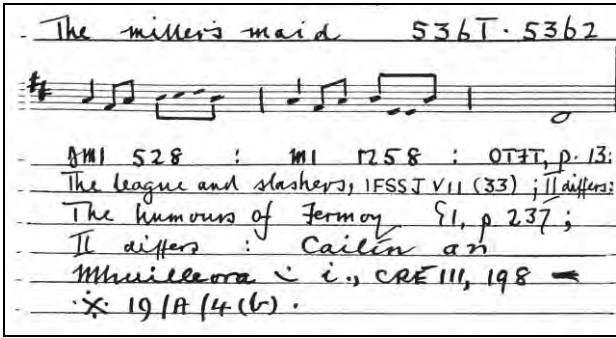
<sup>6</sup> ABC is a convenient text-based format using the actual letters of music notes, which can be quickly displayed in staff notation. It has become the most popular file-sharing medium for music scores of traditional instrumental music. See Figure 1.

<sup>7</sup> *The Session*. <https://thesession.org/>

<sup>8</sup> *Tunepal: a Query-by-Playing Search Engine for Traditional Tunes*. <https://tunepal.org/>

<sup>9</sup> *TuneSearch*. <http://abcnotation.com/search>





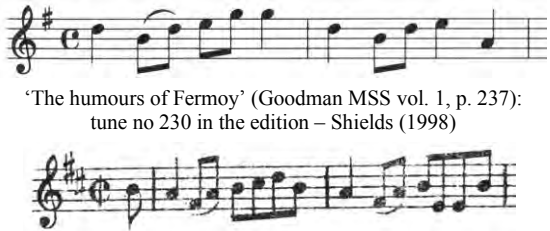
Sample of Breandán Breathnach's index card with references to tunes with the same code.

Unstressed notes are disregarded.

The tune shown above (four stresses per bar) ends in D and its tonic is D.

The number 1 (in the musical scale 1–7) is here assigned to D.

Two tunes with the same code 5361 • 5362



'The humours of Fermoy' (Goodman MSS vol. 1, p. 237):  
tune no 230 in the edition – Shields (1998)

'The miller's maid':  
tune no 528 in O'Neill (1907)

Part of the ABC notation of the second tune:

```
T:The miller's maid [title]
R:Reel [type of tune]
M:C [time signature]
L:1/8 [default note length]
K:D [key]
B|A2(FA) BcdB|A2(FA) BEEB||...
```

**Figure 1.** Examples illustrating Breathnach code and ABC format

## 6.2 How Technology has Extended the Scope of Publications and Enabled them to Reach a Wider Audience

ITMA has included the musical content of the edition (1,051 Goodman tunes) in its free online collection of over 6,000 interactive scores produced by Jackie Small. These have now been integrated into ITMA's remarkable new *PORT*<sup>10</sup> program being developed by Piaras Hoban. This enables interactive scores to be searched simultaneously across many early Irish printed and manuscript music collections.

It is hoped that the annotated online index of the Goodman collection that emerged from the editing of Volume Two will be found valuable as a research tool. It is periodically updated so that, as new search tools are developed and existing ones improved, its usefulness will continue into the future.

## 7. CONCLUSION AND FUTURE DEVELOPMENT

The project was initiated in the pre-digital era. The recent electronic tune-finding and tune-recognition techniques would not have been particularly useful in the production of the first volume of the edition. That is because the selection of tunes was ready-made—confined to those 516 items from local Munster musicians (marked 'K' by Goodman).

Volume Two was another matter, because of the need to identify and exclude music deriving from print. In this case and in other ways modern technological conveniences have definitely been of great assistance to the editors. On the other hand, they have actually increased the work-load by making it possible to extend the scope of the publication in ways not previously possible. The labour involved in the publication and

maintenance of the online supplements and the posting of the music by ITMA on its website was felt to be justified as these enhancements add value to the production and reach a world-wide audience via the internet.

It has been gratifying to learn that ITMA has initiated a collaborative project with TCD to make facsimile digitizations of the whole Goodman manuscript music collection publicly available on the ITMA website [itma.ie](http://itma.ie). For an example of a facsimile page see Figure 2 on the next page. The digital collection will be launched at an ITMA Goodman Symposium in October 2016.



Canon James Goodman

<sup>10</sup> *Port: An ITMA Tune Resource*. <http://port.itma.ie/>



**Figure 2.** The airs of three lyric songs in Irish on the first page of the collection: ‘The bright dawn of day’, ‘The smooth hill where the dark woman lives’, ‘Breens Fort’

## 8. REFERENCES

- Breathnach, B. (1982). Between the jigs and the reels. *Ceol* V(2), 43–48 [an explanation of his system of numerical coding].
- Colonial Music Institute (2002). *Early American Secular Music and its European Sources, 1589–1839: an Index*. <https://www.cdss.org/elibrary/Easmes/Index.htm> (corrected 30 September 2004).
- Gore, C. et al., eds (1994). *Scottish Fiddle Music Index. Tune Titles for the 18th & 19th-century Printed Collections*. Musselburgh, Scotland: Amasing. (electronic version at <http://www.scottishmusicindex.org/about.asp>).
- Kuntz, A. (2003–2012) *The Fiddler's Companion*. <http://www.ibiblio.org/fiddlers/> (This has now been converted to Wiki format, enhanced and renamed *The Traditional Tune Archive*). <http://www.tunearch.org/>
- Ng, A. (2000–2015) *Irish Traditional Music Tune Index*. <https://www.irishtune.info/> (constantly updated).
- Norbeck, H. (1996–2016) *Henrik Norbeck's Abc Tunes*. <http://www.norbeck.nu/abc/>.
- O'Neill, F., ed. (1907). *The Dance Music of Ireland: 1001 Gems*. Chicago: Lyon & Healy.
- Shields, H., ed. (1998). *Tunes of the Munster Pipers: Irish Traditional Music from the James Goodman Manuscripts* [vol. 1]. Dublin: Irish Traditional Music Archive (ITMA). This has long been out of print, but a reprinting is due in summer 2016.
- Shields, H. & L., eds (2013). *Tunes of the Munster Pipers: Irish Traditional Music from the James Goodman Manuscripts* (vol. 2). Dublin: Irish Traditional Music Archive.
- Walshaw, C. (2015). TuneGraph: an online visual tool for exploring melodic similarity. In A. Maragiannis (ed.), *Proc. Digital Research in the Humanities and Arts*, (pp. 55–64). London: University of Greenwich. Retrieved 5 May 2016 from <http://www.drha2014.co.uk>

# CONSTRUCTING PROXIMITY GRAPHS TO EXPLORE SIMILARITIES IN LARGE-SCALE MELODIC DATASETS

Chris Walshaw

Department of Computing & Information Systems,  
University of Greenwich, London SE10 9LS, UK  
c.walshaw@gre.ac.uk

## ABSTRACT

This paper investigates the construction of proximity graphs in order to allow users to explore similarities in melodic datasets. A key part of this investigation is the use of a multilevel framework for measuring similarity in symbolic musical representations. The basis of the framework is straightforward: initially each tune is normalised and then recursively coarsened, typically by removing weaker off-beats, until the tune is reduced to a skeleton representation with just one note per bar. Melodic matching can then take place at every level: the multilevel matching implemented here uses recursive variants of local alignment algorithms, but in principle a variety of similarity measures could be used. The multilevel framework is also exploited with the use of early termination heuristics at coarser levels, both to reduce computational complexity and, potentially, to enhance the matching qualitatively. The results of the matching algorithm are then used to construct proximity graphs which are displayed as part of an online interface for users to explore melodic similarities within a corpus of tunes.

## 1. INTRODUCTION

### 1.1 Background

This paper presents an investigation into constructing proximity graphs using a multilevel melodic similarity metric. The resulting graphs are displayed as part of an online interface for users to identify related tunes, in particular, those found within the abc notation music corpus.

Abc notation is a text-based music notation system popular for transcribing, publishing and sharing music, particularly online. It was formalised and named by the author in 1993 and since its inception he has maintained a website, now at [abcnotation.com](http://abcnotation.com), with links to resources such as tutorials, software and tune collections.

In 2009 the functionality of the site was significantly improved with an online tune search engine which currently indexes over 500,000 abc transcriptions, mostly folk and traditional music, from across the web. Users of the tune search are able to view, listen to and download the staff notation, MusicXML, MIDI representation and abc code for each tune, and the site currently attracts around half a million visitors a year.

In 2014 the search was enhanced with the introduction of TuneGraph, an online visual tool for exploring melodic similarity, [1]. TuneGraph uses a similarity measure to derive a proximity graph representing similarities within the abc notation corpus backing the search engine. From this a local graph is extracted for each vertex, aimed at indicating close variants of the underlying tune represent-

ed by the vertex. Finally an interactive user interface displays each local graph on that tune's webpage, allowing the user to explore melodic similarities.

A typical page display, is shown in Fig. 1, with the tune in standard notation, the MIDI player, the abc notation and the TuneGraph of close variants (top right). One of the close variants has been selected by the user (the vertex is enlarged) and is displayed below by the TuneGraph viewer (bottom right).

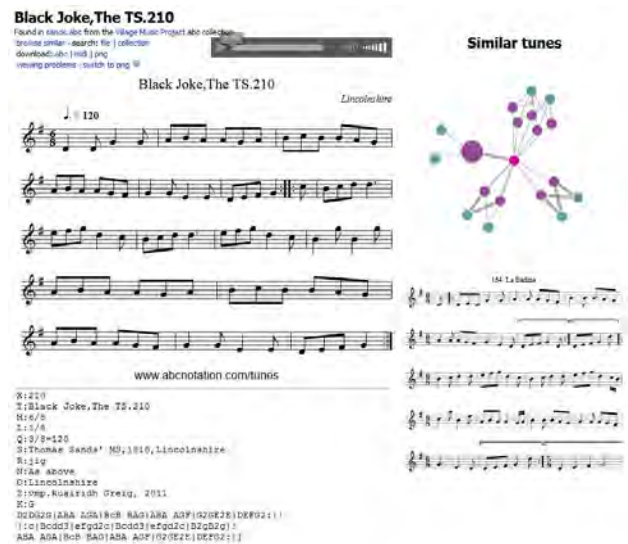


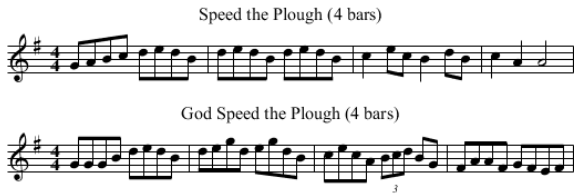
Figure 1. An example of a tune page.

A problem with the initial version of TuneGraph is that the similarity measure used to assess the proximity of variants is based on the incipit only (first three bars, neglecting any anacrusis). Of course not all closely related incipits result from closely related tunes, so this paper considers a different similarity measure which uses a multilevel representation of each tune in its entirety.

The introduction of this new representation has led to an investigation into the construction process for these graphs and a much better understanding of the parameters involved. That investigation is presented here.

### 1.2 Organisation

The rest of the paper is organised as follows. The multilevel paradigm is not (yet!) accepted as a valuable tool in the symbolic music analysis toolkit so section 2 presents a rationale. In section 3 the multilevel matching implementation, and its use in the construction of the proximity graphs, is discussed: this includes two recursive variants of local alignment algorithms and a similarity measure adapted to handle their globalised nature. Experimentation and results follow in section 4 and finally, in section 5, conclusions are presented.



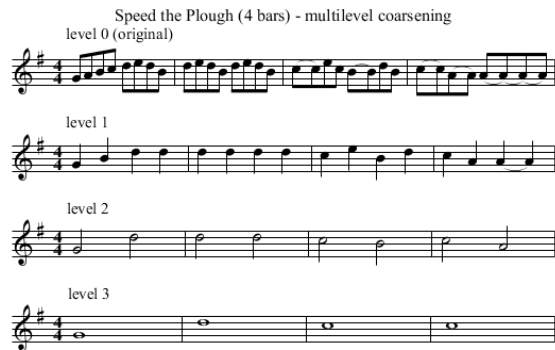
**Figure 2.** Two tune variants for Speed the Plough.

## 2. MULTILEVEL MATCHING: RATIONALE

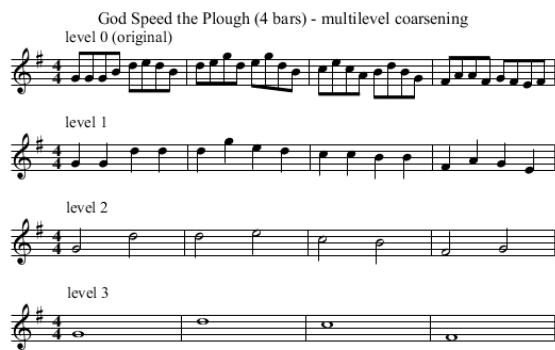
Fig. 2 shows two versions of the first 4 bars of Speed the Plough, a tune well-known across the British Isles (at the time of writing the abcnotation.com tune search has 277 tunes with a title which includes the phrase “Speed the Plough”, of which 157 are exact electronic duplicates. The first version in Fig. 1 is drawn from an English collection and the second, with the title “God Speed the Plough”, from an Irish collection. Clearly these tunes are related but with distinct differences, particularly in the second and fourth bars.

It is typical in tunes like this that the emphasis is placed on the odd numbered notes, and in particular the first note of each beam. The strongest notes of the bar are thus 1 and 5, followed by 3 and 7.

To capture this emphasis when matching tune variants it might be possible to use some sort of similarity metric which weights stress (so that matching 1<sup>st</sup> notes carry more importance than, say, 2<sup>nd</sup> notes, e.g. [2]). However, in this paper the approach is to build a multilevel (hierarchical) representation of the tunes.



**Figure 3.** Multilevel coarsening of Speed the Plough



**Figure 4.** Multilevel coarsening of God Speed the Plough

Figs. 3 & 4 show multilevel coarsened versions of the original tunes, where the weakest notes are recursively replaced by removing them and extending the length of the previous note by doubling it.

At level 0, i.e. the original, the tunes are quantised to show every note as a sixteenth note, thus simplifying the coarsening process. In addition the triplet in bar 3 of “God Speed the Plough” is simplified by representing it as two eighth notes, the first and last notes of the triplet.

To generate level 1, the 2nd, 4th, 6th and 8th notes are removed from each bar; for level 2, the original 3rd and 7th notes (which are now the 2nd and 4th) are removed; for level 3, the original 5th note (now the 2nd) is removed. As can be seen, as the coarsening progresses the two versions become increasingly similar and thus provide a good scope for melodic comparisons which ignore the finer details of the tunes.

## 3. IMPLEMENTATION

This section discusses in detail the construction of the proximity graphs. The implementation is mostly straightforward. Each tune is initially normalised & quantised (section 3.1) and then recursively coarsened down to a skeleton representation with just one note per bar (section 3.2). Melodic matching can then take place at every level (section 3.3) using a melodic similarity measure. A proximity graph is induced by the similarity measure (section 3.4) which is then sparsified (section 3.5). Finally section 3.6 discusses how the multilevel framework is used.

### 3.1 Normalisation

As part of the normalisation process, each tune is cleaned of grace notes, chords and other ornaments. Generally most tunes under consideration from the abc corpus are single-voiced, [1], but if not, only the first voice is used for the matching.

Next, each tune is quantised so that longer notes are replaced with repeated notes (e.g. a half note is replaced with 4 eighth notes); more details can be found in [1].

### 3.2 Coarsening

The coarsening works by recursively removing “weaker” notes from each tune to give increasingly sparse representations of the melody. In the current implementation the coarsening strategy considers that the weaker notes are the off-beats or every other note and it is these which are removed (see Figs. 3 & 4). However, it should be stressed that the multilevel framework is not tied to a particular coarsening strategy and any algorithm that can be used (preferably recursively) to reduce the detail in the melody could be used in principle. For example, it should even be possible to use something as complex as a Schenkerian reduction, [3]; conversely many multilevel algorithms in other fields successfully use randomised coarsenings, [4].

Coarsening progresses until there is one note remaining in each bar; it would be possible to take it further, coarsening down to one single note for a tune, but experimentation suggests that the bar is a good place to stop.

Exceptions to the “remove every other note” rule are handled with heuristics, typically for tunes in compound time. Thus for jigs in 6/8, 9/8 & 12/8, which are normally



written in triplets of eighth notes, the weakest notes are generally the second of each triplet. The same applies for waltzes, mazurkas and polskas in 3/4, so that for 3 quarter notes in a bar, the weakest is generally the second. The heuristics for dealing with these, and other less common time signatures, are discussed in [1].

### 3.3 Similarity Measure

Once the multilevel representation is constructed a variety of methods could be used to compare tunes at each level. This is a strength of the multilevel paradigm which is not reliant on a particular local search strategy, [4].

In a recent comparison study Janssen *et al.*, [5], suggest that one of the best similarity measures for finding melodic segments in a corpus of folk songs is local alignment. Meanwhile in previous work the longest current substring (LCSS) was used successfully within a multilevel context for melodic search, [6] (in fact, LCSS is just a special case of local alignment – see section 3.3.2). Therefore, in this paper recursive versions of both local alignment and LCSS are compared (although unlike Janssen *et al.* local alignment is applied to intervals rather than pitches, making it transposition invariant).

#### 3.3.1 Local alignment (LA)

Local alignment is a well-known technique originating from molecular biology. Given two strings it finds the optimal alignment for two sub-sequences of the originals. The algorithm does not require the aligned sub-sequences to match exactly and makes allowances for gaps and substitutions. For example the strings `***abcde**` and `*acfe****` (where the asterisks represent non-matching entries) could potentially be aligned between a and e with a gap at the b and the substitution of d for f. Gaps, otherwise known as insertions and deletions, and substitutions are penalised with weights.

The algorithm is known as local alignment (LA) because, unlike the global alignment algorithms which preceded it, mismatching sub-strings from either side of the alignment are not penalised (i.e. in the example the string of non-matching entries, indicated by asterisks, could be arbitrarily long without changing the alignment score).

To compute the optimal local alignment for two strings of length  $m$  &  $n$ , an  $(m+1) \times (n+1)$  score matrix  $A$  is constructed with the top row and left hand column initialised to zero. The remainder of the matrix is then filled using

$$A(i, j) = \max \begin{cases} A(i-1, j-1) + s(X_i, Y_j) \\ A(i, j-1) + W_{\text{gap}} \\ A(i-1, j) + W_{\text{gap}} \\ 0 \end{cases}$$

$$s(X_i, Y_j) = \begin{cases} W_{\text{match}} & \text{if } X_i = Y_j \\ W_{\text{substitution}} & \text{if } X_i \neq Y_j \end{cases}$$

where  $W_{\text{match}}$ ,  $W_{\text{substitution}}$  and  $W_{\text{gap}}$  represent the weights for a matching or substituted entry or a gap in the aligned sequences. The implementation discussed here follows Janssen *et al.* and uses  $W_{\text{match}} = 1$ ,  $W_{\text{substitution}} = -1$  and  $W_{\text{gap}} = -0.5$ .

This algorithm was introduced by Smith & Waterman, [7]. In fact their original scheme is a little more computationally involved but the scheme above is widely used and is the variant tested by Janssen *et al.*

To calculate the alignment score, and hence the qualitative similarity, the above scheme suffices. However to determine the aligned sub-sequences a traceback procedure is required. The traceback is implemented by recording a matrix of DIAG, UP or LEFT pointers for every entry of the score matrix indicating where the maximum value originated. If the maximum value is zero an END pointer is stored.

The traceback starts at the pointer matrix entry corresponding to the maximum score found and then tracks back through the pointers, terminating when it reaches an END. Diagonal moves indicate contiguous values in the two aligned sub-sequences whilst left or up moves indicate gap in one of them.

#### 3.3.2 Longest Common SubString (LCSS)

The longest common substring algorithm operates in a similar fashion to local alignment filling in an  $(m+1) \times (n+1)$  matrix of alignment values. However, because there is no need to allow for gaps, no traceback is required: the position of the maximum score in the matrix indicates the end of the longest common substring and the value of this entry gives its length.

In fact it is easy to see that, if the local alignment weights  $W_{\text{substitution}}$  and  $W_{\text{gap}}$  are sufficiently large, so that gaps and substitutions can never occur in an optimal alignment, then the LCSS algorithm is just a special case of local alignment.

From here on, therefore, both algorithms, LA and LCSS, will be referred to collectively as local alignment, the main distinction between the two being that LCSS produces exact matching aligned substrings, is faster to compute and requires less memory (there is no need to use a full matrix and a memory efficient version exists which just repeatedly swaps a pair of arrays, one containing the row under calculation and one containing the previous row). Conversely, LA is more computationally complex and more memory intensive (if the traceback is required to identify the sub-sequences), but will generally match longer sub-sequences. Using  $W_{\text{match}} = 1$ , the similarity measures or alignment scores that either algorithm produces represent the length of the sub-sequences aligned, although in the case of LA there may also be penalty weights for gaps and substitutions so that, for example, the matching of `abcde` with `acfe` has a score of  $1 - \frac{1}{2} + 1 - 1 + 1 = 1\frac{1}{2}$ .

#### 3.3.3 Recursive local alignment = global alignment

A problem with using LCSS, and to a lesser extent LA, is that they are local. For example, using LCSS, `ab**ba` has exactly the same alignment score (of 2) when matched with `**ab` and with `ab**ba`, even though the latter seems a far better match. This is because the second match (`ba`) is not accounted for.

This was less of an issue in the predecessor to this paper, [1], where LCSS was used in a multilevel melodic

search algorithm, since search algorithms are typically trying to find the best matches of a short phrase in a dataset of complete melodies. However for matching it is crucial to distinguish between tunes which match well across their entire length and those which perhaps only match for a short segment.

Interestingly Smith & Waterman touch on this in their original paper where they say “the pair of segments with the next best similarity is found by applying the traceback procedure to the second largest element of [the matrix] not associated with the first traceback”, [7]

Unfortunately, just working from the existing matrix may lead to overlapping local alignments, but instead local alignment may be applied recursively as follows: when applied to two strings, S1 and S2, local alignment splits both into three substrings  $S1 = L1 + A1 + R1$  and  $S2 = L2 + A2 + R2$ , where A1 and A2 are the aligned substrings (exact matches for LCSS or potentially with gaps and substitutions for LA), L1 and L2 are the left hand side unmatched substrings and R1 and R2 are the right hand side unmatched substrings (where any of these unmatched substrings may be of length 0). Thus, having found A1 & A2 and split S1 & S2, local alignment can then be applied to L1 & L2 and to R1 & R2.

This procedure continues recursively, terminating when no alignment is found, or one or both lengths of the substrings being aligned are 0. For example, if the start of S1 is aligned with the end of S2 no further recursion is possible as the lengths of L1 and R2 are 0.

This recursion effectively turns the local alignment algorithms LCSS or LA into a globalised similarity measure, giving an alignment score along the length of both strings being compared. Henceforth these Recursive algorithms will be referred to as RLCSS and RLA.

### 3.3.4 Biased recursive local alignment

An issue that became apparent when using recursive alignment, is that the algorithm makes no distinction between one long aligned sequence and several shorter ones. For example (using RLCSS) `abcd****` has the same alignment score (of 4) when compared with `abcd****` and with `**a**b**c**d**`, even though the former seems a good match and the matching with the latter is essentially noise.

To address this, the similarity measure is biased towards longer aligned sub-sequences by taking the 2-norm (square root of the sum of squares) of the alignment scores found by the recursive local alignment. In the above example this means that the **biased recursive local alignment** score is  $\sqrt{4^2} = 4$  when matching `abcd****` with `abcd****`, whereas when matching with `**a**b**c**d**` it is  $\sqrt{1^2 + 1^2 + 1^2 + 1^2} = 2$ . Space precludes detailed empirical evidence of the effect of this biasing but it made a huge difference to the accuracy of the matching in terms of removing false positives from the results (see also section 3.4 for typical impact).

This biased recursive local alignment thus gives a measure,  $S_{XY}$ , expressing the similarity two arrays of intervals X and Y, each representing a tune.

### 3.4 Constructing the fundamental proximity graph

Neglecting the multilevel framework for now, this similarity measure,  $S_{XY}$ , induces a complete weighted graph on the dataset, where the edge weight between each pair of melodies is given by the similarity. Subsequently, when the graphs are displayed, edge thickness is shown in proportion to the weight with similar vertices joined by thick edges and dissimilar ones by thin edges.

However, most edges in the graph will have very small weights as most melodies in the dataset are only similar to a few others. At this point, therefore, it makes sense to restrict the graph to include only edges for tunes which are reasonably close matches. This graph is referred to henceforth as the **fundamental proximity graph** (FPG). (The FPG has an analogue in search: rather than presenting the whole dataset, ordered by increasing distance, typically search results will be restricted to a subset of “reasonably similar” results with some cut-off after which more dissimilar results are not shown.)

This restriction could be achieved in a variety of ways but here it is assessed by a **fundamental matching threshold**, T, and edges between melodies are only included in the FPG if they match across at least some proportion T of their length. More specifically an edge between vertices  $V_x$  and  $V_y$  is excluded if

$$S_{XY} < \max(\text{length}(X), \text{length}(Y)) * T.$$

As an aside, when calculating using this threshold it is also possible to use the minimum length but this results in very short tunes (such as fragments, included in the dataset as examples) matching with many other tunes and their corresponding vertices having very high degree.

Typical values for T in the experiments are 1/2 (very restrictive, excludes almost all edges), 1/3, 1/4, 1/6 and 1/8 (fairly inclusive, allows a lot of false positives). Note that there is no reason for this to be a simple fraction and T could just as easily be set to, say, 0.40 or 0.317; fractions are simply used as they tend to be more expressive.

Note it is not the intention in this paper to determine a definitive value for T (even if such a value exists). In an ideal world this would be a user chosen parameter and in principle it should be possible to set some range of values, e.g. T in the interval [0.125, 0.5], which the user could adjust according to their needs (provided that the lower value is not too small to make the calculation intractable – if set to 0, every edge is included and the fundamental proximity graph is a complete graph).

Note it is not the intention in this paper to determine a definitive value for T (even if such a value exists). In an ideal world this would be a user chosen parameter and in principle it should be possible to set some range of values, e.g. T in the interval [0.125, 0.5], which the user could adjust according to their needs (provided that the lower value is not too small to make the calculation intractable – if set to 0, every edge is included and the fundamental proximity graph is a complete graph).

The use of *biased* recursive local alignment does obscure what these fractions imply exactly, as it is no longer a case of adding up all the recursively aligned scores. To analyse this further consider that a large proportion of melodies in the dataset are 32 bar tunes in an AABB for-

mat. This is very typical in western European folk music and usually means that the tune is written as 16 bars, AB, with repeat markers at the end of each section. For a reel in common time this would be quantised as 8 eighth notes per bar or a total of  $16 \times 8 = 128$  notes (strictly speaking 127 intervals).

So if  $T$  is set to 0.5 then, when using RLCSS, to be included two tunes would need to match exactly across at least half the tune (8 bars or 64 notes).

If  $T$  is set to 0.25 then they would need to match exactly across one a quarter of the tune (4 bars or 32 notes). Alternatively, again with  $T$  set to 0.25, they could match across four segments, each two bars (16 notes) long (in this case  $S_{XY} = \sqrt{16^2 + 16^2 + 16^2 + 16^2} = \sqrt{1,024} = 32$ ); in other words a total of 64 notes or half the tune.

A similar analysis for  $T = 0.125$  shows that the edge can be included if the tunes match exactly over at least:

- a single 2 bar segment (16 notes or an eighth of the tune); or
- four segments, each 1 bar long (so a total of 32 notes or a quarter of the tune); or
- sixteen segments, each  $\frac{1}{2}$  bar long (so a total of 64 notes, or half the tune).

and obviously many other combinations are possible.

This gives a sense of the impact of the biased recursive local alignment: the matching can occur over a single long phrase or several shorter phrases, but for the latter the total length of the matching substrings will be longer.

Using RLA the picture is more difficult to analyse: for any pair of tunes, the aligned sub-sequences will typically be longer than RLCSS (because of the inclusion of gaps and substitutions) but similarity scores will be lower, because of the penalties. In practice, it seems possible to use higher values of  $T$  (e.g.  $1/2$ ,  $1/3$  and  $1/4$ ) to generate the fundamental proximity graph (see section 4.1.1).

### 3.5 Constructing proximity graphs for users

In fact the fundamental proximity graph is never actually constructed, although a sparsified version is. Ultimately the aim is to create a local proximity graph for each tune showing the closest matching variants. There are practical restrictions on the sizes of graphs that can be easily displayed by the website and assimilated by the user, leading the earlier work on TuneGraph to focus on the size/density of the local graphs and to favour those with no more than 40 vertices, [1].

The use of the FPG does help a great deal towards that end but, as will be seen (later, in Table 1), for some settings of  $T$ , it can still result in some vertices with a large number of neighbours (vertex degree) and consequently some very large local graphs.

To reduce some of these (and simplify the construction algorithm as compared with the previous TuneGraph paper which uses iterative bisection), each vertex is compared with every other vertex and only a fixed number of the closest neighbours which also pass the matching threshold are used to create edges in the **sparsified proximity graph** (SPG). The parameter controlling this is  $D$ , the **maximum included degree**, so that each vertex adds a maximum of  $D$  edges into the graph.

For many vertices there will be no neighbours which pass the matching threshold (i.e. no sufficiently similar tunes) but some will end up with significantly higher degree than  $D$  (since, although a vertex  $V$  may only match with a maximum of  $D$  neighbours, many other vertices could match with  $V$ ). Therefore a further sparsification step takes place (as described in [1]) traversing the list of SPG edges (sorted in decreasing order by combined degree of the incident vertices) and removing any edge if both of its incident vertices have degree greater than a pre-specified **minimum sparsification degree**,  $S$ .

The previous TuneGraph paper focussed heavily on the choice of  $D$  and  $S$  putting the emphasis on the size/density of the local graphs probably at the expense of the data that they contain: potentially the local graphs can be made very rich in structure by matching tunes that are not very similar. Here, instead, by ensuring that the edges of the sparsified proximity graph are a subset of those from the fundamental proximity graph, the aim is to create local graphs that are both visually manageable (by sparsifying those which are not) and which do not contain a lot of spurious edges representing dissimilar tunes. Therefore, although considerable experimentation has been carried out with  $D$  and  $S$  (especially since the introduction of the simplified sparsification algorithm), none of that experimentation is presented here and for all the results they are set to  $D = 6$  and  $S = 4$ .

Finally note that the construction of the SPG is essentially a post-processing cleanup operation which aims to eliminate any vertices of high degree so that the graphs are easy for users to assimilate and understand. In fact, experimentation in section 4.1.1 shows that for the more restrictive settings of  $T$  the FPG could be used in place of the SPG with no cleanup necessary (for example for RLA with  $T = 1/2$  the maximum degree of vertices in the SPG is 37 and for RLCSS with  $T = 1/4$  it is just 16).

### 3.6 Using the multilevel framework

It should be clear by now that constructing the sparsified / fundamental proximity graph is a vast computation. Even for the small test dataset used in the experiments with  $N = \sim 5,000$  tunes, it potentially involves  $\sim 12,500,000$  pairwise comparisons, i.e.  $\frac{1}{2} N(N-1)$  and, if every tune were 16 bars long (128 eighth notes), each comparison involves filling in a  $128 \times 128$  matrix (16,384). So in total 3,200,000,000 calculations and that is without using recursion for the local alignment, which could easily double the total. For the full dataset, which currently has  $N = \sim 187,000$  tunes, the complexity is astronomical.

As previously, [1], a straightforward way to cut this down pragmatically is to segment the dataset according to meter, so that tunes are only compared with others in the same meter. In the small test dataset the largest group (which dominates the calculation) then contains  $\sim 1,500$  tunes in 6/8 resulting in 1,125,000 pairwise comparisons. For the full dataset the largest group contains  $\sim 56,000$  tunes in 4/4 which is close to being intractable, but fortunately the multilevel framework can assist here by computing similarity scores at all levels of the multilevel representation, coarse to fine.

At first sight this might seem to increase the computational complexity but the interval arrays are much smaller at the coarsest level than the original. For a typical 16 bar score of a 32 bar tune the arrays will be 16 entries long at the coarsest level rather than the 128 in the original. If the coarse level matching can detect that a pair of tunes does not match, that edge can be excluded from the SPG at the cost of filling in a 16 x 16 matrix (256 entries) as opposed to the 128 x 128 matrix (16,384 entries), a 64-fold saving.

To that end the multilevel similarity calculation uses **level matching threshold**,  $T^l$ , and the multilevel matching is terminated *at any level* if

$$S'_{XY} < \max(\text{length}(X^l), \text{length}(Y^l)) * T^l$$

where  $X^l / Y^l$  are the interval arrays for tunes  $X$  and  $Y$  at level  $l$  of the multilevel representation and  $S'_{XY}$  is the biased recursive local alignment measured between them.

Obviously some matches which should actually be included in the FPG may be filtered out at a coarse level (i.e. those comparisons which fail the level matching threshold at one or more levels but pass the fundamental matching threshold). Therefore the level matching threshold,  $T^l$ , needs to be used with caution and should be more conservative than  $T$  (obviously there is no point making  $T^l$  larger than  $T$  as it would then take precedence at the finest level). Section 4.1.2 conducts some experiments into how these parameters interact.

This approach is referred to as **multilevel filtering (MLF)**: the multilevel similarity scores,  $S'_{XY}$ , are computed and (as timings show in section 4.1.2) are used extensively to filter out dissimilar matches. However, the  $S'_{XY}$  are discarded for  $l > 0$  (i.e. all but the finest level) and the similarity between a pair of tunes is just the score,  $S_{XY} (= S_{XY}^0)$ , from the original representation.

Another way to use the multilevel framework, alongside the filtering, is to sum the similarity scores,  $S'_{XY}$ , at each level to give a multilevel similarity score,  $\sum_l S'_{XY}$ , and to use this when weighting edges. This approach was used successfully for searching the dataset, [6], and is referred to here as **multilevel weighting (MLW)**. No empirical evidence is presented here that this approach is successful – it is rather a matter of opinion as to whether the multilevel representation is a meaningful reduction of the tune (although the effective use of the technique in search results, [6], and the success of the multilevel filtering in section 4.1.2 suggest that it may be).

Finally, if the multilevel representations are not used the matching framework is referred to as **single level (SL)**.

## 4. EXPERIMENTATION

### 4.1 Results – Test Dataset

The initial experimentation uses a small subset of the full abc corpus consisting of the 5,638 abc transcriptions taken from the Village Music Project<sup>1</sup>, a collection of English social dance music mostly transcribed from handwritten manuscript books in museums and library archives. Of these 30 are removed due to implementation limitations (see [1]) leaving 5,608.

<sup>1</sup> See <http://village-music-project.org.uk/>

#### 4.1.1 Fundamental Proximity Graph

The first experiments are to determine the characteristics of the fundamental proximity graph (FPG). Recall from section 3.4 that the FPG only includes edges between two vertices (tunes),  $V_X$  and  $V_Y$ , if the similarity score for the interval arrays which represent them,  $X$  and  $Y$ , is greater than some fraction,  $T$ , of the length the larger array.

Local alignment	Matching Threshold, $T$	Non-isolated vertices	Degree	
			Avg.	Max.
RLA	1/4	3,907	63.89	738
	1/3	3,206	18.49	441
	1/2	1,923	1.06	37
RLCSS	1/8	4,436	17.26	253
	1/6	2,812	1.8	23
	1/4	1,800	0.86	16

**Table 1.** Characteristics of the fundamental proximity graph for the test dataset.

Table 1 shows the results for different values of  $T$  and both local alignment algorithms, RLA and RLCSS, in terms of the number of non-isolated vertices (those with at least one edge), and the average and maximum degree. Obviously the smaller the value of  $T$ , the more edges are included and so the more dense the graph (i.e. the higher the average degree). As mentioned in section 3.4, ideally the user would be allowed to control the value of  $T$  to determine dynamically the restrictiveness of matching and consequently the size/shape of the local graphs.

No direct comparison between RLA and RLCSS is possible but one feature that is immediately apparent from the table is that they induce somewhat different structures on the dataset. Compare, for example, RLA with  $T = 1/3$  against RLCSS with  $T = 1/8$ : both have similar average degree values (18.49 versus 17.26) and hence a similar number of edges but RLA has fewer non-isolated vertices (3,206 versus 4,436) and consequently a much higher maximum degree (441 versus 253). The same features can be observed for RLA with  $T = 1/2$  as compared with RLCSS with  $T = 1/4$  (both have an average degree close to 1).

This proves nothing but does suggest that at a specific graph density, RLCSS connects up more of the vertices.

Finally the previous work on TuneGraph, [1], suggested that, subjectively, the ideal size for the local graphs displayed to users is a maximum of ~40 vertices with a preferred size of ~20. Local graphs typically include two levels of separation so if the average degree of vertices is 20, say, there could potentially be  $20 \times 20 = 400$  vertices in the average local graph. On the other hand, in reality many vertices in the local graphs are connected (for example, if a vertex of degree 20 is part of a clique then its 20 neighbours will all be connected to each other and so its local graph will only contain 21 vertices). However, this does suggest that the minimum values for the matching threshold should be no less than  $T = 1/3$  for RLA and no less than  $T = 1/8$  for RLCSS, so that the average degree does not rise above 20.



At the opposite end of the scale, the maximum values for  $T$  should not be so large that the FPG contains no edges. If the average degree is around 1 and there are around 2,000 non-isolated vertices then the average degree of non-isolated vertices is  $\sim 5,000 \times 1 / 2,000 = \sim 2.5$  (more accurately 2.79 for RLA with  $T = 1/2$  and 2.42 for RLCSS with  $T = 1/4$ ), leading to average local graphs with 5 – 10 vertices.

In summary, this suggests that a reasonable range of values of  $T$  for the user to control is [0.333, 0.5] for RLA and [0.125, 0.25] for RLCSS.

#### 4.1.2 Multilevel filtering

For small or medium sized datasets, such as the test dataset, computational complexity is not a major issue. However, for the entire corpus it is not practical to run the graph construction process in full, hence the development of the multilevel filtering scheme which aims to filter out dissimilar tunes at coarse representations (when the interval arrays are much shorter and the local alignment much faster). The downside is that the multilevel scheme may mistakenly filter out similar tunes.

Tables 2 and 3 explore this with filtering results for the RLA and RLCSS algorithms and for various combinations of  $T$  and  $T^l$ . For the single level (SL) variants no filtering takes place but, as discussed in section 3.6, for the multilevel filtering variants (MLF), the larger the value of  $T^l$  the more edges will be filtered at coarse levels. Most of these edges would not be included in the fundamental proximity graph (FPG) as the underlying tunes are too dissimilar and so the multilevel filtering speeds up the matching. However, as  $T^l$  increases towards  $T$  the tendency is for it to filter out more FPG edges in error. The aim therefore is to find a suitable value of  $T^l$  which minimises both the runtime and the percentage of FPG edges filtered (although the filtered FPG edges are likely to arise from the weakest matches and might subsequently be removed anyway during sparsification).

	$T$	$T^l$	#edges in FPG	#edges in FPG filtered	%age filtered	runtime (s)
SL	1/3	n/a	51,854	n/a		1,188
MLF		1/16		1,734	3.34%	1,415
MLF		1/12		13,451	25.94%	714
MLF		1/8		35,293	68.06%	235
MLF		1/6		47,790	92.16%	84
SL	1/2	n/a	2,970	n/a		1,001
MLF		1/8		294	9.90%	229
MLF		1/6		597	20.10%	75
MLF		1/4		687	23.13%	52
MLF		1/2		1,347	45.35%	50

**Table 2.** Filtering results for the RLA algorithm.

	$T$	$T^l$	#edges in FPG	#edges in FPG filtered	%age filtered	runtime (s)
SL	1/8	n/a	48,405	n/a		900
MLF		1/16		913	1.89%	740
MLF		1/12		7,119	14.71%	316
MLF		1/8		26,593	54.94%	94
SL	1/6	n/a	5,039	n/a		880
MLF		1/12		153	3.04%	328
MLF		1/8		269	5.34%	96
MLF		1/6		1,304	25.88%	35
SL	1/4	n/a	2,410	n/a		842
MLF		1/8		4	0.17%	90
MLF		1/6		8	0.33%	33
MLF		1/4		90	3.73%	25

**Table 3.** Filtering results for the RLCSS algorithm.

Taking the data as a whole first of all, it can be seen that when the FPG is sparse the filtering is more successful. For example, for RLA with  $T = T^l = 1/2$ , the maximum filtration is 45.35% as compared with 92.16% when  $T = T^l = 1/3$ . Similarly for RLCSS with  $T = T^l = 1/4$  the maximum filtration is just 3.73% as compared with 54.94% when  $T = T^l = 1/8$ .

Comparing RLA with RLCSS, however, it is clear that RLCSS is much more successful at not filtering out FPG edges although it may still filter a lot (say more than 10%) if the FPG is not particularly sparse and  $T^l$  is close to  $T$  (for example when  $T = T^l = 1/8$  or  $T = T^l = 1/4$ ).

It is possible to reduce filtering for RLA down to less than 10% but only for the smallest values of  $T^l$ , specifically  $T^l = 1/16$  for  $T = 1/3$  and  $T^l = 1/8$  for  $T = 1/2$ . This is not so useful as the multilevel filtering doesn't improve the runtime so much: for example MLF actually increases the runtime from 1,188 seconds to 1,415 for  $T^l = 1/16$  and  $T = 1/3$ . The runtime results are better for  $T^l = 1/8$  for  $T = 1/2$  and MLF is over 4 times faster than SL (229 seconds as compared with 1,001) with 9.90% filtering – however, this is at the upper end of the range suggested above for  $T$ .

Conversely for RLCSS there are combinations of  $T$  and  $T^l$  which achieve significantly less than 10% filtering and where  $T^l$  is large enough to dramatically improve runtime. The best example is  $T = 1/6$  and  $T^l = 1/8$  where the MLF runtime is 96 seconds as compared with 880 for SL at the expense of only 5.34% filtering. Fortunately, this is in the middle of the range of values of  $T$  that might be appropriate for a user to control (i.e. [0.125, 0.25] – see above). Even at the bottom end of the range,  $T = 1/8 = 0.125$ , it is possible to use  $T^l = 1/12$  and achieve substantial time savings (316 seconds for MLF as compared with 900 for SL) with only 14.71% filtering. At the top end of the range, where the FPG is very sparse it is possible to use  $T = T^l = 1/4$  and see a huge time saving (25 seconds for MLF as compared with 842 for SL) at the expense of only 3.73% filtering.

It is not totally clear why multilevel filtering does not combine so well with RLA as it does with RLCSS but the likelihood is that the sub-sequences found by RLA at the coarse levels do not necessarily match those found at finer levels. Conversely, provided the coarsening algorithm removes the same entries in both strings, then a longest common substring at a finer level will result in corresponding longest common substrings at coarser levels (for example, if `****abcdefgh****` is coarsened to `**aceg**` and subsequently to `*ae*`).

Note also that this is not an unknown occurrence when using the multilevel paradigm in other fields, [4]. Sometimes the more sophisticated local refinement algorithms interact less well with multilevel coarsening and in fact the best combination is often a smart coarsening algorithm with a relatively simple local refinement scheme.

#### 4.1.3 Sample local graph results

Table 4 shows the characteristics of the local graphs produced for three  $T / T'$  configurations using the three different frameworks (SL, MLF & MLW) and RLCSS as the similarity measure. The characteristics are given in terms of the number of local graphs produced (essentially the number of non-isolated vertices for that value of  $T$ , potentially reduced by filtering and sparsification) plus average and maximum values for the number of vertices and edges in each local graph.

There are not many conclusions that can be drawn from this table but it does indicate that for each value of  $T$  the characteristics are similar for all three frameworks (provided a suitable value of  $T'$  is chosen).

	$T$	$T'$	#graphs	#vertices		#edges	
				avg.	max.	avg.	max.
SL	1/8	n/a	4,436	13.5	32	13.9	36
MLF		1/12	4,381	13.2	29	13.5	32
MLW		1/12	4,381	12.6	26	12.9	32
SL	1/6	n/a	2,812	6.0	20	6.4	26
MLF		1/8	2,745	5.8	22	6.1	28
MLW		1/8	2,745	5.8	22	6.2	28
SL	1/4	n/a	1,800	4.0	15	4.1	26
MLF		1/4	1,742	4.0	15	4.1	26
MLW		1/4	1,742	4.0	13	4.0	24

**Table 4.** Local graph results for the RLCSS algorithm.

#### 4.2 Results – entire abc corpus

The second data set is the entire abc corpus which currently consists of around 509,000 tunes from across the web. Of these 273,000 are exact electronic duplicates which are excluded and another 41,500 are potentially copyright and also ignored. A further 7,500 (3.8% of the remainder) are excluded because of implementation limitations (see [1]), leaving a total of 186,847.

Taking into account the various observations above, it seems that a good configuration is RLCSS as the local matching scheme with  $T = 1/6$  and  $T' = 1/8$ .

Table 5 shows local graph characteristics for MLF and MLW both of which took around 24 hours to run. In contrast the runtime prediction for SL was 2 years! (Indeed if sparser local graphs are acceptable, the multilevel frameworks take only around 8 hours for  $T = T' = 1/4$ .)

	#graphs	#vertices		#edges	
		avg.	max.	avg.	max.
MLF	160,157	9.6	44	12.0	120
MLW	160,157	9.3	40	11.6	116

**Table 5.** Local graph results for the entire corpus.

Again, not many conclusions can be drawn from this table other than the similar characteristics of MLF and MLW. However, the resulting local graphs for MLF can be explored at [abcnotation.com](http://abcnotation.com).

## 5. CONCLUSIONS

This paper presented an investigation into constructing proximity graphs using a multilevel melodic similarity metric. It also discussed the use of two recursive variants of local alignment algorithms (RLA & RLCSS) and a similarity measure adapted to handle their global nature.

The results suggest that multilevel filtering, coupled with RLCSS, works well at building proximity graphs from a corpus of tunes significantly speeding up the runtime without filtering out too many matches.

Although further work remains to eliminate some of the minor limitations in the multilevel matching, the results can be explored at [abcnotation.com](http://abcnotation.com).

## 6. REFERENCES

- [1] C. Walshaw, “TuneGraph: an online visual tool for exploring melodic similarity,” in *Proc. Digital Research in the Humanities and Arts*, 2015, pp. 55–64.
- [2] R. Typke, *Music Retrieval based on Melodic Similarity*, 2007.
- [3] A. Marsden, “Schenkerian Analysis by Computer: A Proof of Concept,” *J. New Music Res.*, vol. 39, no. 3, pp. 269–289, 2010.
- [4] C. Walshaw, “Multilevel Refinement for Combinatorial Optimisation: Boosting Metaheuristic Performance,” in *Hybrid Metaheuristics - An emergent approach for optimization*, C. Blum, Ed. Springer, Berlin, 2008, pp. 261–289.
- [5] B. Janssen, P. van Kranenburg, and A. Volk, “A Comparison of Symbolic Similarity Measures for Finding Occurrences of Melodic Segments,” in *Proc. ISMIR*, 2015, pp. 659–665.
- [6] C. Walshaw, “Multilevel Melodic Matching,” in *5th Intl. Workshop on Folk Music Analysis*, 2015, pp. 130–137.
- [7] T. F. Smith and M. S. Waterman, “Identification of common molecular subsequences,” *Mol. Biol.*, vol. 147, pp. 195–197, 1981.

# NOTE, CUT AND STRIKE DETECTION FOR TRADITIONAL IRISH FLUTE RECORDINGS

Islah Ali-MacLachlan, Maciej Tomczak, Carl Southall, Jason Hockman

DMT Lab, Birmingham City University

islah.ali-maclachlan, maciej.tomczak, carl.southall, jason.hockman  
@bcu.ac.uk

## ABSTRACT

This paper addresses the topic of note, cut and strike detection in Irish traditional music (ITM). In order to do this we first evaluate state of the art onset detection methods for identifying note boundaries. Our method utilises the results from manually and automatically segmented flute recordings. We then demonstrate how this information may be utilised for the detection of notes and single note articulations idiomatic of this genre for the purposes of player style identification. Results for manually annotated onsets achieve 86%, 70% and 74% accuracies for note, cut and strike classification respectively. Results for automatically segmented recordings are considerably lower therefore we perform an analysis of the onset detection results per event class to establish which musical patterns contain the most errors.

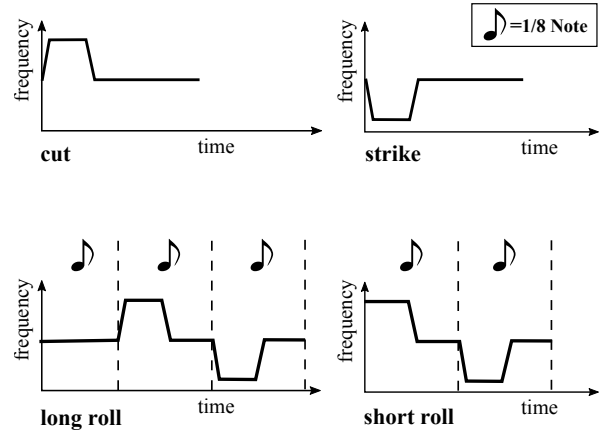
## 1. INTRODUCTION

### 1.1 Background

Irish Traditional Music (ITM) is a form of dance music played on a variety of traditional instruments including the flute. Within the tradition of ITM, players from different backgrounds are individuated based on their use of techniques such as ornamentation, a key factor alongside melodic and rhythmic variation, phrasing and articulation in determining individual player style (McCullough, 1977; Hast & Scott, 2004).

To automatically detect a player's style in audio signals, a critical first step is to detect these notes and ornamentation types. In this paper we evaluate both notes and single note ornaments known as *cuts* and *strikes*. Both ornaments generate a pitch deviation: a cut is performed by quickly lifting a finger from a tonehole then replacing it; a strike involves momentarily covering the tonehole below the note being played. We also analyse the cut and strike elements of multi-note ornaments known as *short roll* and *long roll*.

Figure 1 shows the pitch deviation for cuts and strikes. Long and short rolls are also displayed, showing the inclusion of cut and strike figures. A long roll occupies the same duration as three eighth notes whereas a short roll is equivalent to two eighth notes. In practice, ITM follows a *swing rhythm*—while there is a regular beat, swing follows an irregular rhythm and therefore each eighth-note section may not be of equal duration in normal playing (Schuller, 1991).



**Figure 1:** Frequency over time of *cut* and *strike* articulations showing change of pitch. *Long* and *short rolls* are also shown with pitch deviations (eighth note lengths shown for reference).

### 1.2 Related work

The approach undertaken in this paper utilises onset detection as a crucial first step in the identification of notes and ornaments. There are relatively few studies in the literature that deal specifically with onset detection within ITM, particularly with reference to the flute.

Onsets were found by Gainza et al. (2004) using band-specific thresholds in a technique similar to Scheirer (1998) and Klapuri (1999). A decision tree was used to determine note, cut or strike based on duration and pitch. Kelleher et al. (2005) used a similar system to analyse ornaments on the *fiddle* within Irish music, as bowed instruments also produce slow onsets.

Köküer et al. (2014) also analysed flute recordings through the incorporation of three kinds of information and a fundamental frequency estimation method using the YIN algorithm by De Cheveigné & Kawahara (2002). As in Gainza et al. (2004) a filterbank with fourteen bands optimised for the flute was used. More recently, Jančovič et al. (2015) presented a method for transcription of ITM flute recordings with ornamentation using hidden Markov models.

Unlike the above flute-specific methods, which rely on signal processing based onset detection algorithms, state of the art generalised onset detection methods use proba-

bilistic modelling. The number of onset detection methods using neural networks has substantially risen since Lacoste & Eck (2005). *OnsetDetector* by Eyben et al. (2010) uses bidirectional long short-term memory neural networks, and performs well in a range of onset detection tasks including solo wind instruments.

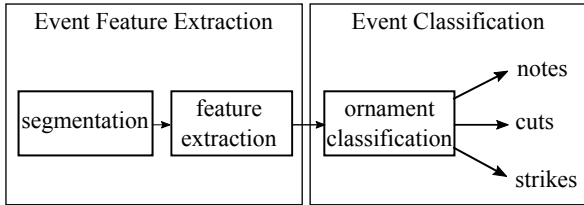
In this paper we perform an evaluation using several modern onset detection algorithms and a dataset comprised of 79 real flute performances of ITM. We then demonstrate how this information may be utilised towards the determination of notes and single note ornamentations.

The remainder of this paper is structured as follows: Section 2 details the method of segmentation, feature extraction and classification. In Section 3 we discuss evaluations of a range of onset detection methods and classification of notes, cuts and strikes. Results of the studies into onset detection and ornament classification are presented in Section 4 and finally conclusions and further work are discussed in Section 5.

## 2. METHOD

Figure 2 shows an overview of the proposed method. We extract features from audio segments representing events (notes, cuts, strikes) and propose an event type classification approach using the segmented event features.

For a fully automated method we use onset detection for segmentation. Event features are then extracted from inter-onset intervals (IOI). These features are used in a supervised learning algorithm to classify the segments as one of three distinct classes: notes, cuts and strikes. For onset detection, we attempt to use the top-performing algorithm from the evaluation presented in Section 3.2, and in the following discuss only the remaining feature extraction and classification stages.



**Figure 2:** Overview of the proposed classification method of notes, cuts and strikes in flute signals. The first phase shows feature extraction from segmented audio events and the second phase shows classification of the events.

### 2.1 Feature extraction

In order to capture the differences between each event type we extract features related to rhythm, timbre and pitch. An important distinction between notes, cuts and strikes is their duration, where notes are significantly longer than the two ornaments. To capture this we use the length *ms* of event segments. We then extract timbral features as

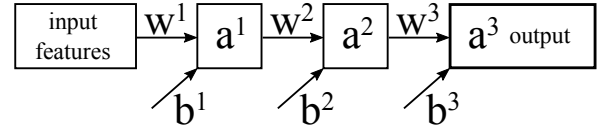
these are also important in class distinction. The change in timbre is caused by player’s fleeting finger motion as a tonehole is temporarily opened or closed. This results in a unique timbre that differs from notes. For that purpose we extract 13 Mel-frequency cepstral coefficients (MFCCs), excluding the first coefficient, and 12 chroma features features to accommodate for timbre and pitch changes in each of the articulations.

To extract features from the audio segments the input audio (mono WAV files) is down sampled to 11,025 Hz. Following the approach in Mauch & Dixon (2010) we calculate the MFCC and chroma features using a Hanning window of 1024 samples with 50% overlap. The extracted features are then normalised to the range [0,1] for every corresponding feature type. Each audio segment is assigned to its class  $\Omega$  (e.g. note). An  $n \times 26$  matrix  $F_\Omega$  is created, where  $n$  represents the number of segments with 26 features (i.e., MFCCs, chroma, durations).

Each  $F_\Omega$  segment appears in the context of musical patterns such as rolls, shakes or just consecutive notes in the recording. To account for the rhythmic, timbral and pitch changes of each event type in the context of these patterns, we concatenate the first derivatives of all features into every  $F_\Omega$  segment.

### 2.2 Neural network classification

Audio segments are then classified into note, cut and strike classes using a feed-forward neural network.



**Figure 3:** Neural network architecture containing two hidden layers  $a^1$  and  $a^2$ , with weights  $w$  and biases  $b$ .

The proposed neural network, shown in Figure 3, consists of two hidden layers containing 20 neurons each. Back propagation is used to train the neural network, updating the weights and biases iteratively using scaled conjugate gradient of the output errors. A maximum iteration limit is set to 10,000 and the weights and biases are initialised with random non-zero values to ensure that training commenced correctly. A validation set is used to prevent over-fitting and cross entropy is used for the performance measure.

The output for each layer of an  $L$  layered neural network can be calculated using:

$$a^{(l)} = f_l(a^{(l-1)}(t)W^l + b^l), \quad (1)$$

where,  $a^l$  is the output at layer  $l$  and  $W$  and  $b$  are the weight and bias matrices. The transfer function is determined by the layer, as shown in Eq. 2.

$$f_l(x) = \begin{cases} 2/(1 + e^{-2x}) - 1, & l \neq L \\ y = e^x / (\sum e^x), & l = L. \end{cases} \quad (2)$$



Classification is performed by finding the index of the maximum value within the output from the neural network.

### 3. EVALUATION

As the performance of the proposed method depends heavily on the accuracy of the chosen onset detection method, the aim of our first evaluation is to determine the best performing onset detection algorithm. We then perform an evaluation of our note and ornament classification.

#### 3.1 Dataset

For both these evaluations, we require a dataset that is representative of a range of respected players with individual stylistic traits. The dataset is comprised of 99 solo flute recordings of between 16 and 81 seconds in length, spanning over 50 years. For the purpose of this study, 79 recordings were selected excluding the excerpts from Larsen (2003), as they contain tutorial recordings not representative of typical ITM performances.

The recordings are 16-bit/44.1kHz WAV files all recorded by professional ITM flute players. Annotations were made using either Sonic Visualiser by Cannam et al. (2010) or Tony by Mauch et al. (2015). The annotation was performed by an experienced flute player. Full details of the annotation methods may be found in Köküer et al. (2014) and Ali-MacLachlan et al. (2015).

Annotations associated with this dataset include the temporal location of onsets and the event type (e.g., note, ornament). Additional classes such as breaths were included in the note class as they contained pitch information from a previous note. The annotated event types are represented by 15,310 notes, 2,244 cuts, and 672 strikes.

#### 3.2 Onset detection evaluation

In this evaluation we measured how well eleven onset detection algorithms were capable of identifying onsets related to notes, cuts and strikes within real-life flute recordings. We reviewed the wind instrument class results from MIREX and examined various studies that concerned detection of soft onsets within these instruments.

Specialised methods for soft onset detection have been proposed in the literature. *SuperFlux* by Böck & Widmer (2013b) calculates the difference between two near short-time spectral magnitudes and is optimised for music signals with soft onsets and vibrato effect in string instruments. *ComplexFlux* by Böck & Widmer (2013a) is based on the *SuperFlux* algorithm with the addition of a local group delay measure that makes this method more robust against loudness variations of steady tones. Similarly, *Log-FiltSpecFlux* introduced in Böck et al. (2012) was designed to deal with onsets of various volume levels but was optimised for real-time scenarios.

In addition, there are several other onset detection methods proposed in the literature that we tested. The *OnsetDetector* by Eyben et al. (2010) processes the input signal both in the forward and backward manner and outputs

peaks that represent the probability of an onset at the detected position. The *Energy* (Masri, 1996), *Spectral Difference* (Foote & Uchihashi, 2001), *Spectral Flux* (Dixon, 2006) and *Kullback-Leibler* (KL) (Hainsworth & Macleod, 2003) represent detection functions solely based in the spectral domain. Brossier (2006) presented a modification to the KL algorithm shown as *Modified Kullback-Leibler* in our evaluation. The *Phase-based* method by Bello & Sandler (2003) looks at phase deviation irregularities in the phase spectrum of the signal. Lastly, the *Complex Domain* approach by Duxbury et al. (2003) combines both the energy and phase information for the production of a complex domain onset detection function. Peak-picking for the evaluate approaches is performed with *Madmom*<sup>1</sup> and *Aubio*<sup>2</sup> MIR toolboxes.

The onset detection results were calculated using the standard precision, recall and F-measure scores that measure performance of each onset detection algorithm. Precision and recall are determined from the detected flute onsets if reported within 25 ms on either side of the ground truth onset times. The mean F-measure is calculated by averaging F-measures across recordings.

#### 3.3 Note and ornament classification evaluation

To assess the performance of our presented note and ornament classification method, we perform two evaluations using the dataset from Section 3.1. In the first evaluation, we attempt to determine the worth of the chosen classification method and selected features alone. In this experiment, we rely on the manually annotated note onsets to segment the audio prior to the feature extraction and classification stages. In the second evaluation, we seek to determine the viability of a fully automated ornament detection approach that relies on onset detection for segmentation. In this evaluation we employ the top performing onset detection algorithm found in the onset detection evaluation detailed in Section 3.2. For the training of the automated method only the true positive onsets will be used to ensure that the neural network is trained with the features corresponding to their correct classes.

To ensure an approximately equal proportion of training examples per class, we reduced the number of notes per recording to 6%, cuts to 30% and left in all strikes due to the proportion of these classes in the dataset. The classification evaluation is then performed using 5-fold cross validation.

## 4. RESULTS

### 4.1 Onset detection results

The results obtained from our experiment are shown in Table 1. The *OnsetDetector* method by Eyben et al. (2010) achieves the highest precision of 83% and F-measure of 78%. The high performance of this approach is in agreement with the results in the literature for the wind instrument class (Böck & Widmer, 2013a,b). While *Spectral*

<sup>1</sup> <https://github.com/CPJKU/madmom>

<sup>2</sup> <http://aubio.org/>

	P	R	F
OnsetDetector2015 Eyben et al. (2010)	<b>0.8306</b>	0.7510	<b>0.7875</b>
ComplexFlux2015 Böck & Widmer (2013a)	0.7414	0.6639	0.6996
SuperFlux2015 Böck & Widmer (2013b)	0.7659	0.6714	0.7144
LogFiltSpecFlux2015 Böck et al. (2012)	0.7597	0.6494	0.6989
Energy Masri (1996)	0.6870	0.5888	0.6270
Complex Domain Duxbury et al. (2003)	0.7548	0.6561	0.6999
Phase-based Bello & Sandler (2003)	0.7206	0.5522	0.6177
Spectral Difference Foote & Uchihashi (2001)	0.7087	0.5928	0.6416
Kullback-Leibler Hainsworth & Macleod (2003)	0.7926	0.4025	0.5265
Modified Kullback-Leibler Brossier (2006)	0.7659	0.1868	0.2890
Spectral Flux Dixon (2006)	0.5854	<b>0.7618</b>	0.6580

**Table 1:** Precision (P), Recall (R) and F-measure (F) for eleven onset detection methods. Maximum values for Precision, Recall and F-measure shown in bold.

class	notes	cuts	strikes
notes	86.97	8.43	8.54
cuts	6.79	70.46	16.96
strikes	6.24	21.07	74.49

**Table 2:** Confusion matrix for classification of notes, cuts and strikes using manually annotated onsets.

class	notes	cuts	strikes
notes	81.57	83.46	82.61
cuts	6.97	6.85	6.48
strikes	11.47	9.70	10.91

**Table 3:** Confusion matrix for classification of notes, cuts and strikes using a fully automated segmentation.

*Flux* achieved the highest recall score of 76% this is likely due to its overestimation of the onset positions thus resulting in a lower precision value. Consequently, in our note, cut and strike detection we use the onsets detected using the *OnsetDetector* as it outperforms other tested methods.

#### 4.2 Note and ornament classification results

Table 2 presents a confusion matrix for note, cut and strike classification using features extracted from the annotated onset boundaries. The results demonstrate the effectiveness of the classification method for all three classes with 86% note, 70% cut and 74% strike detection accuracies. Misclassified notes are equally distributed across the other two classes demonstrating large timbral, pitch and rhythmic differences between note and ornament event types. The cuts and strikes are mostly misclassified as each other, which reflects their similar duration. These findings confirm the importance of duration in identifying the differ-

ence between ornaments (Gainza et al., 2004).

The results for a fully automated system evaluation are presented in Table 3. Here cuts and strikes were overwhelmingly misclassified as notes. These poor results are likely due to the imbalance between the number of annotated onsets and detected onsets. The evaluation using annotated onsets used in 916 notes, 670 cuts and 672 strikes, while the fully automated method used only 691 notes, 503 cuts and 518 strikes.

Training the system with features extracted from annotated segments and testing on automatically found segments did not improve on these results. To investigate the possible reasons for the poor classification results in Table 3, we conducted additional analysis of the onset detection results per event type.

#### 4.3 Note, cut and strike onset detection accuracy

Cuts and strikes are components in multi-note ornaments such as rolls and shakes. To determine where onset detection errors occur we evaluate detection accuracy in relation to events that occurred immediately before and after the detected events. This evaluation allows us to see which event classes are most difficult to detect, and provide insight in the limitations of the real-life application of the proposed method for note, cut and strike detection.

Table 4 presents the onset detection results for each class of musical pattern. The classes consist of three event types where the central event is identified in bold. For example, *note cut note* is a detected cut with a note before and note afterwards, which exists within the event context of short and long roll or a single cut. The number of correctly detected onsets (true positives) is found as a percentage of the overall number of annotated onsets of that pattern.

As can be seen in Table 4, low accuracies were found for notes following ornaments. The largest error was found in the *cut note note*. This pattern exists only in the context of single cuts and shakes and occurred 1579 times with only 574 correctly found instances.

Our proposed note, cut and strike detection method depends on the features extracted from the found inter-onset intervals. The events corresponding to the cut and strike classes are detected with 83% and 82% accuracies respectively. Detecting notes that exist directly after these ornaments in the onset detection stage augments the content of the features describing the ornament event types. This results in training data that does not represent the classes that we intended to capture.

## 5. CONCLUSIONS AND FUTURE WORK

In this paper we present a note, cut and strike detection method for traditional Irish flute recordings. Our chosen approach to this problem is that of inter-onset segment classification using feed-forward neural networks. To evaluate the effectiveness of this approach we first conducted an evaluation of various onset detection algorithms on our dataset with the hope of using this method as a first step in the feature extraction.

Musical pattern			Event context	Accuracy	True positives	Total onsets
note	<b>note</b>	note	<i>single notes</i>	83.36	8651	10378
note	<b>cut</b>	note	<i>short &amp; long rolls &amp; single cuts</i>	83.44	1870	2241
note	<b>note</b>	cut	<i>notes before a roll</i>	84.16	1637	1945
cut	<b>note</b>	note	<i>notes after single cuts &amp; shakes</i>	36.35	574	1579
note	<b>strike</b>	note	<i>short &amp; long rolls &amp; single strikes</i>	82.39	552	670
cut	<b>note</b>	strike	<i>short &amp; long rolls</i>	34.62	180	520
strike	<b>note</b>	note	<i>last notes in rolls</i>	16.22	84	518
cut	<b>note</b>	cut	<i>shakes</i>	31.03	45	145
strike	<b>note</b>	cut	<i>last notes in rolls</i>	20.69	30	145
note	<b>note</b>	strike	<i>notes before single strikes</i>	90.85	129	142

**Table 4:** Onset detection results for each event class (bold) in the context of events happening before and after the detected onset. Accuracy shown as percentage of the accurately detected onsets (true positives) from that pattern

When using ground truth onset annotations, we achieved 86%, 70% and 74% accuracies for note, cut and strike classification respectively. When using detected onsets to train the neural network we achieved poor classification results. We then performed an analysis of the detected onsets and the context in which they appear to establish both the degree of the errors and the musical patterns in which they occur.

In the future we intend to work on improving the automated detection of note events. We will also develop note and ornament classification methods with additional features and other neural network architectures (e.g., recurrent neural networks, networks with long short-term memory) in order to capture trends that appear in time-series data. We also plan to investigate how well the proposed system generalises to other instruments that are characterised by soft onsets such as the tin whistle and fiddle.

## 6. ACKNOWLEDGEMENTS

This work was partly supported by the project Characterising Stylistic Interpretations through Automated Analysis of Ornamentation in Irish Traditional Music Recordings under the Transforming Musicology programme funded by the Arts and Humanities Research Council (UK).

## 7. REFERENCES

- Ali-MacLachlan, I., Köküer, M., Athwal, C., & Jančovič, P. (2015). Towards the identification of Irish traditional flute players from commercial recordings. In *Proceedings of the 5th International Workshop on Folk Music Analysis*, Paris, France.
- Bello, J. P. & Sandler, M. (2003). Phase-based note onset detection for music signals. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 5, (pp. 49–52), Hong Kong.
- Böck, S., Krebs, F., & Schedl, M. (2012). Evaluating the Online Capabilities of Onset Detection Methods. In *Proceedings of the 13th International Society for Music Information Retrieval Conference (ISMIR)*, (pp. 49–54), Porto, Portugal.
- Böck, S. & Widmer, G. (2013a). Local Group Delay Based Vibrato and Tremolo Suppression for Onset Detection. In *Proceedings of the 14th International Society for Music Information Retrieval Conference (ISMIR)*, (pp. 589–594), Curitiba, Brazil.
- Böck, S. & Widmer, G. (2013b). Maximum filter vibrato suppression for onset detection. In *Proceedings of the 16th International Conference on Digital Audio Effects (DAFx)*, (pp. 55–61), Maynooth, Ireland.
- Brossier, P. M. (2006). *Automatic annotation of musical audio for interactive applications*. PhD thesis, Queen Mary, University of London.
- Cannam, C., Landone, C., & Sandler, M. (2010). Sonic visualiser: An open source application for viewing, analysing, and annotating music audio files. In *Proceedings of the 18th international conference on Multimedia*, (pp. 1467–1468), Firenze, Italy. ACM.
- De Cheveigné, A. & Kawahara, H. (2002). YIN, a fundamental frequency estimator for speech and music. *The Journal of the Acoustical Society of America*, 111(4), 1917–1930.
- Dixon, S. (2006). Onset detection revisited. In *Proceedings of the 9th International Conference on Digital Audio Effects (DAFx)*, volume 120, (pp. 133–137), Montreal, Canada.
- Duxbury, C., Bello, J. P., Davies, M., & Sandler, M. (2003). Complex domain onset detection for musical signals. In *Proceedings of the 6th International Conference on Digital Audio Effects (DAFx)*, London, UK.
- Eyben, F., Böck, S., Schuller, B., & Graves, A. (2010). Universal Onset Detection with Bidirectional Long Short-Term Memory Neural Networks. In *Proceedings of the 11th International Society for Music Information Retrieval Conference (ISMIR)*, (pp. 589–594).
- Foot, J. & Uchihashi, S. (2001). The beat spectrum: A new approach to rhythm analysis. In *Proceedings of the 2nd IEEE International Conference on Multimedia and Expo (ICME 2001)*, (pp. 881–884), Tokyo, Japan. IEEE.
- Gainza, M., Coyle, E., & Lawlor, B. (2004). Single-note ornaments transcription for the Irish tin whistle based on onset detection. In *Proceedings of the 7th International Conference on Digital Audio Effects (DAFx)*, Naples, Italy.
- Hainsworth, S. & Macleod, M. (2003). Onset detection in musical audio signals. In *Proceedings of the 29th Inter-*

- national Computer Music Conference (ICMC)*, (pp. 163–166)., Singapore.
- Hast, D. E. & Scott, S. (2004). *Music in Ireland: Experiencing Music, Expressing Culture*. Oxford: Oxford University Press.
- Jančovič, P., Köküer, M., & Baptiste, W. (2015). Automatic transcription of ornamented irish traditional music using hidden markov models. In *Proceedings of the 16th ISMIR Conference*, (pp. 756–762)., Malaga, Spain.
- Kelleher, A., Fitzgerald, D., Gainza, M., Coyle, E., & Lawlor, B. (2005). Onset detection, music transcription and ornament detection for the traditional irish fiddle. In *Proceedings of the 118th Audio Engineering Society Convention (AES)*, Barcelona, Spain.
- Klapuri, A. (1999). Sound onset detection by applying psychoacoustic knowledge. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 6, (pp. 3089–3092)., Phoenix, Arizona.
- Köküer, M., Ali-MacLachlan, I., Jančovič, P., & Athwal, C. (2014). Automated Detection of Single-Note Ornaments in Irish Traditional flute Playing. In *Proceedings of the 4th International Workshop on Folk Music Analysis*, Istanbul, Turkey.
- Köküer, M., Kearney, D., Ali-MacLachlan, I., Jančovič, P., & Athwal, C. (2014). Towards the creation of digital library content to study aspects of style in Irish traditional music. In *Proceedings of the 1st International Workshop on Digital Libraries for Musicology*, London, U.K.
- Lacoste, A. & Eck, D. (2005). Onset detection with artificial neural networks for mirex 2005. *Extended abstract of the 1st Annual Music Information Retrieval Evaluation eXchange (MIREX)*, held in conjunction with ISMIR.
- Larsen, G. (2003). *The essential guide to Irish flute and tin whistle*. Pacific, Missouri, USA: Mel Bay Publications.
- Masri, P. (1996). *Computer modelling of sound for transformation and synthesis of musical signals*. PhD thesis, University of Bristol.
- Mauch, M., Cannam, C., Bittner, R., Fazekas, G., Salamon, J., Dai, J., Bello, J., & Dixon, S. (2015). Computer-aided Melody Note Transcription Using the Tony Software: Accuracy and Efficiency. In *Proceedings of the 1st International Conference on Technologies for Music Notation and Representation (TENOR)*, (pp. 23–30)., Paris, France.
- Mauch, M. & Dixon, S. (2010). Simultaneous estimation of chords and musical context from audio. *Audio, Speech, and Language Processing, IEEE Transactions on*, 18(6), 1280–1289.
- McCullough, L. E. (1977). Style in traditional Irish music. *Ethnomusicology*, 21(1), 85–97.
- Scheirer, E. D. (1998). Tempo and beat analysis of acoustic musical signals. *The Journal of the Acoustical Society of America*, 103, 588.
- Schuller, G. (1991). *The Swing Era: The Development of Jazz, 1930-1945*. Oxford Paperbacks.



# FORMALISING CROSS-CULTURAL VOCAL PRODUCTION

**Polina Proutskova**

Goldsmiths, University  
of London

proutskova@googlemail.com

**Christophe Rhodes**

Goldsmiths, University  
of London

c.rhodes@gold.ac.uk

**Tim Crawford**

Goldsmiths, University  
of London

t.crawford@gold.ac.uk

**Geraint Wiggins**

Queen Mary University  
of London

geraint.wiggins@  
qmul.ac.uk

## 1. INTRODUCTION

How do we speak about the timbre of a singer? How do we compare singers singing the same song? It wouldn't be particularly hard to distinguish a Chinese opera singer from a Western opera singer, but it would be much harder to verbalize how we distinguish them. And when a classical singer performs a rock song, we all hear it is stylistically wrong, but how do we explain to the singer what he needs to change?

All these questions are about vocal production and how it can be captured in words. As it currently stands there is no widely accepted vocabulary to talk about it, not even within a single culture or genre (Garnier, 2007; Mitchell, 2003). Publications in English analysing vocal production in other cultures are rare (Födermayr, 1971; Bartmann 1994). Singing teachers very often use idiosyncratic language based on their subjective perception or learnt from their own teachers, it is hard for teachers from different schools to agree about the terms (McGlashan, 2013). Medical professionals are mainly interested in vocal dysfunction (Little, 2009). Ethnomusicologists focus on the context of music making and rarely mention the sound itself; while for musicologists or music critics it is considered a virtue to use unique terms specific to the particular writer and objectivity of language is not a priority.

We became interested in the subject in the context of MIR, hoping to train a computational model to classify vocal production. Applications would include: differentiating recordings of singing from different cultures; singer recognition; distinguishing originals from covers and covers by different singers; genre classification, etc. All these tasks have been addressed by brute force computational algorithms and by more sophisticated approaches (Tsai 2006, Serra 2010, Holzapfel 2008). Yet there seems to be a glass ceiling of classification accuracy that can be achieved (Karydis 2010, Downie 2008). In MIR it is referred to as “semantic gap” (Wiggins, 2009). If a middle layer could be introduced of more objective categories where further human knowledge is incorporated in the model, that could help improve classification accuracy further.

## 2. MODELS OF VOCAL PRODUCTION

There is no theoretical model of vocal production which could provide the basis for predictions. There are no annotated datasets either. As we have seen above, there isn't even a vocabulary to talk about vocal production. We have found only three approaches to parametrising vocal production that have had a wider reach: one originating in ethnomusicology, another coming from vocal education and one formulated in singing voice science.

Ethnomusicological parametrization was introduced by Alan Lomax in his Cantometrics experiment in which over 5000 recordings from more than 500 cultures were analysed, performance practice was expressed via 36 parameters (Lomax 1976). 13 of these parameters were related to vocal production, including volume, rasp, vocal tension, glottal shake, nasality, vocal pitch, etc. Lomax took an auditory-perceptual approach: human listeners were trained to rate the value of each parameter after listening to an audio recording. Lomax tried to diversify the ratings by getting at least three people to rate each recording. But his raters were mainly US university students with similar life experiences and musical backgrounds. A proper diversification would include people of all ages and professions, from different cultures and with varying musical experience. It is a much bigger undertaking and would have been unworkable in Lomax's circumstances. Only if it were conducted this way though would we be able to say with certainty whether Cantometrics musical parameters are perceived similarly independently of cultural and musical background.

Johan Sundberg, the father of singing voice science, introduced phonation modes describing the voice source aspect of vocal production (Sundberg 1979). They are based on the relationship between subglottal pressure and transglottal airflow. Three of his phonation modes - breathy, neutral and pressed - are widely used by speech and language therapists and in other fields. Sundberg formalised the terms relating them to the aerodynamic processes from which each of the modes originates. He suggests ways to infer phonation mode from an audio recording of singing via inverse filtering. This model works on a milliseconds scale but becomes unmanageable on a seconds scale, which is necessary for humans to recognise music and to feel something about it or deduct its characteristics - the time scale on which the Cantometrics experiment was conducted. Sundberg's phonation model does not include the resonance body aspect, which is crucial for resulting timbre.

Jo Estill was an American singer, teacher and voice researcher, who suggested a physiology-based system for understanding and teaching vocal production. Her idea was to isolate physiological structures, learn to manage them independently and use these building blocks of vocal physiology to construct various kinds of vocal production, ultimately leading to the ability to build any singing style (Estill, 1979; Colton, 1981). While her scientific evidence was partial at best, her work has had a huge influence on contemporary singing education (Sadolin, 2000; Soto-Moretini, 2006; Kayes, 2004).

Since we could not verify the inter-personal and inter-cultural consistency of Cantometrics approach we concentrated on the physiology including phonation. We devised an ontology of vocal production based on Sundberg's and Estill's terminology with some minor additions (Table 1).

physiological dimensions	range	scale	metrics
subglottal pressure	low to high	5-point	interval
transglottal airflow	low to high	5-point	interval
phonation breathy	present/absent	2-point	nominal
phonation pressed	present/absent	2-point	nominal
phonation neutral	present/absent	2-point	nominal
phonation flow	present/absent	2-point	nominal
vocal folds modal vs. falsetto	modal/falsetto	2-point	nominal
vocal folds vibration mode thick to thin	thick/mixed thicker/mixed/mixed thinner/thin	9-point	interval
larynx height	low to high	9-point	interval
thyroid cartilage tilt	vertical/slight tilt/tilted	5-point	interval
cricoid cartilage tilt	vertical/slight tilt/tilted	5-point	interval
velum	low to high	5-point	interval
aryepiglottic sphincter (size of vocal tract)	wide to narrow	5-point	interval
tongue height	low to high	5-point	interval
tongue compression	present/absent	2-point	nominal
position within chest register	low to high	5-point	interval
position within head register	low to high	5-point	interval

**Table 1.** Our ontology of vocal production.

### 3. THE STUDY

The aim of our study is to assess the viability of the physiological approach to modelling vocal production as well as to verify applicability and usefulness of our preliminary ontology of vocal production (Table 1). The study is based on interviews with vocal physiology experts and combines a qualitative and a quantitative approach (Bryman, 2006).

We chose eleven tracks from the Cantometrics dataset (see Chapter on vocal width in Lomax, 1977), all from different musical cultures. Nineteen physiologically stable fragments were extracted from the tracks, which were then used as entities of analysis in the interviews. We recruited 13 participants: otolaryngologists, speech language therapists, singing teachers. Participants' professional involvement with vocal physiology ranged from 10 to over 40 years. Three of them had a non-Western cultural background.

Interviews were structured and lasted from 90 minutes to several hours. Participants were asked to rate physiological dimensions from the preliminary ontology with which they were familiar; they were encouraged to explain their ratings, to point out complexities, to suggest better terms and approaches.

### 4. RESULTS

Participants showed confidence in the majority of terms introduced in the preliminary ontology: only 20% of physiological dimensions were rated by less than 80% of participants. While experts generally supported the ontology, the inter-participant agreement on the ratings was low. Only for two descriptors – position of the larynx and AES – was there a tendency to agreement.

In this talk we shall present the results of the qualitative analysis of the interviews, the analysis of inter-participant (dis)agreement including problem cases and searching for possible causes. We shall demonstrate using the words of our participants how some common themes have emerged from the interviews and how these findings could explain the disagreement. The advantages and disadvantages of physiological vs perceptual approaches to vocal production as well as their possible combinations will be discussed. We shall outline future research directions for this largely understudied area and explain the significance of our findings for academic and applied fields outside MIR.

### 5. REFERENCES

- Bartmann, M. (1994). Rauhgkeiten in der Volksmusik in der Kanarischen Insel El Hierro. In Bockner, M., editor, *Berichte aus dem ITCM- Nationalkomitee Deutschland*, volume 3, Bamberg.
- Bryman, A. (2006). Integrating quantitative and qualitative research: how is it done? *Qualitative research*, 6(1):97–113.
- Colton, R. H. and Estill, J. A. (1981). Elements of Voice Quality: Perceptual, Acoustic, and Physiologic Aspects. In *Speech and language: advances in basic research and practice*, Lass, Norman J. (ed.), volume 5, pages 311–403. Academic Press, New York.
- Downie, J. S. (2008). The music information retrieval evaluation exchange (2005-2007): A window into music information retrieval research. *Acoustical Science and Technology*, 4(29):247–255.
- Estill, J. and Colton, R. (1979). The identification of some voice qualities. *The Journal of the Acoustical Society of America*, 65(S1).
- Födermayr, F. (1971). Zu gesanglichen Stimmgenbung in der außereuropäischen Musik. Ein Beitrag zur Methodik der vergleichenden Musikwissenschaft. Stieglmayr, Wien.
- Garnier, M., Henrich, N., Castellengo, M., Sotiropoulos, D., and Dubois, D. (2007). Characterisation of voice quality in western lyrical singing: from teachers' judgements to acoustic descriptions. *Journal of interdisciplinary music studies*, 1(2):62–91.
- Holzapfel, A. and Stylianou, Y. (2008). Musical genre classification using nonnegative matrix factorization-based features. *Audio, Speech, and Language Processing, IEEE Transactions on*, 16(2):424–434.
- Karydis, I., Radovanovic, M., Nanopoulos, A., and Ivanovic, M. (2010). Looking through the “glass ceiling”: A conceptual framework for the problems of spectral similarity. In *11th International Society for Music Information Retrieval Conference (ISMIR 2010)*.
- Kayes, G. (2004). *Singing and the Actor*. Routledge.

- Little, M., et. al. (2009). Objective dysphonia quantification in vocal fold paralysis: comparing nonlinear with classical measures. *Journal of Voice*.
- Lomax, A. (1968). *Folk Song Style and Culture*. Transaction Books, New Brunswick, New Jersey.
- Lomax, A. (1977). *Cantometrics: A Method of Musical Anthropology* (audio-cassettes and handbook). Berkeley: University of California Media Extension Center.
- McGlashan, J. (2013). What descriptors do singing teachers use to describe sound examples? Presented at *PEVOC 10 (Pan-European Voice Conference)* Prague, Czech Republic.
- Mitchell, H. F., T. Kenny, D., Ryan, M., and Davis, P. J. (2003). Defining ‘open throat’ through content analysis of experts’ pedagogical practices. *Logopedics Phoniatrics Vocology*, 28(4):167–180.
- Sadolin, C. (2000). *Complete vocal technique*. Shout Publishing Copenhagen, Denmark.
- Serra, J., Gomez, E., and Herrera, P. (2010). Audio cover song identification and similarity: background, approaches, evaluation, and beyond. In *Advances in Music Information Retrieval*, pages 307–332. Springer.
- Soto-Morettini, D. (2006). *Popular Singing: A Practical Guide To: Pop, Jazz, Blues, Rock, Country and Gospel*. A&C Black.
- Sundberg, J. (1987). *The science of the singing voice*. Illinois University Press.
- Tsai, W.-H. and Wang, H.-M. (2006). Automatic singer recognition of popular music recordings via estimation and modeling of solo vocal signals. *Audio, Speech, and Language Processing, IEEE Transactions on*, 14(1):330–341.
- Wiggins, G. (2009). Semantic gap?? schemantic schmap!! methodological considerations in the scientific study of music. *Proceedings of IEEE AdMIRE*.

# A METHOD FOR STRUCTURAL ANALYSIS OF OTTOMAN-TURKISH MAKAM MUSIC SCORES

Sertan Şentürk, Xavier Serra

Music Technology Group, Universitat Pompeu Fabra, Barcelona, Spain

{sertan.senturk, xavier.serra}@upf.edu

## ABSTRACT

From a computational perspective, structural analysis of Ottoman-Turkish makam music (OTMM) is a research topic that has not been addressed thoroughly. In this paper we propose a method, which processes machine-readable music scores of OTMM to extract and semiotically describe the melodic and lyrical organization of the music piece automatically using basic string similarity and graph analysis techniques. The proposed method is used to identify around 50000 phrases in 1300 music scores and 21500 sections in 1770 scores, respectively. The obtained information may be useful for relevant research in music education and musicology, and it has already been used to aid several computational tasks such as music score content validation, digital music engraving and audio-score alignment.

## 1. INTRODUCTION

In analyzing a music piece, scores provide an easily accessible symbolic description of many relevant musical components. Moreover they typically include editorial annotations such as the nominal tempo, the rhythmic changes and structural markings. These aspects render the music score a practical source to extract and analyze the melodic, rhythmic and structural properties of the studied music.

Analyzing the structure of a music piece is integral in understanding how the musical events progress along with their functionality within the piece. Automatic extraction of the melodic and lyrical structures, as well as their roles within the composition, might be used to facilitate and enhance tasks such as digital music engraving, automatic form identification and analysis, audio-score and audio-lyrics alignment, music prediction and generation.

Structural analysis is a complex problem which can be approached in different granularities such as sections, phrases and motifs (Pearce et al., 2010). To find such groupings there has been many approaches based on music theory (Jackendoff, 1985), psychological findings and computational models (Cambouropoulos, 2001; Pearce et al., 2010). On the other hand, there are a few studies that have investigated automatic structural analysis of makam musics. Lartillot & Ayari (2009) has used computational models to segment Tunisian modal music and compared the segmentations with the annotations of the experts. Lartillot et al. (2013) has proposed a similar segmentation model for OTMM and also conducted comparative experiments between the automatic segmentations and human annotations. Due to the lack of musicological agreement on how to segment makam music scores, Bozkurt et al. (2014) focused on learning a model from a dataset of music scores

annotated by experts and segmenting larger score datasets automatically using the learned model. They propose two novel culture-specific features based on the melodic and rhythmic properties of OTMM and conduct comparative studies with the features used in the state-of-the-art methods and show that the proposed features improve the phrase segmentation performance.<sup>1</sup> These methods typically focus on finding the segment boundaries and do not study the inter-relations between the extracted segments.

In this study, we propose a method which extracts both the melodic and lyrical organization on phrase-level and section-level using symbolic information available in the music scores of Ottoman-Turkish makam music. The method labels the extracted sections and phrases semiotically according to their relations with each other using basic string similarity and graph analysis. Our contributions are:

- An automatic structural analysis method applied on Ottoman-Turkish makam music scores
- A novel semiotic labeling method based on network analysis
- An open implementation of the methodology extending our existing score parser
- A dataset of sections and phrases automatically extracted from more than 1300 and 1750 music scores, respectively

The structure of the rest of the paper is as follows: Section 2 describes the OTMM score collection we use in our analysis, Section 3 defines the problem and scope of the analysis task, Section 4 presents the proposed methodology, Section 5 explains the experiments and Section 6 discusses our findings, Section 7 gives the use cases where we have already integrated the extracted structure information, finally Section 8 suggests future directions to be investigated and concludes the paper.<sup>2</sup>

## 2. SCORE COLLECTION

In the analysis, we use the release v2.4.1 of the SymbTr score collection (Karaosmanoğlu, 2012).<sup>3</sup> This release includes 2200 scores from the folk and classical repertoires.

<sup>1</sup> For a detailed review of structural analysis applied to OTMM and relevant state of the art we refer the readers to (Bozkurt et al., 2014) and (Pearce et al., 2010), respectively.

<sup>2</sup> The relevant content such as the implementation of the methodology, the score collection, the experiments, the results are also accessible via the companion page <http://compmusic.upf.edu/node/302>.

<sup>3</sup> <https://github.com/MTG/SymbTr/releases/tag/v2.4.1>



It is currently the largest and the most representative machine-readable score collection of OTMM aimed at research purposes (Uyar et al., 2014). The scores typically notate the basic melody of the composition devoid of the performance aspects such as intonation deviations and embellishments. The scores also include editorial metadata such as the composer, the makam, the form, the *usul* (rhythmic structure) of the composition. We use the scores in txt format in our analysis, as they are the reference format from which the other formats are generated.

The content in the SymbTr-txt scores are stored as “tab separated values,” where each row is a note or an editorial annotation (such as *usul* change) and each column represent an attribute such as the note symbol, the duration, the measure marking and the lyrics. The pitch intervals are given according to both the 24 tone-equal-tempered (TET) system defined in the Arel-Ezgi-Uzdilek theory and the 53-TET system.<sup>4</sup> The lyrics are synchronous to the note onsets on the syllable level. The final syllable of each word ends with a single space and the final syllable of each poetic line ends with double spaces (Karaosmanoğlu, 2012). Some columns may be overloaded with additional types of information. For example the lyrics row also includes editorial annotations such as the section names, instrumentation and tempo changes, entered in capital letters.

As will be explained in Section 4.1, we use the explicit section names along with the poetic line ends mentioned above in the section extraction step. However, this set of editorial annotations does not convey the complete information about the section boundaries and the section names. First, the section name (and hence the first note of a section) is only given for the instrumental sections and the final note of these sections are not marked at all. Moreover, the section name does not indicate if there are any differences between the renditions of the same section. Regarding the vocal sections, only the last syllable of a poetic line is marked as explained above. This mark does not typically coincide with the actual ending of the vocal section since a syllable can be sung for longer than one note or there might be a short instrumental movement in the end of the vocal section. Out of 2200, 1771 txt-scores in the SymbTr collection has some editorial section information. The remaining 429 scores either lack the editorial section information or they are very short such that they do not have any sections.

### 3. PROBLEM DEFINITION

As explained in Section 1, symbolic structural analysis is a complex problem that can be approached from different perspectives and granularities. For our initial work in the topic, we assume that the structural elements of the same type are non-overlapping and consecutive (e.g. the last note of a section is always adjacent to the first note of the next section). Consecutiveness restriction also implies that any transitive interactions between two consecutive structural elements are ignored.

<sup>4</sup> The unit interval of the 53-TET, which is simply the 1/53th of an octave, is called a Holderian comma (Hc).

Given the note sequence  $N := \{n_1, n_2, \dots\}$  and the measure sequence  $M := \{m_1, m_2, \dots\}$  in the score, our aim is to extract the sections  $S := \{s_1, s_2, \dots\}$  and the phrases  $P := \{p_1, p_2, \dots\}$  (which we call as structural elements collectively, throughout the text) along with their boundaries, and the melodic and lyrical relationship with other structural elements of the same type. We assume each poetic line as a section.

Remark that each subsequence<sup>5</sup> might cover or overlap with subsequences of different types, e.g. the note sequence in a section would be a subsequence of  $N$  or a phrase might start in the middle of a measure and end in another. We denote the index of the first note and the index of the last note of an score element  $x$  in the note sequence  $N$  as  $\beta(x)$  and  $\gamma(x)$ , respectively. For example, the start of an arbitrary section  $s_i$ , phrase  $p_j$  and measure  $m_k$  are denoted as  $\beta(s_i)$ ,  $\beta(p_j)$  and  $\beta(m_k)$ , respectively.

## 4. METHODOLOGY

We first extract the section boundaries from the score using a heuristic process taking the editorial structure labels in the score as an initial reference (Section 4.1). In parallel, we automatically segment the score into phrases according to a model learned from the phrases annotated by an expert (Section 4.2). Next, we extract the synthetic pitch and the lyrics of each section and phrase (Section 4.3). Then, a melodic and a lyrical similarity matrix are computed between the extracted phrases and the sections separately. A graph is formed from each similarity matrix and the relation between the structural elements in the context of the similarity (melodic or lyrical) is obtained (Section 4.4). Finally semiotic labeling is applied to the computed relations (Section 4.5).

### 4.1 Section Extraction

We infer section boundaries using the explicit and implicit boundaries given in the lyrics column of the SymbTr-txt scores (Section 2). As a preprocessing step to distinguish the instrumental section labels from other editorial annotations in the lyrics column, we extract the unique strings in the lyrics column of all SymbTr scores. We only keep the strings, which are written in capital letters and obtain the set of all editorial annotations in the SymbTr-scores. Then, we pick the section annotations manually.<sup>6</sup>

Given a score, we first search the set of instrumental section names in the lyrics column. The matched note indices mark the actual beginning  $\beta(s_i)$ s of the instrumental sections  $s_i \in S \mid \lambda(s_i) = \emptyset$ . Next, the lyrics column is searched for syllables ending with double spaces. The index of the matched notes are assigned to the final note  $\gamma(s_i)$ s of the vocal sections  $s_i \in S \mid \lambda(s_i) \neq \emptyset$ . As explained in Section 2, the index  $\gamma(s_i)$ s may not coincide

<sup>5</sup> or element, which can also be regarded as a subsequence composed of a single element

<sup>6</sup> [https://github.com/sertansenturk/symbtrdataextractor/blob/master/symbtrdataextractor/makam\\_data/symbTrLabels.json](https://github.com/sertansenturk/symbtrdataextractor/blob/master/symbtrdataextractor/makam_data/symbTrLabels.json)

with the actual ending and it may be moved to a subsequent note.

Up to here we have found the section sequence  $S := \{s_1, s_2, \dots, s_I\}$ , where  $I$  is the total number of sections. The first note of the vocal sections and the last note of the instrumental sections are unassigned at this stage. We proceed to locate the section boundaries using a rule-based scheme iterating though all sections starting from the last one.

If a section  $s_i$  is instrumental, the  $\beta(s_i)$  is already assigned. If a section  $s_i$  is vocal and the previous section  $s_{i-1}$  is instrumental, we find the last instrumental measure,  $m_k \in M \mid \lambda(m_k) = \emptyset$ , before the last note  $\gamma(s_i)$  of the section  $s_i$ . We then assign the first note  $\beta(s_i)$  to the first note  $\beta(m_{k+1})$  of the next measure  $m_{k+1}$ . If both the current section  $s_i$  and the previous section  $s_{i-1}$  are vocal, we assign  $\beta(s_i)$  to the index of the first note with lyrics after the last note  $\gamma(s_{i-1})$  of  $s_{i-1}$ . If  $\beta(s_i)$  and  $\gamma(s_{i-1})$  are not in the same measure, we reassign  $\beta(s_i)$  to the first note of its measure, i.e.  $\beta(m_k) \mid \beta(s_i) \in m_k$ . Finally the last note  $\gamma(s_i)$  of the section is moved to the index of the first note  $\gamma(s_{i+1})$  of the next section  $s_{i+1}$  minus one. The pseudocode of the procedure is given in Algorithm 1. Note that the start of the first section and the end of the final section are assigned to 1 and  $|N|$ , respectively, where  $|N|$  is the number of notes in the score. This detail omitted from the pseudocode for the sake of brevity.

---

**Algorithm 1** Locate section boundaries

---

```

for  $i := L \rightarrow 1$  do           ▷ from the last index to the first
  if  $s_i$  is vocal then         ▷ find  $\beta(s_i)$  of the vocal section
    if  $s_{i-1}$  is instrumental then
       $\beta(s_i) \leftarrow \arg_k \min (\beta(m_k) > \beta(s_{i-1}) \wedge$ 
         $\lambda(m_k) \neq \emptyset)$ 
    else                         ▷  $s_{i-1}$  is vocal
       $\beta(s_i) \leftarrow \arg_k \min (k > \gamma(s_{i-1}) \wedge$ 
         $\lambda(n_k) \neq \emptyset)$ 
      if  $\beta(s_i) \in m_k \wedge \gamma(s_{i-1}) \notin m_k$  then
         $\beta(s_i) \leftarrow \beta(m_k)$ 
       $\gamma(s_i) \leftarrow \beta(s_{i+1}) - 1$    ▷ sections are consecutive

```

---

Having located the boundaries, the sections are extracted by simply taking all information (i.e. rows in the SymbTr-txt score) between these note boundaries. Figure 1 shows the section boundaries obtained on a mock example.

## 4.2 Automatic Phrase Segmentation

In our method we use the only automatic phrase segmentation methodology proposed by Bozkurt et al. (2014) (Section 1). The source code and the training dataset (Karaosmanoğlu et al., 2014) are open and available online.<sup>7</sup>

In order to train the segmentation model, we use the annotations of Expert 1, who annotated all the 488 scores in the training dataset (Karaosmanoğlu et al., 2014). There are a total of 20801 training phrases annotated by the first expert. Using the trained model, we apply automatic phrase

segmentation on the score collection (Section 2) and obtain the phrase boundaries  $\beta(p_k)$  and  $\gamma(p_k)$  for each phrase  $p_k \in P := \{p_1, p_2, \dots\}$ , where  $P$  is the automatically extracted phrase sequence. In Figure 5 (Appendix A), the vertical red and purple lines shows the phrase boundaries extracted from the score “Kimseye Etmem Şikayet.”<sup>8</sup>

## 4.3 Synthetic Pitch and Lyrics Extraction

We use the information in the lyrics column to determine the boundaries of the vocal sections in Section 4.1. Later, the lyrics of each structural element are extracted in Section 4.4 and a lyrical similarity is computed between each structural element of the same type using the extracted. The lyrics associated with a sequence or an element  $x$  is a string denoted as  $\lambda(x)$ , simply obtained by concatenating the syllables of the note sequence  $\{\beta(x), \dots, \gamma(x)\}$  of  $x$ . The editorial annotations (Section 2) and the whitespaces in the lyrics column are ignored in the process. Then the characters in the obtained string are all converted to lower case. Trivially,  $\lambda(n_i)$  of a note  $n_i$  is the syllable associated with the note  $n_i$  in the lyrics column.

Given a subsequence or element  $x$  in the score, the synthetic pitch  $\rho(x)$  is computed by sampling each note in  $x$  according to the note symbol and the duration, and then concatenating all of the samples (Şentürk et al., 2014). The synthetic pitch is used in melodic similarity computation parallel to the lyrics (Section 4.4). Figure 2 shows the lyrics and the synthetic pitch extracted from an excerpt of the SymbTr-score of the composition “Gel Güzelim”.

## 4.4 Melodic and Lyrical Relationship Computation

Given the structure sequence  $F := \{f_1, f_2, \dots\}$  (which is either the section sequence  $S$  or the phrase sequence  $P$ ) extracted from the score, we first compute the synthetic pitch and extract the lyrics of each structural element (Section 4.3). Then, we compute a melodic similarity and lyrical similarity between each element using a similarity measure based on Levenshtein distance (Levenshtein, 1966). The similarity measure  $\hat{L}(x, y)$  is defined as:

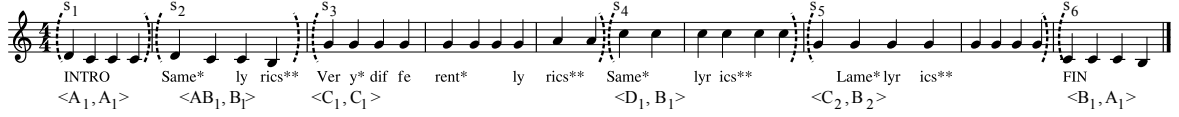
$$\hat{L}(x, y) := 1 - \frac{L(x, y)}{\max(|x|, |y|)} \quad (1)$$

where  $L(x, y)$  is the Levenshtein distance between the two “strings”  $x$  and  $y$  with the lengths  $|x|$  and  $|y|$ , respectively and  $\max()$  denotes the maximum operation. In our case,  $x$  and  $y$  are the synthetic pitch or the lyrics of two structural elements. The similarity yields a result between 0 and 1. If the strings of the compared structural elements are exactly the same, the similarity will be one. Similar strings (e.g. the melodies of two instances of the same section with volta brackets) will also output a high similarity.

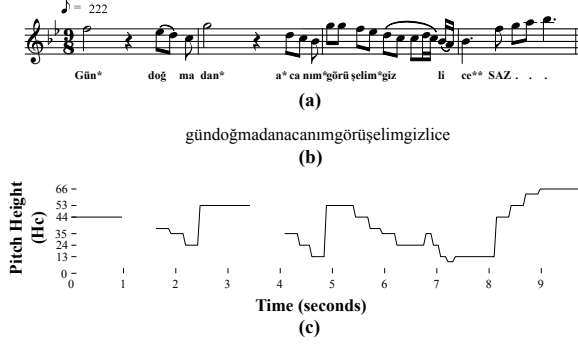
From the melodic and lyrical similarities, we build two separate graphs, in which the nodes are the structural elements and the elements are connected to each other with

<sup>7</sup> <http://www.rhythmos.org/shareddata/turkishphrases.html>

<sup>8</sup> [https://github.com/MTG/SymbTr/blob/a50a16ab4aa2f30a278611f333ac446737c5a877/txt/nihavent--sarki--kapali\\_curcuna--kimseye\\_etmem--kemani\\_sarkis\\_efendi.txt](https://github.com/MTG/SymbTr/blob/a50a16ab4aa2f30a278611f333ac446737c5a877/txt/nihavent--sarki--kapali_curcuna--kimseye_etmem--kemani_sarkis_efendi.txt)



**Figure 1:** Section analysis applied to a mock example. The section labels (“INTRO” and “FIN”) are given in the lyrics written in capital letters, The spaces in the end of the syllables are visualized as \*. The semiotic  $\langle \text{Melody}, \text{Lyrics} \rangle$  label tuples of each section are shown below the lyrics. The similarity threshold in the similar clique computation step is selected as 0.7 for both melody and lyrics.

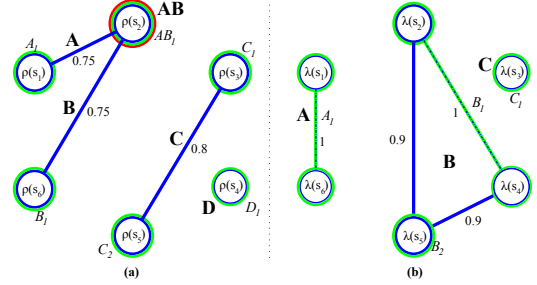


**Figure 2:** A short excerpt from the score of the composition, *Gel Güzelim*. **a)** The score, **b)** the lyrics, **c)** the synthetic pitch computed from the note symbols and durations. The spaces in the end of the syllables are displayed as \*s.

undirected edges. The weight of an edge connecting two structural elements  $f_i$  and  $f_j$  is equal to  $\hat{L}(\rho(f_i), \rho(f_j))$  in the melodic relation graph and  $\hat{L}(\lambda(f_i), \lambda(f_j))$  in the lyrics relation graph, respectively. Next, we remove the edges with a weight less than a constant similarity threshold  $w \in [0, 1]$ . In Section 5, we will investigate the effect of using different  $w$  values.

Given the graph, we obtain the groups of structural elements having similar strings by finding the maximal cliques in the graph (Tomita et al., 2006). A maximal clique is a subgraph, which has its each node connected to each other and it cannot be extended by including another node. We denote these cliques as  $v_j \in V$ , where  $V$  is the set of “similar cliques.” We additionally compute the maximal cliques of the graph only considering the edges with zero weight. These cliques show us the groups of structural elements, which have exactly the same string. We call each of these cliques as “unique clique”  $u_k \in U$ , where  $U$  is the set of the unique cliques. Note that two or more similar cliques can intersect with each other. Such an intersection resembles all the relevant similar cliques. We denote these “intersections” as  $w_l \in W$ , where  $W$  is the set of intersections between different similar cliques. Also,  $\eta(x)$  denotes the nodes of an arbitrary graph  $x$ . Here we would like to to remark a few relations:

- A unique clique is a subgraph of at least one similar clique, i.e.  $\forall u_k \in U, \exists v_j \in V \mid \eta(u_k) \subseteq \eta(v_j)$ .
- A unique clique cannot be a subgraph of more than one intersection, i.e.  $\forall u_k \in U, \nexists \{w_l, w_m\} \subseteq W \mid \eta(u_k) \subseteq \eta(w_l) \wedge \eta(u_k) \subseteq \eta(w_m)$ .



**Figure 3:** The graphs, the cliques and the semiotic labels obtained from the mock example (Figure 1) using an edge weight threshold of 0.7 for both melody and lyrics. The circles represent the nodes and the lines represent the edges of the graphs, respectively. The edge weights are shown next to the lines. Green, blue and red colors represent the unique cliques, the similar cliques and the intersection of similar cliques, respectively. The semiotic label of each similar clique and each intersection is shown in bold and the semiotic label of each unique clique is shown in italic, respectively.

- A structural element belongs to only a single unique clique, i.e.  $\forall f_i \in F, \exists! u_k \in U \mid \eta(f_i) \subseteq \eta(u_k)$ .

Figure 3 shows the graphs computed from the sections of the mock example introduced in Figure 1. In the melodic relations graph, each section forms a unique clique since the melody of each section is not exactly the same with each other. Using a similarity threshold of 0.7, we found four similar cliques formed by  $\{s_1, s_2\}$ ,  $\{s_2, s_6\}$ ,  $\{s_3, s_5\}$ ,  $\{s_4\}$ . Notice that  $\{s_4\}$  is not connected to any clique. so it forms both a unique and a similar clique. Moreover,  $s_2$  is a member of both the first and the second similar cliques and hence it is the intersection of these two cliques. For the lyrics, there are four unique cliques, formed by the sections  $\{s_1, s_6\}$  (aka. instrumental sections),  $\{s_2, s_4\}$ ,  $\{s_3\}$  and  $\{s_5\}$ . The lyrics of  $s_5$  is very similar to  $\{s_2, s_4\}$  and they form a similar clique composed of these three nodes and the relevant edges.

#### 4.5 Semiotic Labeling

After forming the cliques, we use semiotic labeling explained in Bimbot et al. (2012) to describe the structural elements. First we label similar cliques with a base letter (“A”, “B”, “C”, ...). Then we label the intersections by concatenating the base letters of the relevant similar cliques (e.g. “AB”, “BDE”, ...). We finally label each

unique clique with the label of the relevant intersection, if exists and with respect to the relevant similar clique otherwise, plus a number according to the occurrence order of the clique in the score. Right now, we only use the simple labels (e.g. “ $A_1$ ”, “ $A_2$ ”, “ $AB_2$ ”) as termed by Bimbot et al. (2012) to label the unique cliques.

The pseudocode of the process is given in Algorithm 2. During labeling,  $V$ ,  $W$  and  $U$  are sorted with respect to the index of their first occurrence in the score. We denote the label of an arbitrary element  $x$  as  $\Lambda(x)$ . In the algorithm, we also use iterators  $\#(v_j)$  and  $\#(w_l)$  for each similar clique  $v_j$  and each intersection  $w_l$ , which are used to assign the numerical index to each unique clique  $u_k \in U$  according its relation with the relevant similar clique or intersection.

---

**Algorithm 2** Semiotic labeling

---

```

 $\lambda \leftarrow "A"$  ▷ Start the base letter from "A"
 $\#(v_j) \leftarrow 1, \forall v_j \in V$  ▷ Init. the iterators for all  $v_j$ 
 $\#(w_l) \leftarrow 1, \forall w_l \in W$  ▷ Init. the iterators for all  $w_l$ 
for  $v_j \in \text{sort}(V)$  do ▷ Label similar cliques
   $\Lambda(v_j) \leftarrow \lambda$ 
for  $w_l \in \text{sort}(W)$  do ▷ Label intersections
   $\Lambda(w_l) \leftarrow \text{concat. } \Lambda(v_j), \forall (v_j) \mid \eta(w_l) \subseteq \eta(v_j)$ 
for  $u_k \in \text{sort}(U)$  do ▷ Label unique cliques
  if  $\exists w_l \mid \eta(u_k) \subseteq \eta(w_l)$  then ▷ e.g. " $ACD_1$ "
     $\Lambda(u_k) \leftarrow \Lambda(w_l)\#(w_l)$ 
     $\#(w_l) \leftarrow \#(w_l) + 1$ 
  else ▷ e.g. " $C_2$ "
     $\Lambda(u_k) \leftarrow \Lambda(v_j)\#(v_j) \mid \eta(u_k) \subseteq \eta(v_j)$ 
     $\#(v_j) \leftarrow \#(v_j) + 1$ 
for  $f_i \in F$  do ▷ Label structural elements
   $\Lambda(f_i) \leftarrow \Lambda(u_k) \mid \eta(f_i) \subseteq \eta(u_k)$ 

```

---

The label of each section of the mock example is shown below the staff in Figure 1. The same semiotic labels are also shown on the computed graphs in Figure 3. Notice that the melodic semiotic label of  $s_6$  is  $B_1$  because the first occurrence of the relevant similar clique is at  $s_2$ .

By extracting the relations in the graphs computed from the melodic and lyrics similarity matrices (Section 4.4) and then applying semiotic labeling to each section and phrase according to its relation, we obtain a  $\langle \text{Melody}, \text{Lyrics} \rangle$  tuple for each section and phrase (Section 4.5). For each phrase we additionally mark the sections, which enclose and/or overlap with the phrase. Appendix A shows the results of the structural analysis applied to the score “Kimseye Etmem Şikayet.” We leave the examination of the analysis to the readers as an exercise.

## 5. EXPERIMENTS

In (Bozkurt et al., 2014) report the evaluation of the phrase segmentation method (Section 4.2) on an earlier and slightly smaller version of the annotations that we use to compute the segmentation model. We refer the readers to (Bozkurt et al., 2014) for the evaluation of the training data. Furthermore, the labels of the automatic phrase segmentations

need to be validated by musicologists parallel to the discussions brought by Bozkurt et al. (2014). For this reason, we leave investigating the effects of the similarity threshold  $w$  in phrase analysis as future research.

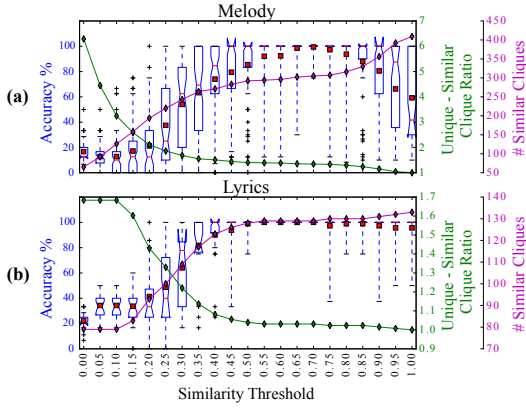
To observe the effect of the similarity threshold in the melodic and lyrical relationship extraction (Section 4.4), we have collected a small dataset from the SymbTr collection. The test dataset consists of 23 vocal compositions in the şarkı form and 42 instrumental compositions in peşrev and sazsemaisi forms. These three forms are the most common forms of the classical OTMM repertoire. Moreover their sections are well-defined within music theory; the two instrumental forms typically consists of four distinct “hane”s and a “teslim” section, which follow a verse-refrain-like structure; the sections of the şarkıs typically coincide with the poetic lines. In our initial experiments we focused on şarkıs with the poetic organization “zemin, nakarat, meyan, nakarat,” which is one of the most common poetic organization observed in the şarkı form. Using the automatically extracted section boundaries (Section 4.1) as the ground-truth, the first author has manually labeled the sections in the scores with the same naming convention explained in Section 4.5.<sup>9</sup> Due to lack of data and concerns regarding subjectivity, we leave the evaluation of section boundaries as future research.

We have conducted section analysis experiments on the test dataset by varying the similarity threshold from 0 to 1 with a step size of 0.05. After the section labels are obtained, we compare the semiotic melody and lyrics labels with the annotated labels. We consider an automatic label as “True,” if it is exactly the same with the annotated label and “False,” otherwise. For each score, we compute the labeling accuracy for the melody and the lyrics separately by dividing number of correctly identified (melody or lyrics) labels with the total number of sections. We additionally mark the number of similar cliques and its ratio to the unique cliques obtained for each score. For each experiment, we find the average accuracy for the similarity threshold  $w$  by taking the mean of the accuracies obtained from each score.

Figure 4 shows the notched boxplots of the accuracies, the total number of similar cliques and the ratio between the number of unique cliques and the number of similar cliques obtained for each similarity threshold. For the melody labels, the best results are obtained for the similarity threshold values between 0.55 and 0.80 and the best accuracy is 99%, when  $w$  is selected as 0.70. For lyrics labeling, any similarity value above 0.35 yields near perfect results and 100% accuracy is obtained for all the values of  $w$  between 0.55 and 0.70. In parallel, the number of similar cliques and the ratio between the unique cliques and the similar cliques gets flat in these regions. From these results we select the optimal  $w$  as 0.70 for both melodic and lyrical similarity.

---

<sup>9</sup> The experiments and results are available at [https://github.com/sertansenturk/otmm-score-structure-experiments/releases/tag/fma\\_2016](https://github.com/sertansenturk/otmm-score-structure-experiments/releases/tag/fma_2016)



**Figure 4:** The notched boxplots of the accuracies, number of similar cliques and the ratio between the number of unique cliques and similar cliques obtained for **a)** the melody labels and **b)** the lyrics labels (only for vocal compositions) using different similarity thresholds. The squares in the boxplots denote the mean accuracy.

## 6. DISCUSSION

As shown in Section 5, the similarity threshold  $w$  has a direct impact on the structure labels. A high threshold might cause most of the similar structural elements regarded as different, whereas a low threshold would result in many differences in the structure disregarded. In this sense the extreme values of  $w$  (around 0 or 1), would not provide any meaningful information as  $w = 0$  would result in all the structures being labeled similar and  $w = 1$  would be output all the structures as unique. We also observe that the melodic similarity is more sensitive to value of  $w$  than lyrics similarity. This is expected as the strings that make up the lyrics are typically more diverse than the note symbols used to generate the synthetic pitch. In our experiments we found the optimal value of  $w$  as 0.7 for the small score dataset of compositions in the peşrev, sazsemai and şarkı forms. Moreover we observe that the curves representing the number of similar cliques and the ratio between the unique cliques and the similar cliques are relatively flat around the same  $w$  value, where we obtain the best results (Figure 4). This implies that there is a correlation between decisions of the annotator and our methodology.

Nevertheless, we would like to emphasize that the  $w$  value found above should not be considered as a general optimal. First of all, the sections were annotated by a single person and therefore our evaluation does not factor in the subjectivity between different annotators. Second, the section divisions in different forms are much different from the forms we have experimented upon, which might influence the structure similarity. For example, we expect many vocal compositions of OTMM with “terennüm”s (repeated words with or without meaning such as “dost,” “aman,” “ey”) need a lower similarity threshold in the lyrics relationship computation step. Moreover the poetic lines might not coincide with melodic sections in many vocal compositions especially in folk music genre. Third, the threshold can be different in different granularities. For example, the

phrases are much shorter than the sections as can be seen in Appendix A. The human annotators might perceive the intra-similarity between sections and phrases differently.

## 7. APPLICATIONS

We have implemented the structural analysis methodology in Python and integrated it to the *symbtrdataextractor* package, a SymbTr-score parser written by us.<sup>10</sup> We have also forked the open automatic phrase segmentation package by Bozkurt et al. (2014), which is written in MATLAB scripting language. The fork modularizes the code and packages it into a standalone binary so it can be integrated to other tools without the need of a MATLAB proprietary license. Moreover, the code is optimized such that it performs considerably faster than the original code.<sup>11</sup> We have been using the information extracted from the structural analysis in several applications:

**Score collection analysis:** Using the optimal similarity threshold ( $w = 0.7$ ), we applied structural analysis on the latest release of the SymbTr collection (Section 2). We have extracted and labeled 49259 phrases from 1345 scores, which have both their makam and usul covered in the phrase segmentation training model. Because there is no training data for the usul variants “Yürüksemai II”, “Devrihindi II”, “Müsemmen II”, “Raksaksağı II”, “Devrituran II” and “Kapalı Curcuna,” we treat them as the most common variant of the same usul, namely “Yürüksemai”, “Devrihindi”, “Müsemmen”, “Raksaksağı”, “Devrituran” and “Curcuna”. In parallel, 21569 sections are extracted from 1771 scores.<sup>12</sup> The data can be further used to study the structure of musical forms of OTMM.

**Automatic score validation:** Structural analysis, along with the other functionalities of the *symbtrdataextractor* package are used in unittests applied to SymbTr collection in a continuous integration scheme to automatically validate the contents of the scores.<sup>13</sup>

**Score format conversion:** We are currently developing tools in Python to convert the SymbTr-txt scores to the MusicXML format<sup>14</sup> and then to the LilyPond format<sup>15</sup> to improve the accessibility of the collection from popular music notation and engraving software. The converters use the information obtained from *symbtrdataextractor* to add the metadata and the section names in the converted scores.

**Audio-score alignment:** In the performances of OTMM compositions, the musicians occasionally insert, repeat and omit sections. Moreover they may introduce musical passages, which are not related to the composition (e.g. im-

<sup>10</sup> <https://github.com/sertansenturk/symbtrdataextractor/>

<sup>11</sup> The fork is hosted at <https://github.com/MTG/makam-symbolic-phrase-segmentation>

<sup>12</sup> The data is available at [https://github.com/sertansenturk/turkish\\_makam\\_corpus\\_stats/tree/66248231e4835138379ddec970eabf7dad2c7f8/data/SymbTrData](https://github.com/sertansenturk/turkish_makam_corpus_stats/tree/66248231e4835138379ddec970eabf7dad2c7f8/data/SymbTrData)

<sup>13</sup> <https://travis-ci.org/MTG/SymbTr/>

<sup>14</sup> <https://github.com/burakuyar/MusicXMLConverter>

<sup>15</sup> <https://github.com/hsercanatli/makam-musicxml2lilypond>



provisations). In (Şentürk et al., 2014), we have proposed a section-level audio-score alignment methodology proposed for OTMM, which considers such structural differences. In the original methodology the sections in the score are manually annotated with respect to the melodic structure. Next, the candidate time intervals in the audio recording are found for each section using partial subsequence alignment. We replaced the manual section annotation step with the automatic section analysis part of our alignment method, where we use the melody labels to align relevant audio recordings and music scores. Using the modified method we have aligned the related audio and score pairs in the CompMusic Turkish makam music corpus (Uyar et al., 2014) and linked 18.770 sections performed in 1767 pairs of audio recordings and music scores. The aligned audio-score pairs are accessible via *Dunya makam*, our prototype web application for the discovery of OTMM (Şentürk et al., 2015).<sup>16</sup> In the application, the audio can be listened synchronous to the related music score(s) on the note-level and the sections are displayed on the audio timeline.

We have additionally conducted experiments using the melodic relations of the extracted phrases. Our preliminary results suggest that phrase-level alignment may provide better results than section-level alignment.

## 8. CONCLUSION

We proposed a method to automatically analyze the melodic and lyrical organization of the music score of OTMM. We applied the method on the latest release of the SymbTr collection. We extracted 49259 phrases from 1345 scores and 21569 sections from 1771 scores. We are also using the extracted structural information in automatic score validation, score engraving and audio-score alignment tasks.

In the future, we would like to test other string matching and dynamic programming algorithms (Serrà et al., 2009; Şentürk et al., 2014) in general, for similarity measures with different constraints and select the optimal similarity threshold  $w$  automatically according to the melodic and lyrical characteristics of the data. We would also like to solidify our findings by working on a bigger dataset annotated by multiple experts and cross-comparing the annotated and the automatically extracted boundaries as done in (Bozkurt et al., 2014). Our ultimate aim is to develop methodologies, which are able to describe the musical structure of many music scores and audio recordings semantically and on different levels.

## 9. ACKNOWLEDGEMENTS

We are thankful for Dr. Kemal Karaosmanoğlu for his efforts in the SymbTr collection, Dr. Barış Bozkurt for his suggestions on the phrase segmentation and Burak Uyar and Hasan Sercan Atlı for developing the score conversion tools in Python. This work is partly supported by the European Research Council under the European Union's Seventh Framework Program, as part of the CompMusic project (ERC grant agreement 267583).

## 10. REFERENCES

- Bimbot, F., Deruty, E., Sargent, G., & Vincent, E. (2012). Semi-otic structure labeling of music pieces: Concepts, methods and annotation conventions. In *Proceedings of the 13th International Society for Music Information Retrieval Conference (ISMIR)*, (pp. 235–240)., Porto, Portugal.
- Bozkurt, B., Karaosmanoğlu, M. K., Karaçalı, B., & Ünal, E. (2014). Usul and makam driven automatic melodic segmentation for Turkish music. *Journal of New Music Research*, 43(4), 375–389.
- Cambouropoulos, E. (2001). The local boundary detection model (lbdm) and its application in the study of expressive timing. In *Proceedings of the International Computer Music Conference*, (pp. 17–22).
- Jackendoff, R. (1985). *A generative theory of tonal music*. MIT Press.
- Karaosmanoğlu, K. (2012). A Turkish makam music symbolic database for music information retrieval: SymbTr. In *Proceedings of 13th International Society for Music Information Retrieval Conference (ISMIR)*, (pp. 223–228)., Porto, Portugal.
- Karaosmanoğlu, M. K., Bozkurt, B., Holzapfel, A., & Doğrusöz Dişiaçık, N. (2014). A symbolic dataset of Turkish makam music phrases. In *Proceedings of 4th International Workshop on Folk Music Analysis*, (pp. 10–14)., Istanbul, Turkey.
- Lartillot, O. & Ayari, M. (2009). Segmentation of Tunisian modal improvisation: Comparing listeners' responses with computational predictions. *Journal of New Music Research*, 38(2), 117–127.
- Lartillot, O., Yazıcı, Z. F., & Mungan, E. (2013). A pattern-expectation, non-flattening accentuation model, empirically compared with segmentation models on traditional Turkish music. In *Proceedings of the 3rd International Workshop on Folk Music Analysis*, (pp. 63–70)., Amsterdam, Netherlands.
- Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, (pp. 707–710).
- Pearce, M. T., Müllensiefen, D., & Wiggins, G. A. (2010). Melodic grouping in music information retrieval: New methods and applications. In *Advances in music information retrieval* (pp. 364–388). Springer.
- Şentürk, S., Ferraro, A., Porter, A., & Serra, X. (2015). A tool for the analysis and discovery of Ottoman-Turkish makam music. In *Extended abstracts for the Late Breaking Demo Session of the 16th International Society for Music Information Retrieval Conference (ISMIR)*, Málaga, Spain.
- Şentürk, S., Holzapfel, A., & Serra, X. (2014). Linking scores and audio recordings in makam music of Turkey. *Journal of New Music Research*, 43(1), 34–52.
- Serrà, J., Serra, X., & Andrzejak, R. G. (2009). Cross recurrence quantification for cover song identification. *New Journal of Physics*, 11(9).
- Tomita, E., Tanaka, A., & Takahashi, H. (2006). The worst-case time complexity for generating all maximal cliques and computational experiments. *Theoretical Computer Science*, 363(1), 28 – 42.
- Uyar, B., Atlı, H. S., Şentürk, S., Bozkurt, B., & Serra, X. (2014). A corpus for computational research of Turkish makam music. In *1st International Digital Libraries for Musicology Workshop*, (pp. 57–63)., London, United Kingdom.

<sup>16</sup> <http://dunya.compmusic.upf.edu/makam>

## A. EXAMPLE ANALYSIS

Kimseye Etmem Şikâyet  
Nihâvent ŞarkıUsul: K. Curcuna  
♩ = 204 ⇒ 3 Dk 29 SnBeste: Kemânî Sarkis Efendi (1885 - 12/12/1944)  
Güfte: ?

**Section 1 (Blue):** ARANAĞME ...  
 $s_1 \leftarrow \langle A_1, A_1 \rangle$   
 $s_2 \leftarrow \langle A_1, A_1 \rangle$   
 $p_1 \leftarrow \langle A_1, A_1 \rangle$   
 $p_2, p_6 \leftarrow \langle B_1, A_1 \rangle$   
 $p_4 \leftarrow \langle AD_1, A_1 \rangle$   
 $p_5 \leftarrow \langle E_1, A_1 \rangle$   
 $p_3 \leftarrow \langle C_1, A_1 \rangle$   
 $p_7 \leftarrow \langle C_2, B_1 \rangle$

**Section 2 (Red):** Kim se ye et mem şi kâ yet  
 $s_3 \leftarrow \langle B_1, B_1 \rangle$   
 $s_4 \leftarrow \langle B_2, B_1 \rangle$   
 $p_8, p_{12} \leftarrow \langle F_1, C_1 \rangle$   
 $p_9, p_{13} \leftarrow \langle G_1, D_1 \rangle$   
 $p_{10}, p_{14} \leftarrow \langle DH_1, E_1 \rangle$   
 $p_{11} \leftarrow \langle I_1, B_1 \rangle$   
 $p_{15} \leftarrow \langle J_1, A_1 \rangle$

**Section 3 (Yellow):** Tit re rim mûc rim gi bi bak  
 $s_5 \leftarrow \langle C_1, C_1 \rangle$   
 $s_6 \leftarrow \langle C_2, C_1 \rangle$   
 $s_9 \leftarrow \langle C_1, C_1 \rangle$   
 $s_{10} \leftarrow \langle C_3, C_1 \rangle$   
 $p_{16}, p_{20}, p_{31}, p_{35} \leftarrow \langle K_1, F_1 \rangle$   
 $p_{17}, p_{21}, p_{32}, p_{36} \leftarrow \langle L_1, G_1 \rangle$   
 $p_{18}, p_{33} \leftarrow \langle H_1, H_1 \rangle$   
 $p_{19}, p_{34} \leftarrow \langle J_1, A_1 \rangle$   
 $p_{23} \leftarrow \langle N_1, I_1 \rangle$   
 $p_{22}, p_{37} \leftarrow \langle M_1, H_1 \rangle$

**Section 4 (Pink):** Per de i zul met çe kil miş  
 $s_7 \leftarrow \langle D_1, D_1 \rangle$   
 $s_8 \leftarrow \langle D_2, D_1 \rangle$   
 $p_{24}, p_{28} \leftarrow \langle O_1, J_1 \rangle$   
 $p_{25} \leftarrow \langle P_1, K_1 \rangle$   
 $p_{26} \leftarrow \langle Q_1, L_1 \rangle$   
 $p_{27} \leftarrow \langle N_2, I_1 \rangle$

**Section 5 (Purple):** kor ka rım ik ba li me SAZ  
 $p_{29} \leftarrow \langle DH_2, M_1 \rangle$   
 $p_{30} \leftarrow \langle J_1, A_1 \rangle$

Kimseye etmem şikâyet ağlarım ben halime  
 Titerim mücrim gibi baktıkça istikbalime  
 Perdeyi zulmet çekilmiş korkarım ikbalime  
 Titerim mücrim gibi baktıkça istikbalime

**Figure 5:** The results of the automatic structural analysis of the score “Kimseye Etmem Şikâyet.” The sections are displayed in colored boxes with the volta brackets colored with a darker shade of the same color. The section labels and their semiotic  $\langle \text{Melody}, \text{Lyrics} \rangle$  label tuple is shown on the left. The phrase boundaries are shown as red lines for the first and as purple for the second pass. The phrases and their semiotic labels are shown on top of the relevant interval and on the bottom, when there are differences in the boundaries in the second pass. Note that  $s_5, s_6, s_9$  and  $s_{10}$  are the repetitive poetic lines (tr: “Nakarat”). “[Son]” in the end of the “Nakarat” marks the end of the piece. The similarity threshold is taken as 0.7 for both melody and lyrics. The usul of the score is *Kapalı Curcuna*, which we treat as *Curcuna* in the phrase segmentation step.

# AFTER THE *Harmonie Universelle* BY Marin MERSENNE (1636), WHAT FINGERING FOR THE *CHABRETTE* IN 2016?

Philippe RANDONNEIX

Musician – Researcher – Teacher – [philippe.randonneix@gmail.com](mailto:philippe.randonneix@gmail.com)

## 1. SCIENTIFIC CONTEXT

In 1636, Marin Mersenne publishes his *Harmonie Universelle, Contenant la Théorie et la Pratique de la Musique, ou est traité de la Nature des Sons, & des Mouvements, des Consonances, des Dissonances, des Genres, des Modes, de la Composition, de la Voix des Chants, de toutes sortes d'instruments harmoniques*.

In Book Five “On Wind Instruments”, two sets of bagpipes are illustrated on separate plates.

The first one, *Cornemuse des bergers* (i.e. the *shepherd's bagpipe*) has a monoxyle chanter (made from one piece of wood) with two drones [Figure 1].

The second one, *Cornemuse de Poitou* (in reference to the french province of Poitou) has a composite chanter and one drone [Figure 2].

These two types of chanter are reproduced several times in the book.

The *chabrette* borrows from these two bagpipes: the chanter is the same as the *Cornemuse de Poitou* but a small drone of medium length is set on a box while the bass drone is bigger and borne on the arm. Based on a known corpus of tens of ancient pieces, this instrument has been rebuilt since the seventies.

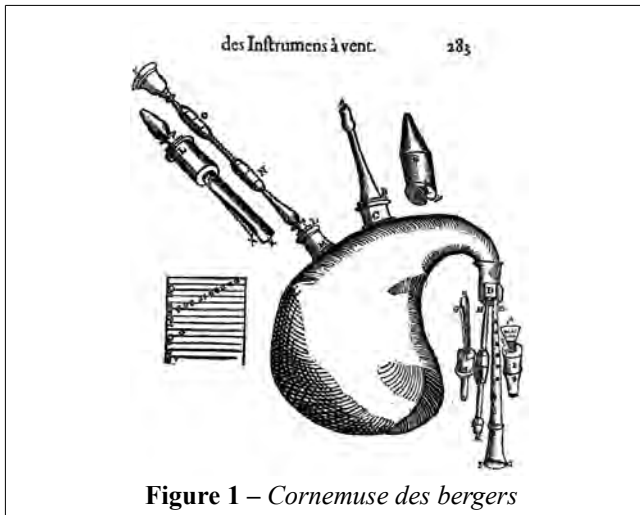


Figure 1 – *Cornemuse des bergers*

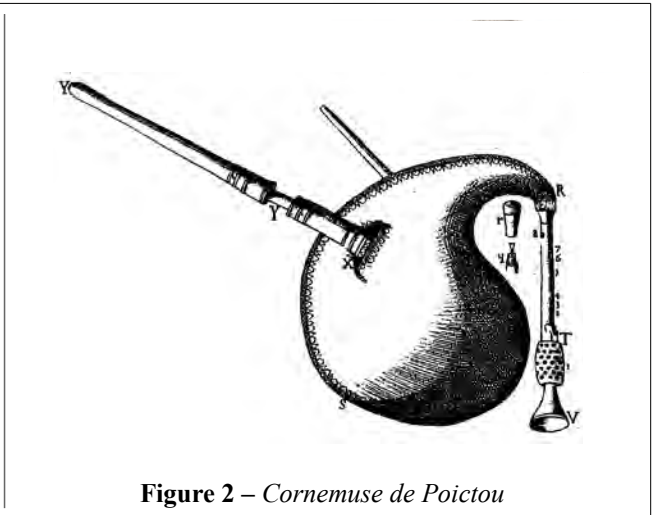


Figure 2 – *Cornemuse de Poitou*

## 2. OBJECTIVES OF THE WORK

What fingering for the *chabrette*?

This bagpipe is often played in Limousin<sup>1</sup> and compared to its two geographical neighbours, the *cabrette* (in *Auvergne*) and the *musette du Centre* (in *Berry-Bourbonnais-Nivernais*). Naturally, the fingerings of these three instruments are often confused because of their proximity.

The chanters of the *cabrette* and the *musette* are made from one piece of wood whose internal bore is continuous. The employed fingers are relatively well known and established. They are semi-closed and each note is composed of an ensemble of holes, open or closed. The general rule is that the lower hand closes when the upper hand opens.

Identical to the *Hautbois de Poitou* presented by Mersenne [Figure 3], the melody pipe of the *chabrette* is composed of several pieces, with its flaring bell quite detached from the main body on which it is set [Figure 4].

A keywork (covered by a *fontanelle*) enables the player to reach the subtonic or leading tone of the oboe.

All these characteristics point to a specific functioning.

There is a need to define a comprehensive, precise and reliable fingering for the *chabrette*, just like for any instrument.

<sup>1</sup> <https://fr.wikipedia.org/wiki/Chabrette>

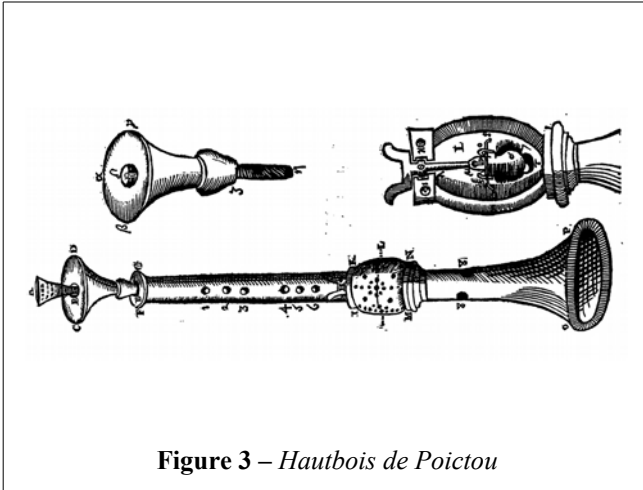


Figure 3 – Hautbois de Poitou

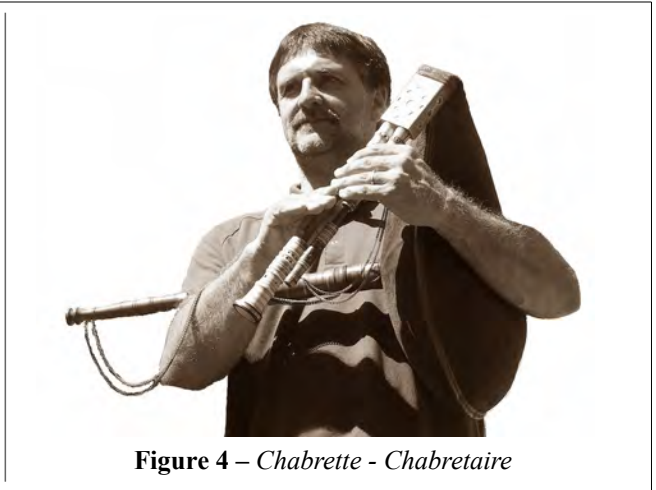


Figure 4 – Chabrette - Chabretaire

### 3. METHODS

Initially, the *cabrette*'s semi-closed fingering was used for the *chabrette* after the former replaced the latter in Limousin during the twentieth century. Marin MERSENNE's work has been long known for its representations of these bagpipes of the seventeenth century, identical to our *chabrettes*. If the lecture and interpretation of the drawings are immediate, the same is not true for the text. The author describes the flutes in a relatively complete manner and relies on them for the oboes and bagpipes.

"*Tout ouvert*" (i.e. *everything open*) is mentioned several

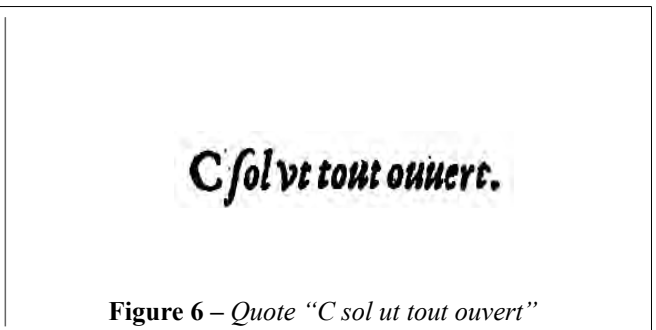
times in addition to the "*tout fermé*" (i.e. *everything closed*).

The first mention "*tout ouvert*" clearly means that the highest note can be obtained by keeping all the holes open, both on the flutes and oboes.

The second indicates the fundamental note of the oboe, with all the holes plugged. Likewise, many tablatures clearly show a "*trou par trou*" opening (i.e. *hole by hole*), where the following note is obtained by lifting the finger on the next hole. Thus, the highest note can be obtained by lifting all the fingers.



Figure 5 – MERSENNE's Tablature

Figure 6 – Quote "*C sol ut tout ouvert*"

Marc ECOCHARD<sup>2</sup> furthers our understanding of MERSENNE's description of the fingerings at that time, when he uses the term "*doigté naturel*" (i.e. *natural fingering*) for this playing technique.

<sup>2</sup> *Les hautbois dans la société française du XVII<sup>e</sup> siècle, une approche par l'Harmonie universelle de Marin MERSENNE et sa correspondance - 2001*

#### 4. RESULTS

As said before, the *cabrette*'s technique of combining semi-closed holes was used by many musicians. But, as I went along in an empirical but nevertheless musical way, I realized that opening *hole by hole* was much more satisfying, as shown by the illustration on the right.

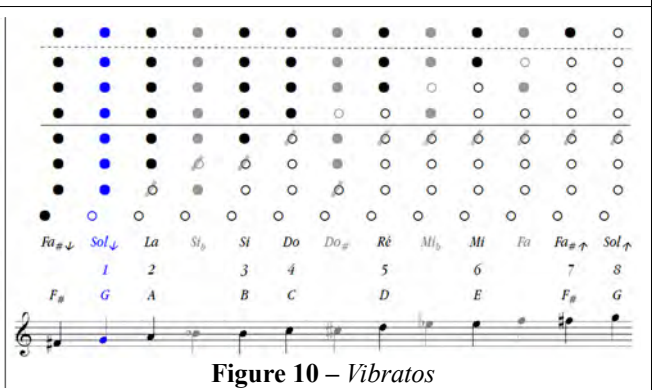
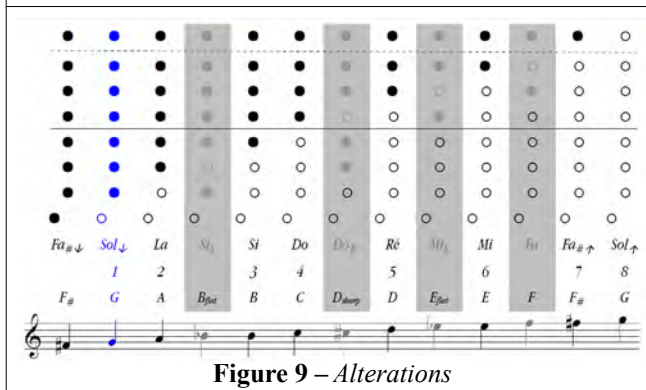
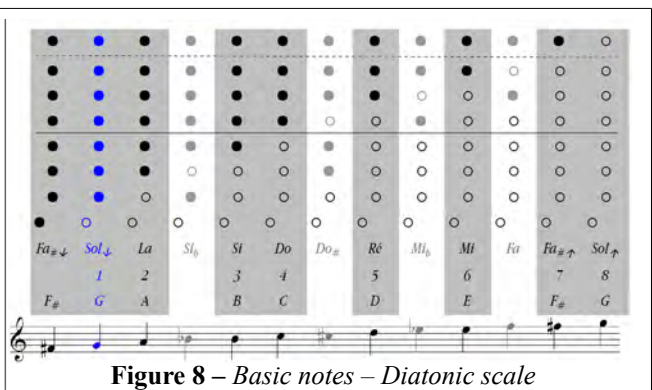
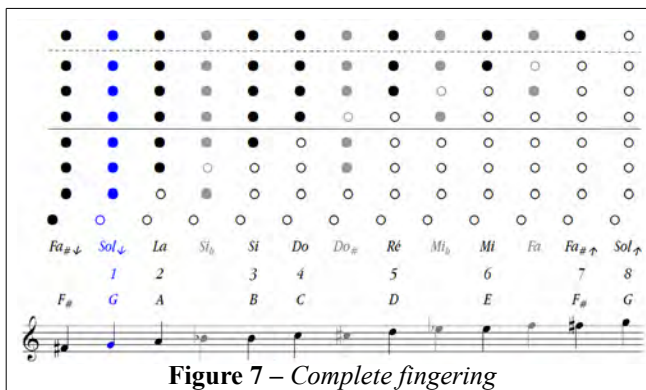
While inventing new playing techniques for the

*chabrette*, I wonder about what fingering to employ. As research progressed, as instruments were remanufactured, after many reed-making attempts, after many confrontations, I realized that the initial fingering was borrowed from other traditions of bagpipes adopted in the region, and thus remained unsatisfying.

The complete fingering [Figure 7] can be read in 3 steps:

- First, we are interested in the basic notes forming a diatonic scale, starting from the fundamental note (*G* highlighted in blue) [Figure 8];

- The alterations, with fork fingering, in grey [Figure 9];
- The vibratos, represented on the edge of the used hole [Figure 10].



#### 5. DISCUSSION

As we all know very well, and to cite but two examples, the *Uilleann pipes* as well as the *Great Highland Bagpipe* have their own fingerings and no one would even think of questioning these fundamentals...

Therefore, it appears possible, important, and necessary to build the same principles for the *chabrette*. This new technique is detailed in an online publication<sup>3</sup>.

#### 6. CONCLUSION

The *chabrette* has been rebuilt and played again for forty years now. But it is only in the last ten years that reflections on its use have empirically led to precise functional elements.

Marin MERSENNE's representations of instruments have always been a reference for me. It is quite remarkable to find (again) consistent elements after almost four centuries. Some of them derive from a technique that was presumably already well established and perfected in the 17<sup>th</sup> century.

<sup>3</sup> <http://philippe.randonneix.free.fr/DuJeudeChabrette.pdf>



# NeoMI : a new environment for the organization of musical instruments

**Carolien Hulshof**

Musical Instruments  
Museum, Brussels

c.hulshof@mim.be

**Xavier Siebert**

Mathematics and Operational  
Research, University of Mons

xavier.siebert@umons.ac.be

**Hadrien Mélot**

Computer Science Department,  
University of Mons

hadrien.melot@umons.ac.be

## 1. INTRODUCTION

The current system to classify musical instruments, (Hornbostel-Sachs), is conceptually and practically outdated, because it has a reducing effect by only considering morphological features (Weisser et al., 2011). Our research project NeoMI aims at developing a new environment for the organization of musical instruments that takes into account their many aspects. The aim is to develop an environment consisting of an integrated, un-hierarchical and flexible tool to organize the musical instruments. Without reducing the complexity and the richness of these multifaceted objects, it includes the manifold aspects of musical instruments into a unique environment. To that end, the system is based on temporary grouping of instruments among their “peers”, according to user-based criteria. This allows an important variability in the precision level: it can be used to group instruments according to a single-criterion (such as the presence on the instrument of an anthropomorphic decoration), or to constitute a corpus of very specific instruments (for example, instruments equipped with devices contributing to provide buzzing sounds), or, on the contrary, to constitute a group of similar instruments made by the same maker, at the same place, over time. NeoMI aims at providing a flexible and pertinent tool for managing museum collections, as well as a fruitful and innovative conceptual framework for research. It explores three different axes: (1) the instrument as an artefact (production time and place, maker, morphological features, etc.); (2) the instrument in its social/cultural context; (3) the instrument as a tool for music. In this paper we focus on the latter, and study the sound-based classification (Fourer et al., 2014; Dupont et al., 2010) of one family of instruments: the fiddles, or bowed chordophones.

## 2. METHODS

To form a sound-based classification of fiddles, many sound recordings of different fiddle types were gathered from libraries, personal archives and online sources. Effort has been made to ensure that fiddles are included with diverse geographic provenances. The recordings were edited in the Musical Instruments Museum using SoundStudio<sup>1</sup> to get smaller samples of 2 to 4 seconds with minimal environmental noise. Representative

samples -referred to as the MIM database from now on- have thus been created for the following fiddle types (number of sound samples between parentheses):

*Endingidi* (10), a one-string spike tube fiddle from the Baganda people in Uganda;

*Erhu* (14), a two-string spike tube fiddle from China;

*Haegum* (9), a two-string spike tube fiddle from Korea;

*Hardingfele* (20), a folk violin with 4 playing strings and 4 sympathetic strings from Norway;

*Imzad* (15), a one-string spike bowl fiddle from the Touareg people in Northern Africa

*Izeze* (17), a spike fiddle from the Wagogo people in Tanzania with one to four strings;

*Kamanche* (9), a spike bowl fiddle from Iran with four strings;

*Kiiki* (31), a half-spike bowl fiddle with one string from Chad;

*Mamokhorong* (10), a one-string fiddle with a tin can resonator from Lesotho;

*Masenqo* (11), a one string spike fiddle with a rhombus-shaped resonator from the Amhara in Ethiopia;

*Morin khuur* (18), a two-string fiddle with a horsehead scroll from Mongolia;

*Njarka* (15), a one-string spike bowl fiddle from the Songhay people in Mali;

*Orutu* (10), a spike tube fiddle with one string from the Luo people in Kenya;

*Ruudga* (10), a one-string spike bowl fiddle from the Mossi people in Burkina Faso;

*Sarangi* (9), an classical Indian fiddle with three playing strings and up to 35-37 sympathetic strings.

The timbre of the MIM instruments was studied using a set of 22 sound features from MirToolbox (Lartillot et al., 2008). Two other databases were also used to test the relevance of the proposed methods as well as to select a subset of discriminating features:

1. MIS: recorded in standardized conditions by the Electronic Music Studios of the University of Iowa, USA<sup>2</sup>.

2. PHIL: recorded by musicians from the Philharmonic Orchestra of London, UK<sup>3</sup>.

<sup>1</sup> <http://felttip.com/ss/>

<sup>2</sup> <http://theremin.music.uiowa.edu/MIS.html>

<sup>3</sup> [http://www.philharmonia.co.uk/explore/make\\_music](http://www.philharmonia.co.uk/explore/make_music)

Several classification algorithms (K-nearest neighbors (kNN), naïve Bayes, Support Vector Machines (SVM)) were applied to each database.

We started with the MIS and PHIL databases, for which 30% of the sounds were used as a test set to estimate the percentage of correct classifications, while the other 70% were used as a training set.

Those results were compared with a complete exploration of all the combinations of 22 features from MirToolbox: a set of 13 MFCC coefficients, centroid, spread, skewness, kurtosis, brightness, flatness, entropy, roll frequencies, and the mean of the signal's envelope. This feature selection allowed us to select a subset of 14 features that gives a better classification performance.

Afterwards, the MIM database (15 fiddle types) was grouped into classes using either all features or the subset of features identified by feature selection. Because the MIM database is too small to allow 30% of the sounds to be kept aside, we performed an n-fold cross-validation, with a stratified scenario to preserve the percentage of samples for each class and  $n=9$ , which corresponds to number of samples in the smallest class.

A multidimensional scaling approach was then used to represent the results in two dimensions.

### 3. RESULTS

#### 3.1 MIS and PHIL databases

The confusion matrices for the MIS and PHIL databases are shown in Figures 1 and 2, respectively, using one representative classifier (kNN with  $k=3$ ). Confusion matrices with other classifiers (kNN with  $k=1,5$ ; Naïve Bayes; SVM) are similar.

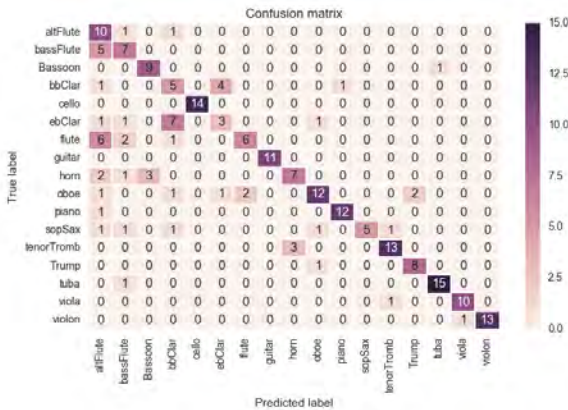


Figure 1. Confusion matrix for the MIS database

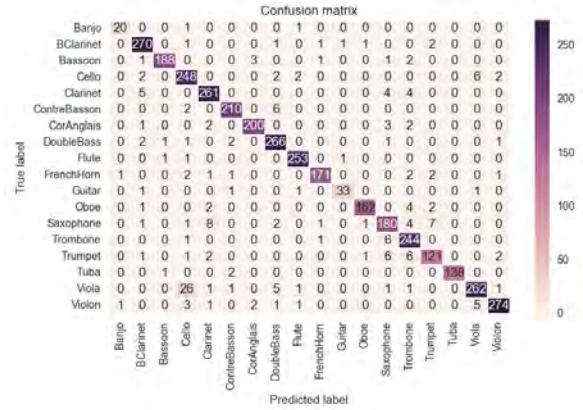


Figure 2. Confusion matrix for the PHIL database

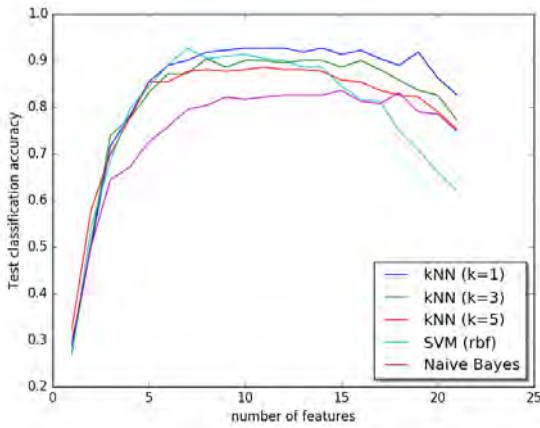
In Figures 1 and 2, the numbers in the diagonal indicate a correct classification, while the off-diagonal ones reflect a confusion between the true and predicted labels.

For the MIS database, the precision is 77%, while the recall is 73%. Some confusion occurs for example among the different types of flutes (altFlute, bassFlute and flute) or among clarinets. This indicates some difficulty to distinguish between instruments of the same family or whose timbre is similar.

For the PHIL database, precision and recall are both around 95%. This reflects the fact that the PHIL database is bigger, but mostly that it contains shorter recordings, each producing a specific note, which simplifies the task of the classifier. Some confusion occurs for example between Cello and Violin, which makes sense considering the proximity of these instruments.

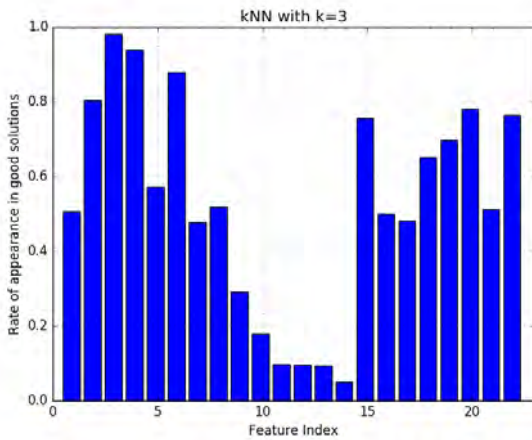
As mentioned in Section 2, these performances have been measured on the test set composed of 30% of the samples.

To improve these results, we performed feature selection, starting from the observation that not all 22 features from MIRTOOLBOX were contributing efficiently to the classification. We thus performed a complete combinatorial analysis to find the best combinations among the 22 descriptors from the MIRTOOLBOX, by comparing the best results obtained with several classifiers:  $k$  nearest neighbours (kNN) with  $k$  values ranging from 1 to 5, naïve Bayes and SVM. The results in Figure 3 show indeed that the classification rate reaches a maximum between 10 to 15 features, before decreasing progressively when increasing the number of features until 22.



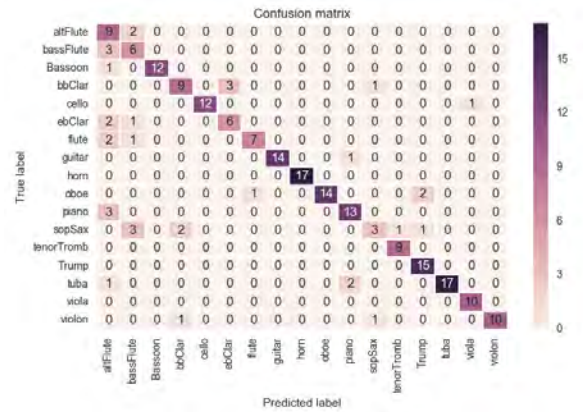
**Figure 3.** Number of features and accuracy

A study of the frequency of appearance of each feature in the most accurate combinations (i.e., more than 85% accuracy) of features is shown in Figure 4, which shows that features with indices 9 to 14 are less efficient.

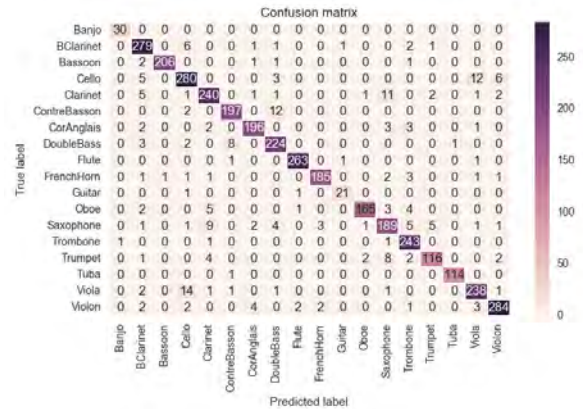


**Figure 4.** Efficiency of features, measured by the frequency of appearance of each feature in the solutions with more than 85% accuracy in the MIS database.

Removing the features 9 to 14 from the set of features used for the classification leads to the confusion matrices shown in Figures 5 and 6, for the MIS and PHIL databases, respectively.



**Figure 5.** Confusion matrix for the MIS database, with a subset of features.



**Figure 6.** Confusion matrix for the PHIL database, with a subset of features.

For the MIS database, the precision has now increased to 86%, and the recall to 84%. However, for the PHIL database, the precision and recall remain stable around 94%.

The slight variations in the PHIL database upon feature selection (95% to 94%) are probably caused by the fact



that a different subset of 30% of sounds is chosen each time.

### 3.2 MIM database

The confusion matrix for the MIM database is shown in Figure 7, with one representative classifier (kNN with  $k=3$ ).

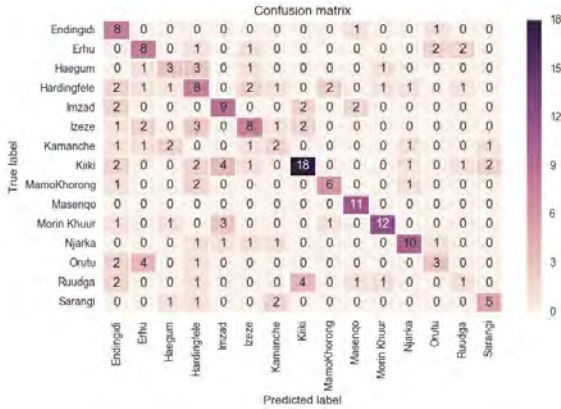


Figure 7. Confusion matrix for the MIM database

Considering the proximity of the instruments involved - the fiddle family- it is not surprising that the confusion matrix is less accurate than for the MIS and PHIL databases.

Some tendencies can be extracted but have to be interpreted with caution. For example, the Kiiki family seems to be fairly homogeneous. However, it is also the most populated (31 instruments), which has a tendency to bias the classification by attracting other instruments (such as Imzad, Izeze or Ruudga) in this category. Another class that appears quite homogeneous is Masenqo. Endingidi, on the contrary, has a high recall (most Endingidi have indeed been classified as Endingidi) but a low precision (several instruments from the Hardingfele, Imzad, Izeze, Kamanche, Kiiki, Mamokhorong, Orutu and Ruudga types have been misidentified as Endingidi).

We also tried the feature selection to classify the MIM sounds with the subset of features, giving us a confusion matrix as shown in Figure 8.

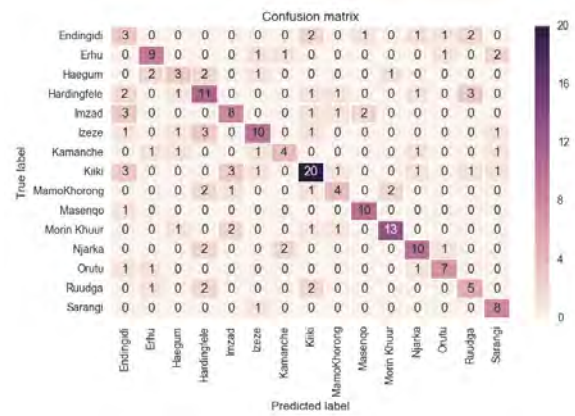


Figure 8. Confusion matrix after feature selection

The new confusion matrix shows a slight overall improvement; all fiddle types have a higher recall, except Endingidi and Mamokhorong.

To visualize and to be able to interpret the results, we computed the distance matrices between predicted classes of instruments (Figure 9), and represented them using a multidimensional scaling (MDS) approach (Cox et al., 2000), as shown in Figure 10.

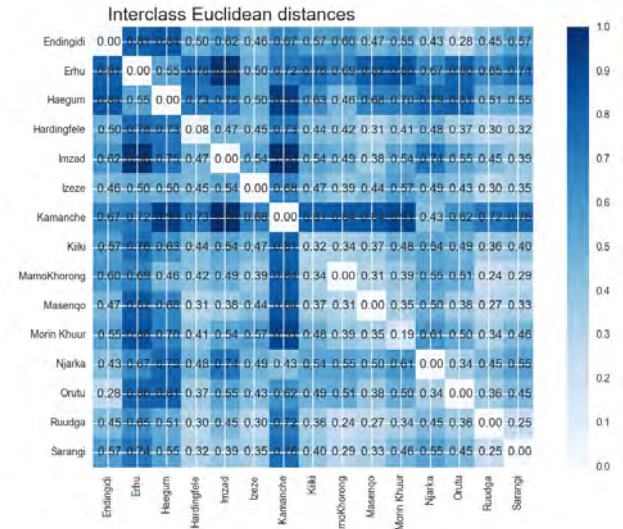
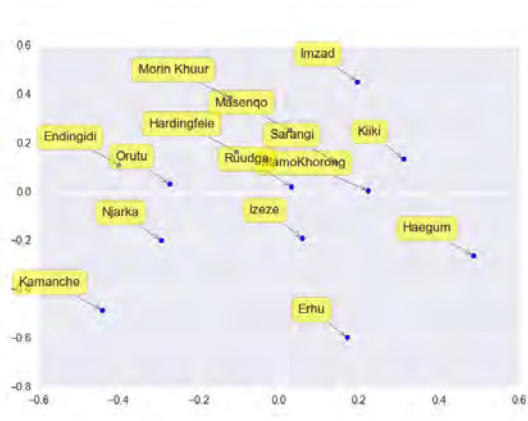
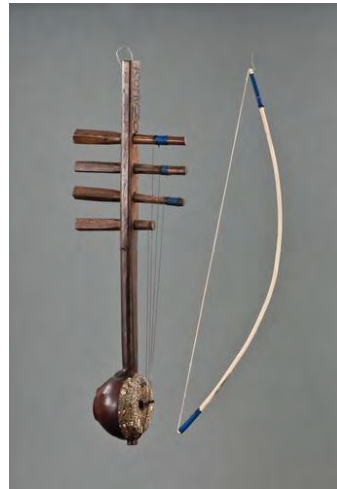


Figure 9. Interclass Euclidean distances



**Figure 10.** Distances between fiddle types



**Figure 12.** Tanzanian izeze, mim inv. 2014.273.001. © mim, photo Simon Egan

#### 4. DISCUSSION

A visual representation using an MDS approach leads to some interesting questions. For example, based on their morphology and geographic distribution one would not expect a close proximity between the Indian Sarangi (Figure 11) and the Tanzanian Izeze (Figure 12):



**Figure 11.** Indian sarangi, mim inv. 1972.003. © mim, photo Simon Egan

However, as shown in Figure 10, these two fiddle types are quite close to each other. This leads to new questions: is it because they both possess sympathetic strings? Does the playing technique play a role in their similarity? Another question arises when looking at the Imzad, a fiddle from the Touareg people in Northwest Africa (Figure 13), and the Njarka, a fiddle from the Songhai people in Mali; both are single string spike fiddles with a calabash resonator, played with a horsehair bow (Figure 14):



**Figure 13.** Touareg imzad, mim inv. 2009.002. © mim, photo Simon Egan





**Figure 14.** Njarka from Mali, RMCA inv. MO.1967.63.777. © RMCA Tervuren

However, apparently there are certain qualities that make them appear far from each other in Figure 10. How can we explain this distance? Not all distances between the different fiddle types are surprising, though - to the human ear, the Endingidi and Orutu sound very much alike, and they are indeed quite close to each other in the graph in Figure 10.

## 5. CONCLUSION

Confusion matrices show that a classification based on sound features is efficient for two databases (MIS and PHIL) containing various kinds of instruments.

Our results indicate that it is also feasible with the MIM database, containing only various fiddle families.

The interest of the sound-based classification is that it allows us to discover possible new links between certain instruments, for example between different fiddle types, as shown on the visualization using an MDS approach. Furthermore, at the dawn of the 21st century, the persistent use of a conceptual framework designed in the 19th century is a problem. Indeed, classificatory systems are not a mere way to sort objects: they are also (and often implicitly) a conceptual ground and a basis for research. The NeoMI project aims therefore to induce an important change of scientific paradigm: from a linear thought to a truly multidimensional one, in which the relative importance of features is adjusted according to the needs of the research.

## 6. REFERENCES

- Cox T.F. and Cox M.A.A. (2000). *Multidimensional scaling*. CRC Press.
- Dupont S., Frisson C., Siebert X., and Tardieu D. (2010). Browsing sound and music libraries by similarity. *128<sup>th</sup> Audio Engineering Society (AES) Convention*, London UK.
- Fourer D., Rouas J.-L., Hanna P. and Robine, M. (2014). *Automatic timbre classification of ethnomusicological audio recordings*, International Society for Music Information Retrieval Conference (ISMIR)
- Lartillot O., Toivainen P., and Eerola T. (2008). A matlab toolbox for music information retrieval. *Data Analysis, Machine Learning and Applications*. Berlin: Springer, pp. 261-268.
- Weisser, S. and Quanten, M. (2011). *Rethinking Musical Instrument Classification: Towards a Modular Approach to the Hornbostel-Sachs System*, Yearbook for Traditional Music 43, pp. 122—146.

# CLOSED PATTERNS IN FOLK MUSIC AND OTHER GENRES

**Iris Yuping Ren**

University of Rochester

yuping.ren.iris@gmail.com

## 1. INTRODUCTION

In this extended abstract, we would like to present the concept of the ‘closed pattern’ from computer science and use it to investigate patterns in folk music. We also show how the quantity of patterns can be different comparing to other genres. We use three symbolic music databases: The Essen Folksong Collection (Schaffrath & Huron, 1995), The Jazz Tune Collection (Rodríguez López et al., 2015), and Bach’s chorales (Sapp, 2005).

There have been lots of quantitative analyses on the Essen dataset (Huron, 1996; Bodet al., 2002; Toivainen & Eerola, 2001; Bod, 2002; Von Hippel & Huron, 2000). One central topic that appears in many of these analyses is the discovery of patterns. The difficulty of pattern discovery in music lies in the ambiguity of the term ‘pattern’. With a rigid definition of what is a ‘pattern’, the process of extracting a pattern is comparatively easy. Here, we use the definition of pattern from MIREX (2015): a sequence which appears at least twice in a corpus is called a pattern.

Such a definition is very broad. For example, in a sequence of letters ‘ABC ABCDE ABCDE’ (the spaces are not included in the sequence), omitting the single letters which appear twice, we have ‘patterns’: AB, ABC, ABCD, ABCDE, BC, BCD, BCDE, CD, CDE, CDE, DE. However, from intuition, we can tell that, within these patterns, there are more important sequences: ABC and ABCDE. To capture this intuition, we borrow the definition of the ‘closed pattern’ developed in the computer science and data mining community. A closed pattern is the type of pattern which is more significant in terms of its length and repetitiveness, first proposed in (Pasquier et al., 1999). Intuitively, they are the patterns with the longest length and repeated the most frequently. Formally, a closed pattern is a pattern that is not included in another pattern which has the same support (or the number of sequences which contain the sequence in consideration).

In fact, people have used the closed pattern for analysing music in multiple occasions (Lartillot, 2005; MIREX, 2015), but as far as we know, there has not been research which systematically investigated the closed pattern of the Essen dataset and the Jazz Tune dataset.

In the case of music, we can treat each piece of music as a sequence of pitch-duration pairs. Nevertheless, such an arrangement is not able to capture the translation of pitches and the self-similarity of durations. For example, a pitch pattern of ‘C4, D4, E4’ and a pitch pattern of ‘G4, A4, B4’, in a general sense, should be considered as the same pattern since they have the same interval structure; a duration

pattern of ‘crochet, quiver, quiver’ and a duration pattern of ‘minim, crochet, crochet’, similarly, should be treated the same. Therefore, we use the pairs of pitch differences and duration ratio as the input music pattern sequence, but not the simple absolute values of pitch-duration pairs.

Using the above definition of pattern and closed pattern, we present the number of closed patterns in three datasets of different genres. We also take one folk song from the Essen dataset and look at the specific closed patterns which were extracted. All the extracted patterns are available in .mid format per request.

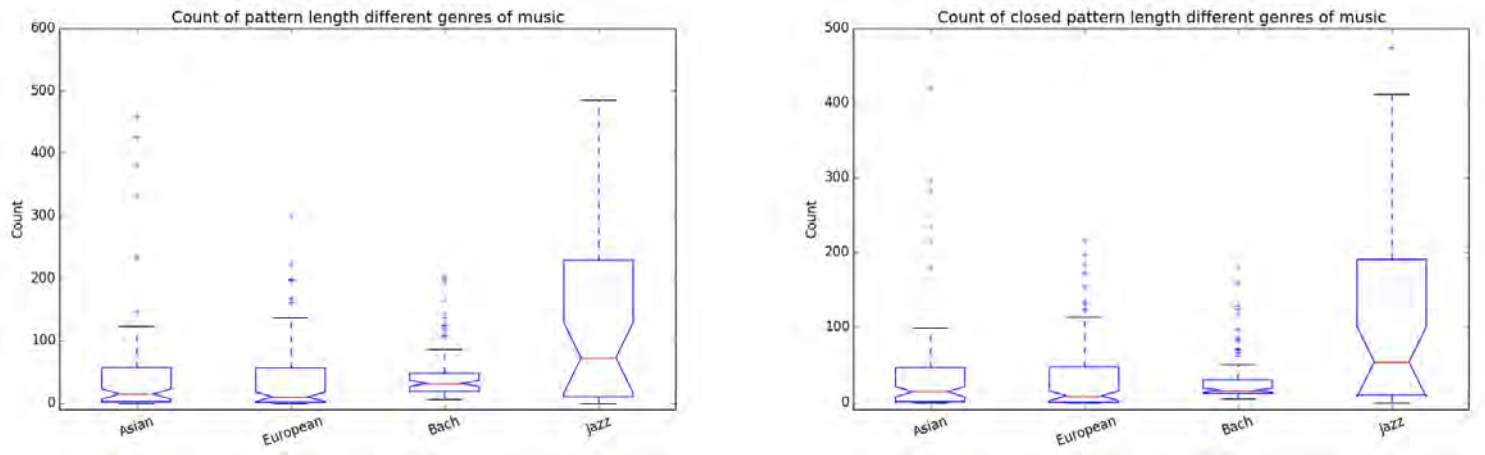
## 2. RESULTS

Figure 1 shows the number of patterns and closed patterns we extracted using music from different genres: folk, jazz and classical. We can see that, although the ranges of the numbers of the patterns are similar, there is a big difference in the group variances. The classical Bach’s chorales show a steady count of the number of patterns and closed patterns; the jazz pieces have the most uncertain amount of closed patterns; for folk music, we split the dataset into European and Asian groups to see if we could find any regional differences, but they are very similar both in range and variances, which are larger than the classical variance and smaller than the jazz variance.

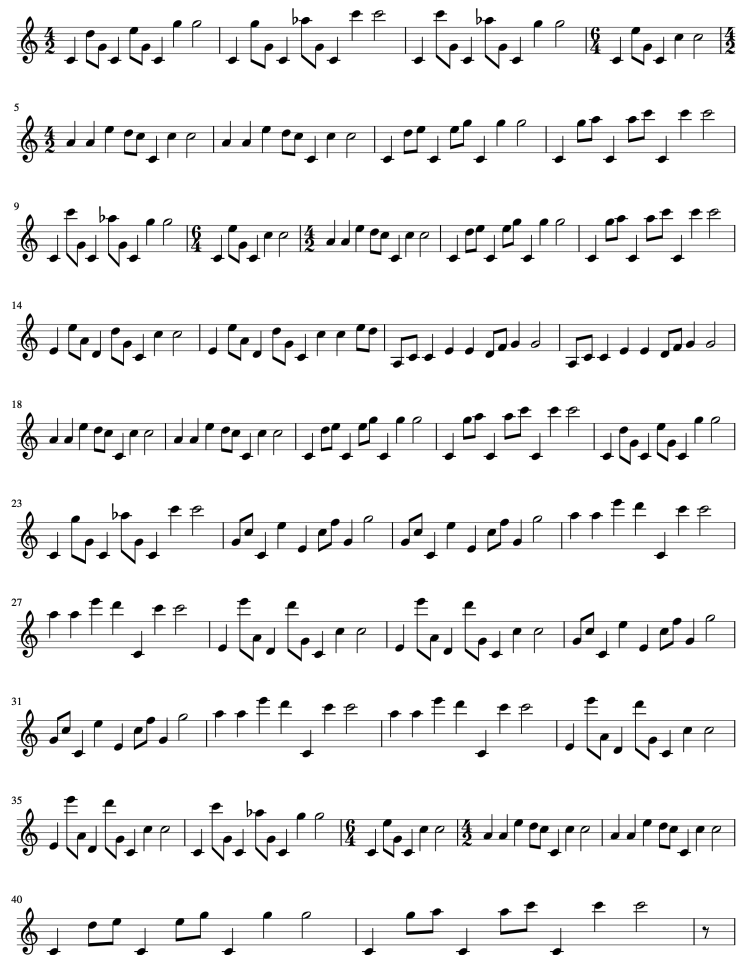
Comparing the quantity of patterns and closed patterns, it is clear that using the definition of the closed pattern eliminates a certain amount of patterns. In addition, the variance differences of different genres are preserved regardless of patterns or closed patterns.

These observations on the ranges and variances help us establish the fact that the abundance of patterns and closed patterns is universal across the three datasets of different genres.

Figure 2 is the example of a Chinese folk song. Figure 3 and Figure 4 show the closed patterns extracted from this specific Chinese song. As described in Section 1, we use the pitch difference and duration ratio pairs for the pattern extraction, so we do not have the absolute values of the patterns. Therefore, to re-construct the melody, we use the midi number 60, which is the note C4, as our first pitch, and a minim as our first duration. The information in the sequence of pitch differences and duration ratios is then used to generate the rest of the melody. We can see that the extracted melodies in Figure 3 and Figure 4 do have musical meaning.



**Figure 1:** The number of patterns and closed patterns of different genres. For each x label, we use a hundred songs to calculate the number of the closed patterns. The y axis gives the count of how many patterns or closed patterns there are. The red line in the box plot shows the median of the distribution of the number of patterns across the hundred pieces. The four boundaries in the box plot indicate the Q1, Q2, Q3, Q4 of the distribution. The plus sign markers indicate outliers. The figure on the left shows the results of patterns, and the figure on the right is the results of closed patterns.



**Figure 2:** The example of a Chinese folk song.



**Figure 3:** Closed pattern extracted from the song in Figure 3.



**Figure 4:** Closed pattern extracted from the song in Figure 3.

### 3. DISCUSSION AND FUTURE WORKS

We used a rigid definition of the ‘pattern’ and the ‘closed pattern’ to investigate the patterns in folk music, and compared the results with other genres. We also showed some musically meaningful melodies extracted from the closed pattern definition.

With limited space, we could not show every pattern and closed pattern we extracted as they are numerous. Most musically meaningful patterns are covered in this definition of the pattern and the closed pattern, but there are ones which are less important. In the future, we hope to devise further conditions to restrict the amount of patterns, and make cross-genre and cross-region comparison.

### 4. REFERENCES

- Bod, R. (2002). Memory-based models of melodic analysis: Challenging the gestalt principles. *Journal of New Music Research*, 31(1), 27–36.
- Bod, R. et al. (2002). A unified model of structural organization in language and music. *Journal of Artificial Intelligence Research*, 17(2002), 289–308.
- Huron, D. (1996). The melodic arch in western folksongs. *Computing in Musicology*, 10, 3–23.
- Lartillot, O. (2005). Efficient extraction of closed motivic patterns in multi-dimensional symbolic representations of music. In *Web Intelligence, 2005. Proceedings. The 2005 IEEE/WIC/ACM International Conference on*, (pp. 229–235). IEEE.
- MIREX (2015). Discovery of Repeated Themes and Sections Results.
- Pasquier, N., Bastide, Y., Taouil, R., & Lakhal, L. (1999). Discovering frequent closed itemsets for association rules. In *Database Theory ICDT99* (pp. 398–416). Springer.
- Rodríguez López, M., Bountouridis, D., & Volk, A. (2015). Novel music segmentation interface and the jazz tune collection. In *Proceedings of the 5th International Workshop on Folk Music Analysis*, (pp. 99–105). CNRS.
- Sapp, C. S. (2005). Online database of scores in the humdrum file format. In *ISMIR*, (pp. 664–665).
- Schaffrath, H. & Huron, D. (1995). The essen folksong collection in the humdrum kern format. *Menlo Park, CA: Center for Computer Assisted Research in the Humanities*.
- Toiviainen, P. & Eerola, T. (2001). Self-organizing map of the essen collection. *ISSCM 2001*.
- Von Hippel, P. & Huron, D. (2000). Why do skips precede reversals? the effect of tessitura on melodic structure. *Music Perception: An Interdisciplinary Journal*, 18(1), 59–85.

# A GRAPH-THEORETICAL APPROACH TO THE HARMONIC ANALYSIS OF GEORGIAN VOCAL POLYPHONIC MUSIC

**Frank Scherbaum**

Institute of Earth- and  
Environmental Sciences,  
University of Potsdam  
fs@geo.uni-  
potsdam.de

**Simha Arom**

UMR 7206, CNRS-MNHN, Paris,  
simha.arom@gmail.com

**Frank Kane**

Kane.frank@gmail.com

## 1. INTRODUCTION

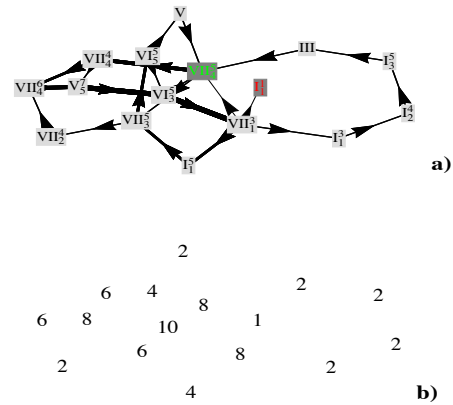
The present paper proposes a computational approach to the comparative analysis and visualization of the harmonic structure of three-voiced vocal music. The dataset which has been used in this study is the same as in Scherbaum et al. (2015), a corpus of polyphonic songs from Svaneti (Akhobadze, 1957). Similar to the earlier work, a song is treated as a discrete temporal process in which harmonic or melodic states change according to unknown rules which are implicitly contained in the song itself. In contrast to the prior study, however, there are no assumptions regarding their probabilistic or deterministic nature.

## 2. METHODOLOGICAL FRAMEWORK

In the preprocessing phase of the analysis described in Scherbaum et al. (2015), each score was analysed for its mode type. It turned out that 99% of the usable songs where in mode La (75%), Sol (21%), and Re (3%). These modes differ only in the size of the 3rds and 6ths being sung as minor or major. Based on the observation of recent recordings of authentic Svan singers (Scherbaum, 2016) which suggest that 3rds and 6ths in traditional Svan music are neither sung as minor nor as major intervals, it was concluded that the separation of the Akhobadze corpus into different modes is not sufficiently supported by the data. For the subsequent analysis it was therefore provisionally assumed that all songs belong to a single 7-step mode in which the distinction between minor and major intervals is dropped, but for which the particular scale does not have to be specified.

In the main part of the analysis, each song is represented as a directed graph (e. g. Chartrand, 1985)  $G = (V, E)$  which consists of a set of vertices  $V$  (representing the harmonic states of a song) and a set of edges  $E$  which represent all the chord transitions in a song (Figure 1). Fig. 1 shows the harmonic structure of the song "Tamar Mepla" in a very efficient graphical way, which also contains some statistical information regarding the harmonic structure of the song. Fig. 1b) for example shows the number of times chords - the positions of which correspond to the positions of the edges in Fig. 1a) - are used. It can be seen that the most common chord is  $VI_3^5$  which is used 10 times, followed in frequency by the chords

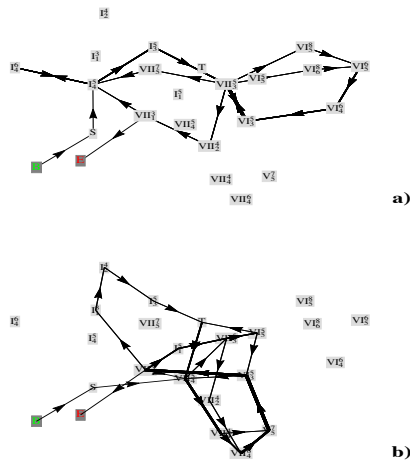
$VII_4^5$  (the starting chord),  $V_5^7$ , and  $VII_1^3$  each of which are used 8 times. From the edge thicknesses in Fig. 1a) one can see for example that the sequence  $V_5^7, VI_3^5, VII_1^3$  is the most often used chord progression in the whole song.



**Figure 1.** Graph representation of the song "Tamar Mepla" in which the line thickness of each edge indicates how many times the corresponding chord progression is used in the song. The vertex positions in the graph are calculated such that the number of edge crossings of the graph is minimized (Tutte embedding). Fig. 1. a) shows the graph while Fig. 1. b) shows the number of times the corresponding chord is used in the song.

Within the graphical framework, an individual song is simply a "path" (called "song path") in a "landscape of chords" which will be referred to as "chordscape". The thickness of the individual segments of a song path reflects how often the particular segment is "travelled". The interpretation of songs as directed graphs might require some training on the side of a musicologist but its advantages become obvious in the context of analysing a whole set of songs together. Naturally, the concept of song paths is easily expanded to a larger group of songs by simply adding new chords (as vertices) to the chordscape and recalculating their optimum positions so that the number of path crossings of all song paths is minimized, using the principle of Tutte embedding. This is illustrated in Fig. 2 for the combination of two songs (Akhobadze song number 5 "Mgzavruli" Fig. 2a with Akhobadze song number 9 "Tamar Mepla" Fig. 2b.

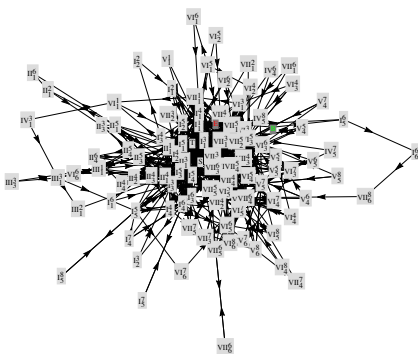




**Figure 2.** Chordscape and song paths for the combination of two songs, a) "Mgzavruli" and b) "Tamar Mepla".

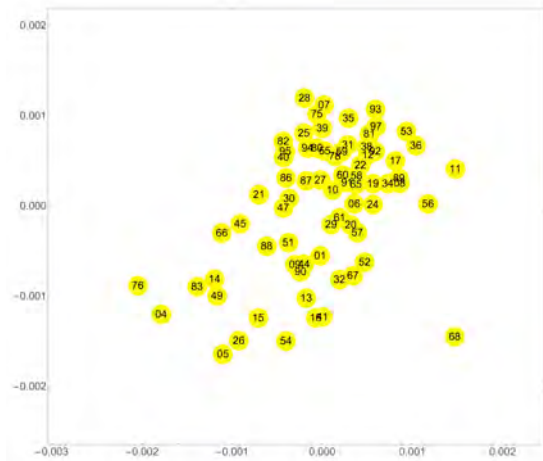
### 3. RESULTS

The representation of songs as song paths in a chordscape offers interesting ways of graphically analysing the harmonic organization of songs and their relationships. Due to the space constraints in this abstract, the potential of this framework can only be highlighted through some selected features. If for example one displays all song paths in a single graph on the full chordscape of the whole corpus (Fig. 3), the chords in the outer locations of the chordscape reveal those chords which are less often used (maybe even only once) while the most used chords in the whole corpus are found in the center of the cluster.



**Figure 3.** Joint song paths for all songs in the corpus plotted on top of the chordscape for the complete corpus. The complete chordscape consists of a total of 102 different chords.

One can now also quantitatively calculate the relationship of songs in terms of their harmonic organisation, e. g. by calculating the Sammon's map (Sammon, 1969) for the song path images (Fig. 4), just to mention another example.



**Figure 4.** Sammon's map for the song paths images of all analysed songs. The two-dimensional mutual distances between the individual points, each representing a song, are reasonably good approximations of the mutual distances of the dissimilarity of the corresponding song paths.

### 4. CONCLUSIONS

The representation of songs as directed graphs allows the quantitative analysis of the harmonic organization of individual songs in a graphical, transparent and reproducible way. It also provides a framework to quantitatively compare the similarity of songs in a whole corpus e. g. by using techniques such as Sammon's maps. The resulting neighborhood relations from the latter analysis can be displayed in ways which can be used for further musicological studies even by non-mathematically inclined analysts.

### 5. REFERENCES

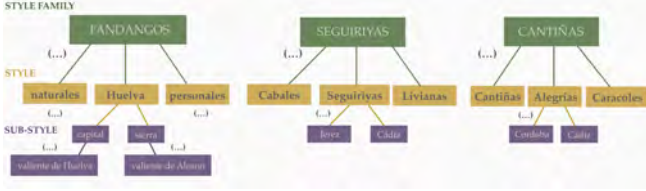
- Akhobadze, Vladimir. (1957). *Collection of Georgian (Svan) Folk Songs*. Tbilisi: Shroma da Teknika. (In Georgian. Foreward in Georgian and Russian).
- Chartrand, G. (1985). "Directed Graphs as Mathematical Models." §1.5 in *Introductory Graph Theory*. New York: Dover, pp. 16-19.
- Scherbaum, F., Arom, S., & Kane, F. (2015). "On the feasibility of Markov Model based analysis of Georgian vocal polyphonic music". In *Proceedings of the 5th International Workshop on Folk Music Analysis, June 10-12, 2015, University Pierre and Marie Curie, Paris, France* (pp. 94–98).
- Scherbaum, F. (2016). "On the benefit of Larynx-microphone field recordings for the documentation and analysis of polyphonic vocal music". In *Proceedings of the 6th International Workshop on Folk Music Analysis, June 15-17, Dublin, 2016* (accepted).
- Sammon, J. W. (1969). „A nonlinear mapping for data structure analysis“. *IEEE Transactions on Computers*, C-18(5), 401–409.

# TOWARDS FLAMENCO STYLE RECOGNITION: THE CHALLENGE OF MODELLING THE AFICIONADO

Nadine Kroher, José-Miguel Díaz-Bañez

University of Seville

{nkroher, dbanez}@us.es



**Figure 1:** Hierarchical organisation of flamenco style families, styles and sub-styles.

## 1. INTRODUCTION

Flamenco is a rich music tradition from the southern Spanish province of Andalucía. Having evolved from an oral tradition, the singing voice remains the central musical element, typically accompanied by a guitar and rhythmic hand-clapping. Since its existence, flamenco songs have been transmitted orally throughout generations and only manual transcriptions are the rare exception. Consequently, performances are highly improvisational and not bound to a musical score. Despite its improvisational character, flamenco music is based on a hierarchical structure of style families, styles and sub-styles (Kroher, Díaz-Bañez, Mora & Gómez, 2015), each of which is defined by a set of melodic, rhythmic and harmonic concepts (Figure 1). Based on these characteristics and their experience, flamenco *aficionados* can identify a flamenco style in a matter of seconds.

In the relatively new field of computational flamenco analysis, automatic style recognition is considered a key challenge. So far, approaches have been limited to the discrimination of two styles belonging to the *tonás* family, the *deblas* and *martinetes* (Cabrera, Díaz-Bañez, Escobar-Borrego, Gómez & Mora, 2008). In this particular case, performances of the same style share a common melodic skeleton which is subject to strong melodic ornamentation. Based on this knowledge, previous approaches have focused on classification solely based on melodic similarity (Díaz-Bañez, Kroher & Rizo, 2015; Mora, Gómez, Escobar-Borrego & Díaz-Bañez, 2010; Gómez, Mora, Gómez & Díaz-Bañez, 2015). Even though promising results have been obtained, this particular task represents only a small sub-problem of automatic flamenco style classification.

In a first step towards the development of a generic system for automatic style categorisation of flamenco recordings, we demonstrate the particular challenges and difficul-

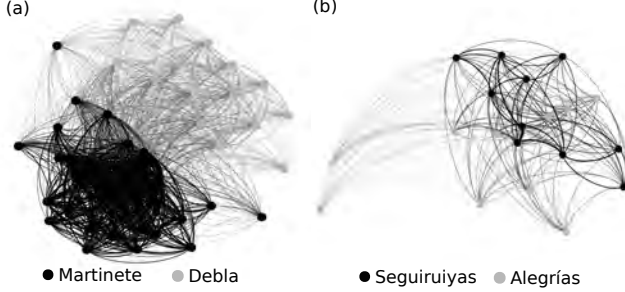
ties of this task on 78 recordings belonging to three styles: *fandangos de Huelva*, *seguiriyas* and *alegrías*. We investigate in how far the melody-based approach generalises to these three styles (Section 2) and furthermore explore the domains tonality (Section 3) and rhythm (Section 4) as potential features for style classification.

## 2. MELODY

It has been demonstrated in Díaz-Bañez et al. (2015) that for the particular case of discriminating styles from the *tonás* family, a classification based on melodic similarity yields nearly perfect accuracies. In order to evaluate in how far this concept holds for the three styles investigated in the scope of this study, we follow the method proposed by Díaz-Bañez et al. (2015) to compute pair-wise similarities of automatic melody transcriptions (Kroher & Gómez, 2016) of the first sung verse. The resulting similarity matrix  $S$  holding the pair-wise similarity values can be represented as a graph  $G(V, E)$ , which we visualise using the *Gephi* software (Bastian, Heymann & Jacomy, 2009). We furthermore evaluate the discriminate power of the obtained representation by computing the *cluster quality*  $q$  as the ratio of intra and inter cluster edges, where a cluster is formed by all instances belonging to the same style.

The cluster qualities for different style combinations (Table 1) and the graph visualisations (Figure 2) indicate a poor class discrimination among *fandangos de Huelva*, *seguiriyas* and *alegrías* compared to the task of discriminating among two members of the *tonás* family: *deblas* and *martinetes*. We identify conceptual as well as methodological causes for this behaviour: Contrary to the particular case of members of the *tonás* family, not all styles necessarily share a single common melodic skeleton, but may encompass a large set of characteristic melodies or melodic patterns. In other words, the degree of intra-style melodic similarity is highly style dependent. Further experiments show that a reliable discrimination based on the melodic contour is achieved in lower hierarchical structures, e.g. among sub-variants of a style which tend to share the same melody. We furthermore observed that in particular in the *alegrías*, melodies exhibit structural differences, i.e. repetitions of a phrase or sub-phrase, which cause a high local alignment cost resulting in low melodic similarity values.

styles	cluster quality $q(S)$
Martinete vs. Debla	3.19
Alegrías vs. Seguiriyas	1.15
Fandangos de Huelva vs. Seguiriyas	1.06
Alegrías vs. Fandangos de Huelva	1.07

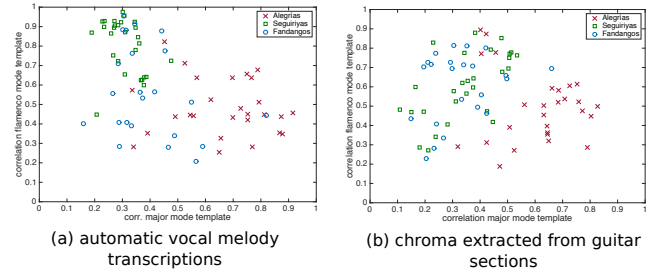
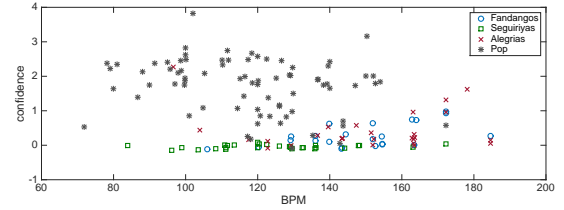
**Table 1:** Cluster quality for various style combinations.**Figure 2:** Graph visualisations of melodic distances.

### 3. TONALITY

In flamenco music, apart from major and minor, we encounter a third scale, the *flamenco mode*: While its diatonic structure is identical to the *phrygian* mode, the dominant is located on the second and the subdominant on the third scale degree. Among the three considered styles, the *alegrías* are set in major mode, *seguiriyas* in flamenco mode and the *fandangos de Huelva* are bimodal in a structural sense, where the guitar plays in flamenco mode during its solo sections and modulates to major when the vocals set in.

In order to detect and investigate tonality across styles, we analyse the distribution of occurring pitch classes (Gómez, 2006; Temperley & Marvin, 2008): We extract pitch class profiles from automatic vocal transcriptions and chromagrams of guitar sections and compute the correlation with pitch class templates for the major mode taken from Temperley & Marvin (2008) and for the *flamenco mode*, which we have estimated by analysing 40 flamenco recordings from this tonality.

Displaying the resulting correlation values obtained from the vocal melody across styles (Figure 3 (a)), clearly reflects the mode affinity of *alegrías* and *seguiriyas*. The *fandangos de Huelva* seem to be spread across both tonalities, which indicates a weak tonal identity. This is an interesting finding, since vocal melodies of the *fandangos de Huelva* are in literature referred to as being sung in major mode (Fernández-Martín, 2011). Further studies indicate a typical pitch class distribution in the *fandangos de Huelva* which differs clearly from the major mode known from Western music. When analysing the same illustration for pitch histograms extracted from guitar sections, we identify a clear separation tendency between the *alegrías* which are played in major mode and the *fandangos de Huelva* and *seguiriyas* in flamenco mode.

**Figure 3:** Histogram correlation with major and flamenco mode templates.**Figure 4:** Estimated tempo and confidence.

### 4. RHYTHM

Flamenco is based on a complex accentuation of style-dependent metric structures: While the *fandangos* are set in a 3/4 meter, both *alegrías* and *seguiriyas* are based on a 12/8 pattern. *Seguiriyas* are performed in slow tempo with weak rhythmic accentuation and tempo fluctuations. The faster *alegrías* are characterised by a complex accentuation shifting between on- and off-beat, which is often emphasised by hand-clapping. In the case of *fandangos*, the tempo and its stability can vary strongly among performances.

We apply a beat tracking algorithm proposed by Zapata & Gómez (2014) to estimate the tempo value in BPM and together with confidence value. We compare the tempo estimates obtained from the three considered styles to the estimate for pop recordings taken from the *Jamendo*<sup>1</sup> dataset.

The results in Figure 4 indicate that the flamenco recordings yield overall lower confidence values than the pop recordings, probably due to the irregular accentuation. Among the styles, the *seguiriyas* obtain the lowest confidence values. Both *alegrías* and *fandangos de Huelva* are on average estimated to have a faster tempo and return a higher beat confidence.

### 5. DISCUSSION

We have introduced the task of automatic flamenco style detection and have shown the limitations of existing problems. Based on the findings of this study we identify a need to develop novel descriptors related to melodic, harmonic and rhythmic content targeting style-specific characteristics. In particular, we aim to develop systems capable of extracting chord progressions, characteristic melodic patterns and the underlying metric structures.

<sup>1</sup> <http://www.mathieura.coma.com/wp/data/jamendo/>

## 6. REFERENCES

- Bastian, M., Heymann, S., & Jacomy, M. (2009). Gephi: An open source software for exploring and manipulating networks. In *Proceedings of the International AAAI Conference on Weblogs and Social Media*.
- Cabrera, J. J., Díaz-Báñez, J. M., Escobar-Borrego, F. J., Gómez, E., & Mora, J. (2008). Comparative melodic analysis of a cappella flamenco cantes. In *Proceedings of the 4th Conference on Interdisciplinary Musicology (CIM08)*.
- Díaz-Báñez, J. M., Kroher, N., & Rizo, J. C. (2015). Efficient algorithms for melodic similarity in flamenco singing. In *Proceedings of the International Workshop on Folk Music Analysis (FMA)*.
- Fernández-Martín, L. (2011). La bimodalidad en las formas del fandango y en los cantes de levante: Origen y evolución. *La Madrugá. Revista de Investigación sobre Flamenco*, 5(1), 37–53.
- Gómez, E. (2006). Tonal description of polyphonic audio for music content processing. *INFORMS Journal on Computing*, 18(3), 294–304.
- Gómez, F., Mora, J., Gómez, E., & Díaz-Báñez, J. M. (In Press). Melodic contour and mid-level global features applied to the analysis of flamenco cantes. *Journal of New Music Research*.
- Kroher, N., Díaz-Báñez, J. M., Mora, J., & Gómez, E. (In Press). Corpus cofla: A research corpus for the computational study of flamenco music. *ACM Journal on Computing and Cultural Heritage*.
- Kroher, N. & Gómez, E. (2016). Automatic transcription of flamenco singing from polyphonic music recordings. *IEEE Transactions on Audio, Speech and Language Processing*, 24(5), 901–913.
- Mora, J., Gómez, F., Escobar-Borrego, F., & Díaz-Báñez, J. M. (2010). Melodic characterization and similarity in a cappella flamenco cantes. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*.
- Temperley, D. & Marvin, E. W. (2008). Pitch-class distribution and the identification of key. *Music Perception: An Interdisciplinary Journal*, 25(3), 193–212.
- Zapata, J. & Gómez, E. (2014). Multi-feature beat tracker. *IEEE/ACM Transactions on Audio, Speech and Language Processing*, 22(4), 816–825.

# A SEARCH THROUGH TIME: CONNECTING LIVE PLAYING TO ARCHIVE RECORDINGS OF TRADITIONAL MUSIC

**Bryan Duggan, Jianghan Xu**

Dublin Institute of Technology  
Ireland

bryan.duggan@dit.ie,  
chrisxue815@gmail.com

**Lise Denbrok, Breandan Knowlton**

Historypin  
United Kingdom

lise.denbrok@historypin.org,  
breandan.knowlton@historypin.org

## 1. INTRODUCTION

This poster describes new developments in the popular Tunepal project. Tunepal is a query-by-playing music score search engine used primarily by musicians on smartphones in traditional music sessions and classes. Using Tunepal, a musician can quickly identify the name of a melody being played and download the score for later study. Since 2009, there has also been a version of Tunepal that runs in a web browser that allows a musician to play a tune extract and find the name of the tune. Over the summer of 2015, we embarked on a project to redevelop the Tunepal website in HTML5. Additionally we aimed to connect Tunepal searches which normally return music scores, to recordings of those scores, through the Europeana Sounds project. Finally, we aimed to make the core Tunepal technology open source and provide API access to the Tunepal corpus and search engine so that others could build on our work<sup>1</sup>.

## 2. BACKGROUND

Tunepal is predominantly used on IOS and Android smartphones and it allows users to search for a music score by playing a 12 second extract on a traditional instrument. The transcription is then sent to the Tunepal server where it is matched against over 23K music scores and the results are returned to the user in order of similarity to the audio search query. Tunepal has in excess of 20K users in over 40 countries who submit around 1K music searches per day. For a more detailed description of the functionality and impact of Tunepal see (Duggan and O'Shea, 2011). Since 2009 there has been a browser hosted version of the Tunepal search engine that used a Java applet to record and transcribe audio. However by 2015, it was clear that this needed to be redeveloped given that browsers were increasingly dropping support for Java applets.

The Europeana Sounds project unifies access to artifacts stored by digital libraries and museums across Europe through a common API. It aims to provide one million audio recordings by January 2017 whilst improving access and promoting the creative reuse of these recordings (Europeana Sounds, 2016).

## 3. GOALS

Over the summer of 2015 we embarked on an ambitious project to redevelop the Tunepal website in HTML5 and add the ability to find matching audio artifacts from the archives of Comhaltas Ceoltoiri Eireann, the Irish Traditional Music Archive and Tobar an Dualchais through the Europeana Sounds API.

Our goals were as follows:

- Replace the Java applet with record and transcription functionality implemented in HTML5
- Return recordings of music, not just music scores.
- Make all the functionality of the Tunepal, including query-by-playing work similarly across all devices including smartphones.
- Open-source Tunepal and make an API server available to other projects.

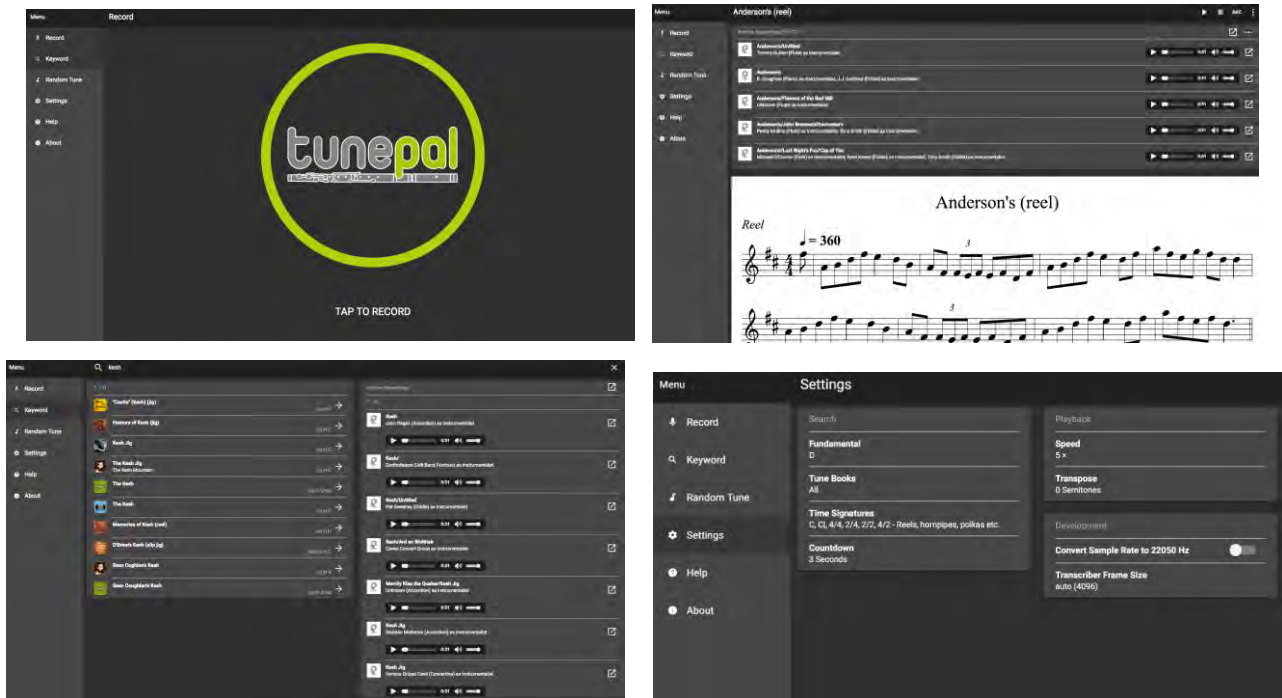
## 4. IMPLEMENTATION

Our team consisted of a back end developer, a front end developer and two experts in music archiving and digital libraries who provided leadership and support from Historypin. Three of the team were based in London whilst the back end developer and project coordinator was based in Dublin. The team communicated regularly using Google Hangouts and Slack, while all the coding was managed in git repositories.

Over the summer, the back end technology that performs Tunepal searches was redeveloped as a JSON API using Jersey (Jersey, 2016). The front end of the project was redeveloped using Materialize and AngularJS, open source frameworks that allows web applications to developed in HTML/CSS and Javascript that conform to Google's Material Design principles (Materialize, 2016). Emscripten was used to cross-compile the ABC2MIDI library on which Tunepal depends from C to Javascript (Emscripten, 2016; Shlien, 2011). To display music scores in the browser, the ABCJS library is used (Rosen and Dyke, 2016).

<sup>1</sup> See <http://github.com/skooter500>





**Figure 1:** Screenshots of the new Tunepal web application

In the new Tunepal web application, whenever a user makes a query-by-playing search for a tune, we also perform a Europeana API search for the title of the closest matching Tune returned by Tunepal. When a user searches for a title, we also search for that title in Europeana Sounds. Also when a user loads a specific tune in Tunepal, we also show search results from Europeana Sounds. We limit our searches to those collections in Europeana Sounds we know to have most traditional music content. Figure 1 illustrates some of these workflows.

## 5. EVALUATION

We evaluated our work by running user trials for the week of Feadh Ceoil na hEireann in various locations in Sligo in August 2015. In total 40 users tested the new version of Tunepal.

From our user trials, we established that users valued the provision of archive recordings greatly, though we did discover that transcription in the browser version of Tunepal was not as accurate as the version implemented in the apps. We are still investigating this and hope to provide improved accuracy in future a version. Also, although Comhaltas provides the majority of archive recordings, these are currently limited to 30 seconds extracts. Often these extracts are from recordings of sets of tunes and sometimes the tune being searched for is not in the first 30 seconds of the recording. We are currently working with Comhaltas to resolve this issue.

Typically we are handling around 2K music searches per month through the new, browser hosted version of Tunepal. This compares to around 20K music searches that originate in the native app versions of Tunepal.

## 6. CONCLUSIONS & FUTURE WORK

We achieved our goal of redeveloping the Tunepal website using modern technologies and also integrating search results from Europeana. We have made a number of enhancements and bug fixes since launch including defaulting to HTTPS connections which was necessary to support access to the microphone on the Chrome browser. When the functionality works, the experience is compelling. It is possible to start an interaction by playing an unknown melody and conclude with the music score from several manuscript collections in addition to recordings of the tune played by iconic musicians on a variety of instruments and contexts. We are also happy to report that the core Tunepal technology is now being integrated into other projects including thesession.org. We aim to build on our work by improving transcription accuracy, including key invariant searches and improving the utility of the archive recordings returned. We are hopeful that as more people become aware of the functionality of the new version of Tunepal, that it will broaden access to a wealth of cultural heritage available through Europeana Sounds.

## 7. REFERENCES

- Duggan, B., O'Shea, B., 2011. Tunepal: searching a digital library of traditional music scores. *OLC Syst. Serv.* 27, 284–297.
- Emscripten, 2016. kripken/emscripten [WWW Document]. GitHub. URL <https://github.com/kripken/emscripten> (accessed 4.18.16).
- Europeana Sounds, 2016. About. Eur. Sounds.
- Jersey, 2016. Jersey [WWW Document]. URL <https://jersey.java.net/> (accessed 4.18.16).
- Materialize, 2016. Materialize [WWW Document]. URL <http://materializecss.com/> (accessed 4.18.16).
- Rosen, P., Dyke, G., 2016. abcjs [WWW Document]. URL <https://abcjs.net/> (accessed 4.18.16).
- Shlien, S., 2011. The ABC Music project - abcMIDI [WWW Document]. URL <http://abc.sourceforge.net/abcMIDI/> (accessed 1.27.11).

# HUMAN PATTERN RECOGNITION IN DATA SONIFICATION

**Charlie Cullen**

Dublin Institute of Technology  
charlie.cullen@dit.ie

**William Coleman**

Dublin Institute of Technology  
d15126149@mydit.ie

## ABSTRACT

Computational music analysis investigates the relevant features required for the detection and classification of musical content, features which do not always directly overlap with musical composition concepts. Human perception of music is also an active area of research, with existing work considering the role of perceptual schema in musical pattern recognition. Data sonification investigates the use of non-speech audio to convey information, and it is in this context that some potential guidelines for human pattern recognition are presented for discussion in this paper. Previous research into the role of musical contour (shape) in data sonification shows that it has a significant impact on pattern recognition performance, whilst investigation in the area of rhythmic parsing made a significant difference in performance when used to build structures in data sonifications. The paper presents these previous experimental results as the basis for a discussion around the potential for inclusion of schema-based classifiers in computational music analysis, considering where shape and rhythm classification may be employed at both the segmental and supra-segmental levels to better mimic the human process of perception.

## 1. INTRODUCTION

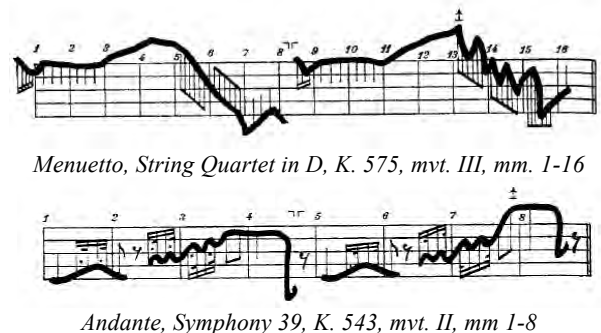
The innate audio processing capability of all humans (and indeed most animals (Kaas, Hackett, & Tramo, 1999)) is amply demonstrated by the ability of infants to discriminate between pitches (Olsho, Koch, & Halpin, 1987), melodic contour (Trehub, Bull, & Thorpe, 1984) and rhythm (Trehub & Thorpe, 1989) as well as an adult can. This ability even extends to the segmenting of melodies (Thorpe & Trehub, 1989) into smaller phrases, and the association of music with other events (Fagen et al., 1997) in a similar manner to adults. The mechanism for such processing is musically specific, with certain neurons directly responsible for pitch perception, rhythm and melodic contour (Johnsrude, Penhune, & Zatorre, 2000; Weinberger & McKenna, 1988) being found only in the right hemisphere of the brain (Trehub et al., 1984).

Perception is a subjective manner of assessment, as by definition differences in perception account for subjective opinion and hence do not easily conform to standardisation. The pitch, loudness or location of sounds can help define their similarity- as can their individual timbres. Also the temporal variations of sounds (such as modulations over time or even their initial onset), can lead to sounds being perceived as grouped or separate- relative to their occurrence and subsequent change (Bregman, 1993). Physically, the fundamental frequency of a sound (and its associated harmonic series) is important in distinguishing between separate sources, as sounds of different fundamental frequency can be detected as separate rather than fused. The rhythmic components of a source also play a major role in its detection (Deutsch, 1980) and recognition, and different rhythmic patterns allow sounds of often similar timbre and pitch to be perceived as separate rather than fused (Bregman, 1993).

Some studies of the mechanics of human audio perception suggest that the requirements for detection and recognition of melodic patterns are different (Hébert & Peretz, 1997), where long-term memory pattern recognition is biased more towards melodic factors than the rhythmic elements required by pattern detection. Although not an arrhythmic condition by any means, a preference is exhibited for melodic criteria when testing the ability of participants to recognise previously introduced patterns. For this reason, the work presented in this paper distinguishes between recognition using contour (shape) and detection using rhythm, aiming to illustrate the crucial role of both criteria in human perception of sound and music.

## 2. CONTOUR PATTERN RECOGNITION

Melodic contour has been considered by many musicologists as a means of defining relative changes in pitch (Toch, 1948) (with respect to time), rather than the definition of absolute values. In this manner, the shape, direction and range of a melody can all be summarised by its overall contour. Graphical contour representations were considered by composers such as Schoenberg (Schoenberg & Strang, 1967) as a means of supplementing a musical score (Figure 1):



**Figure 1.** Contour Graphs of Selected Mozart Compositions, taken from Schoenberg (Schoenberg & Strang, 1967)

Contour can be considered an important part of musical memory. Dowling (Dowling, 1978) suggests that contour information functions separately and independently from scalar information in memory. Experiments by Edworthy (Edworthy, 1983) showed that single pitch alterations in a melody could be detected by participants as changes in contour- even when they were unable to define what pitch had been actually altered in the pattern. This capability is believed to be present in infancy (Chang & Trehub, 1977) (around 5 months), at a stage of development where changes in pitch cannot be recognised. It has also been shown that different brain cells are used in the processing of melodic contour (Weinberger & McKenna, 1988) than are used in the detection of temporal or harmonic (Sutter & Schreiner, 1991) components of music. This aspect of neural activity would again suggest that different parts of the brain are used (Zatorre, 1999) in

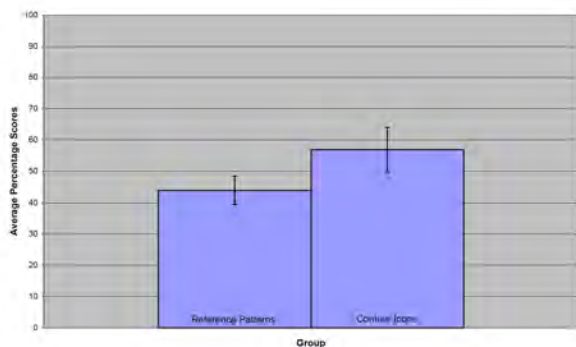
the detection and recognition of musical events: rhythmic factors being paramount in detection, while melodic contour and range (Dowling, 1991; Massaro, Kallman, & Kelly, 1980) and being more important in the recognition of familiar and recently learned melodies.

In previous research into the use of contour (Cullen & Coyle, 2005, 2006), multimodal patterns defined as contour icons were developed to exploit gestalt concepts of good continuation and belongingness (Bregman, 1993) (Figure 2):



**Figure 2.** Example Up and Down Contour Icons, with associated musical score representations

Testing was then performed to assess whether contour icons were more memorable than low-level earcon pattern designs (Hankinson, John & Edwards, 2000) within a data sonification, to determine the effect of shape on pattern recognition (Figure 3):



**Figure 3.** Graph showing overall average percentage scores for recognition of low-level patterns and contour icons in a data sonification, showing standard deviations

Results showed that performance had improved from 44% in the low-level (earcon) reference pattern condition to 56.87% in the contour icon condition, a significant improvement ( $T(20) = -3.68$ ,  $p = 0.0007$ ) that suggests contour icons are more memorable than low level reference patterns that do not employ shape as a melodic feature. Post-test Task Load Index testing (Hart, Sandra, 2006) that examines participant workload during a task showed a significant reduction ( $T(20) = 4.53$ ,  $p < 0.0001$ ) in overall workload from 50.33 to 36.25 for the contour icon condition.

Though no reduction was significant in any individual category, the scores were lower for the contour icon condition in each case. Having said this, higher data to pattern combinations had proven less effective, and it was observed on several occasions that whilst participants could recognise a particular contour icon they were subsequently unable to remember its data mapping. This suggests that the abstract nature of the mapping between value and contour icon was difficult to remember for some participants, though this may not necessarily interfere with the use of shape as an aid to recognition.

Although significant for data sonification, the role of contour in musical pattern recognition requires further investigation in relation to its potential role in computational music analysis. Some consideration has been given to the concept of stream analysis of musical segments (Rafailidis et al., 2008), whilst Karydis et al (Karydis, Nanopoulos, A., Papadopoulos, & Cambouropoulos, 2007) define a computational model of the musical score that includes the concept of a perceptual 'voice' within the overall auditory stream. It is argued that contour may play a significant role within such models, given its demonstrable effect on human musical pattern recognition.

### 3. RHYTHMIC PATTERN DETECTION

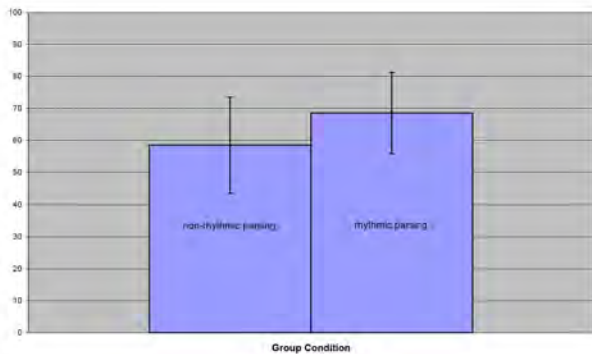
Rhythm is a fundamental building block of musical composition (Taylor, 1989) that serves to group various sonic events within a piece for aesthetic purposes. In sonification research, rhythm can be employed to group patterns used to represent data for analysis so that they may be more efficiently processed by the listener. It is argued that rhythm is a fundamental component of all human interactions (Jones, 1976), and so is similarly fundamental to the communication of effective musical patterns to a listener.

In the case of infants, the role of rhythm is the most fundamentally important aspect of early cognitive development (Zentner & Eerola, 2010), and is believed to begin in the womb (where the child is often observed to move in response to rhythms in speech or music). Infants display several common rhythms (Fridman, 1991), which are used to seek attention from their parents or other adults. This use of rhythmic patterns is both frequent and essential (Kempton, 1980) in the communication between infant and adult, communication that is dictated by a pulse common to all parties. Indeed, the variation or absence of such rhythmic components is observed to engender disinterest and negative responses from the child involved (Drake, Jones, & Baruch, 2000).

Rhythm dictates the structure of a piece of music, from the individual sequence of notes to the hierarchical groupings of different musical phrases or passages. The ability of musicians to detect and convey complex structures (Jongsma, Desain, & Honing, 2004) within a piece is a direct result of training and experience, the lack of which effectively reduces rhythmic patterns to sequential processes. This means of structuring music relies heavily on the metrical organization (Essens, 1995) of such rhythmic patterns into regular frameworks, utilising the time signature of the piece to define different sections. Thus rhythm allows a piece of music to be organised into sections- sections of differing levels of complexity. By defining the bar (or measure) in terms of the beat, the basic organisational structure of a piece of music is decided. When this bar structure is then further organised into sections (such as the simple verse and chorus of popular music) it allows differing pieces of related musical information to be conveyed in a structured manner.

In previous research (Cullen & Coyle, 2003, 2006), rhythm was investigated as part of a strategy to sonify data, and the specific role of rhythmic parsing was subsequently investigated in the sonification of (fictitious) exam results (Cullen, Coyle, & Russell, 2005). Test participants were informed they would be asked questions on a sonification of 20 exam results, which contained 4 distinct course groups (with 5 members each) in sequential order. The test used rest notes between course groups in the parsing condition, compared to a single grouping of musical events in the control condition, as a means of using rhythm to delineate groupings (or structures) within the data. Participants were asked questions that compared the data of each group (e.g.

which group had a higher pass rate) to determine the effect of adding rhythmic gaps to the processing of information in the sonification (Figure 4):



**Figure 4.** Graph showing overall average percentage scores (by test condition) for rhythmic parsing of a data sonification, showing standard deviations

Overall results showed performance improved to 75.3% in the rhythmic parsing condition from 67.6% in the non-rhythmic parsing condition. This improvement was significant ( $T(20) = -2.79$ ,  $p=0.008$ ), suggesting that rhythmic parsing had a positive effect on performance in multiple stream sonification. In addition, post-test TLX questions relating to the workload involved in analyzing a data sonification showed a significant reduction in overall workload from 60.75 to 41.33 in the rhythmic parsing condition ( $T(20) = 7.45$ ,  $p<0.001$ ), with significant reductions in temporal demand (16.33 to 7.65,  $T(20) = 6.236$ ,  $p<0.001$ ), effort (9.95 to 4.583,  $T(20) = 4.435$ ,  $p<0.001$ ), and frustration (7.983 to 4.05,  $T(20) = 2.966$ ,  $p=0.005$ ).

These results suggested that participants had found the rhythmic parsing condition a more effective method of representing sub-groups in a data sonification, though the use of a rest note to parse the data arguably serves only to indicate a change in the current context within the sonification. A more effective method of rhythmic parsing could employ features such as markers and labels (Smith & Walker, 2002), in combination with rest notes to better mimic the compositional use of rhythm as a means of grouping motifs and patterns into distinct structures within a larger piece (Barry, Gainza, & Coyle, 2007).

#### 4. DISCUSSION & FUTURE WORK

This section is still to be completed, but will consider the following 3 areas:

- Hierarchical models for short-term/long-term structures- Contour?
- Measuring relevance of different musical properties and structure principles- Dan & Mikel (Barry et al., 2007)
- Developing taxonomies/ontologies for structure annotation- Rhythm & Contour.

#### 5. REFERENCES

Barry, D., Gainza, M., & Coyle, E. (2007). Music Structure Segmentation using the Azimugram in conjunction with Principal Component Analysis. In *Audio Engineering Society, 123rd Convention* (pp. 1–8).

Bregman, A. S. (1993). Auditory scene analysis: hearing in

complex environments. *Thinking in Sound: The Cognitive Psychology of Human Audition*.

- Chang, H. W., & Trehub, S. E. (1977). Auditory processing of relational information by young infants. *Journal of Experimental Child Psychology*, 24(2), 324–331. doi:10.1016/0022-0965(77)90010-8
- Cullen, C., Coyle, D. E., & Russell, D. N. (2005). *The Sonic Representation of Mathematical Data*. Faculty of Engineering and Faculty of Applied Arts. Dublin Institute of Technology.
- Cullen, C., & Coyle, E. (2003). Rhythmic Parsing of Sonified DNA and RNA Sequences. *Irish Signals and Systems Conference, ISSC 2003*.
- Cullen, C., & Coyle, E. (2005). Musical Pattern Design Using Contour Icons. *Eleventh Meeting of the International Conference on Auditory Display (ICAD 05)*.
- Cullen, C., & Coyle, E. (2006). Harmonically Combined Contour Icons for Concurrent Auditory Display. In *Proc. IET Irish Signals and Systems Conf* (pp. 501–506).
- Deutsch, D. (1980). The processing of structured and unstructured tonal sequences. *Perception & Psychophysics*, 28(5), 381–389. doi:10.3758/BF03204881
- Dowling, W. J. (1978). Scale and contour: Two components of a theory of memory for melodies. *Psychological Review*, 85(4), 341–354. doi:10.1037/0033-295X.85.4.341
- Dowling, W. J. (1991). Tonal strength and melody recognition after long and short delays. *Perception & Psychophysics*, 50(4), 305–313. doi:10.3758/BF03212222
- Drake, C., Jones, M. R., & Baruch, C. (2000). *The development of rhythmic attending in auditory sequences: Attunement, referent period, focal attending*. *Cognition* (Vol. 77). doi:10.1016/S0010-0277(00)00106-2
- Edworthy, J. (1983). The Acquisition of Symbolic Skills. In D. Rogers & J. A. Sloboda (Eds.), (pp. 263–271). Boston, MA: Springer US. doi:10.1007/978-1-4613-3724-9\_30
- Essens, P. (1995). Structuring temporal sequences: Comparison of models and factors of complexity. *Perception & Psychophysics*, 57(4), 519–532. doi:10.3758/bf03213077
- Fagen, J., Prigot, J., Carroll, M., Pioli, L., Stein, A., & Franco, A. (1997). Auditory context and memory retrieval in young infants. *Child Development*, 68(6), 1057–1066. doi:10.1111/j.1467-8624.1997.tb01984.x
- Fridman, R. (1991). Proto-rhythms: Basis for the birth of musical intelligence and language expression. *Journal of Prenatal & Perinatal Psychology & Health*.
- Hankinson, John, C., & Edwards, A. D. N. (2000). Musical Phrase-Structured Audio Communication. *Proceedings of the 6th International Conference on Auditory Display, Atlanta, GA, USA, 2000*.
- Hart, Sandra, G. (2006). NASA-task load index (NASA-TLX); 20 years later. *Human Factors and Ergonomics Society Annual Meeting*, 904–908. doi:10.1037/e577632012-009
- Hébert, S., & Peretz, I. (1997). Recognition of music in long-term memory: are melodic and temporal patterns equal partners? *Memory & Cognition*, 25(4), 518–533. doi:10.3758/BF03201127
- Johnsrude, I. S., Penhune, V. B., & Zatorre, R. J. (2000).



- Functional specificity in the right human auditory cortex for perceiving pitch direction. *Brain: A Journal of Neurology*, 123 ( Pt 1, 155–163. doi:10.1093/brain/123.1.155
- Jones, M. R. (1976). Time, our lost dimension: toward a new theory of perception, attention, and memory. *Psychological Review*, 83(5), 323–355. doi:10.1037/0033-295X.83.5.323
- Jongsma, M. L. A., Desain, P., & Honing, H. (2004). Rhythmic context influences the auditory evoked potentials of musicians and nonmusicians. *Biological Psychology*, 66(2), 129–152. doi:10.1016/j.biopsycho.2003.10.002
- Kaas, J. H., Hackett, T. A., & Tramo, M. J. (1999). Auditory processing in primate cerebral cortex. *Current Opinion in Neurobiology*, 9(2), 164–170. doi:10.1016/S0959-4388(99)80022-1
- Karydis, I., Nanopoulos, A., Papadopoulos, A., & Cambouropoulos, E. (2007). Visa : the Voice Integration / Segregation Algorithm, (April).
- Kempton, W. (1980). The rhythmic basis of interactional micro-synchrony. *The Relationship of Verbal and Nonverbal Communication*, 67–75.
- Massaro, D. W., Kallman, H. J., & Kelly, J. L. (1980). The role of tone height, melodic contour, and tone chroma in melody recognition. *Journal of Experimental Psychology: Human Learning and Memory*, 6(1), 77–90. doi:10.1037/0278-7393.6.1.77
- Olsho, L. W., Koch, E. G., & Halpin, C. F. (1987). Level and age effects in infant frequency discrimination. *The Journal of the Acoustical Society of America*, 82(2), 454–464. doi:10.1121/1.395446
- Rafailidis, D., Nanopoulos, A., Cambouropoulos, E., Manolopoulos, Y., Science, C., & Studies, M. (2008). Detection of stream segments in symbolic musical data. In *The International Society of Music Information Retrieval (ISMIR 2008)* (pp. 83–88).
- Schoenberg, A., & Strang, G. (1967). *Fundamentals of music composition*. St. Martin's Press.
- Smith, D. R., & Walker, B. N. (2002). Tick-marks, Axes, and Labels: The Effects of Adding Context to Auditory Graphs. *International Conference on Auditory Display*, 1–6.
- Sutter, M. L., & Schreiner, C. E. (1991). Physiology and topography of neurons with multi-peaked tuning curves in cat primary auditory cortex. *J Neurophysiology*, 65(5), 1207–1226. Retrieved from <http://jn.physiology.org/content/65/5/1207>
- Taylor, E. (1989). *The AB Guide to Music Theory: Part 1*. Associated Board of the Royal Schools of Music.
- Thorpe, L. A., & Trehub, S. E. (1989). Duration illusion and auditory grouping in infancy. *Developmental Psychology*, 25(1), 122–127. doi:10.1037/0012-1649.25.1.122
- Toch, E. (1948). *The shaping forces in music: An inquiry into harmony, melody, counterpoint, form*. Criterion Music Corporation.
- Trehub, S. E., Bull, D., & Thorpe, L. A. (1984). Infants' perception of melodies: The role of melodic contour. *Child Development*, 55(3), 821–830. doi:10.1016/S0163-6383(84)80430-0
- Trehub, S. E., & Thorpe, L. A. (1989). Infants' perception of rhythm: categorization of auditory sequences by temporal structure. *Canadian Journal of Psychology/Revue Canadienne de Psychologie*, 43(2), 217–229. doi:10.1037/h0084223
- Weinberger, N. M., & McKenna, T. M. (1988). Sensitivity of Single Neurons in Auditory Cortex to Contour: Toward a Neurophysiology of Music Perception. *Music Perception: An Interdisciplinary Journal*, 5(February), 355–389. doi:10.2307/40285407
- Zatorre, R. J. (1999). Brain imaging studies of musical perception and musical imagery. *Journal of New Music Research*, 28(3), 229–236. doi:10.1076/jnmr.28.3.229.3112
- Zentner, M., & Eerola, T. (2010). Rhythmic engagement with music in infancy. *Proceedings of the National Academy of Sciences of the United States of America*, 107(13), 5768–5773. doi:10.1073/pnas.1000121107

# A PATTERN MINING APPROACH TO STUDY A COLLECTION OF DUTCH FOLK-SONGS

**Peter van Kranenburg**

Meertens Instituut  
Amsterdam, Netherlands, and  
Utrecht University  
Utrecht, Netherlands

peter.van.kranenburg@meertens.knaw.nl

**Darrell Conklin**

University of the Basque Country UPV/EHU  
San Sebastian, Spain, and  
IKERBASQUE, Basque Foundation for Science  
Bilbao, Spain

darrell.conklin@ehu.es

## 1. INTRODUCTION

In the ethnomusicological study of oral music cultures, the question what are the units of music has been of particular interest. Bohlman (1988) regards the song as the most basic unit. To better understand a given song culture, a possible next question would be what is the *smallest* unit of music. Nettl (2005, p.117) observes that folk musicians making field recordings are not always willing, or even unable to perform individual phrases, or motifs in isolation. Nevertheless, these units can to a certain extent have an independent existence, recurring in different pieces. This observation was first elaborated on by Tappert (1890), who entitled his study *Wandernde Melodien* (Wandering Melodies), employing the metaphor of traveling.

An important ethnomusicological concept we use in our study, is the concept of *tune family*, which has been introduced by Bayard (1950) to group together a set of folk song melodies that supposedly descend from one original tune through the process of oral or semi-oral transmission.

In a previous study on the way in which human collection specialists categorize Dutch folk song melodies into tune families (Volk & Van Kranenburg, 2012), it was found that the recurrence of short characteristic motifs is most relevant for the perception of similarity between songs belonging to the same tune family. Therefore, in the current work, we set out to analyse tune families in terms of shared melodic motifs.

In our approach, the set of melodies is divided into a *corpus* and an *anticorpus* (Conklin, 2010). The algorithm is capable of discovering recurring patterns that are statistically over-represented in the corpus with respect to the anticorpus. In all cases described in this paper, the corpus consists of all members of a given tune family, while the anticorpus consists of members of other tune families.

The question we ask is how to employ an existing sequential pattern mining algorithm (Conklin, 2010) to discover recurring patterns in a collection of Dutch folk tunes that can be considered building blocks for the melodies, and that characterize a melody as member of a tune family. In the following, we outline the method, the first results we obtained, and some open questions we want to address in our future work.

## 2. DATA

The pre-existing data set MTC-ANN 2.0, which is part of the Meertens Tune Collections (MTC) (Van Kranenburg et al., 2014)<sup>1</sup>, contains 360 digitized vocal folk songs in 26 tune families from Dutch oral tradition, made available in symbolic encoding (\*kern). These songs have been collected through ethnological field work in the Netherlands and from written sources such as song books. The collection specialists at the Meertens Instituut grouped the songs into tune families based on melodic similarity.

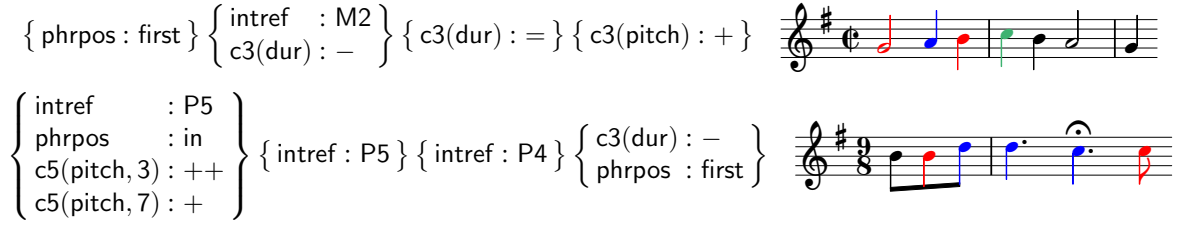
The small sample of 360 songs in 26 tune families has carefully been selected from a larger collection of thousands of songs. The sample is claimed to be representative for the larger collection concerning the kinds of variety that occur among variants of a tune family (Volk & Van Kranenburg, 2012). MTC-ANN 2.0 is provided with several sets of human annotations including a tune family label for each melody, but also 1,657 motif occurrences in 102 motif classes. Each of these motif classes represents an abstract melodic motif that has a number of concrete occurrences in songs within a tune family. These motifs are considered characteristic for the tune family in which they occur by the expert annotators. Therefore, we would expect an algorithmic pattern discovery method to find patterns that correspond to some of these annotated motifs.

## 3. METHOD

A melody is represented as a sequence of *events*, each a tuple comprised of basic attributes such as pitch, duration, and onset time. A *viewpoint* is a function that computes a value for each event in a sequence. Viewpoints can be *basic*: simply returning the basic attribute of an event; *derived* from other viewpoints; or *constructed*. For example, the derived level viewpoint, computed from the prevailing time signature and event onset time, describes the metric level of the event (0 being the highest metric level); and another derived viewpoint intref computes the diatonic interval from the reference pitch (the tonic) to the given pitch.

The choice of viewpoints is crucial for our study. The

<sup>1</sup> <http://www.liederenbank.nl/mtc>. Accessed: 5 June 2016.



**Figure 1:** Patterns discovered in tune family Koopman (top), and Stad (bottom), with one example occurrence. The colored notes constitute the occurrence, red indicating a note-event that is determined by non-pitch features only, green indicating the presence of pitch contour in the feature set, and blue indicating the presence of scale degree. The Koopman pattern describes a note that is the start of a phrase, followed by a note that is a major second above the tonic, and has shorter duration than the previous note, followed by a note of equal duration, and concluded with a note that has a higher pitch than the previous note. The Stad pattern describes a note somewhere in the middle of a phrase that is the fifth of the scale, and is approached by a leap of a third or fourth from the previous note, followed by, again, the fifth of the scale, then by the fourth of the scale, and concluded by a note of shorter duration, which is the first of a new phrase.

abstraction level of the viewpoints should be high enough to capture variability in the melodies as caused both by the process of oral transmission and by variations in choices that were made in the process of transcription into music notation. To achieve a suitable level of abstraction, we measure *relative* values for all viewpoints derived from pitch or duration.

For the current study we define the following viewpoints: *phrpos*, which records whether the note is the first in a phrase, the last in a phrase, or inside a phrase; *intref*, which represents the scale degree of the note given the key of the song; *c3i(level)*, which records whether the metric level of a note is higher, lower or equal with respect to the previous note; *c3(dur)*, which records whether the note is shorter, equal, or longer in duration than the previous note; *c3(pitch)*, which records whether the note is higher, equal, or lower in pitch than the previous note; *c5(pitch, 3)*, which records whether the note was approached by a leap (three semitones or larger), a step (smaller than a three semitones), or a unison, with distinction between ascending and descending intervals; and *c5(pitch, 7)*, which records whether the note was approached by a leap (seven semitones or larger), a step (smaller than seven semitones), or a unison, with distinction between ascending and descending intervals.

A *feature* is a tuple  $\tau : v$  comprised of a viewpoint name  $\tau$  paired with a value  $v$ . A *feature set* is a set of features, for example the feature set

$$\left\{ \begin{array}{l} c3(pitch) : - \\ intref : M2 \end{array} \right\}$$

contains two features, expressing that the pitch of the corresponding note is lower than that of the previous note, and is the major second (M2) of the scale. An event *instantiates* a feature set if all features in the set are true for the event.

A *feature set pattern* is a sequence of feature sets, and a song instantiates a pattern (or, stated equivalently, the pattern *occurs* in the song) if the successive feature sets of the pattern instantiate successive events in the song in at least one place. For example, the patterns shown in Figure 1

have four feature sets, with different features in each of them.

Following the method presented by Conklin (2010), a *one vs. all* strategy (Neubarth & Conklin, 2016) is used for mining patterns that contrast between groups of data. The method is designed to discover maximally general distinctive patterns (MGDPs), meaning that for each reported discovered pattern there is no more general pattern that is also distinctive. Each tune family is mined individually for distinctive sequential patterns, using each tune family  $F$  as a positive corpus and the rest of the pieces ( $\neg F$ ) as the anticorpus.

In this work a statistical approach is used to measure the distinctiveness of a pattern: it is the probability  $p$  of finding at least the observed number of pieces of family  $F$  when taking a single random sample of pieces from the entire corpus  $F \cup \neg F$ . A pattern is then considered distinctive if its p-value falls below some specified significance level  $\alpha$  (see Conklin, 2013, for details).

The MGDP set may contain overlapping patterns, so for the tune family mining task this set is further reduced by a greedy pruning strategy. Proceeding from the best (lowest p-value) pattern, a pattern is placed in the final set if it does not overlap, in any piece, with any pattern already in the final set. Thus none of the patterns in the final set will overlap in any piece with any other pattern.

#### 4. RESULTS

The mining algorithm was applied repeatedly with each of the tune families in MTC-ANN in turn as corpus, while the other 25 tune families constitute the anticorpus. For this initial study, to obtain only a few highly distinctive patterns, we set the p-value threshold at the very low value of  $\alpha = 10e-15$ . The resulting set of discovered patterns contains 22 patterns in 14 tune families, showing that the algorithm is capable of discovering various kinds of melodic patterns that are significantly over-represented in the tune family.

We compare the discovered patterns with the manually annotated motifs as provided in MTC-ANN 2.0. These an-

notated motifs show what parts of the melodies are considered characteristic for the tune family according to human specialist annotators. We compute the establishment precision and recall<sup>2</sup> with a similarity function that considers an overlap of a discovered pattern occurrence with at least half of the notes of an annotated motif a hit, provided that the discovered pattern is not much longer than the annotated motif occurrence. We obtain an establishment precision of 0.86 and an establishment recall of 0.23, showing that the discovered patterns do correspond quite well with annotated motifs, but that the algorithm discovers much less patterns than human annotators did annotate. The low recall is caused by the very conservative p-value that we set. We only discover 22 patterns in 14 tune families, while the annotations consist of 102 motif classes in 26 families.

It is an open question what exactly this evaluation means. The motifs as provided in MTC-ANN 2.0 seem to be a highly subjective choice of the annotators. It is questionable to take this as *ground truth* for pattern discovery. Nevertheless, the high establishment precision suggests at least that the algorithm is able to find parts of the melodies that are considered stable within the tune family by human specialists. Further study of the interaction between the algorithmic results and the human annotations is needed.

As an example of a pattern that does not correspond with an annotated motif, we show a distinctive pattern that was discovered in tune family Koopman (adopting the abbreviations of the tune family names of Volk & Van Kranenburg, 2012). This pattern comprises an ascending contour starting from the tonic, which may seem trivial. However, the current results show that this particular way of starting a phrase is in fact rare outside Koopman.

The second example that is presented in Figure 1 is interesting because the fourth feature set of the pattern contains { phrpos : first }, which indicates a phrase break as part of the pattern. Such a phrase-crossing pattern would not be considered a motif in traditional hierarchical conceptualization of motifs in music theory. However, in the context of oral transmission, this seems a very meaningful piece of information, stating that this particular way of phrase transition, as part of the pattern, is specific for the tune family. For a singer generating a version of this tune, this might be crucial knowledge to properly sing the song.

## 5. CONCLUDING REMARKS

In this study, we present a first step towards a computational model of a given folk song culture as constituting of recombinations of a (possibly very large) number of melodic motifs. The occurrences of these motifs establish the identity of a song as member of a tune family. Since motifs may reoccur in a more or less varied appearance, the current approach in which not all notes of a motif are necessarily described with the same set of features, is very appropriate. The current study shows that the employed MGDp discovery method is capable of discovering parts

of the melodies that are stable within the variants of a tune family. Furthermore, it shows that the algorithmic mining results in patterns that very likely would not occur in traditional analysis, but that are meaningful in the context of understanding oral transmission of melodies.

There are several questions that should be addressed in future work when pursuing this approach. The relation of discovered patterns to experts' annotations is still poorly understood. Furthermore, there is still a gap between traditional musicological conceptualizations of motifs and tune families, and the kinds of patterns that are discovered by automatic discovery as presented in our study. We are convinced that a proper confrontation between the two domains will be beneficial for both, enriching traditional folk song analysis with objective methods, and enriching the algorithmic approach with knowledge of oral transmission of melodies.

## 6. ACKNOWLEDGMENTS

This research is partially supported by the project Lrn2Cre8 which is funded by the Future and Emerging Technologies (FET) programme within the Seventh Framework Programme for Research of the European Commission, under FET grant number 610859. Peter van Kranenburg is supported by the Computational Humanities Programme of the Royal Netherlands Academy of Arts and Sciences, under the auspices of the Tunes & Tales project. Thanks to Kerstin Neubarth for assistance with the manuscript.

## 7. REFERENCES

- Bayard, S. (1950). Prolegomena to a study of the principal melodic families of British-American folk song. *Journal of American Folklore*, 63(247), 1–44.
- Bohlman, P. (1988). *The Study of Folk Music in the Modern World*. Bloomington: Indiana University Press.
- Conklin, D. (2010). Discovery of distinctive patterns in music. *Intelligent Data Analysis*, 14(5), 547–554.
- Conklin, D. (2013). Antipattern discovery in folk tunes. *Journal of New Music Research*, 42(2), 161–169.
- Nettl, B. (2005). *The Study of Ethnomusicology: Thirty-one Issues and Concepts* (2nd ed.). Urbana and Chicago: University of Illinois Press.
- Neubarth, K. & Conklin, D. (2016). Contrast pattern mining in folk music analysis. In D. Meredith (Ed.), *Computational Music Analysis* (pp. 393–424). Springer.
- Tappert, W. (1890). *Wandernde Melodien: Eine musikalische Studie*. Leipzig: List und Francke.
- Van Kranenburg, P., De Bruin, M., Grijp, L. P., & Wiering, F. (2014). The Meertens Tune Collections. Meertens Online Reports 2014-1, Meertens Institute, Amsterdam.
- Volk, A. & Van Kranenburg, P. (2012). Melodic similarity among folk songs: An annotation study on similarity-based categorization in music. *Musicae Scientiae*, 16(3), 317–339.

<sup>2</sup> As defined at: [http://www.music-ir.org/mirex/wiki/2015:Discovery\\_of\\_Repeated\\_Themes\\_&\\_Sections](http://www.music-ir.org/mirex/wiki/2015:Discovery_of_Repeated_Themes_&_Sections). Accessed: 5 June 2016.

# THE GEORGIAN MUSICAL SYSTEM

**Malkhaz Erkvanidze**

maxokulasheli@gmail.com

## 1. INTRODUCTION

The present paper discusses selected results from my dissertation work on “The History of Georgian Chant Notation and the Georgian Musical System” (Erkvanidze, 2014). One of the main challenges in the context of trying to understand the Georgian Musical System is the fact that no theoretical treatise has survived to this day. The main sources of our information today are the audio recordings of professional chanter-singers made over 100 years ago. Theoretical and acoustic analysis of this material are the only means to understand the old Georgian musical system and the underlying systemic thinking, or modal thinking as we call it. This has been the focus of my research since I was a Conservatoire student.

Naturally the question arises how the musical system and systemic thinking are preserved and maintained in these recordings? For years, together with my students, I conducted experiments in the direction of singing and thinking in the original Georgian scale. I believe that both the musical system and systemic thinking are invariably preserved in the available recordings, primarily implying modal thinking. The magnificent traditions of the old Georgian schools of chant and song had been imprinted and settled in the minds and memories of the recorded performers.

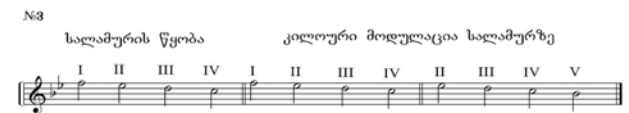
## 2. GEORGIAN MUSICAL THINKING

It is well known among musicologists studying traditional Georgian music that Georgian musical thinking is based on tetrachords, pentachords, as well as seven- and eighth-note scales. The largest challenges in Georgian ethnomusicology today, and a topic of intense research, are related to the attempts to understand the modal structure of the music. Although the importance of the modal structure is basically undisputed among musicologists, there is still controversial discussion on several aspects. In this context, critical points from my perspective are e. g.: a) modes and steps are often counted bottom-up and not top-down ; b) the bass is often considered as determinant for mode, i.e. the mode is ascertained according to the bass either at the beginning or at the end of the song; c) the bass is often considered to direct modulations; d) conclusions are often drawn from transcriptions of the songs into a five-line notation system, which often does not correspond to the original sound; e) theory is often distanced from practice.

Based on the results of my research which will be further detailed below, I have come to the conclusion that the Georgian system conforms to descending thinking. It

resembles the Ancient Greek descending modal system, where the basic mode is obtained from two ways of binding two tetrachords: “interlocked” and “separated”. Modal formations of higher quality came into existence as tetrachord combinations. There were two principles of combination: interlocked – with the coincidence of adjacent sounds in tetrachords and separate – with adjacent sounds distanced by a whole tone” (Kholopov, 1975: 30). In this context, see also the discussion of the descending nature of Georgian modes by Kakhi Rosebashvili discussed (Rosebashvili, 1988).

The discovery of the 3500-year-old tongueless *salamuri* in Mtskheta is the earliest evidence for the existence of a musical system in Georgia. The tongueless *salamuri* is a unique instrument from the perspective of musical acoustics; different tetrachords are formed by inclining the instrument. Particularly interesting is the modulation on the *salamuri*. The basic scale of the instrument is the descending tetrachord; it should be noted that this tetrachord is the same as the basic upper tetrachord of the general scale discovered by us, but with a distinctive character: with the elevated II and lowered III steps. The eight-step scale manifested in the polyphonic modal system could have been formed based on the descending tetrachord (Ex. 1).



**Example 1.** Scale of Salamuri and modal modulation on the Salamuri.

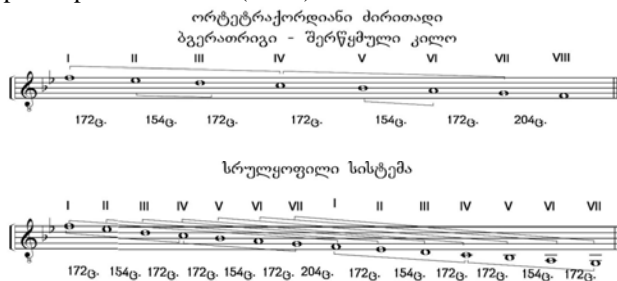
Particular mention should be made of several circumstances. In Georgian musical thinking, modes, as well as the principle of harmonization, are conceptualized downward. As a consequence of this, the mode should naturally be determined top-bottom and not vice versa. The bass functionally is derived from the top voice, but at the same time the bass establishes the stability of the mode. During modulation (apart from rare exceptions, especially in folk songs), it does not initiate the process but acts according to the upper voices, particularly the top voice-part. In chanting, it operates strictly from the top voice, but in folk music either from *mtkmeli* or *modzakhili*, depending on the case. I would also like to add that the bass is always ready for modulation, as it always knows in advance when the modulation is going to happen, as the bass singer also knows both top and middle voice-parts (in the old times every performer knew all voice-parts).



The analysis of the audio material revealed that Georgian chant thinking is based on an eight-degree scale according to the two methods for binding tetrachords into a scale, which I refer to as merged and split scales. Otherwise, this is one scale which implies two in itself. In this context two modes can coexist: one of them is basic, the other one auxiliary. There are cases when only one scale with one mode is standing out. I have observed that both modes are equally present in chants.

In a large number of Georgian chants and songs, with the two ways of tetrachord binding (merged and split), Georgian scales outwardly resemble old Greek ones. When discussing the old Greek modal system, Kholopov and Herzman correctly consider modal steps from Nete to Hipate (Kholopov, 19 : 306; Herzman, 1986: 29). Sulkhan-Saba Orbeliani provides similar explanation: “Hipatoi and Nita are strings tuned from *zili* to *bokhi*” (Orbeliani, 1991: 594). Such an explanation by the lexicographer underlines the descending nature of musical thinking, i.e. from *zili* to *bokhi* (*zili* is the top string of the instrument, while *bokhi* is the bottom one).

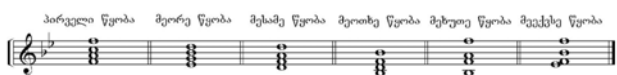
What is the principal difference between the old Greek and Georgian descending modes? As a consequence of the fact that Georgian musical thinking is polyphonic and Georgian polyphony is constructed on fourth-fifth-octave parallelism, at one time tetrachords were evidently divided with the consideration of this type of polyphony, namely: a) by avoiding the tritone (the existence of a tritone as augmented fourth or diminished fifth was inadmissible at the time), at the same time by the maintenance of interval features by fourth and fifth; b) to achieve modal diversity by displacing the centers of upper steps of the mode (Ex. 2).



**Example 1.** Top staff: Basic mode with two tetrachords –interlocked mode. Bottom staff: Accomplished system.

See also the chart illustrating interval conformity in the Appendix (Fig. A1). If the distance between the sounds of the Georgian systemic scale changes by a few cents, everything will become mixed together. It is remarkable that during singing the old masters accurately follow this modal thinking and system.

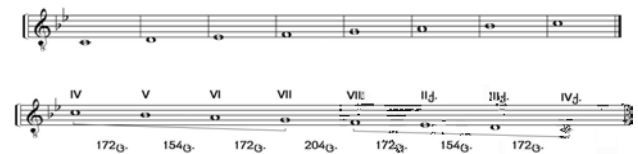
Based on the studied material we suppose that there are six ways for tuning the *chonguri* (Ex. 3).



**Example 3.** Six different tunes of the *chonguri*.

### 3. EVIDENCE FOR THE TOP-DOWN STRUCTURE OF THE MODAL SYSTEM

As mentioned above, the Georgian mode is descending. There is a huge difference between ascending and descending modes, expressed in conceptual categories. For instance, if we compare the ascending Dorian mode with the Georgian descending double-tetrachord split mode we can see that both have a similar modal character (Ex. 4).



**Example 4.** Top staff: ascending Dorian mode. Bottom staff: descending double-tetrachord split mode.

The main aspect here is that the two modes have different reference points: the bottom degree I for the ascending and the top degree I for the descending type. The ascending Dorian mode is generally considered to be of a minor character and indeed this is so; as for the Georgian descending split mode, which resembles the Dorian mode at first glance, the picture is absolutely different. In the descending Georgian mode songs and chants are mostly interpreted as of major character with the consideration of the central I degree. This is determined by the existence of the upper degree as centre (Ex. 5).



**Example 5.** The Georgian Major character of Dorian mode in double-tetrachord separated mode.

For more clarity, all examples in the paper are presented in one mode-tonality with the consideration of corresponding degrees.

In support of the conclusion that the mode is determined not by the bass or lower support, but by degree I, below are three examples of *Kriste aghdga* (Ex. 6).



**Example 6a.** The chant *Kriste Aghdga* (*Christ Is Risen*) in Kartli-Kakhetian simple mode (Erkvanidze, 2014a: 267).



**Example 6b.** The chant *Kriste Aghdga* (*Christ Is Risen*) as performed by Artem Erkomaishvili. Preserved at the Archive of the Georgian Musical Folk Laboratory of Tbilisi State Conservatoire.



**Example 6c.** The chant *Kriste Aghdga* (*Christ Is Risen*) in the version of Benia Mikadze (Erkvanidze, 2004: 147).

Let us look carefully. The first part is the same in all three cases, but in the beginning of the first variant the bass enters the fifth, whilst in the second variant it enters the oc-

tave, and in the third variant, the ninth. Does this not mean that in all three cases we have different modes? Indeed, this would be the conclusion based on the current perspective of many in ethnomusicologists. However, I believe that the determination of the mode from the bass will not lead to logical conclusions. In the presented examples it seems obvious that the mode is determined from by the top voice with its centre on degree I which directs and determines the mode.

Here we have an example of the old method of vocal tuning according to which in chanting and singing the bass tunes with the first voice, sometimes in a fifth, sometimes in an octave, but in West Georgia, particularly in the Gurian tradition, in a ninth. This is a usual occurrence. If in the three afore-mentioned cases of *Kriste aghdga*, we would determine the modes according to bass we would come to the conclusion that there are three different modes on the same melody. I definitely believe that this is wrong. In this particular case we have a double-tetrachord split mode with the consideration of the centrality of the upper step and different ways for the harmonization of bass.

#### 4. MODULATIONS

Some remarks on modulations within the Georgian modal system. If F is conditionally considered as a basic mode-tonality, then G and E flat are considered closely kindred tonalities (Ex. 7).



**Example 7.** Basic mode-tonality and related closely and distant kindred tonalities. Modulations in closely kindred tonalities (from F to G and Es).

Here is one obvious example of modulation activity in the Georgian modal system, the chant *Tsmindano motsameno*, where two mode-tonalities have only one shared pitch (notably, while this single pitch is common, its function is not), but the rest are different (Ex. 8).



**Example 8a.** The chant *Tsmindano motsameno* (*Ye holy martyrs*) as performed by Artem Erkomaishvili. Preserved at the Archive of the Georgian Musical Folk Laboratory of Tbilisi State Conservatoire.



**Example 8b.** The tones of the F mode-tonalities.



**Example 8c.** The tones of the G mode-tonalities.



**Example 8d.** The correlation between the steps during the modulation.

This is exactly what Ioane Batonishvili discusses in “Kalmasoba”: “The *kankledi*, also called *shinpardi*, only slightly differs from mode; it is used the same way with an instrument as with chant in the eight-tone system” (Bagrationi, 1991: 524). I believe that *kankledi* and *shinpardi* imply modulation.

Noteworthy here is that in the European functional system, modulation is realized via a modulating chord, whilst in the Georgian modal system this happens either via a modulating sound or direct transition. In the case of mode-tonal modulation one tone upward or downward, only one sound remains common, while the other 7 change. As an example of modulation into distant tonalities related to F mode-tonality, here I provide the modulation schemes from F to A flat mode-tonalities via E flat mode-tonality (Ex. 9).



**Example 9.** Modulation to the distant mode-tonalities (from F to As and Es) and the tones of the As and Es mode-tonalities (bottom staves).

For an example of such modulation, see the chant *Ganatldi* and the Rachan song *Maqruli* (Ex. 10)



**Example 10a.** The Chant *Ganatldi, ganatldi* (Shine, Shine O New Jerusalem) as performed by Artem Erkomashvili. Preserved at the Archive of the Georgian Musical Folk Laboratory of Tbilisi State Conservatoire.



**Example 10a.** The Rachan folk song *maqruli*, from the field expedition of Chikhikvadze-Grimo, 1967. Transcribed by M. Erkvanidze. Preserved at the Archive of the Georgian Musical Folk Laboratory of Tbilisi State Conservatoire.

Please note that of these two examples one is a chant and the other one is a song. The first stanzas of both examples are identical in terms of modulation and structure, which emphasizes the unity of modal thinking of chant and song.

## 5. MODAL THINKING

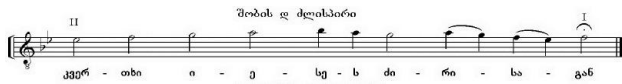
In Georgian polyphonic chant practice, diversity is created not by different modes with corresponding centers (as in the Greek eight-tone system), but by stanzas with different modal characters similar to eight-tone. For instance, the stanza starting on degree I and ending on degree IV has a major character, but the one starting on degree I and ending on degree III has a minor character. As an example, we provide several five-step phrases on different degrees within one mode-tonality (Ex. 11).



**Example 11a.** The chant *Daghatatu Nabsit Tvisit* (Though thou didst descend) as performed by Artem Erkomashvili. Preserved at the Archive of the Georgian Musical Folk Laboratory of Tbilisi State Conservatoire.



**Example 11b.** The chant *Mtsa zeda per itsvale kriste* (*Thou wast transfigured*). Preserved at the Archive of the Georgian Musical Folk Laboratory of Tbilisi State Conservatoire.



**Example 11c.** The Christmas IV Heirmos *Kvertkhilesdzirisagan* (*The Rod of the Root of Jesse*) as performed by Artem Erkomaishvili. Preserved at the Archive of the Georgian Musical Folk Laboratory of Tbilisi State Conservatoire.



**Example 11d.** The IX Heirmos *Saidumlo utskho* (*A Mystery strange*) as performed by Artem Erkomaishvili. Preserved at the Archive of the Georgian Musical Folk Laboratory of Tbilisi State Conservatoire.

As we see, these stanzas have different modal characters, but this does not necessarily mean that they belong to different modes and mode-tonalities. Here we have one F (in this case) mode-tonality and an arrangement of stanzas on different steps within this mode-tonality.

Similar stanzas are on the same degrees of the mode, which determines stability and order in musical material. A number of chant and song examples are bound not according to a particular melody, but mostly by the unity of different stanzas. Thus, for example, the same chant may consist of different stanzas in East and West Georgia.

## 6. CONCLUSIONS AND OUTLOOK

The chief goals of my research and practical work are:

- to understand three-part and polyphonic thinking in general
- to master modal thinking
- to sing in traditional scale and to revive medieval sound
- to think with parallel stanzas in chant and song
- to understand how to differentiate stylistic-harmonious peculiarities
- to master a diversity of traditional performance manners
- to revive systemic thinking in general

- to solve the enigma of *dasdebeli*

The focus (and novelty) of my research is on the deciphering of the Georgian polyphonic musical system, allowing us to conceptualize our ancient music. I hope that in the long run this will lead to

- a revival of the original Georgian modes and scales, allowing us to perform thousands of chants in the original way. In this context, singing modally is particularly important, as this contains the treasures of the Georgian musical language.
- the revival of the musical system, allowing us to make correct and valuable scientific conclusions on issues such as mode, modulations, scales, instrument tunings, eight-tone system, etc.
- the revival of musical systems, implying the revival of modal thinking i.e. the performer knows in which mode the chant or song was started and finished (in the case of modulation), what kind of modulation was applied, and where the musical construction was moved.
- the preparation of text-books of Georgian solfeggio and harmony, allowing the students to become accustomed to correct Georgian musical thinking; such manuals have never yet been prepared.
- enabling the revival of the oral tradition of learning chants and song. I apply this method practically at the High School; after completing the preparatory period in vocal tuning, students can study all three parts of a chant in an oral way with the help of neumes in 15-20 minutes. This practically means that the old method of teaching has been achieved.

One of the most significant issues of Georgian musical thinking is performance in the correct manner. This has survived invariably in old authoritative records. They allowed reviving the sound system, guaranteeing the mastery of correct performance manner.

## 7. REFERENCES

- Batonishvili, Ioane. (1991). *Kalmasoba*. Vol. II, Tbilisi: Merani.
- Erkvanidze, Malkhaz. (2014). *Kartuli galobis notebze gadaghebis istoria da kartuli samusiko sistema* (The History of Georgian Chant Notation and the Georgian Musical System). Doctoral dissertation (manuscript copyright). Preserved at the library of Georgian Folk Music Department of Tbilisi State Conservatoire.
- Gertsman, Evgeny. (1986). *Antichnoe muzikalnoe mishlenie* (*Antique Musical Thinking*). Leningrad: Muzika. (In Russian)
- Orbeliani, Sulchan-Saba. (1991). *Leksikoni kartuli* (*Georgian Lexicon*); vol. 1. Tbilisi: Merani. (In Georgian)
- Kholopov, Yuri. (1975). *Drevnegrecheskie ladi* (*Ancient Greek Modes*). In: *Muzikalnaya entsiklopediya* (*Music*

*Encyclopedia*), vol. II. Pp. 306\_307. Moskva: Sovetskaya entsiklopedia.

Rosebashvili, Kakhi. (1988). “Kartuli khalkhuri simgheris kilos gansazghvis sakitkhisatvis (aghmosavlet sakartvelos khalkhuri simgherebis magalitze)” (“On the Definition of the Mode in Georgian Folk Song (On the Example of East Georgian Folk Songs)”). In: *Kartuli musikis poliponiis sakitkhebi (Issues of Georgian Polyphony)*; pp.: 42\_61. Editors: Gabunia, Nodar, Tsursumia, Rusudan and at al. Tbilisi: Tbilisi State Conservatoire. (In Georgian)

## 8. NOTATED COLLECTIONS

Erkvanidze, Malkhaz (comp.). (2006). *Kartuli saeklesio galoba, t. IV (Georgian Church Chant, vol. IV)*. Tbilisi: Center for Chant of the Georgian Patriarchy.

Erkvanidze, Malkhaz (comp.). (2008). *Kartuli saeklesio galoba, t. V (Georgian Church Chant, vol. V)*. Tbilisi: Center for Chant of the Georgian Patriarchy.

Erkvanidze, Malkhaz (comp.). (2014a). *Georgian Church Chant, East Georgian School (Karbelashvili Mode), Hymns of the Twelve Feasts of Our Lord, the Immovable Feasts, the Lenten Triodion, and the Pentecostarion*. Vol. VII. Tbilisi: Center for Chant of the Georgian Patriarchy.

Koridze, Philimon (comp.). (1895). *Kartuli galoba, partitura 1 (Georgian Chant, Score 1)*. Tbilisi: M. Sharadze's publishing and typography. (In Georgian)

Shughliashvili, Davit. (2002). *Kartuli saeklesio galoba. Shemokmedis skola (Georgian Sacred Chant. Shemokmedi School)*. Tbilisi: Center for Chant of the Georgian Patriarchy.

## 9. APPENDIX



ინტერვალები საკლესიო საგარეო საკლესიო	1	2	3
ინტერვალები საკლესიო	216 ცენტები	172 ცენტები	154 ცენტები
ინტერვალები საკლესიო	376 ცენტები	336 ცენტები	344 ცენტები
ინტერვალები საკლესიო	496 ცენტები	436 ცენტები	444 ცენტები
ინტერვალები საკლესიო	616 ცენტები	536 ცენტები	544 ცენტები
ინტერვალები საკლესიო	736 ცენტები	636 ცენტები	644 ცენტები
ინტერვალები საკლესიო	856 ცენტები	736 ცენტები	744 ცენტები
ინტერვალები საკლესიო	976 ცენტები	836 ცენტები	844 ცენტები
ინტერვალები საკლესიო	1096 ცენტები	936 ცენტები	944 ცენტები
ინტერვალები საკლესიო	1216 ცენტები	1036 ცენტები	1044 ცენტები

Fig. A1. Interval conformity in cents.



# ON THE BENEFIT OF LARYNX-MICROPHONE FIELD RECORDINGS FOR THE DOCUMENTATION AND ANALYSIS OF POLYPHONIC VOCAL MUSIC

Frank Scherbaum

UP Transfer,  
University of Potsdam  
fs@geo.uni-potsdam.de

## 1. INTRODUCTION

In a previous study Scherbaum et al. (2015) have demonstrated that recordings of body vibrations during singing contain all the essential information of a singer's voice regarding pitch, intonation, and voice intensity, but are practically unaffected by the voices of other singers (except for extreme situations). This allows the recording of the contribution of each singer while they are singing together. Because of these characteristics, Scherbaum et al. (2015) proposed the utilization of body vibrations recorded as an additional source of information for the documentation and analysis of traditional polyphonic vocal music. Questions remained, however, regarding the applicability of this approach under field recording conditions and if it indeed provides useful information not obtainable by other means. These questions were at the focus of an exploratory field trip to Upper Svaneti/Georgia during the summer of 2015. Here I report on selected results of the analysis of recordings (larynx microphone and audio) of 20 Svan songs sung by two different trios in Lakhushdi and Ushguli in Svaneti/Georgia recorded during this pilot study.

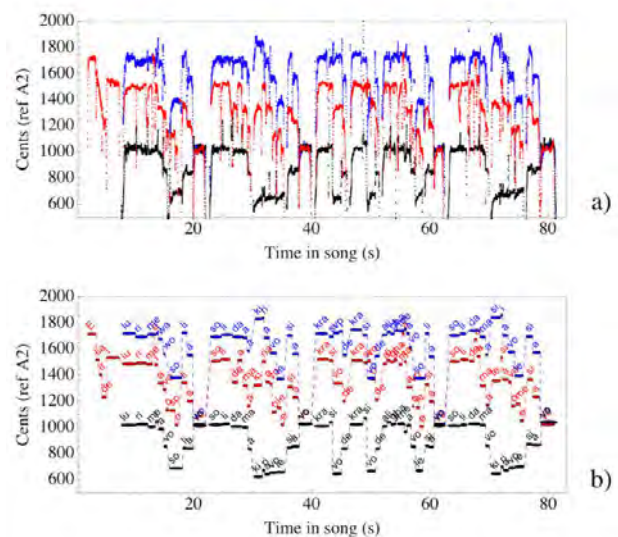
## 2. FIELD EXPERIMENT

The region of Svaneti, located on the southern slopes of the high Caucasus Mountains in Northwestern Georgia, is the home of a highly distinctive musical heritage. Svan songs represent a living part of ancient traditions and are believed to be one of the oldest forms of Georgian vocal polyphony (e. g. Araqishvili, 2010). During the field trip of 2015, a total of ten singers in three villages in Upper Svaneti/Georgia, were willing to take part in the experiment and have themselves recorded in four different trio combinations with a combination of conventional stereo microphones, a video camera but in particular also with larynx microphones tied around their necks.

The analysis of body vibration recordings allows to address a number of interesting musicological problems from a new perspective. For the following illustration, three topics have been selected: documentation, intonation and interaction of singers, and the tuning of traditional Georgian vocal music.

## 3. DOCUMENTATION OF MICROTONALITY

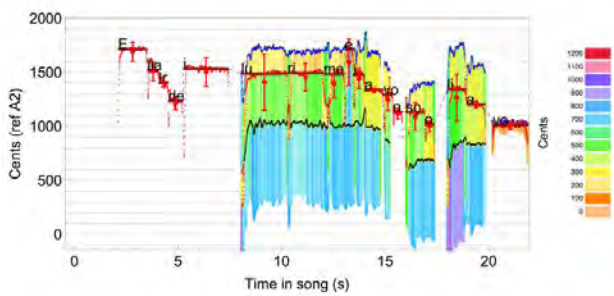
One of the most obvious uses of body vibration recordings is in the context of documentation. In contrast to conventional field recordings, larynx-microphone recordings capture the contribution of each singer separately and in a way which allows automatic pitch recognition and note estimation with high precision. Fig. 1 shows an example of the individual pitch and note tracks for the song *Elia Lrde*, recorded by three larynx microphones. The pitch and note estimation was done with the TONY software (Mauch et al., 2015) on each of the recordings and subsequently combined in Fig. 1. The lyrics were manually added to the output file.



**Figure 1.** Pitch tracks (a) and annotated note tracks (b) for the song *Elia Lrde* (singers: Islam Pilpani (red), Gigo Chamgeliani (blue), Murad Pirtskhelani (black)). Pitches are given in cent relative to A2 (110 Hz).

Several advantages of documenting oral tradition music this way come to mind. First, the process captures all microtonal details (naturally limited to the precision permitted by the sampling process and the subsequent analysis) of the music and does not force it into a possibly inappropriate (tempered) notation system. It is completely transparent and reproducible. Furthermore, it documents the music in a digital form which allows subsequent processing in a multitude of new ways. To illustrate this further, Fig. 2 shows the beginning of the song *Elia Lrde* displayed in a way which allows to see both the complete

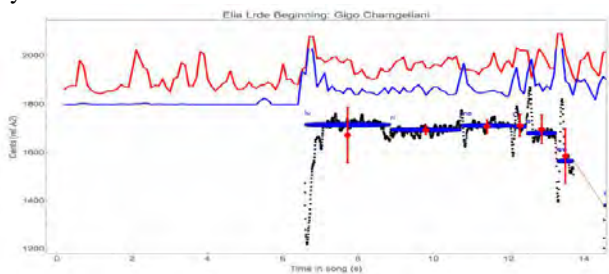
melodic and harmonic content including microtonal details in a single plot.



**Figure 2.** Melodic and harmonic content of the beginning of the song *Elia Lrde*. The black, red and blue dotted lines show the pitch tracks for the bass, middle and top voice respectively. The spaces between the middle and top voice and the bass and middle voice are color coded according to the corresponding interval sizes between the voices. The space below the bass voice is shaped and color coded according to the interval between bass and top voice.

#### 4. INTONATION AND COMMUNICATION BETWEEN SINGERS

One of the most fascinating aspects of the intonation process in polyphonic a-cappella music is how the individual singers find and maintain their pitches and timbres and how their perception of their own and the other voices influences them in this process (e.g. Mauch et al., 2014). Larynx-microphone recordings in combination with regular microphones can help to monitor the intonation process in an interesting way. Fig. 3 shows again the song *Elia Lrde*, but only for the beginning of the polyphonic part and only for the top voice. The dense dotted line at the bottom part of Fig. 3 shows the sequence of pitches (determined by the TONY pitch tracking algorithm) for the first few seconds. The horizontal blue lines show pitches of the determined notes with the red error bars indicating their corresponding standard deviation. The blue and red traces in the top part of the figure show the sensory roughness values (Vassilakis, 2007) for the top voice and the mix of all voices, respectively, which before the onset of the polyphonic part is only determined by the contribution of the middle voice.



**Figure 3.** Voice track of the top voice onset together with sensory roughness track for top voice and mix of all voices.

From Fig. 3 it can be noted that the voice slides to the target pitch from below. Interestingly, this “sliding

phase” is so soft that it is not really audible on the acoustical microphone but is clearly detected on the larynx microphone. It coincides with a short time of increase of sensory roughness (blue trace) which is also observed on the mix of all voices (red trace). In the present context, change of sensory roughness is seen as a simple proxy for change of timbre. Roughly speaking, while tuning in to the other singers, the singer of the top voice adjusts both pitch and timbre at the same time. Interestingly, the other singers do the same, which points to a strong mutual interaction. This feature was consistently observed for all voices during intonation. Further analysis of these records in this direction (intended to be addressed in a future analysis) might provide interesting information regarding the factors controlling the intonation process (pitch or interval precision, sensory roughness, etc.) in a polyphonic a-cappella setting.

#### 5. TUNING OF TRADITIONAL GEORGIAN VOCAL MUSIC

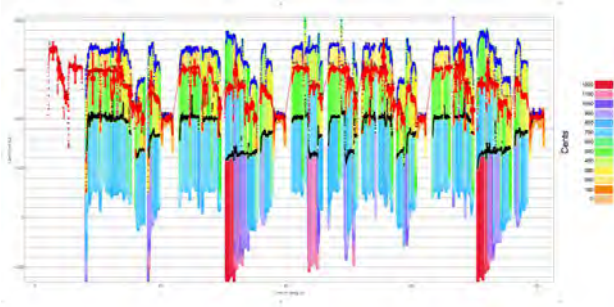
Part of the fascination and archaic beauty of Georgian vocal polyphonic music in general, but Svan music in particular, stems from the abundant use of chords which to ears trained on western music sound “unusual”. In addition, part of the distinctiveness of this music is the fact that the scale(s) from which the pitches for these chords are drawn are not tuned to the 12 tone equal temperament scale (12-TET scale) on which most western music nowadays is based. While the non-tempered nature of traditional Georgian vocal music can be considered consensus amongst musicologists, the particular nature(s) of the Georgian sound scale(s) is an ongoing topic of intense and controversial discussion (Erkvanidze, 2002; Gelzer, 2002; Westman, 2002; Kawai et al, 2010; Tsereteli and Veshapidze, 2014). Complicating the evaluation of the different propositions on what could be called “the Georgian sound-scale controversy” is the fact that it is hard to judge if at least part of the controversy is actually caused by methodological differences or by fundamental disagreement. The analysis of the present set of recordings might be able to contribute to this discussion from a completely new perspective. Since synchronized pitch information from all voices can be derived unambiguously, the analysis of larynx-microphone recordings can help to shed some light on some of the principal questions behind this issue.

Sound-scale and tuning analysis can be done in many different ways, possibly leading to very different results even if the same sound recordings were used. When we listen to polyphonic music, we will perceive melodies and chords. In piano music for example, the intervals which we can hear in a melody and the intervals which we can hear in a chord both draw from the same “interval inventory”, namely the set of all intervals which can be played on the piano. With vocal a-cappella music interval perception can interfere with the intonation which can lead to pitch drifts of the whole ensemble (e. g. Howard, 2007; Mauch et al., 2014). In such a situation, the interval sizes in a melody (horizontal perspective) might differ from the interval sizes in a chord (vertical perspective) which in

turn would make the results of a tuning analysis dependent on the way the intervals are determined. Larynx-microphone recordings provide a very convenient way to quantitatively analyse the magnitude of this effect and if it might affect the determination of sound scale(s).

### 5.1 The harmonic interval set

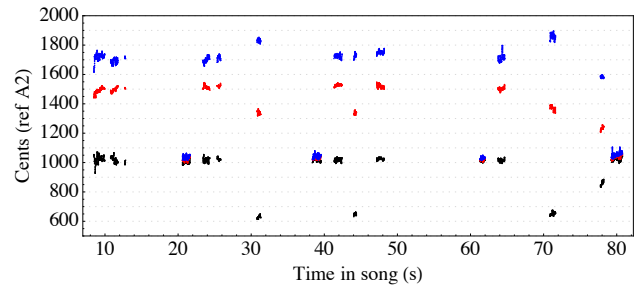
In order to obtain a first impression of the harmonic interval set, in other words the set of concomitantly perceived intervals, in the song *Elia Lrde*, Fig. 4 jointly displays the melodic and harmonic content of the complete song.



**Figure 4.** Melodic and harmonic content of the complete song *Elia Lrde*. The black, red and blue dotted lines show the pitch tracks for the bass, middle and top voice, respectively. The spaces between the middle and top voice and the bass and middle voice are color coded according to the corresponding interval sizes between the voices. The space below the bass voice is shaped and color coded according to the interval between bass and top voice.

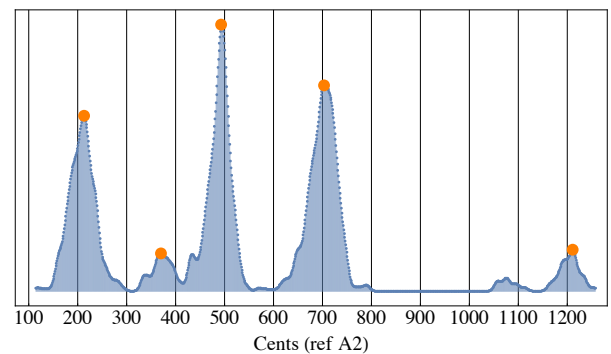
From Fig. 4 it can be seen that the colors representing the intervals between the bass and the top voice are mostly light blue which corresponds to pitch differences of around 700 cents (a fifth), interrupted once in a while by red colors, which corresponds to 1200 cents (an octave). The color codes for the pitch differences between the bass and the middle voice indicate values of approximately 500 cents (a fourth), once in a while interrupted by a difference of 700 cents (a fifth). Consequently the differences between the middle and the top voice correspond to values around 200 cents (a major second), once in a while interrupted by values of approximately 500 cents (a fourth). At times all three voices approach the same pitch value (unison). So in a single glance, Fig. 4 reveals the harmonic character of the song *Elia Lrde*.

For the subsequent analysis, only those pitch samples from the complete pitch tracks were selected which belong to stable notes (as determined by TONY) which have a minimum duration of 1 sec of which the first 0.4 and the final 0.25 sec are discarded for the analysis. The purpose of these restrictions is to discard sliding phases at the beginning (as e. g. seen in Fig. 3) and the end of a note and to focus on intervals which could be called "stably established" by all three singers. The resulting pitch-track sample sets are shown in Fig. 5.



**Figure 5.** Pitches for concomitantly perceived intervals of at least 1 sec duration for the song *Elia Lrde* (singers: Islam Pilpani (red), Gigo Chamgeliani (blue), Murad Pirtskhelani (black)). Pitches are given in cent relative to A2 (110 Hz).

Yet another way to look at the harmonic interval set is by plotting the statistical frequency distribution of the stable concomitant intervals shown in Fig. 5, which results in the distribution shown in Fig. 6.



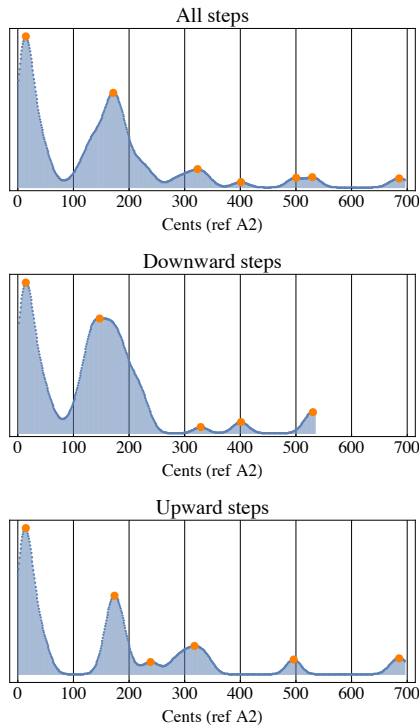
**Figure 6.** Frequency distribution of the stable concomitant intervals. The peaks (orange disks) occur at values of 213, 370, 494, 704, and 1212 cents. For comparison, in Pythagorean tuning the fourth and the fifth correspond to 498 and 702 cents.

The most prominent intervals visible from this perspective appear at 213, 370, 494, 704, and 1212 cents. This corresponds to a slightly sharp major second, a "neutral" third, a fourth, a fifth and a slightly sharp octave. For comparison, in Pythagorean tuning, which can be build up from a series of fifth and which was already described in Babylonian artifacts (West, 1994), the fourth and the fifth correspond to values of 498 and 702 cents which is pretty close to what is observed here.

### 5.2 The melodic interval set

In contrast to the harmonic interval set, the determination of the melodic interval set requires the estimation of the pitch step sizes of successive notes in each of the voices. In this context, all note durations (not only the long ones) where considered. The resulting statistical frequency distribution is shown in Fig. 7.





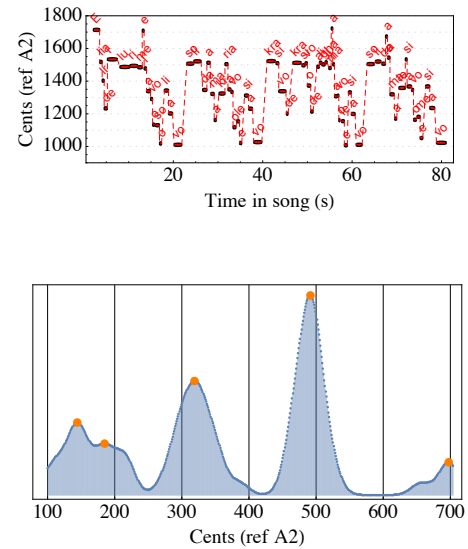
**Figure 7.** Melodic interval distribution in the song Elia Lrde obtained from the pitch differences of successive notes in all three voices. The peaks, marked by the orange discs, of the overall step size distribution (top panel) appear at 14, 172, 323, 401, 500, 529 and 685 cents. The peaks in the distribution of the downward steps (middle panel) appear at 15, 147, 328, 401 and 530 cents. The peaks in the distribution of the upward steps (bottom panel) appear at 14, 174, 239, 318, 496 and 685 cents.

Fig. 7 reveals a number of interesting features. The peak at approximately 15 cents corresponds to the small amplitude fluctuation discussed above but are not really seen as a real feature of the melody. The most prominent deliberate melodic pitch step shows up at 172 cents for all steps combined but appears to be smaller (147 cents) for downward steps than for upwards steps (174 cents). The peaks in the distribution for all steps combined are not very far from the integer multiples of the most prominent melodic interval at 172 cents (which would be at 344, 516 and 688 cents) which could therefore be interpreted as the basic building block of the melodies. Interestingly, this value coincides very well with the value Tsereteli and Veshapidze (2014) determined as basic distance for their proposed equidistant Georgian sound scale.

### 5.3 The analysis of a single voice

As far as I know, Tsereteli and Veshapidze (2014) derived their sound scale model essentially by analysis of individual voices, in other words by melodic and not by harmonic analysis. In order to investigate the consequences of this approach on the current records, the middle voice of the song Elia Lrde was selected (Fig. 8 top panel) and the corresponding pitches of the note set as determined by TONY were determined. The melodic intervals were calculated with respect to the mean value of

the lowest notes in the song (at 1017 cents in the top panel). The resulting statistical frequency distribution is shown in the bottom panel of Fig. 8.



**Figure 8.** Frequency distribution of melodic intervals obtained by analysis of the middle voice of the song Elia Lrde. The peaks marked by orange discs appear at 144, 185, 319, 491, and 697 cents.

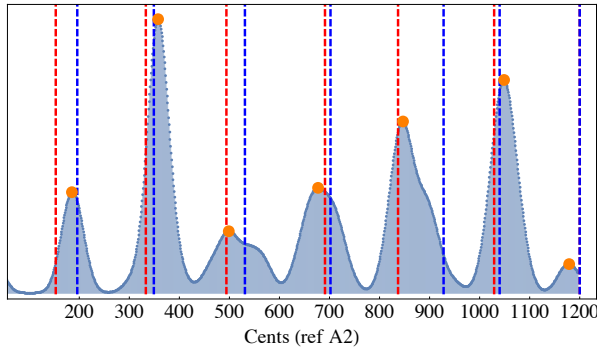
The location of the peaks of the interval frequency distribution reasonably well matches the melodic interval distribution shown in Fig. 7. This would be in line with the hypothesis that it is the basic melodic step size which will control the resulting sound scale model. The double peak below 200 cents might be due to the difference in the upward and downward melodic step size.

### 5.4 Melodyne's scale detective

Using larynx microphones, the melodic and harmonic interval set of a song can be precisely determined since the individual voices are already separated during recording but time synchronisation is kept. With traditional audio recordings, however, the situation becomes blurred because polyphonic pitch determination is still subject to considerable technical challenges. A few commercial software packages exist which have tried to attack this problem with mixed success. One of those, the recently released Melodyne 4 (Celemony GmbH), contains polyphonic pitch tracking and a feature called direct note access (DNA) which claims to allow to access the properties of the individual notes detected in an audio record. In addition, it contains a feature called “scale detective” which allows the determination of a sound scale corresponding to the analysed audio material. Since the underlying algorithms are unknown, it is impossible to test the performance of these tools in a scientific way, but a comparison of algorithms in the present case might provide some information regarding their applicability for tuning analysis in cases where only audio material is available, e.g. historical phonograph records.

For this comparison, the notes for all voices determined from the individual larynx microphone recordings by the TONY algorithm were jointly used for the analysis of the

frequency distribution of intervals determined from the pitch differences of note pairs. In this case, the distinction between melodic and harmonic intervals is lost because some pitch pairs may belong to the same time and hence be harmonic while the majority will be correspond to different times and hence has to be considered melodic. It is suspected that this setup best matches the situation of the scale detective in Melodyne which faces the additional challenge of polyphonic pitch determination. The resulting distribution is shown in Fig. 9. The vertical dashed lines correspond to the pitch values found for a seven degree scale using the mix of all larynx microphone recordings (red lines) and the conventional audio stereo record (blue lines).

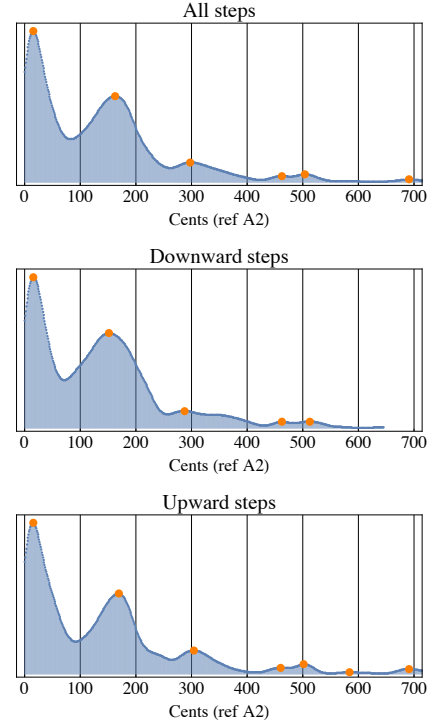


**Figure 9.** Frequency distribution of intervals obtained by analysis of all three voice of Elia Lrde. The peaks marked by orange discs appear at 187, 358, 499, 678, 847, 1049, and 1179 cents and 144, 185, 319, 491, and 697 cents. The application of Melodyne's (release 4) scale detective on the mix of all larynx microphone recordings and the audio stereo recording results in sets of pitch values of {196, 349, 531, 702, 928, 1040, 1200} (blue lines) and {153, 333, 494, 691, 837, 1029, 1200} (red lines), respectively.

The results of applying Melodyne's scale detective to the mix of larynx microphone recordings results in pitch values which are reasonably close to the peaks of the frequency distribution of intervals shown by the values indicated by the orange discs. Except for the first one, these values are reasonably close to the integer multiples of the basic melodic pitch step size of 172 cents which would be at 172, 344, 516, 688, 860, and 1032 cents. The fact that the first peak appears closer to 200 cents than for the analysis of the individual voice could be due to the fact that in particular the seconds in this mixed data set are a mixture of harmonic and melodic intervals as already discussed above. The results of applying Melodyne's scale detective to the audio stereo signal are similar except for the pitch value at 928 cents. Since the algorithm is not known, the reasons for this remain unknown.

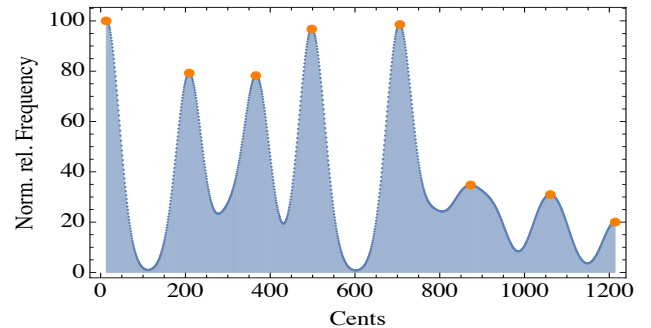
### 5.5 How robust are these features?

In order to test the robustness of the observed features, the analysis was extended in two ways. First, all voices in the recordings of five songs sung by Islam Pilpani, Gigo Chamgeliani, and Murad Pirtskhelani in Lakhushdi were analysed regarding the containing melodic and harmonic interval sets. The results are shown in Figs. 10 and 11, respectively.



**Figure 10.** Frequency distribution of all melodic intervals obtained by analysis of all voices in the songs Elia Lrde, Jragish, Kviria, Lile and Riho sung by Islam Pilpani, Gigo Chamgeliani, and Murad Pirtskhelani. The peaks, marked by the orange discs, of the overall step size distribution (top panel) appear at 15, 163, 298, 462, 504, and 691 cents. The peaks in the distribution of the downward steps (middle panel) appear at 16, 152, 287, 463, and 513 cents. The peaks in the distribution of the upward steps (bottom panel) appear at 15, 169, 304, 460, 502, 584, 691 cents.

The results in Fig. 10 are similar to the ones for the single song Elia Lrde in that the dominant melodic interval for all voices and all songs is still on the order of 150 - 170 cents. In addition, the feature that the pitch steps downward are systematically smaller than the upward steps is also observed for all songs.

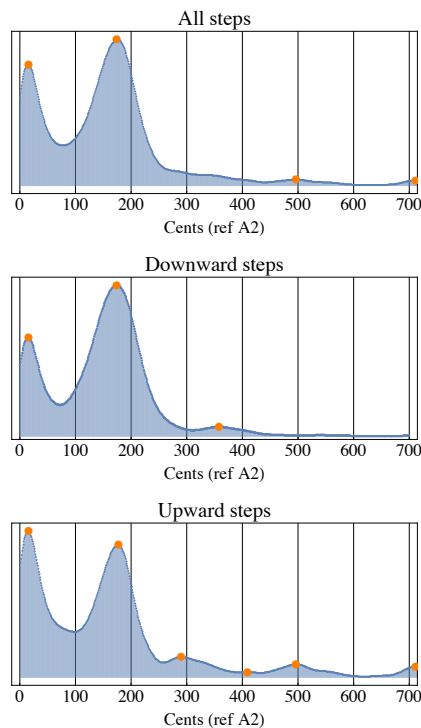


**Figure 11.** Frequency distribution of all harmonic intervals obtained by analysis of all voices in the songs Elia Lrde, Jragish, Kviria, Lile and Riho sung by Islam Pilpani, Gigo Chamgeliani, and Murad Pirtskhelani. The peaks (orange disks) occur at values of 13, 209, 366, 498, 706, 873, 1061, and 1214 cents.

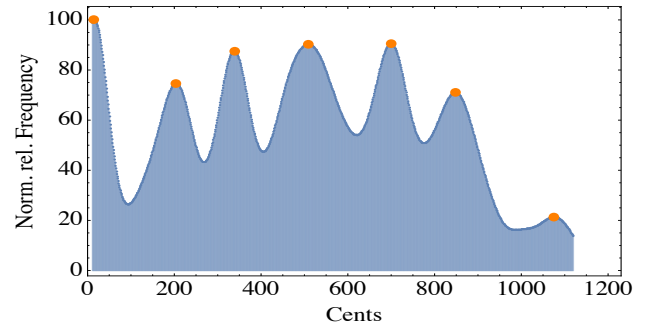


The harmonic interval set derived from the analysis of all songs sung by Islam Pilpani, Gigo Chamgeliani, and Murad Pirtskhelani turns out to consist of 7 steps and is clearly not equidistant. The “major 2nd” at 209 cents, which would not exist if the scale were equidistant, is clearly present in all songs and all voices. The fourth and the fifth at 498 and 706 cents are very close to just tuning as will be further discussed below.

The second test to check the robustness of the observed features was to analyse the recordings of 15 songs sung by Jano Charkseliani, Zoia Charkseliani, Lola Nizharadze in Ushguli in the same way. The corresponding melodic and harmonic interval distribution from all voices and all songs are displayed in Figs. 12 and 13, respectively.



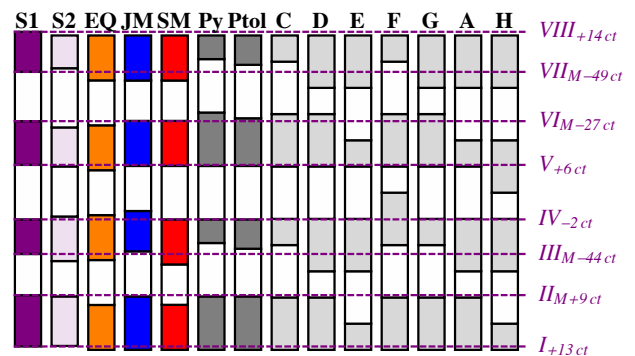
**Figure 12.** Frequency distribution of all melodic intervals obtained by analysis of all voices in 15 songs sung by Jano Charkseliani, Zoia Charkseliani, Lola Nizharadze in Ushguli. The peaks, marked by the orange discs, of the overall step size distribution (top panel) appear at 15, 174, 496, 711, and 814 cents. The peaks in the distribution of the downward steps (middle panel) appear at 15, 174, and 358 cents. The peaks in the distribution of the upward steps (bottom panel) appear at 15, 177, 290, 409, 496, 711, and 814 cents.



**Figure 13.** Frequency distribution of all harmonic intervals obtained by analysis of all voices in 15 songs sung by Jano Charkseliani, Zoia Charkseliani, Lola Nizharadze in Ushguli. The peaks (orange disks) occur at values of 15, 204, 339, 509, 700, 849, and 1075 cents

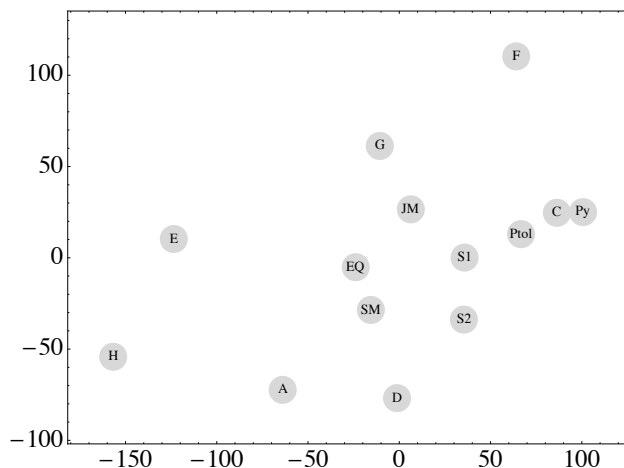
Again, the harmonic interval set is clearly different from the melodic one. The melodic 2nd is again at approximately 170 cents while the harmonic one is at 204 cents. The harmonic interval set derived from the Ushguli recordings is very similar to the one derived from the Lakhushdi recordings while the melodic interval sets differ in that the Ushguli recordings do not show a difference between downward and upward movements.

From the analysis so far, it looks like there is a significant difference between the interval set for the chords and the one for the melodic movements. Below, the harmonic interval sets derived from the analysis of the recordings in Lakhushdi and Ushguli are compared to the ones which were proposed by Erkvanidze (2002) as models for Georgian tunings, to the one proposed by Tsereteli and Veshapidze (2014) but also to the Pythagorean scale and Ptolemy’s diatonic scale as well as the modal scales derived from 12 tone equal tempered (12-TET) tuning (Fig. 14).



**Figure 14.** Comparison of the interval sizes of the harmonic interval sets derived from the analysis of the recordings in Lakhushdi (S1) and Ushguli (S2) with the equidistant scale suggested by Tsereteli and Veshapidze (2014) (EQ), the joined (JM) and split (SM) mode scales suggested by Erkvanidze (2002), the Pythagorean scale (PY), the Ptolemy diatonic scale (Ptol) and the modal scales derived from 12 tone equal tempered tuning (modes C- H).

Obviously, the harmonic interval set derived from the analysis of the recordings in Lakhushdi and Ushguli (S1 and S2) are quite different from the rest of the modal scales derived from the 12 TET based scales, but over all it is difficult to interpret the mutual relationships simply from the visual appearance in Fig. 14. One way to visually represent the information contained in Fig. 14 in a more intuitive way is by making use of methods from high-dimensional visualization and multi-dimensional scaling analysis. For this purpose one can view each interval set as a feature vector in a high-dimensional space the dimensions of which are given by the number of different intervals present in (here 7). What would actually be interesting to "see" is the set of the individual points (to which the feature vectors point to) in this seven dimensional space, which of course can only be calculated but not "seen". However, one can try to project the points from the high-dimensional space onto a map (similar to the way the three-dimensional surface of the Earth is projected onto a two-dimensional map) in such a way, that the neighbor-relations between nearby points in the high-dimensional space is preserved. One of the techniques by which this can be achieved is the non-linear Sammons map (Sammon, 1969). The resulting distribution of scales (or interval sets) is shown in Fig. 15.



**Figure 15.** Sammon's map for the scales in Fig. 13. The mutual distances between the scales quantitatively reflect their similarity in a Euclidean sense.

Each labeled disc in Fig. 15 corresponds to one of the scales (or interval sets) in Fig. 14. The proximity of the scales in Fig. 15 corresponds to the similarity of the corresponding interval vectors in a Euclidean sense. Fig. 15 shows that the harmonic interval sets obtained from the recordings in Lakhushdi (S1) and Ushguli (S2) are most similar to each other, followed in similarity by the scale suggested by Erkvandize (JM) and Ptolemy's diatonic scale (Ptol).

## 6. DISCUSSION AND CONCLUSIONS

Based on the material presented above it seems justified to say that field recordings of body vibrations can provide new and very valuable information on the tunings and

intonation of traditional singers which would be difficult to obtain by only conventional audio recording setups. Most importantly, larynx microphone recordings capture the contributions of the individual singers undisturbed by the other singers and therefore offer the possibility to investigate the melodic and harmonic interval inventory of a song separately. The results of the analysis of 20 Svan songs sung by two different trios in Lakhushdi and Ushguli suggest a clear and significant difference between the melodic and the harmonic interval set. For one of the trios, the melodic interval set even showed differences in step sizes between downward and upward movements. It is worth noting, that this is not a feature unique to Svan songs, as I learned from S. Arom (pers. comm., Arom, 2016). I cite from his comments: "Both observations absolutely corroborate what I experienced when, years ago, I was transcribing (only by ear, alas...) the polyphonic songs of the Aka Pygmies: when listening to the isolated recording of a singer's voice and to its combination with another voice, the perception of the intervals is different!".

The sizes in which melodic movements in the analysed songs happen occur in multiples of roughly 170 cents which would be consistent with an interval set with intervals of equal size, while the concomitantly perceived intervals are related to sets in which the different degrees are clearly non-equidistant. In the context of the discussions on the Georgian sound scale(s), a quote which is attributed to the German physicist Werner Heisenberg, nobel laureate and one of the fathers of quantum mechanics comes to mind: "We have to remember that what we observe is not nature in itself, but nature exposed to our method of questioning.". The results of the present analysis seem to suggest that the analysis of individual voices or monodic segments will result in an equidistant scale model, while the analysis of concomitant intervals will result in a non-equidistant scale model.

In conclusion, it seems worth to investigate further if the different propositions regarding the "Georgian sound scale" could be reconciled by assuming that the difference between the melodic and the harmonic interval set is a general and robust feature of traditional Georgian vocal polyphony. At present this is admittedly only a speculation based on a very limited data set but as an hypothesis it seems worth to be tested further. The consequences of such a model would be that during melodic movements of a song the singers would continuously readjust the tunings of their intervals to the desired values which is similar to what a brass instrument player is doing when playing in an orchestra (pers. comm. Arom, 2016). This might actually also explain some of the seemingly random pitch fluctuations observed in the individual pitch tracks of these highly skilled traditional singers.

## 7. ACKNOWLEDGMENTS

I am extremely thankful to Nana Mzhavanadze for organising the recording sessions for this project. Without her help these recordings simply would not exist. I feel honoured and very grateful to the singers (in alphabetical order) Ana Chamgeliani, Eka Chamgeliani, Gigo Chamgeliani, Madonna Chamgeliani, Jano Charkseliani, Zoia Charkseliani, Lola Nizharadze, Islam Pilpani, Murad Pirtskhelani, and Givi Pirtskhelani for participating in this study. I am also greatly indebted to Malkhaz Erkvanidze for many stimulating discussions on the Georgian scale problem and for generously sharing his perspective on Georgian music and its theory with an enthusiastic geophysicist. Last but not least, I want to express my gratitude to Frank Kane who got me interested in the topic of body vibrations and Georgian vocal music in the first place and to Simha Arom for his comments on an earlier version of the manuscript.

## 8. REFERENCES

- Araqishvili, D. (2010), Svan Folk Song, in *Echoes from Georgia: seventeen arguments on Georgian polyphony*, (eds. Tsurtsumia, R. and Jordania, J.), p. 35-56.
- Erkvanidze, M. (2002). On georgian scale system. In *The First International Symposium on Traditional Polyphony: 2-8 September, 2002, Tbilisi, Georgia* (pp. 178–185).
- Gelzer, S. (2002). Testing a scale theory for Georgian folk music. In *The First International Symposium on Traditional Polyphony: 2-8 September, 2002, Tbilisi, Georgia* (pp. 194–200).
- Howard, D. M. (2007). Intonation Drift in A Capella Soprano, Alto, Tenor, Bass Quartet Singing With Key Modulation. *Journal of Voice*, 21(3), 300-315. doi:10.1016/j.jvoice.2005.12.005
- Kawai, N., Morimoto, M., Honda, M., Onodera, E., & Oohashi, T. (2010). Study on sound structure of Georgian traditional polyphony. Analysis of its temperament structure. In *The Fifth International Symposium on Traditional Polyphony, 4-8 October, 2010, Tbilisi, Georgia* (Vol. 1, pp. 532-537).
- Mauch, M., Cannam, C., Bittner, R., Fazekas, G., Salamon, J., Dai, J., Dixon, S. (2015). Computer-aided Melody Note Transcription Using the Tony Software: Accuracy and Efficiency. In *Proceedings of the First International Conference on Technologies for Music Notation and Representation* (p. 8). Retrieved from <https://code.soundsoftware.ac.uk/projects/tony/>.
- Mauch, M., Frieler, K., & Dixon, S. (2014). Intonation in unaccompanied singing: Accuracy, drift, and a model of reference pitch memory. *The Journal of the Acoustical Society of America*, 136(1), 401–411.
- Sammon, J. W. Jr, (1969), A nonlinear mapping for data structure analysis, *IEEE Transactions on Computers*, vol. C-18, no. 5, pp. 401-409.
- Scherbaum, F., Loos, W., Kane, F., & Vollmer, D. (2015). Body vibrations as source of information for the analysis of polyphonic vocal music. In *Proceedings of the 5th International Workshop on Folk Music Analysis, June 10-12, 2015, University Pierre and Marie Curie, Paris, France* (Vol. 5, pp. 89–93).
- Tsereteli, Z., & Veshapidze, L. (2014). On the Georgian traditional scale. In *The Seventh International Symposium on Traditional Polyphony: 22-26 September, 2014, Tbilisi, Georgia* (pp. 288-295).
- Vassilakis, P. N. (2007). SRA: A web-based research tool for spectral and roughness analysis of sound signals. *Proceedings SMC'07, 4th Sound and Music Computing Conference* (pp. 319-325).
- West, M. L. (1994). The Babylonian musical notation and the Hurrian melodic texts. *Music & Letters*, 75, 161–179.
- Westman, J. (2002). On the problem of the tonality in Georgian polyphonic songs: The variability of pitch, intervals and timbre. In *The First International Symposium on Traditional Polyphony: 2-8 September, 2002, Tbilisi, Georgia* (pp. 212-220).

# AUTOMATIC ALIGNMENT OF LONG SYLLABLES IN A CAPPELLA BEIJING OPERA

Georgi Dzhambazov, Yile Yang, Rafael Caro Repetto, Xavier Serra

Music Technology Group, Universitat Pompeu Fabra, Barcelona, Spain

{georgi.dzhambazov, yile.yang, rafael.caro, xavier.serra}@upf.edu

## ABSTRACT

In this study we propose how to modify a standard approach for text-to-speech alignment to apply in the case of alignment of lyrics and singing voice. We model phoneme durations by means of a duration-explicit hidden Markov model (DHMM) phonetic recognizer based on MFCCs. The phoneme durations are empirically set in a probabilistic way, based on prior knowledge about the lyrics structure and metric principles, specific for the Beijing opera music tradition. Phoneme models are GMMs trained directly on a small corpus of annotated singing voice. The alignment is evaluated on a cappella material from Beijing opera, which is characterized by its particularly long syllable durations. Results show that the incorporation of music-specific knowledge results in a very high alignment accuracy, outperforming significantly a baseline HMM-based approach.

## 1. INTRODUCTION

The task of lyrics synchronization (also known as lyrics-to-audio alignment) has as an aim to find in an automatic way a match between two representations of a musical composition: the singing voice and the corresponding lyrics. Lyrics-to-audio alignment may be used in various applications: for example to automatically match structural sections from lyrics (verse, chorus) to a recording of a particular singer. This facilitates navigation and can thus be beneficial for musicologists or singing students.

The problem of lyrics-to-audio alignment has inherent relation to text-to-speech alignment. Text-to-speech alignment has been a research field for more than 20 years and thus yielded established successful ways for modeling phonemes (Anguera et al., 2014). However, compared to speech, singing voice has some substantially different characteristics including harmonics, pitch range, pronunciation, vibrato, etc. In particular, unlike speech, for singing voice, durations of vocals have on average somewhat higher variation (Kruspe, 2014). This suggests that applying an approach from speech recognition out of the box might not lead to satisfactory results. Traditional music, characterized by frequent local tempo changes, poses an additional challenge: Singers might prolong substantially certain syllables, as a way to emphasize them or as an expressive singing element.

Furthermore, current approaches on modeling lyrics are confined by the necessity of a large speech corpus, on which phoneme models are typically trained (Fujihara & Goto, 2012). Such corpora might not be present for every language or not freely available, as is the case for Mandarin. Recent work has shown that training on singing

voice instead might be a viable alternative (Hansen, 2012).

In this paper we propose a lyrics-to-audio alignment method, which relies on some of the specificities of lyrics structure of Beijing opera as an additional cue to an approach adopted from speech alignment. One of the goals of the study is to show that enhancing computational tasks with music-specific knowledge might improve accuracy.

## 2. BACKGROUND ON JINGJU MUSIC PRINCIPLES

Lyrics in Jingju (also known as Beijing opera or Peking opera) come from poetry and are thus commonly structured into couplets: each couplet has two lyrics lines. A line is usually divided into 3 syllable groups: a group is called *dou* and consists of 2 to 4 written characters (Wichmann, 1991, Chapter III)<sup>1</sup>. To emphasize the semantics of a phrase or according to the plot, an actor has the option to sustain the vocal of the *dou*'s final syllable. In this work we will refer to the final syllable of a *dou* as *key syllable*.

In addition to that, each aria from Jingju can be arranged into one or more metrical pattern (called *banshi*): it indicates the mood of singing and is correlated to meter and tempo (Wichmann, 1991). Usually an aria starts with a slow *banshi*, which gradually changes a couple of times to a faster one, to express more intense mood. The language of Jingju is standard Mandarin with some slight dialect.

## 3. RELATED WORK

Current lyrics-to-audio alignment is mostly based on an approaches, adopted from text-to-speech alignment (Mesaros & Virtanen, 2008; Fujihara et al., 2011): A phonetic recognizer is built from speech corpus, whereby a hidden Markov model (HMM) is trained for each phoneme. The acoustics of phonemes are described by mel frequency cepstral coefficients (MFCCs). In an example of such an approach, polyphonic Japanese and English pop music is aligned (Fujihara et al., 2011). The authors propose to adapt the speech phoneme models to the specific acoustics of singing voice by means of Maximum Likelihood Linear Regression. This is necessary because of the lack of a big enough singing voice corpus for training. Further, an automatic segregation of the vocal line is performed, in order to reduce the spectral content from background instruments.

<sup>1</sup> We use the term *syllable* as equivalent to one written character.

HMMs, being originally applied to model spoken phonemes, have the drawback that, in general, are not capable to represent well vowels with long and highly-variable durations. This is because the waiting time in a state in traditional HMMs cannot be unlimitedly long (Rabiner, 1989). Durations can be modeled instead by a duration-explicit hidden Markov model (DHMM) (also known as hidden semi-Markov model). In DHMMs the underlying process is allowed to be a semi-Markov chain with variable duration of each state (Yu, 2010). DHMMs have been applied to detect keywords from a cappella English pop songs (Kruspe, 2015). The author showed that accuracy of detection increases if the duration of each phoneme is learned from a singing dataset. In addition, DHMMs have been shown to be successful for modeling other problems from the domain of music information retrieval: They have been, for example successful in representing chord durations in automatic chord recognition (Chen et al., 2012).

To our knowledge, very few studies of lyrics-to-audio alignment have been conducted on songs with Chinese language (Wong et al., 2007).

#### 4. APPROACH OVERVIEW

To model phoneme durations, we rely on a DHMM<sup>2</sup>. A general overview of the proposed approach is presented in Figure 1. First an audio recording of an aria is manually divided into audio segments corresponding to lyrics lines as indicated in the lyrics script of the aria, whereby instrumental-only sections are discarded. All further steps are performed on each line segment of audio. If we had used automatic segmentation instead, potential erroneous lyrics and features could have biased the comparison of a baseline system and DHMM. As we focus on evaluating the effect of DHMM, manual segmentation is preferred.

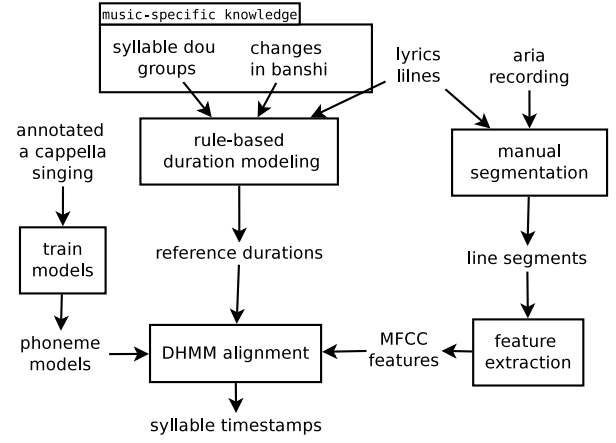
Then each lyrics line is expanded to a sequence of phonemes, whereby reference syllable durations guide the alignment process. The main contribution of this work is twofold: 1) the application of music-specific rules for the creation of reference durations and 2) training phonemes on singing voice.

##### 4.1 Rule-based duration modeling

The idea of the duration modeling is that the actual duration of a phoneme can be seen as being generated by a statistical distribution with highest probability at an expected reference duration. The reference durations can be assigned using any prior knowledge like for example structure of lyrics segments, as has been done by Wang et al. (2004). In this work they are derived as follows:

Firstly, each *key syllable* in a *dou* is assigned longer reference duration according to empirically found ratios, while the rest get equal durations. Additionally, we observed in the dataset that usually the last *key syllable* of the last line in a *banshi* is prolonged additionally. Thus

<sup>2</sup> For brevity in the rest of the paper the proposed alignment scheme will be referred to as DHMM.



**Figure 1:** Approach Overview: The middle column shows how reference durations are derived based on music-specific knowledge.

we lengthened additionally the reference syllable duration of these last *key syllables*. Figure 2 depicts an example. According to *dou* groups the 3rd, 6th and last syllable are expected to be prolonged. Note that for the example this expectation does not hold for the 3rd syllable.

Then, to form a sequence of phoneme reference durations  $R_i$ , the reference durations of syllables are divided among their constituent phonemes, according to the initial-middle-final division of syllables in Mandarin (Duanmu, 2000). A syllable has a middle part (nucleus) being a simple vowel, a diphthong, or triphthong. An initial part (a consonant) or a final part (a group of consonants) is optional. We assign consonants a fixed reference duration  $R_c = 0.3$  seconds, while the rest of the syllable is distributed equally among vowels. The reference durations  $R_i$  are linearly scaled to a reference number of frames according to the ratio between the number of phonemes in a lyrics line and the duration of its corresponding audio segment.

##### 4.2 Phoneme models

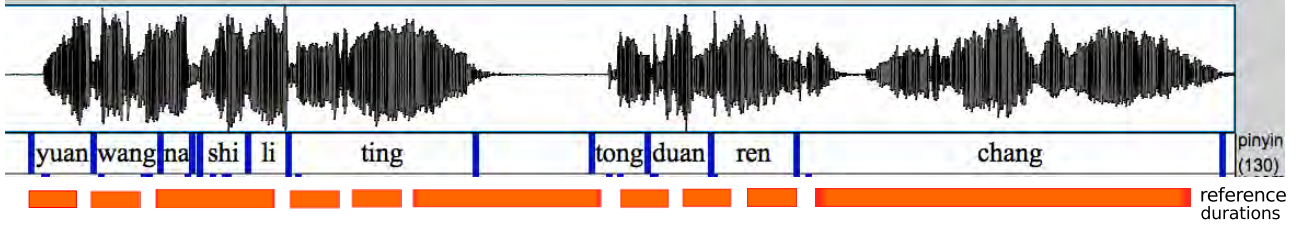
For each phoneme a GMM is trained on annotated a cappella singing. The first 13 MFCCs and their  $\Delta$  and  $\Delta\Delta$  are extracted from 25ms audio frames with the hop size of 10ms. The extracted features are then fit into a phoneme GMM with 40 components: a number of components usually proved as sufficient in speech recognition. A model for silent pause *sp* is added at the end of each syllable, which is optional on decoding. This allows to accommodate the frequent for Jingju regions of pauses after some syllables.

##### 4.3 DHMM alignment

The syllables for a line are expanded to a sequence of phonemes based on grapheme-to-phoneme rules<sup>3</sup>. Then the trained GMMs are concatenated into a phonemes network, represented by a HMM, where each GMM is a state.

<sup>3</sup> We built a pinyin-to-X-Sampa mapping available at <https://github.com/georgid/AlignmentDuration/blob/noteOnsets/jingju/syl2ph.txt>





**Figure 2:** An example of 10-syllable line, being last in a *banshi* (before the *banshi* changes). Actual syllable durations are in pinyin, whereas reference durations are in orange parallelograms (below).

The HMM is aligned to the MFCC features, extracted from the aria, being aligned. The most likely state sequence is found by means of a forced alignment with Viterbi decoding.

We have adopted the idea of Chen et al. (2012) not to represent durations by an additional counter state in the HMM, but instead to modify the Viterbi decoding stage. Let us define

$\delta_t(i)$  : probability for the path with highest probability ending in state  $i$  at time  $t$  (comply with the notation of Rabiner (1989, III. B)))

Now maximization is carried over the most likely duration for each state, instead of over different states:

$$\delta_t(i) = \max_d \{ \delta_{t-d}(i-1) P_i(d) [B_t(i, d)] \} \quad (1)$$

where  $B_t(i, d)$  is the observation probability of staying  $d$  frames in state  $i$  until frame  $t$ . The duration  $d$  of a phoneme is modeled as a normal distribution  $\mathcal{N} \sim (R_i; \sigma)$ , with a peak at  $R_i$ . Thus, we chose to restrict the domain of  $d$  to  $(\max\{R_i - \sigma, 1\}, R_i + \sigma)$ . Note that in forced alignment the source state could be only the previous state  $i - 1$ . More details on the inference with DHMM can be found in our previous work Dzhambazov & Serra (2015). In comparison to our previous work, we opted for dividing the global standard deviation  $\sigma$  into  $\sigma_c$  for consonants and  $\sigma_v$  for vowels. Proper values for  $\sigma_c$  and  $\sigma_v$  assure that a phoneme sung longer or shorter than the expected  $R_i$  can be adequately handled. Another modification we did is that *sp* models are assigned an exponential distribution, because the duration of inter-syllable silences cannot be predicted.

## 5. DATASET

The dataset has been especially compiled for this study and consists of excerpts from 15 arias of two female singers, chosen from a *CompMusic* corpus of Jingju arias (Repetto & Serra, 2014). For a given aria were present two versions: a recording with voice plus accompaniment and an accompaniment-only one. Thus a cappella singing was generated by subtracting the instrumental accompaniment from the complete version<sup>4</sup>. Table 1 presents the average values for lines and syllables.

<sup>4</sup> The resulting monophonic singing is as clean as if it were a cappella, having slightly audible artefacts from percussion on the non-vocal regions

	dataset	'canonical' dataset
<b>duration (minutes)</b>	67	27
<b>#lines per aria</b>	9.2	9.9
<b>#syllables per line</b>	10.7	10.3
<b>line duration (seconds)</b>	18.3	23.4
<b>syllable duration (seconds)</b>	2.4	3.1

**Table 1:** Line and syllable averages about the dataset

Each aria is annotated on the phoneme level by native Chinese speakers and a Jingju musicologist. The phoneme set has 29 phonemes and is derived from Chinese pinyin, and represented using the X-sampa standard<sup>5</sup>. To assure enough training data for each model, certain phonemes are grouped into phonetic classes, based on their perceptual similarity.

Further, we selected a 'canonical' subset of the dataset, consisting of lines, according to the assumptions we made: *key syllables* should be prolonged. Thus, we kept only these audio segments, for which at most one *key syllable* is not prolonged and discarded the rest. We considered a syllable as being prolonged if it is longer than 130% of the average syllable duration for the current line.

## 6. EXPERIMENTS

Alignment accuracy is evaluated as the percentage of duration of correctly aligned syllables from total audio duration (see Fujihara et al. (2011, figure 9) for an example). Accuracy is measured for each manually segmented line and accumulated on total for all the recordings<sup>6</sup>.

### 6.1 Experiment 1: oracle durations

To define a glass ceiling accuracy, alignment was performed considering phoneme annotations as an oracle for acoustic features. Looking at phoneme annotations, we set the probability of a phoneme to 1 during its time interval

<sup>5</sup> Annotations are made available at <http://compmusic.upf.edu/node/286>

<sup>6</sup> To encourage reproducibility of this research an efficient open-source implementation together with documentation is available at <https://github.com/georgid/AlignmentDuration/tree/noteOnsets/jingju>. Further, a script for building the models is available at <https://github.com/elitrou/lyrics/blob/master/code/htk/buildModelHTKSave.py>

	baseline	DHMM	oracle
<b>overall</b>	56.6	89.9	98.5
<b>'canonical'</b>	57.2	96.3	99.5

**Table 2:** Comparison of accuracy on oracle, baseline and DHMM alignment on total and selected arias. Accuracy is reported as accumulate correct duration over accumulate total duration over all lines from a set of arias.

and 0 otherwise. We found that the accuracy per line of lyrics is close to 100%, which means that the model is generally capable of handling the highly-varying vocal durations of Jingju singing. Most optimal results were obtained with  $\sigma_c = 0.7$  seconds;  $\sigma_v = 2.0$  seconds, which are used in experiment 2.

## 6.2 Experiment 2: comparison with baseline

As a baseline we employ a standard Viterbi decoding, run with the *htk* toolkit (Young, 1993). For both baseline and DHMM, to assure good generalization of results, evaluation is done by cross validation on 3 folds with approximately equal number of syllables: Phoneme models are trained on 10 of the arias using the phoneme-level annotations and evaluated on a 5-aria hold-out subset. We have further evaluated on the 'canonical' selected subset of lyrics lines, introduced in Section 5. Table 2 shows that the proposed duration model outperforms significantly the baseline alignment. The improved accuracy for 'canonical' lyric lines can be attributed to the increased degree, to which prior duration expectations are met.

## 7. CONCLUSION

In this work we evaluated the behavior of a HMM-based phonetic recognizer for lyrics-to-audio alignment in two settings: with and without utilizing lyrics duration information. Using probabilistic duration-explicit modeling of phonemes for the former setting outperformed the latter on recordings of a cappella Beijing opera. It has incorporated prior expectations of syllable durations, based on knowledge specific for this music genre. In particular, the proposed DHMM aligns remarkably well a selected set of lyrics lines, which comply more precisely with these music-specific principles.

**Acknowledgements** We are thankful to Wanglei from *Doreso* for providing a dictionary of pinyin syllables. This work is partly supported by the European Research Council under the European Union's Seventh Framework Program, as part of the CompMusic project (ERC grant agreement 267583) and partly by the AGAUR research grant.

## 8. REFERENCES

- Anguera, X., Luque, J., & Gracia, C. (2014). Audio-to-text alignment for speech recognition with very limited resources. In *INTERSPEECH*, (pp. 1405–1409).
- Chen, R., Shen, W., Srinivasamurthy, A., & Chordia, P. (2012). Chord recognition using duration-explicit hidden markov models. In *Proceedings of the 13th International Society for Music Information Retrieval Conference*, (pp. 445–450).
- Duanmu, S. (2000). *The Phonology of Standard Chinese*. Clarendon Studies in Criminology. Oxford University Press.
- Dzhambazov, G. & Serra, X. (2015). Modeling of phoneme durations for alignment between polyphonic audio and lyrics. In *Sound and Music Computing Conference*, Maynooth, Ireland.
- Fujihara, H. & Goto, M. (2012). Lyrics-to-audio alignment and its application. *Multimodal Music Processing*, 3, 23–36.
- Fujihara, H., Goto, M., Ogata, J., & Okuno, H. G. (2011). Lyric-synchronizer: Automatic synchronization system between musical audio signals and lyrics. *IEEE Journal of Selected Topics in Signal Processing*, 5(6), 1252–1261.
- Hansen, J. K. (2012). Recognition of phonemes in a-cappella recordings using temporal patterns and mel frequency cepstral coefficients. In *Proceedings of the 9th Sound and Music Computing Conference*, (pp. 494–499), Copenhagen, Denmark.
- Kruspe, A. M. (2014). Keyword spotting in a-cappella singing. In *Proceedings of the 15th International Society for Music Information Retrieval Conference*, (pp. 271–276), Taipei, Taiwan.
- Kruspe, A. M. (2015). Keyword spotting in singing with duration-modeled hmms. In *Signal Processing Conference (EUSIPCO), 2015 23rd European*, (pp. 1291–1295). IEEE.
- Mesaros, A. & Virtanen, T. (2008). Automatic alignment of music audio and lyrics. In *Proceedings of the 11th Int. Conference on Digital Audio Effects (DAFx-08)*.
- Rabiner, L. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2), 257–286.
- Repetto, R. C. & Serra, X. (2014). Creating a corpus of jingju (beijing opera) music and possibilities for melodic analysis. In *Proceedings of the 15th International Society for Music Information Retrieval Conference*, (pp. 313–318).
- Wang, Y., Kan, M.-Y., Nwe, T. L., Shenoy, A., & Yin, J. (2004). Lyrically: automatic synchronization of acoustic musical signals and textual lyrics. In *Proceedings of the 12th annual ACM international conference on Multimedia*, (pp. 212–219). ACM.
- Wichmann, E. (1991). *Listening to theatre: the aural dimension of Beijing opera*. University of Hawaii Press.
- Wong, C. H., Szeto, W. M., & Wong, K. H. (2007). Automatic lyrics alignment for cantonese popular music. *Multimedia Systems*, 12(4-5), 307–323.
- Young, S. J. (1993). *The HTK hidden Markov model toolkit: Design and philosophy*.
- Yu, S.-Z. (2010). Hidden semi-Markov models. *Artificial Intelligence*, 174(2), 215–243.

# Analysis of Tahreer in traditional Iranian singing

Parham Bahadoran

Queen Mary University of London  
p.bahadoran@se15.qmul.ac.uk

## ABSTRACT

Iranian tradition singing is based on a rich musical heritage and contains styles and techniques distinct to the region, which differentiate it from other styles of Middle Eastern singing. In this paper I aim to highlight the specific characteristics of a traditional Iranian vocal technique called Tahreer by analysing its features using computational tools and methods.

## 1. INTRODUCTION

The song of nightingale is regarded as the symbol of musical beauty in Persian/Iranian<sup>1</sup> visual arts, literature and poetry (A'lam, Clinton, 1989). Iranian traditional singing, *āvāz*, is often enriched by a vocal ornamentation called *Tahreer*<sup>2</sup> which is regarded to be inspired by the song of nightingale too. Miller (1999) quotes an Iranian master about different styles of Tahreer and mentions one of the main styles of Tahreer being the *nightingale Tahreer*<sup>3</sup>. Tahreer is a quick alternation between laryngeal mechanisms producing a frequency jump during a very short time interval, typically 50 to 70 ms (Castellengo, 2006; Caton, 1974). During the performance of Tahreer, each consecutive pair of notes of the melody (primary notes) are bridged by a higher pitched note (secondary note) in between with a quick transition. The secondary note is also referred to as Tekiyeh<sup>4</sup> note, which translates to the note on which to lean and in fact a single unit of Tahreer is called Tekiyeh which when performed twice or more becomes Tahreer (Caton, 1974; Fereydooni 2015). Due to the fast nature of the technique, it is perceived as an abrupt break in a continuous melody but the secondary note is not heard. Tahreer is generally used in multiples at the end of singing a phrase or while emphasising a part of a phrase. It is performed on most vowels but typically while uttering an /a/ or /o/ phoneme. Tahreer could be called Iranian yodelling while unlike a yodel melody expanding successively in both the modal register(M1) and the Falsetto(M2), Tahreer melody stays completely in M1 with short ornamental excursions in M2 (Castellengo, 2006; Roubeau, 2007). The current literature on Tahreer is mainly focused on its contextual use in Iranian music as well as high level characteristics with regards to ethnomusicology. What follows in this paper is an analysis of the fundamental building blocks of Tahreer and its different features, to help expose more information about its characteristics at a granular level.

## 2. DATASET, TOOLS AND METHODS

The dataset used consists of 50 excerpts of Avaz from a selection of five renowned Iranian singers of the 20<sup>th</sup> century (1920s-present) to represent all eras of recorded Avaz available. The singers selected also represent a good variety of different schools/Maktabas of avaz, which do vary to a large extent (Simms & Koushkani, 2012). I chose 10 different vocal segments from the repertoire of each singer with the conditions that there must be a good presence of Tahreer in the Avaz, instrumental accompaniment to be minimal and subordinate to the voice and the Avaz to be in its most characteristic free time, non-rhythmic form. These conditions allow for better focus on the Tahreer itself and reduce the effect of other parameters for the purpose of this analysis.

The selected excerpts were annotated using the Sonic Visualizer software (Cannam et al., 2010) which was also used for some analysis. Time constraints of manual calculations in Sonic Visualizer resulted in reduced dataset for parts of the analysis. Some melodic transcriptions were performed using the software Tony (Mauch et al. 2015) and for other aspects of the analysis, MATLAB was used with the whole dataset in the form of audio excerpts as well as exported data from Sonic Visualizer and Tony.

## 3. SPECTRAL, TEMPORAL AND PHONEMIC ANALYSIS

The following section presents findings regarding different characteristics of Tahreer. The aim was to automate as much of the process as possible by creating recognition mechanism to handle the large amount of data. However, in this study some measurements have still been performed by manual annotation and calculation. Some visual evidence based on temporal and spectral views of the data have also been used as evidence, only when the findings have been clear, almost expected and further measurements were not deemed required.

### 3.1 Distinction from Vibrato

The spectrogram of a Tahreer at the first glance looks similar to that of a Vibrato due to the visible oscillation in pitch. There are however major differences between the spectral characteristics of the two techniques. Since Vibrato is also used in Iranian Avaz, I selected instances of

<sup>1</sup> The word Iranian will be used in this text as it relays a broader meaning of the word, which is particularly required when discussing Iranian music as it's not limited to the borders of Iran and shares plenty with neighbouring countries such as Azerbaijan and Iraq.

<sup>2</sup> Tahreer, also spelled Tahrir in other literature translated to any form of ornamentation but in the context of traditional music it refers to this particular technique

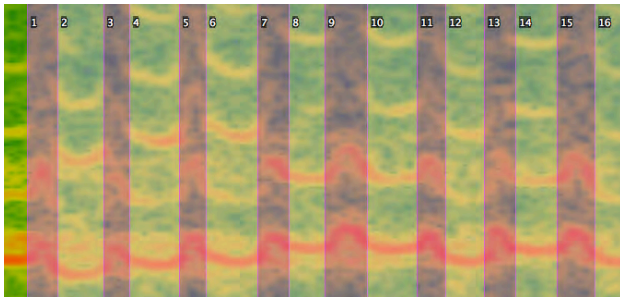
<sup>3</sup> Tahreer-e Bolboli

<sup>4</sup> Tekiyeh/Takiyeh means either to lean or the object/place to lean onto

each technique both from within each piece to ensure similar quality on other musical aspects such as timbre or background music. A visual comparison of the spectrogram of Tahreer and Vibrato excerpts within the same piece demonstrated their distinctions. The most important difference between these techniques is that Tahreer comprises of a transition between a primary dominant note and a secondary higher pitched note while Vibrato is a bidirectional oscillation around the primary note. The pitch rise in Tahreer has larger deviation/step from the main note compared to that of a Vibrato which is typically within 1 semitone in each direction (Hakes, Shipp & Doherty 1988). The sinusoidal shape of a vibrato dictates gradual rise and fall and a sustain on each secondary note while Tahreer reaches the peak of the secondary note and also returns to primary with a sharp rise and fall. Unlike the audible sound on each secondary pitch of a vibrato, the duration of the secondary note of a Tahreer is very short which makes it inaudible. Tahreer and Vibrato are often used back to back in one breath. Tahreer usually ends on a prolonged note which may have Vibrato accompanying its sustain.

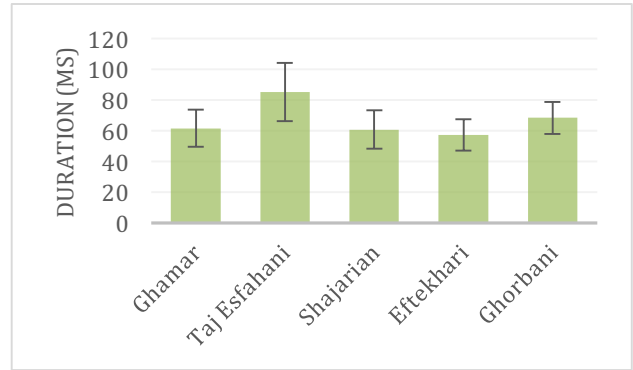
### 3.2 Overall Duration in time

In order to calculate the duration of a single Tahreer or Tekiyeh, the boundaries of a single instance were defined to be between when the pitch of the primary note starts ascending and when the descend back to primary is completed. I then calculated the distance in time for several instances of Tahreer per excerpt. Each instance was marked at its boundaries and durations were calculated in milliseconds.



**Figure 1.** A selection of annotated Tahreer durations

The mean and standard deviation were calculated for all Tahreer instances for each singer independently and also for the overall dataset. The overall mean duration was calculated as 66 ms with standard deviation of 13 ms. One particular singer, Taj Esfahani, was found to have the mean duration at 85 ms and Standard deviation of 19 ms which lifted the overall numbers. The standard deviation shows a range of durations for each singer independently as well which indicates different scenarios affecting the duration. However, the majority of excerpts used are free time and therefore it is not easy to judge how much of the speed of performance is bound by glottal characteristics of the technique, as opposed to traditional or personal stylistic touches.

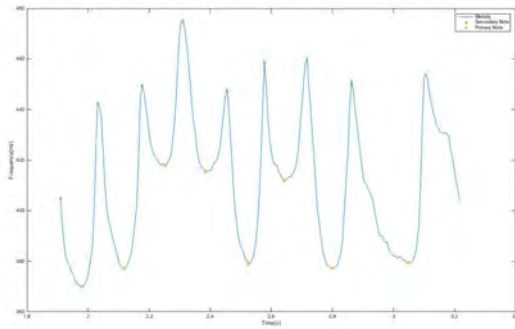


**Chart 1.** Tahreer Duration Mean and SD for each singer

### 3.3 Interval distance

Due to the speed of the transition, the higher (secondary) note is not audible in normal playback speed. An essential part of this analysis was to determine the pitch of the secondary note and find if a relationship exists with the primary note. Using the software Tony, I transcribed the melody line, identified the Tahreer segments and isolated them. The transcribed Tahreer portions were imported into MATLAB and a peak-picking algorithm was used to find the local minima and maxima in the segments. Due to the continuous oscillation between the primary and secondary notes, it was possible to assume every minimum and maximum found (apart from where transcription errors had occurred) within a segment of continuous Tahreer are guaranteed to be the primary and secondary respectively. The algorithm output for each individual instance of Tahreer is the difference between the fundamental frequency of the two notes in Hertz. These intervals were not always equivalent to a discrete number of steps on the chromatic scale and therefore were rounded to the closest number of semitones. The intervals used, range from 2 to 5 semitones overall. All singers but Taj Esfahani, used a 2-3 semitone interval. This number was 4-5 semitones for Him. The actual interval used by each singer and across singers covered the whole range between 2 and 3 semitones. At first this seemed possibly related to the frequent use of quarter-tones in Iranian scales in that the secondary note could itself be a quartertone. However most of the motifs selected for these experiments were not sung in scales or sections of the scales using quartertones and therefore the argument could not musically justify this. It could however be related to the difficulty of landing on a note accurately in the high speed of the transition.

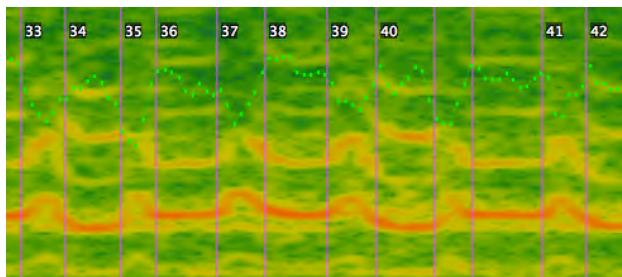
Looking at the findings across all singers allows for making a few musical observations too which conform with instrumental techniques of traditional Iranian music. If a primary note is repeated in the melody, the secondary notes following each would stay the same too. This however does not always hold for the last repeated primary note before a new one is introduced. If the melody is about to descend, the last secondary could be of a lower pitch and if the melody is about to ascend, the secondary could be of a higher pitch.



**Figure 2.** A segment containing 8 instances of Tahreer with identified primary and secondary notes

### 3.4 Change in intensity

During the performance of Tahreer a Tremolo type of effect can be heard which is also easily identifiable by looking at the time-domain representation of a recording. By comparing the time-domain and spectrogram view of the same segment in Sonic Visualiser I found that the amplitude drops correspond to the peak point of each Tahreer. To support this assumption further I calculated the raw power of the signal over time using the Mazurka Power-Curve plugin for Sonic Visualiser (Sapp, 2006). It was revealed that the power has a sharp decrease at the point in time when the secondary note is being voiced. This is due to the intensity drop which is associated with moving from the M1 mechanism to M2 (Henrich et al., 2005).



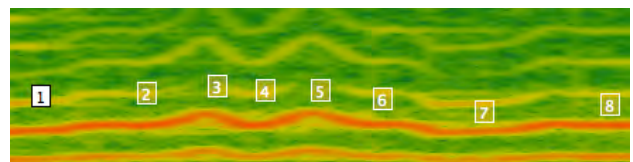
**Figure 3.** The power Curve in bright green laid onto the spectrogram of a segment of Tahreer

### 3.5 The phoneme “h”

In majority of cases the performance of Tahreer is accompanied with an audible phonation of “h” after the secondary note is voiced. For instance, the word “Jaan”, a popular lyrical word used as a base for Tahreer, would be heard as “Jaahaahaahaahan” after using 4 instances of Tahreer. It was not easily possible to capture this added phenomenon in time or frequency domain analysis due to the voiceless nature of the sound. It is however very audible and an important characteristic of Tahreer to the listener.

## 4. COMPARISON WITH OTHER VOCAL ORNAMENTATIONS

In order to get a better perspective about Tahreer and its use in traditional music of Iran, I looked at other singing styles of the region, all of which include some form of similar vocal ornamentation. Pop music styles of Iranian singing which developed in the second half of the 20th century, use a considerable amount of vocal ornamentation. Pop music singing also utilised ornamentations in fast melodic transitions which could be perceived as similar to the traditional Tahreer. Spectral analysis of excerpts from a few popular songs reveal short breaks in the melody similar to that of Tahreer but the transitional secondary note is a lot less visible or non-existent. Many pop music singers are influenced by traditional music of Iran and this influence could have led to the development of the vocal technique used by many pop singers. Arabesque<sup>1</sup> music influenced by Arabic music gained popularity during the 1960s and 1970s in Iran. Arabesque singers used Persian lyrics but the instrumentation, phrasing and ornamentations used resembled that of Arabic pop music. Analysis of some of these examples revealed melodic transitions via a higher note but these transitions are smooth and in a prolonged audible form and include no change of vocal mechanism. Therefore, they are heard as part of the melody unlike the secondary notes of Tahreer. The figure below is a short excerpt from a prominent Egyptian singer, Mohammed Abdel Wahab, singing 8 notes in the short duration of 1.4 seconds. Despite the fast performance of these notes, they are all clearly audible, have a smooth transition between each pair and all sung in the modal M1 register. A detailed analysis of the characteristics of these other techniques was not performed in this study and the findings are limited to evidence from looking at the spectrogram of a handful of excerpts.



**Figure 4.** A short excerpt of an Arabic vocal ornamentation marking the position of 8 consecutive notes

The form of Tahreer discussed in this paper is unique to Iranian, Kurdish and Azeri music and practised mainly in Iran and Azerbaijan (Miller, 1999). The ornamentations used in neighbouring Iraqi and Turkish music are different from Tahreer and a lot closer to the above arabesque examples and what may be more widely regarded as Middle Eastern style of singing.

## 5. DISCUSSION

The limited published work on Tahreer has been focused on high level analysis of this form as a musical ornament

<sup>1</sup> A genre of popular music in Iran known as the “Kucheh Bazari” music during the 1960s-70s which was inspired by popular Arabic music of the time.



and its use with respect to other aspects of music. The primary aim of this research was to depict a better representation of this unique vocal technique by analysing it with respect to the temporal and spectral features of a single instance of Tahreer. Traditional Iranian music is an ancient art form which is still predominantly taught via a direct teacher-student relationship and mostly holds an oral form to this date. Music notation and available literature in this field have not seemed to capture the depth and subtleties of the techniques that would allow for more accessible methods of independent learning without compensating quality and detail. In the case of singing in particular, it is more difficult to refer to written text or sheet music for any form of practical learning. Additionally, the human voice unlike musical instruments, isn't explicitly accessible to allow use of visual aid in training. The findings of this research propose a bottom up approach to learning this particular technique. This method of analysis introduces new forms of transferable knowledge, and provides more accessible ways to learn the techniques and ornamentations of this kind. This approach can directly help preserve the subtle and often complex technicalities of ornamentations which otherwise may disappear in the near future. Furthermore, detailed comparison of Tahreer with other styles of vocal ornamentation clarifies similarities and differences which may not be evident to non-native listeners. These subtle characteristic differences may facilitate ethnomusicology research to expand on the variety, contrast and depth of different musical techniques used in each region, culture or country.

Future goals of this research are to expand on Tahreer analysis in two different aspects. The first goal is to develop automatic segmentation of individual Tahreer excerpts to easily analyse larger quantities of data. The calculation of Tahreer durations in the current study was performed on a reduced number of excerpts due to the manual nature of segmentation. This could be further expanded following the development of automatic segmentation. The second goal is to increase the dimensions of comparison in order to expose differences within various forms of Tahreer itself. This study was focused on characterising Tahreer as a whole and contrasting this form with other styles of singing. However, one interesting finding was the statistical difference in temporal and spectral characteristics of one singer, Taj Esfahani, compared to other singers in the study. An explanation for this could be the characteristics of the school of Avaz he practiced. However, confirming the correlation requires further comparison with other singers of the same school. The dataset used in the current study consisted of one female and four male singers. Adding more female singers to the dataset could help pinpoint potential gender-specific characteristics of Tahreer. Finally, the influence of time and era in which these singers lived is another dimension worth exploring.

## 6. REFERENCES

- A'lam, H., Clinton, J., "nightingale", *Encyclopædia Iranica*, Vol. IV, Fasc. 3-4, pp. 336-338
- Cannam, C., Landone, C., & Sandler, M. (2010). Sonic visualiser. *Proceedings of the International Conference on Multimedia - MM '10*.
- Castellengo, M, J. D. The Iranian tahrir: Acoustical analysis of an ornamental vocal technique. *Cim07*.
- Caton, M. (1974). The Vocal Ornament Takiyah in Persian Music. *UCLA Selected Reports in Ethnomusicology*. 2:1. p. 42-53
- Fereydooni, N. (2015). Fundamentals of instrumental technique - Takiyeh and Tahreer. Retrieved June 12, 2016, from <https://www.youtube.com/watch?v=GQsw9HC6Kqw>
- Hakes, J., Shipp, T., & Doherty, E. T. (1988). Acoustic characteristics of vocal oscillations: Vibrato, exaggerated vibrato, trill, and trillo. *Journal of Voice*, 1(4), 326-331.
- Henrich, N., D'Alessandro, C., Doval, B., & Castellengo, M. (2005). Glottal open quotient in singing: Measurements and correlation with laryngeal mechanisms, vocal intensity, and fundamental frequency. *The Journal of the Acoustical Society of America J. Acoust. Soc. Am.*, 117(3), 1417.
- Miller, L. (1999). *Music and song in Persia: The art of āvāz*. Salt Lake City: University of Utah Press.
- Mauch, M., Cannam, C., Bittner, R., Fazekas, G., Salamon, J., Dai, J., Bello J., & Dixon, S., "Computer-aided Melody Note Transcription Using the Tony Software: Accuracy and Efficiency", in *Proceedings of the First International Conference on Technologies for Music Notation and Representation*, 2015.
- Roubeau, B., Henrich, N., & Castellengo, M. (2009). Laryngeal Vibratory Mechanisms: The Notion of Vocal Register Revisited. *Journal of Voice*, 23(4), 425-438.
- Simms, R., & Koushkani, A. (2012). *The art of āvāz and Mohammad Reza Shajarian: Foundations and contexts*. Lanham: Lexington Books.
- Sapp, C. S. (Ed.). (2006, June). Manpage for SV Mazurka Plugin: MzPowerCurve. Retrieved June 12, 2016, from <http://sv.mazurka.org.uk/MzPowerCurve/>

# SEGMENTATION OF FOLK SONGS WITH A PROBABILISTIC MODEL

**Ciril Bohak, Matija Marolt**

University of Ljubljana,

Faculty of Computer and Information Science

{ciril.bohak,matija.marolt}@fri.uni-lj.si

## 1. INTRODUCTION

Structure is an important aspect of music. Musical structure can be recognized in different musical modalities such as rhythm, melody, harmony or lyrics and plays a crucial role in our appreciation of music.

In recent years many researchers have addressed the problem of music segmentation, mainly for popular and classical music. Some of the more recent approaches are Mauch et al. (2009), Foote (2000), Serrà et al. (2012) and McFee & Ellis (2014). Last three are included in the music structure analysis framework MSAF Nieto & Bello (2015). None of the mentioned approaches however, addresses the specifics of folk music.

While commercial music is performed by professional performers and recorded with professional equipment in suitable recording conditions, this is usually not true for folk music field recordings, which are recorded in everyday environments and contain music performed by amateur performers. Thus, recordings may contain high levels of background noise, equipment induced noise (e.g. hum) and reverb, as well as performer mistakes such as inaccurate pitches, false starts, forgotten melody/lyrics or pitch drift throughout the performance.

One of the most recent approaches which addressed folk music specifics was presented by Müller et al. (2013). The approach was designed for solo singing and was evaluated on a collection of Dutch folk music by Müller et al. (2010).

In our paper, we present a novel folk music segmentation method, which also addresses folk music specifics and is designed to work well with a variety of ensemble types (solo, choir, instrumental and mixtures).

## 2. METHOD

The proposed method processes the input audio recording in several steps and returns a list of segment boundaries. The method assumes that songs consist of similar repetitions of a single part (stanza).

### 2.1 Feature extraction

The method averages the input audio to a single channel and normalizes it. To find repetitions in a melodic/harmonic space, we use harmonic chroma features to represent the contents of recordings, more specifically we use 24-dimensional HPCP features presented in Gómez (2006).

### 2.2 Finding similarity

Our aim is to find segment boundaries that separate repetitions of a segment in a song. We do not know how long individual repetitions are, how many repetitions there are in a song nor how similar they are. To bootstrap the segment finding process, we randomly select a number of 10 second long parts in a song and calculate their distances to the entire song. We use dynamic time warping (DTW) to calculate the distances, as it can tolerate tempo variations well, the technique was already presented by Müller et al. (2009).

Besides rhythm and tempo variations, we also have to take into the account pitch drifting, which occurs when intonation of performers changes upwards or downwards over the course of a song. Ignoring pitch drift would result in inaccurate distance curves and thus poor segmentation. We thus calculate several distance curves for each selected segment, where we shift the intonation of the selected part before calculating the distance. As drifting occurs gradually, we obtain the final distance curve by minimizing distances across all curves, and at the same time restricting the number of intonation changes over the course of a song. An example of an obtained pitch drift curve is presented in Figure 1 (a).

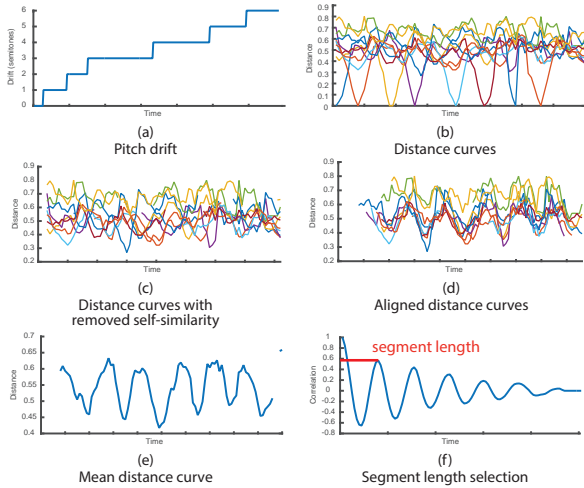
The process results in a series of distance curves, describing the distance of each randomly selected part to the entire song, where tempo and intonation variances are taken into consideration. An example is given in Figure 1 (b). Local minima in these curves represent repetitions of a chosen part in the song. We then remove the self-similar parts of the distance curves, and the resulting curves are shown in Figure 1 (c).

### 2.3 Alignment and length

The set of distance curves (Figure 1 (c)) is not time aligned, since the parts used for their calculations were randomly chosen. To perform alignment, we select a reference distance curve, which is the one that has the highest correlation (is the most similar) to all other curves, thus we may say that it is very representative of the song. Alignment is performed by time-shifting each curve according to its closest local minimum to the part the reference curve was calculated for. From aligned curves we calculate the average distance curve shown in Figure 1 (e).

We also calculate the approximate segment length from

the average distance curve with auto-correlation, as shown in Figure 1 (f).



**Figure 1:** Segmentation steps.

## 2.4 Segmentation

Segmentation is performed with a probabilistic framework similar to hidden Markov models. The model has a state for every possible segment beginning (placed at each second of a song). Segmentation is calculated as an optimal path through the model, defined by state and transition probabilities.

State probabilities are proportional to the likelihood of placing a segment boundary at a certain time. We assume that this likelihood is larger if the boundary is preceded by a region of low-amplitude: for singing, this often corresponds to breathing pauses, while for instrumental music this often corresponds to end of phrases. The longer this region is, the higher is the probability of a segment boundary.

Transition probabilities represent the probability of placing a segment boundary at certain time  $i$  if the previous was located at some other time  $j$ . We consider three restrictions in calculation of transition probabilities: (a) two segments beginning at times  $i$  and  $j$  should be similar; (b) the segments should be separated by approximately the estimated segment length and (c) only forward transitions are allowed.

To find an optimal path through states of this model, we use Viterbi algorithm, whereby we allow the starting state to occur within first 6 seconds of a song and enforce the ending in the last state. The resulting sequence of states represents the set of found segment boundaries, as the states are directly mapped to time.

The detailed description of the method and its individual steps can be found in Bohak & Marolt (2016).

## 3. EVALUATION AND RESULTS

We have evaluated the methods on a collection of folk music from the Ethnomuse presented in Strle & Marolt (2007)

archive and part of the Dutch folk music collection presented in Müller et al. (2010). The EthnoMuse collection consists of different ensemble types: solo singing, two- and three-voice ensembles, choirs, instrumental and mixed singing and instrumental ensembles. We chose 206 songs of different types and recording quality for our collection with a total duration of 534 minutes. The collection was manually annotated, placing segment boundaries with  $\pm 100$  ms precision.

We calculated precision, recall and F1 measure values per song for each ensemble type and for the entire collection. The estimated segment boundary was considered as correct (true positive) if it was located within a  $\pm 3$  second window around an annotated boundary (the same window size as in MIREX evaluations).

The proposed approach significantly outperforms compared methods for non-instrumental music, while for instrumental it is comparable to the best performer. The overall results are presented in Table 1.

Results are also comparable with current state-of-the-art segmentation method for folk music presented in Müller et al. (2013), with an F1 measure of 0.87 on a collection of solo Dutch folk songs - our method achieves an F1 measure of 0.85 on the same collection.

**Table 1:** Evaluation results.

Method	P	R	F1
Mauch et al. (2009)	0.74	0.40	0.4
Foote (2000)	0.39	<b>0.81</b>	0.52
McFee & Ellis (2014)	0.41	0.59	0.48
Serrà et al. (2012)	0.41	0.56	0.47
Proposed method	<b>0.78</b>	0.80	<b>0.76</b>

## 4. CONCLUSION

We presented a novel approach to segmentation of folk music. The method takes into account folk music specifics and significantly outperforms current state-of-the-art segmentation methods for segmenting commercial music and is on par with a state-of-the-art method for solo singing segmentation.

As part of our future work we can envision several improvements of the method, especially for segmentation of instrumental music. We also plan to further specialize the method for better performance with individual ensemble types, by first automatically detecting ensemble type and then choosing an appropriate set of method parameters. We also aim to extend the method for hierarchical musical structure discovery.

## 5. ACKNOWLEDGMENT

This work would not have been done without field recordings provided by the Institute of Ethnomusicology at Research Centre of Slovenian Academy of Sciences and Arts.

## 6. REFERENCES

- Bohak, C. & Marolt, M. (2016). Probabilistic segmentation of folk music recordings. *Mathematical Problems in Engineering*, 2016, Article ID 8297987.
- Foot, J. (2000). Automatic audio segmentation using a measure of audio novelty. In *2000 IEEE International Conference on Multimedia and Expo. ICME2000. Proceedings. Latest Advances in the Fast Changing World of Multimedia (Cat. No.00TH8532)*. IEEE.
- Gómez, E. (2006). *Tonal Description of Music Audio Signals*. PhD thesis, Universitat Pompeu Fabra.
- Mauch, M., Noland, K. C., & Dixon, S. (2009). Using Musical Structure to Enhance Automatic Chord Transcription. In *Proceedings of the 10th International Conference on Music Information Retrieval (ISMIR 2009)*, (pp. 231–236).
- McFee, B. & Ellis, D. P. W. (2014). Analyzing Song Structure With Spectral Clustering. In *Proceedings of 15th International Society for Music Information Retrieval Conference (ISMIR 2014)*, (pp. 405–410).
- Müller, M., Grosche, P., & Wiering, F. (2009). Robust Segmentation and Annotation of Folk Song Recordings. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, (pp. 735–740)., Kobe, Japan.
- Müller, M., Grosche, P., & Wiering, F. (2010). Automated analysis of performance variations in folk song recordings. In *Proceedings of the International Conference on Multimedia Information Retrieval (MIR)*, (pp. 247–256)., Philadelphia, Pennsylvania, USA.
- Müller, M., Jiang, N., & Grosche, P. (2013). A Robust Fitness Measure for Capturing Repetitions in Music Recordings With Applications to Audio Thumbnailing. *IEEE Transactions on Audio, Speech & Language Processing*, 21(3), 531–543.
- Nieto, O. & Bello, J. P. (2015). Msaf: Music structure analysis framework. In *Proceedings of 16th International Society for Music Information Retrieval Conference (ISMIR 2015)*.
- Serrà, J., Müller, M., Grosche, P., & Arcos, J. L. (2012). Unsupervised Detection of Music Boundaries by Time Series Structure Features. In *Twenty-Sixth AAAI Conference on Artificial Intelligence*, (pp. 1613–1619). AAAI Press.
- Strle, G. & Marolt, M. (2007). Conceptualizing the ethnomuse: Application of cidoc crm and frbr. In *Proceedings of CIDOC2007*.