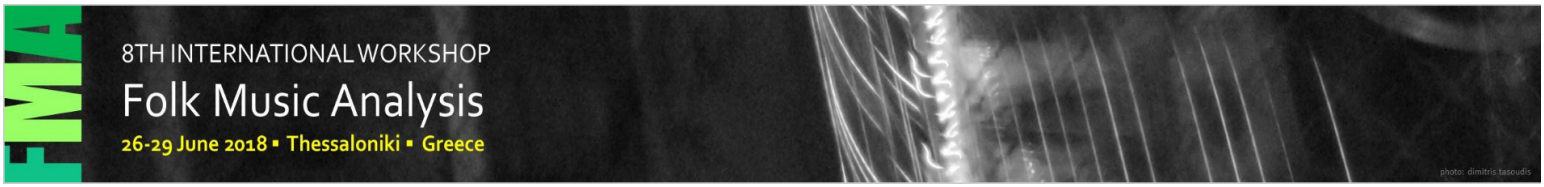
AAWM&FMA

Eighth Folk Music Analysis International Workshop FMA2018

PROCEEDINGS

Hosted by the
School of Music Studies
Aristotle University of Thessaloniki

26-29 June 2018
Pireaus Bank Conference Center, Katouni 12-14, Thessaloniki



Proceedings of the 8th International Workshop on Folk Music Analysis (FMA2018)

26-29 June, 2018

Thessaloniki, Greece

Editors:

Andre Holzapfel & Aggelos Pikrakis

Chairs:

Emilios Cambouropoulos & Costas Tsougras

Title Proceedings of the 8th International Workshop on Folk Music Analysis (FMA2018)

ISBN 978-960-99845-6-0

Editors Andre Holzapfel & Aggelos Pikrakis

Publishers Aristotle University of Thessaloniki, Greece

Program Committee

Chairs

- PIKRAKIS Aggelos (University of Piraeus, Greece)
- HOLZAPFEL Andre (KTH Royal Institute of Technology, Stockholm, Sweden)

Members

- ANAGNOSTOPOULOU Christina (University of Athens, Greece)
- ATHANASOPOULOS George (Aristotle University of Thessaloniki, Greece)
- BEAUGUITTE Pierre (DIT, Dublin, Ireland)
- BENETOS Emmanouil (Queen Mary University, London, UK)
- BOZKURT Baris (Universitat Pompeu Fabra, Spain)
- CAMBOUROPOULOS Emilios (Aristotle University of Thessaloniki, Greece)
- CARROLL David (DIT, Dublin, Ireland)
- CONKLIN Darrell (University of the Basque Country UPV/EHU, Donostia - San Sebastián, Spain)
- DÍAZ-BÁÑEZ Jose-Miguel (University of Seville, Spain)
- DUGGAN Bryan (DIT, Dublin, Ireland)
- KALIAKATSOS-PAPAKOSTAS Maximos (Aristotle University of Thessaloniki, Greece)
- KROHER Nadine (University of Seville, Spain)
- MAROLT Matija (University of Ljubljana, Slovenia)
- PINQUIER Julien (IRIT, Toulouse, France)
- TSOUGRAS Costas (Aristotle University of Thessaloniki, Greece)
- VAN KRANENBURG Peter (Meertens Institute, Amsterdam, The Netherlands)
- VOLK Anja (Utrecht University, The Netherlands)
- WALSHAW Chris (Old Royal Naval College, London, UK)

Organizing Committee

- CAMBOUROPOULOS Emilios, Co-Chair (Aristotle University of Thessaloniki, Greece)
- TSOUGRAS Costas, Co-Chair (Aristotle University of Thessaloniki, Greece)
- KALAITZIDOU Stamatia (Aristotle University of Thessaloniki, Greece)
- ATHANASOPOULOS George (Aristotle University of Thessaloniki, Greece)
- PENINTA Katerina (Aristotle University of Thessaloniki, Greece)

Preface

The 8th International Workshop on Folk Music Analysis was held from 26 to 29 June 2018 in Thessaloniki, Greece. This International Workshop brings together researchers from the fields of ethnomusicology, musicology and music information retrieval (MIR). It provides a forum that encourages sharing of ideas, needs, research methods and discoveries, among ethnomusicologists, musicians, librarians, students, museum curators, computer science experts and music information retrieval researchers. The aim is to foster cross-disciplinary collaborative networks and the development of new interdisciplinary tools and techniques that promote an enriched understanding of traditional musics and the preservation/dissemination of world musical cultural heritage.

We believe that the discussions we had during this workshop were diverse and encouraging, also thanks to the Analytic Approaches to World Music Conference (AAWM) that was hold at the same venue in parallel to the FMA. We see that the year-long effort to continue the FMA resulted in a highly interdisciplinary community between Humanities and Sciences, and we hope to see a continuation of this throughout the next years.

We would like to thank the local team in Thessaloniki, all students and administrative staff who helped us. We would like to thank Emmanouil Benetos for the keynote talk that built bridges to the AAWM community. We would like to thank the musicians who created a great atmosphere during the evenings. And, last but not least, all participants and reviewers who support the FMA.

Thessaloniki, September 2018.

The organizers

Contents

Keynote

Speaker: Emmanouil Benetos
Title: Automatic transcription of world music collections

Emmanouil Benetos is Lecturer and Royal Academy of Engineering Research Fellow at the Centre for Digital Music, Queen Mary University of London (QMUL). He holds a BSc and MSc in Informatics from the Aristotle University of Thessaloniki. After receiving his PhD in Electronic Engineering at QMUL (2012), he joined City, University of London as University Research Fellow (2013-14). His research interests include signal processing and machine learning methods for audio analysis, as well as applications of these methods to music information retrieval, environmental sound analysis, and computational musicology, having authored/co-authored over 80 papers in the aforementioned fields. His ongoing research on automatic music transcription has been highly cited and he was author of top ranking software on the same topic (MIREX 2013, 2015).

Website:http://www.eecs.qmul.ac.uk/∼emmanouilb/

Abstract

Automatic music transcription refers to the process of converting a music recording into some form of human- or machine-readable music notation. It is considered to be a fundamental problem in the field of music information retrieval, with several potential uses in the fields of digital musicology and ethnomusicology. However, it still remains an open problem, especially in the context of polyphonic and heterophonic music. Another challenge facing automatic music transcription methods and music informatics methods in general is the so-called "Western bias": most such computational methods are not directly applicable to music styles outside the purview of Western/Eurogenetic Music. In this talk I will first present the state-of-the art on automatic music transcription, with a focus on world, traditional and folk music. I will illustrate it with our own research on automatic music transcription for specific music styles, including Turkish makam music and Cretan dance tunes. I will describe the challenges regarding modelling, evaluation and adoption of such tools, and on ongoing efforts towards pitch and tuning analysis on a large corpus of audio recordings from the British Library's World & Traditional Music collections. In the final part of the talk I will outline future directions in the intersection between computational ethnomusicology and music information retrieval, and on ways of carrying out mutually beneficial research between the two communities.

Full papers

PLAYER RECOGNITION FOR TRADITIONAL IRISH FLUTE RECORDINGS

Islah Ali-MacLachlan, Carl Southall, Maciej Tomczak, Jason Hockman
DMT Lab, Birmingham City University
islah.ali-maclachlan, carl.southall, maciej.tomczak, jason.hockman
@bcu.ac.uk

ABSTRACT

Irish traditional music (ITM) is a form of folk music that developed alongside dancing over hundreds of years to become an integral part of Irish culture. The wooden flute is widely played in this tradition and mastery in performance is judged by personal stylistic interpretation. Automatic player recognition allows for musicological analysis in an environment where players are individuated based on their interpretation of a common set of melodies. This paper presents two player recognition methods based on convolutional neural networks (CNN). We implement two evaluation contexts for both methods, using a new *ITM-Flute-Style6* dataset alongside our existing *ITM-Flute-79* dataset. The results demonstrate that in both simplified and realistic scenarios, the proposed system is capable of high performance in recognising individual musicians playing melodies with individual stylistic traits that are idiomatic of the genre.

1. INTRODUCTION

Irish traditional music (ITM) is a solo and collective instrumental tradition with roots in social dance music (Vallely, 2011). Playing of the wooden simple system flute in ITM was historically linked to the west and northwest of Ireland (Williams, 2010) and traditional flute players are individuated based on their use of techniques such as ornamentation, phrasing and articulation (McCullough, 1977; Larsen, 2003; Hast & Scott, 2004; Keegan, 2010) alongside idiosyncratic timbral differences (Widholm et al., 2001; Ali-MacLachlan et al., 2013, 2015).

1.1 Related work

Musical genre classification is a widely studied area of music information retrieval (Fu et al., 2011) and an overview, including state of the art techniques, is presented by Sturm (2013). A subset of this field is musician recognition involving the definition of timbral, rhythmic and pitch content. Studies in flute acoustics have found that individual players produce markedly different timbres while changes in manufacturing material make very small spectral differences (Backus, 1964; Coltman, 1971; Widholm et al., 2001). Previous methods of player detection in ITM have used signal processing methods (Ali-MacLachlan et al., 2013, 2015).

Convolutional neural networks (CNN) have been successfully applied not only to image processing, but also to various audio analysis tasks, where the assumption is that auditory events can be recognised by analysing their time-frequency representations. To this end, CNNs provide multiple advantages to the task of musician recognition that other neural network models constitute impractical. The first benefit lies in the shared weights over the input that enable CNNs to process a greater number of features at a lower computational cost. This is achieved by applying the same function (filter) on sub-regions of the input images (spectrograms). This convolution operation is capable of feature translation that preserves the spatial information of the input, and can be used to learn musical features where the target musician's events can appear at any time or occupy any frequency range. CNNs have been implemented successfully with input features derived from spectrograms (Lee et al., 2009) and mel-frequency cepstral coefficients (MFCCs) representing timbre, tempo and key variations (Li et al., 2010). A CNN was trained to perform artist and genre recognition on the Million Song Dataset (Bertin-Mahieux et al., 2011) using segments related to note onsets and feature vectors containing timbre and chroma components (Dieleman et al., 2011). Lidy & Schindler (2016) used CNN to classify genre, mood and composer and achieved the highest results in MIREX 2016 using a 40-band Mel filter. Costa et al. (2017) reinforced the effectiveness of CNNs for music genre classification, performing analysis on three genres: Western, Latin and African music. Individual instrument classification is discussed in Park & Lee (2015) and as part of an ensemble in Han et al. (2017).

1.2 Motivation

In order to determine stylistic differences between players, we must first develop methods to recognise different players in audio signals. Earlier studies in player recognition for flute in ITM have relied upon existing collections of recordings where musicians do not play the same pieces of music (Ali-MacLachlan et al., 2015). We collect and evaluate recordings of six accomplished traditional flute players, all playing music from a predetermined corpus offering a range of typical modes and rhythms.

The use of deep learning, in particular CNN, is an important step in more accurate player identification. In order to identify flute players in ITM, we propose a CNN system in order to make use of their ability to efficiently process large datasets.

The remainder of this paper is structured as follows:

Section 2 details CNNs and their implementation. In Section 3 we discuss the evaluation strategies and the datasets used. Results of the studies into player recognition are presented in Section 4 and finally conclusions and further work are discussed in Section 5.

2. METHOD

We utilise a deep learning model to classify musician-specific features for the task of player recognition. More concretely, first the audio waveforms are pre-processed to create a desired signal representation in a form of MFCCs. Then the signal is split into 5-second segments that represent different timbral and rhythmic characteristics of each performer. Given these characteristics, our model aims to recognise the player of previously unseen audio segments. The next subsections discuss the single blocks of the system in more detail.

2.1 Feature Extraction

First, a downsampled 22.05 kHz 16-bit mono audio signal is split into five second segments. These audio segments offer enough information to capture rhythm patterns as well as result in a large number of observations. A magnitude spectrogram is then calculated using a 1024-sample window size and a resulting frame rate of 100 Hz. Finally, the frequency bins are transformed into 40 MFCCs in a frequency range from 32 Hz to 4,000 Hz.

2.2 Convolutional Neural Network

Figure 1 gives an overview of the implemented CNN architecture. The convolutional layers are constructed using two different building blocks that process the input features: Block A consists of a layer with 10 5x5 filters with 1x5 stride lengths and block B consists of a layer with 20 5x5 filters with 1x1 stride lengths; both are followed by max pooling ((2x2) and (2x5)), dropout layers (Srivastava et al., 2014) and batch normalisation (Ioffe & Szegedy, 2015). A fully connected layer with 100 neurons and a softmax output layer of size c (number of player classes) follows the convolutional blocks. This results in roughly 200,000 total parameters with slight variations depending on c.

2.3 Training

The Adam optimiser is used with a learning rate of 0.003 to train the model. Stochastic gradient descent is performed (batch size = 250) with the cross entropy loss function. Training is stopped when two criteria have been met: 1) 50 epochs have commenced and 2) validation set loss has not increased between epochs. The weights are initialised using a scaled uniform distribution (Sussillo, 2014) and biases are initialised to zero.

3. EVALUATION

Our proposed method of player recognition relies on the accuracy of recordings in representing the playing style of each musician. We implement four evaluation strategies, utilising 5-second audio segments from recordings to assess the ability of the proposed system to recognise different players.

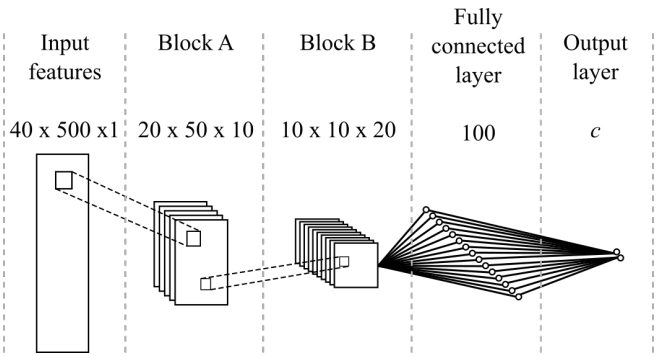


Figure 1: Overview of the proposed CNN. The input data flows between different network layers from left to right. The input data size at each computational block is presented above each layer.

3.1 Datasets

For the purposes of our evaluation we introduce a new *ITM-Flute-Style6* dataset and include a part of a previously used *ITM-Flute-99* dataset (Köküer et al., 2014; Ali-MacLachlan et al., 2015; Jančovič et al., 2015; Ali-MacLachlan et al., 2016, 2017).

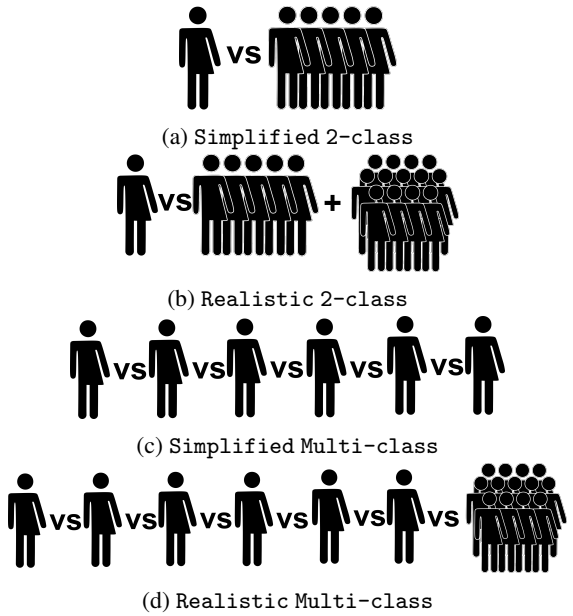
3.1.1 ITM-Flute-Style6

The new dataset consists of 28 recordings by each of 6 players (168 total) and is freely available on Github. [1] All tracks include only flute recordings with no accompaniment. The set covers a range of melodies or *tunes* that are common in the ITM community. The average duration of all tracks is approximately 43 seconds. The total duration of the dataset is 2 hours.

This dataset differs from the existing ITM flute datasets in that it targets multiple player traits and playing contexts that can substantiate further player style research. The presented tune types (i.e., *reels*, *jigs* and *hornpipes*) correspond to an informal online survey conducted among a group of experienced ITM players. The tune names can be seen in Table 1, where the last two represent individually chosen *wild* tracks by each player. The tune type category covers the three most popular tune types in ITM. Five categories are used to structure the dataset by: 1) player, 2) tune name, 3) tune type, 4) timed (*i.e.,* played to metronome) and 5) first or second repeat. All recorded flute players have substantial experience in playing and performing in the style of ITM. The timing category segregates the tracks into timed using a metronome, and untimed. All melodies were recorded twice in segue (first and second repeat) with and without metronome except wild tracks, which were only recorded without metronome.

The recordings were collected as 16-bit/44.1kHz WAV files using a Thomann MM-1 measurement microphone

[1] https://github.com/izzymaclachlan/datasets



(a) `Simplified 2-class`

(b) `Realistic 2-class`

(c) `Simplified Multi-class`

(d) `Realistic Multi-class`

Figure 2: Four evaluation strategies consisting of two player recognition approaches (`2-class` and `Multi-class`) and two contexts (`Simplified` and `Realistic`).

No.	Tune Title	Type	Scale	Ends on
1	Maids of Mount Cisco	Reel	G	Ray
2	The Banshee	Reel	G	Soh
3	Cooley's Reel	Reel	G	Lah
4	Banish Misfortune	Jig	G	Doh
5	Morrison's Jig	Jig	D	Ray
6	The Home Ruler	Hornpipe	D	Doh
7	Players choice 1	Wild	n/a	n/a
8	Players choice 2	Wild	n/a	n/a

Table 1: Corpus recorded by all players detailing tune type, scale and ending note.

connected to an Audient ID14 audio interface. The microphone was positioned above the middle of the flute in order to minimise wind noise caused by blowing.

3.1.2 ITM-Flute-99

ITM-Flute-99, includes 79 released recordings of 9 professional players detailed in Ali-MacLachlan et al. (2016). The remaining 20 recordings belong to a set of tutorial files by Larsen (2003) and were discarded due to the recordings being developed for teaching purposes rather than a true representation of the player. In our evaluation we treat the *ITM-Flute-99*, from now referred to as *ITM-Flute-79*, as representative of professionally played and recorded ITM flute performances.

3.1.3 Dataset experimental setup

The audio segments, described in section 2.1, are used to evaluate our player recognition accuracy. There are a total of 1885 segments comprising of 1438 from the *ITM-Flute-Style6* and 447 from *ITM-Flute-79*.

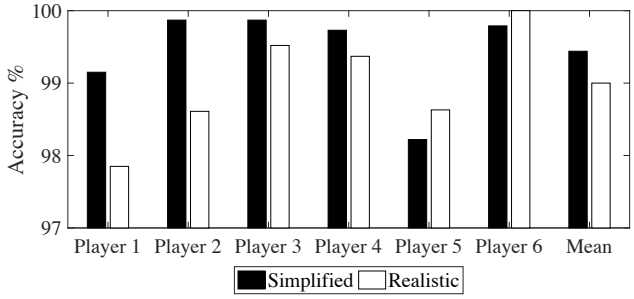


Figure 3: The `2-class` individual and mean player accuracies for both `Simplified` and `Realistic` contexts.

3.2 Evaluation Strategies

We implement four different evaluation strategies, consisting of two player recognition approaches (`2-class` and `Multi-class`) and two contexts (`Simplified` and `Realistic`) to test the system performance. An overview of the strategies is given in Figure 2. It is expected that the `Realistic` context will return lower accuracies as the larger dataset is representative of a wider range of player styles.

3.2.1 `Simplified 2-class`

In the first evaluation strategy, termed `Simplified 2-class` we use a `2-class` approach (using a softmax output layer with 2 neurons, *c=2*) to identify a single player from a mixed corpus. The first class (1) corresponds to the observed player and the second class (0) represents all other players. Due to there being only 6 players, the difference in total class observations should not cause significant bias during training. We test the `2-class` approach in a `Simplified` case using just the 6 player data from *ITM-Flute-Style6*. All recording variations of six tracks are used for training and all recording variations of the other two tracks are split evenly into validation and test sets. Four fold cross validation is performed so that each track appears in the test set. This is repeated 6 times with the player class corresponding to a different player each time.

		Pred	
		P	**O**
GT	**P**	99.7	0.6
	O	0.3	99.4

		Pred	
		P	**O**
GT	**P**	97.5	0.8
	O	2.5	99.2

Table 2: `2-class` confusion matrices where Pred is the predicted class, GT is the ground truth class, P is the player class and O the other class. The `Simplified` context is on the left and the `Realistic` context is on the right.

3.2.2 `Realistic 2-class`

In the second evaluation strategy, termed `Realistic 2-class`, we use the same `2-class` approach as Section

	1	2
M	99.7	100
N	100	100

	1	2
M	95.0	96.7
N	97.3	98.8

Table 3: `2-class` subgroup accuracies where M is metronome, N is no metronome, 1 is first repeat and 2 is second repeat. The `Simplified` context is on the left and the `Realistic` context is on the right.

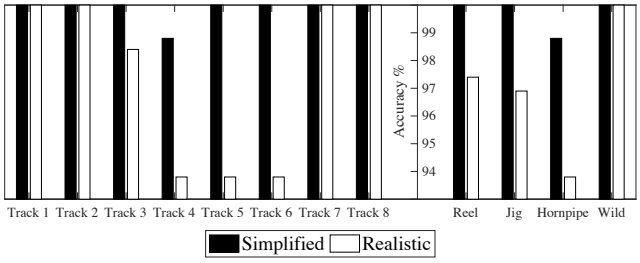


Figure 4: `2-class` results per track and the mean for the different track types. The `Simplified` context is on the left and the `Realistic` context is on the right.

3.2.1 but include *ITM-Flute-79* to create a more realistic evaluation. We expect that the inclusion of the additional data will reduce performance. All tracks from the added dataset are labelled as the other player class (0) and the dataset is divided by track into training, validation and testing respectively 75%, 12.5% and 12.5%.

3.2.3 `Simplified Multi-class`

In the third evaluation strategy, termed `Simplified Multi-class`, we aim to be able to classify an audio segment as one of multiple players. To do this a separate class is used for each of the 6 players (1,2,3,4,5,6) using a softmax output layer with 6 neurons (c=*6*). We then test the `Multi-class` approach using the same evaluation methodology as the `Simplified` context used in Section 3.2.1.

3.2.4 `Realistic Multi-class`

In the final evaluation strategy, termed `Realistic Multi-class` we test the `Multi-class` approach in a more realistic situation using the same two datasets and evaluation methodology from Section 3.2.2. However, all of audio segments from *ITM-Flute-79* are given their own new label (7).

4. RESULTS AND DISCUSSION

4.1 `2-class`

Figure 3 presents the results of each player and the mean across players for both `2-class` evaluation strategies (Figure 2(a) and 2(b)). As expected, a higher mean player accuracy is achieved in the `Simplified` context than the `Realistic` context. Player 5 achieves the lowest classification accuracy whereas player 6 achieves the highest. This could be due to player 6 having a much harder *tone* and higher harmonic energies whereas player 5 has a softer tone similar to the other four players.

A confusion matrix for the mean player results of the two `2-class` contexts are presented in Table 2. For the `Simplified` context (Figure 2(a)) there is approximately the same amount of misclassified player segments as there are misclassified other player segments. In the `Realistic` context there is a significantly higher amount of misclassified player segments and the greatest decrease in performance occurs for player 1 and player 2 (Figure 3), because these players show less individual stylistic traits like use of ornamentation or changes in timbre.

Table 3 presents the `2-class` dataset category distributions of correctly classified audio segments. For both contexts, first repeat with metronome (top left) achieves the lowest accuracy and second repeat without metronome (bottom right) achieves the highest accuracy. This makes sense as an artist is generally more reserved when they are trying to stay in time with a metronome or are playing a melody for the first time.

Figure 4 presents the percentages of the correctly classified player segments (same as Table 3) for each of the tracks. Also presented are the mean accuracies for the track types (Figure 4). The highest accuracies are achieved on the wild tracks. This could be due to the fact that the wild tracks are chosen by the players and suit their preferred playing technique. The lowest accuracies were achieved in the Jig and Hornpipe tracks (Tracks 4, 5 and 6) and they also see the largest decrease in accuracy between the `Simplified` and `Realistic` contexts. Reels are the most common tunes in ITM. As jigs and hornpipes are less common, musicians may be less comfortable playing this type of melody.

4.2 `Multi-class`

Table 4 presents confusion matrices for the `Multi-class` evaluation strategies. Again, as expected, a higher mean accuracy is achieved in the `Simplified` context than the `Realistic` context with 99.6% and 96.6% achieved respectively. While both approaches achieve a similar accuracy in the `Simplified` context the `Multi-class` approach achieves a lower accuracy than the `2-class` approach in the `Realistic` context. In order to recognise the work of a single player, the `2-class` approach is more accurate. As in the `2-class` evaluations, the highest overall accuracy is achieved when recognising player 6 and the lowest when recognising player 1. This is likely due to player 6 having a more individual style. The majority of the errors within the `Realistic` context are misclassified other players. Again, the few examples that are misclassified are of players with similar playing characteristics like timbre and amount of ornamentation.

In this study high accuracies are achieved suggesting that individual players can be recognised using spectral features, however the data consists of only a small number of flute players. It is expected that extending the *ITM-Flute-Style6* dataset would result in lower accuracies

		Prediction					
		1	**2**	**3**	**4**	**5**	**6**
Ground Truth	**1**	98.8	0	0	0	0	0
	2	0	100	0	0	0	0
	3	0	0	100	0	0.6	0
	4	0	0	0	100	0.9	0
	5	1.2	0	0	0	98.5	0
	6	0	0	0	0	0	100

		Prediction						
		1	**2**	**3**	**4**	**5**	**6**	**O**
Ground Truth	**1**	96	0.8	0	1.7	0	0	0
	2	0.7	97.8	0	0	0	0	0.8
	3	0.8	0	97.9	0	0.6	0	3.4
	4	0	0	0	98.3	0	0	2.3
	5	2.5	0	2.1	0	98.8	0	0.4
	6	0	0	0	0	0	100	1.3
	O	0	1.4	0	0	0.6	0	91.8

Table 4: `Multi-class` confusion matrices. `Simplified` context is on the left and the `Realistic` context is on the right.

as new players would be similar to those represented by existing recordings.

5. CONCLUSIONS AND FUTURE WORK

In this paper we have presented a flute player recognition method using a CNN trained on five-second solo excerpts. We evaluated the method using four strategies consisting of two approaches (`2-class` and `Multi-class`) and two contexts (`Simplified` and `Realistic`). For single player recognition, results from the evaluation show that the `2-class` method is more efficient. A player can be more easily recognised in second repeats and when not playing to a metronome. This suggests that players are less characteristic and more self-restricting when they play a tune for the first time or to a metronome. The highest accuracies are achieved when musicians play their own choice of melody (*i.e.,* wild tracks).

In future research, we aim to develop single note and ornament classification methods with additional features. We also aim to gain a deeper understanding of how the network is differentiating between stylistic features. We plan to implement other neural network architectures in order to compare the accuracy of different methods. In order to determine whether accuracies decrease with the addition of other players, we plan to extend the *ITM-Flute-Style6* dataset by recording additional flute players. We will follow the same methodology to record other traditional Irish instruments in order to compare stylistic traits across a range of instruments.

6. 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*, (pp. 13–17)., Paris, France.

Ali-MacLachlan, I., Köküer, M., Jančovič, P., Williams, I., & Athwal, C. (2013). Quantifying Timbral Variations in Traditional Irish Flute Playing. In *Proceedings of the 3rd International Workshop on Folk Music Analysis*, (pp. 7–13)., Amsterdam, Netherlands.

Ali-MacLachlan, I., Southall, C., Tomczak, M., & Hockman, J. (2017). Improved onset detection for traditional Irish flute recordings using convolutional neural networks. In *Proceedings of the 7th International Workshop on Folk Music Analysis.*, (pp. 73–79)., Malaga, Spain.

Ali-MacLachlan, I., Tomczak, M., Southall, C., & Hockman, J. (2016). Note, cut and strike detection for traditional Irish flute recordings. In *Proceedings of the 6th International Workshop on Folk Music Analysis*, Dublin, Ireland.

Backus, J. (1964). Effect of wall material on the steady-state tone quality of woodwind instruments. *The Journal of the Acoustical Society of America*, *36*(10), 1881–1887.

Bertin-Mahieux, T., Ellis, D., Whitman, B., & Lamere, P. (2011). The million song dataset. In *Proceedings of the 12th International Society for Music Information Retrieval Conference*, Miami, FL, USA.

Coltman, J. (1971). Effect of material on flute tone quality. *The Journal of the Acoustical Society of America*, *49*(2B), 520.

Costa, Y., Oliveira, L., & Silla, C. (2017). An evaluation of convolutional neural networks for music classification using spectrograms. *Applied Soft Computing*, *52*, 28–38.

Dieleman, S., Brakel, P., & Schrauwen, B. (2011). Audio-based music classification with a pretrained convolutional network. In *Proceedings of the 12th International Society for Music Information Retrieval Conference*, (pp. 669–674)., Miami, FL, USA.

Fu, Z., Lu, G., Ting, K. M., & Zhang, D. (2011). A Survey of Audio-Based Music Classification and Annotation. *IEEE Transactions on Multimedia*, *13*(2), 303–319.

Han, Y., Kim, J., & Lee, K. (2017). Deep convolutional neural networks for predominant instrument recognition in polyphonic music. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *25*(1), 208–221.

Hast, D. & Scott, S. (2004). *Music in Ireland: Experiencing Music, Expressing Culture*. Oxford, UK: Oxford University Press.

Ioffe, S. & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, (pp. 448–456)., Lille, France.

Jančovič, P., Köküer, M., & Baptiste, W. (2015). Automatic transcription of ornamented Irish traditional music using Hidden Markov Models. In *Proceedings of the 16th International Society for Music Information Retrieval Conference*, (pp. 756–762)., Malaga, Spain.

Keegan, N. (2010). The Parameters of Style in Irish Traditional Music. *Inbhear, Journal of Irish Music and Dance*, *1*(1), 63–96.

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.

Larsen, G. (2003). *The essential guide to Irish flute and tin whistle*. Pacific, Missouri, USA: Mel Bay Publications.

Lee, H., Pham, P., Largman, Y., & Ng, A. (2009). Unsupervised feature learning for audio classification using convolutional deep belief networks. In *Proceedings of the 23rd Annual Conference on Neural Information Processing Systems*, (pp. 1096–1104)., Vancouver, Canada.

Li, T., Chan, A., & Chun, A. (2010). Automatic musical pattern feature extraction using convolutional neural network. In *Proceedings of the International MultiConference of Engineers and Computer Scientists*, (pp. 546–550)., Kowloon, Hong Kong.

Lidy, T. & Schindler, A. (2016). Parallel convolutional neural networks for music genre and mood classification. *MIREX*.

McCullough, L. E. (1977). Style in traditional Irish music. *Ethnomusicology*, *21*(1), 85–97.

Park, T. & Lee, T. (2015). Musical instrument sound classification with deep convolutional neural network using feature fusion approach. *CoRR*, *abs/1512.07370*.

Srivastava, N., Hinton, G. E., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, *15*(1), 1929–1958.

Sturm, B. L. (2013). Classification accuracy is not enough. *Journal of Intelligent Information Systems*, *41*(3), 371–406.

Sussillo, D. (2014). Random walks: Training very deep nonlinear feed-forward networks with smart initialization. *CoRR*, *abs/1412.6558*.

Vallely, F. (2011). *The Companion to Irish Traditional Music*. Cork, Ireland: Cork University Press.

Widholm, G., Linortner, R., Kausel, W., & Bertsch, M. (2001). Silver, gold, platinum-and the sound of the flute. In *Proc. Int. Symposium on Musical Acoustics*, (pp. 277–280)., Perugia, Italy.

Williams, S. (2010). *Irish Traditional Music*. Abingdon, Oxon: Routledge.

RHYTHM INFERENCE FROM AUDIO RECORDINGS OF IRISH TRADITIONAL MUSIC

Pierre Beauguitte
Dublin Institute of Technology
pierre.beauguitte@mydit.ie

Bryan Duggan
Dublin Institute of Technology
bryan.duggan@dit.ie

John D. Kelleher
Dublin Institute of Technology
john.d.kelleher@dit.ie

ABSTRACT

A new method is proposed to infer rhythmic information from audio recordings of Irish traditional tunes. The method relies on the repetitive nature of this musical genre. Low-level spectral features and autocorrelation are used to obtain a low-dimensional representation, on which logistic regression models are trained. Two experiments are conducted to predict rhythmic information at different levels of precision. The method is tested on a collection of *session* recordings, and high accuracy scores are reported.

1. INTRODUCTION

Our goal is to automatically extract rhythmic information from an audio recording of Irish traditional music (ITM). The large majority of the repertoire can be categorized in a small number of tune types, often related to dance forms (Vallely, 2011). Metres used are:

- simple duple: $\frac{4}{4}$ (reel, hornpipe, fling, barndance) and $\frac{2}{4}$ (polka)
- simple triple: $\frac{3}{4}$ (waltz, mazurka)
- compound duple: $\frac{6}{8}$ (jigs) and $\frac{12}{8}$ (slides)
- compound triple: $\frac{9}{8}$ (slip jigs)

Simple and *compound* refer to the beat subdivision, *duple* and *triple* refer to the grouping of beats. No assymetric metres such as $\frac{5}{8}$ or $\frac{7}{8}$ are found in this musical tradition. Rather than focusing on the metre, we are interested in the tune type. Indeed, inferring a $\frac{4}{4}$ metre would not allow us to differentiate between a reel and a hornpipe, although their rhythm is noticeably different, the latter being interpreted with a clear swing. Melodies in ITM have a lot of repetition, and the majority of the notes have the duration of a quaver. This rhythmic stability makes it possible to extract useful information locally, from short excerpts. Slight tempo deviations can occur in live performances, and the use of a short sliding window will allow us to accomodate for these.

The method we introduce in this article first computes an onset detection function, then uses autocorrelation to extract its periodicity. No prior knowledge such as beat location or tempo is required. Rather than relying on handcrafted decision criteria or predefined templates reflecting musical knowledge, we will use a statistical approach to learn decision functions from a novel representation summarizing the autocorrelation function (ACF). As a first step in this study, we will attempt to predict the beat subdivision (*simple* or *compound*). Then, the same method will be used to predict the tune type of an audio recording.

In Section 2 we present some related work on rhythm inference, not restricted to ITM. We introduce the dataset of recordings used in this study in Section 3. Then we present in Section 4 our proposed method. Results are reported and discussed in Section 5. Section 6 contains closing remarks and ideas for future work.

2. RELATED WORK

Brown (1993) is an early example of using autocorrelation to determine the metre of a piece of music from its score. Decision criteria on the ACF are explicitly defined. Also focusing on symbolic music, Toiviainen & Eerola (2006) use discriminant function analysis to predict the metre of folk tunes. Two experiments are conducted, first to distinguish duple and triple metre, then the actual time signature. As stated in Section 1, our first experiment concerns the distinction of simple vs. compound metre instead. Indeed for the musical genre considered here, it is more natural to keep jigs and slip jigs (both compound) in a same category than *e.g.* jigs and polkas (both duple).

Pikrakis et al. (2004) and Fouloulis et al. (2013) determine the metre of Greek traditional music recordings, including assymetric metres, by hand-crafted decision criteria or template matching on the auto similarity matrix.

In Coyle & Gainza (2007), the time signature is also detected using self-similarity matrix, but the method is based on a prior knowledge of the tempo. The method presented in Gouyon & Herrera (2003) relies on beats extracted in a semiautomatic manner, and uses hand-crafted decision criteria to infer the metre. Gainza (2009) and Varewyck et al. (2013) first extract the beats from the raw audio, then determine the metre by analysing inter-beat similarity.

3. DATASET

We will use as our dataset for this study the collection of recordings accompanying the Foinn Seisiún books published by the Comhaltas Ceoltóirí Éireann organisation. They offer good quality, homogeneous examples of the heterophony inherent to an Irish traditional music session, although some solo recordings are also present. Instruments in the recordings include flute, tin whistle, uillean pipes (Irish bagpipes), accordion, concertina, banjo, piano, guitar, bodhran (drum). The whole collection consists of 3

CDs, representing 326 unique recordings. The first 2 CDs (273 tunes) are available under a Creative Commons Licence, while the third is commercially available.

We label each recording by the type of tune played. In most cases the type can be easily identified, but two notable exceptions need mentioning: *Fanny Power* was written as a jig by Turlough O'Carolan ; *Brian Boru's march* is written as a $\frac{6}{8}$ march. However, in the recordings they are arguably played as waltzes, and we decide to label them as such. Two songs are present, both with a duple simple metre. The distribution of tunes per tune type is given in Table 1, as well as the beat subdivision of the metre.

Type	number of tunes	beat subdivision
reel	139	simple
jig	104	compound
polka	28	simple
hornpipe	18	simple
slide	14	compound
barndance	6	simple
waltz	5	simple
mazurka	4	simple
slip jig	3	compound
fling	3	simple
song	2	simple
	205	simple
	121	compound

Table 1: Distribution of tunes per type

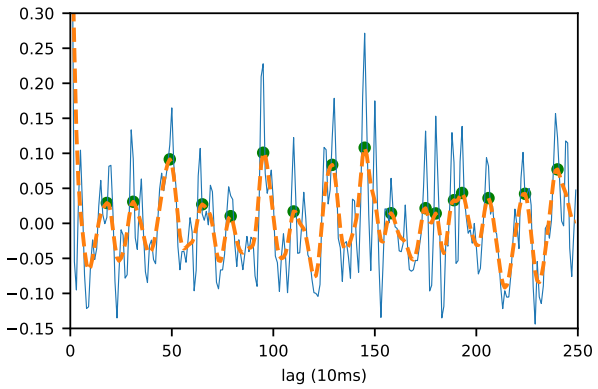


Figure 1: Peak picking of the ACF function. Solid line: ACF function. Dashed line: smoothed function

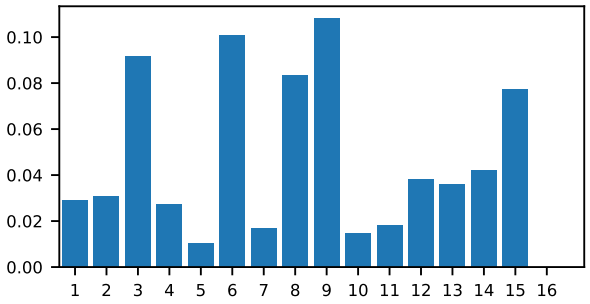


Figure 2: Quantized lag vector

4. METHOD

We now give the details of our proposed approach. First we explain how the audio files are processed to obtain *quantized lag vectors*, then how we train a logistic regression model to infer rhythm features from these vectors.

4.1 Audio processing

The audio files we consider are sampled at 44100Hz. A magnitude spectrogram is generated, with window size of 2048 and step size of 10ms, or 441 samples. This spectogram is then filtered through a filter bank of triangular filters centered at Bark frequencies, resulting in a Bark spectrogram $X_k(t)$ where $1 \leq k \leq 24$ is the Bark index. Following Bello et al. (2005), we obtain an onset detection function by a method of spectral difference:

$$SD(t) = \sum_{k=1}^{24} (H(X_k(t) - X_k(t-1)))^2 \quad \text{for } t > 0$$

where the rectifier $H(x) = (x + |x|)/2$ has the effect of ignoring decreases of energy, because it is equal to zero for negative values. Thus it emphasises onsets more than offsets. As the energy difference is computed in each spectral band before being summed, the presence of percussive instruments is not required to detect onsets.

The autocorrelation functions is then computed on a 5s window of the *SD* function $(w_t) = (SD(t_0+t))_{0 \leq t < N=500}$ (where t_0 is the start of the window) using Pearson correlation coefficient. The autocorrelation for a lag l is:

$$ACF(l) = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y} \text{ where } \begin{cases} X = (w_t)_{0 \leq t < N-l} \\ Y = (w_t)_{l \leq t < N} \end{cases}$$

where cov is the covariance and σ designates standard deviation. We smooth this function by Gaussian filtering with a standard deviation of 20ms, and find the local maxima of this smoothed curve, ignoring the always present peak at $l = 0$. Figure 1 shows and example of the peak picking procedure on a window of a jig. Each peak p has a lag and a value, represented by p_l and p_v respectively. For the goal of this study, what matters is not the actual locations of the peaks p_l, but their relative positions from each other. By abstracting our representation from the actual lag values, we will obtain a form of tempo invariance. The quaver duration will be extracted from the peaks locations and then used to compute a *quantized* representation.

The quaver duration q is determined by the *fuzzy histogram* algorithm, introduced in Duggan (2009), and given in Algorithm 1. The intervals, or lag differences, between the peaks are grouped into bins, allowing for a deviation of a fraction of the bin center, set to $1/3$. The centers of the bins are adjusted for each new interval added. The quaver length is taken as the center of the largest bin.

Knowing the quaver length will now allow us to obtain a tempo-invariant representation of the peaks of the ACF. The following step involves summing peak values.

```
Data: P, list of peaks of the ACF (size l)
Result: quaver length
bins ← {};
max ← 0;
for i ← 1 to l do
    if i = l then
        dur = P[i]_l;
    else
        dur = P[i]_l − P[i − 1]_l;
    end
    found ← false;
    for b in bins do
        bin_start ← b.center∗(1 − 1/3);
        bin_end ← b.center∗(1 + 1/3);
        if dur ≥ bin_start and dur ≤ bin_end then
            found ← true;
            b.center
              ← (b.center∗b.count+dur)/(b.count+1);
            b.count += 1;
            break;
        end
    end
    if found = false then
        newBin.center ← dur;
        newBin.count ← 1;
        bins.add(newBin);
    end
end
for b in bins do
    if b.count > max then
        maxBin ← b;
        max ← b.count;
    end
end
return maxBin.center;
```

Algorithm 1: Fuzzy histogram algorithm, adapted from (Duggan, 2009)

We now introduce the *quantized lag vector* $(ql_i)_{1\leq i\leq 16}$ [1] obtained by first grouping the peaks as:

$$P_i = \{p \in P \text{ where } \textit{round}(p_l/q) = i\}$$

and averaging across these sets:

$$ql_i = \begin{cases} \left(\sum_{p\in P_i} p_v\right)/|P_i| & \text{if } P_i \neq \emptyset \\ 0 & \text{otherwise} \end{cases}$$

An example of such a vector is plotted in Figure 2, computed from the ACF peaks shown of Figure 1. The ratio of the first nine peaks is preserved, but the absolute durations of the lags have been discarded, making this representation tempo-invariant. Some of the subsequent peaks are grouped together by the rounding operations. More prominent peaks appear at multiples of 3, as is to be expected from the compound metre of that tune type (jig).

Each 5s window produces a 16 valued vector, and we slide the window with a step size of 0.5s. Choosing such a small step size results in a large amount of examples, which is an advantage for the machine learning methodology we present in the next section.

[1] The number of 16 quavers was chosen empirically. Experiments with alternative values did not lead to significantly different results.

4.2 Model training

Regression analysis in general attempts at modelling the relationship between independent variables x (here the *ql* vectors) and a dependent variable y (here the rhythm information). We will use logistic regression models, or classifiers, because our dependent variables are categorical, *i.e.* they can only take one of a given set of values. A similar methodology will be used to predict, in a first experiment, the beat subdivision and, in a second one, the tune type.

4.2.1 Experiment A: beat subdivision prediction

The dataset consists of pairs (x, y), where x is a *ql* vector and y a label in $\{$*simple*, *compound*$\}$. We use 10-fold cross validation as a way to evaluate how well the models generalize (Kelleher et al., 2015). Each fold is, in turn, kept as a test set, and a binary classifier is trained on the remaining 9. When preparing the folds, we make sure to keep all *ql* vectors from one tune in only one of the folds. This way, the models will be tested on recordings that have not been used during training, thus avoiding a form of cheating.

To account for the fact that the classes *simple* and *compound* are not balanced in the dataset, during training the error on an instance is weighted by the inverse of the relative frequency of the output class of the instance in the training set; *i.e.*, errors on *compound* instances are given a higher weighting than errors on *simple* instances in the calculation of the loss function to account for the fact that *compound* instances are less frequent.

4.2.2 Experiment B: tune type prediction

In this second experiment, we attempt to predict the tune type from the *ql* vector. Because some tune types are too rare in the dataset, we limit ourselves to the 5 types at the top of Table 1, namely reel, jig, polka, hornpipe and slide. There are only 14 slides in the collection, hence using 10-fold cross validation would only result in one or two of them in each fold. To avoid this problem, we use 4-fold validation instead. For each fold, a multinomial logistic regression classifier is trained in a *one-versus-all* manner, meaning that the model actually consists of a set of binary classifiers. As in experiment A, during the training phase, errors are weighted by the inverse of the relative frequency of the output class.

5. RESULTS AND DISCUSSION

We now report the results of our 2 experiments. Accuracy scores are given for aggregate matrices resulting from the k-fold cross validation methodlogy described above.

The models of both experiments predict a label for a 5s window. In addition to the window-level scores, we are also interested in predictions across a span of several consecutive windows. The reason we are interested in this is that rhythm is not as easily identifiable on all 5s sections of a tune. Thus we hope to reach better accuracy by gathering predictions on a longer segment. The prediction over a span of s windows is simply defined as the most frequent of the s predictions. We report performances at window-level, across s windows, and finally over whole tunes.

	simple	compound
simple	26910	1105
compound	2292	15515
overall (%)	92.2	93.4

Table 2: Aggregate confusion matrix at window-level for experiment A (column: reference, line: prediction)

Type	Accuracy (%)
reel	96.5
jig	95.1
polka	88.6
hornpipe	86.5
slide	79.1
barndance	93.4
waltz	65.4
mazurka	68.2
slip jig	99.2
fling	85.0
song	96.1

Table 3: Window-level accuracy score per tune type for experiment A

5.1 Experiment A

The aggregate confusion matrix resulting from the 10-fold cross validation is given on Table 2. The overall accuracy score at the 5s window level is 92.6%. The prediction accuracy is slightly lower on the *simple* class than on the *compound* class. A possible explanation for this is that there are more distinct tune types included in the *simple* class (reel, hornpipe, polka, waltz,...) than in the *compound* class (only jig, slip jig and slide), as can be seen on Table 1. Looking at the score per tune type on Table 3, we see that it is particularly low for waltz and mazurka, both in simple triple metre $\frac{3}{4}$. Mazurkas are also interpreted with a noticeable swing.

When considering spans of successive overlapping windows, the accuracy increases up to 99.3%, as is shown on Figure 3. We can only compute this up to a span size of 87 windows, corresponding to the duration of the shortest tune in the collection.

Lastly, for each tune, we consider the prediction over the span of all windows of its recording. At this tune-level, the prediction only fails on 3 tunes, all of type *slide*. The overall prediction accuracy is of 99.1%.

Although the task tackled in this first experiment is arguably easy, these near-perfect scores are very encouraging and suggest that our *ql* vector representation does capture some useful rhythmic information.

5.2 Experiment B

The aggregate confusion matrix resulting from the 4-fold cross validation is given on Table 4, and the overall accuracy score at the 5s window level is 82.7%. The accuracy per window span length is shown on Figure 4, and reaches a maximum of 92.9% at $s = 87$.

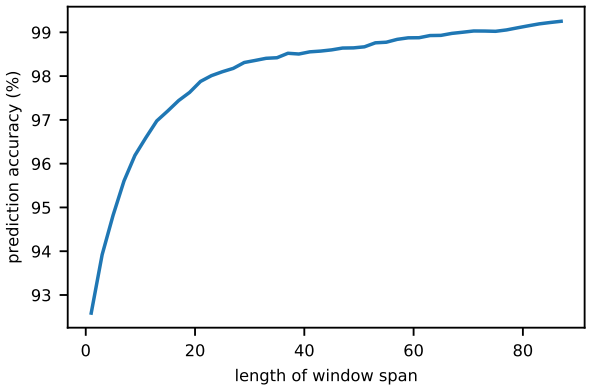


Figure 3: Prediction accuracy by span length for experiment A

	reel	jig	polka	hornpipe	slide
reel	16421	406	617	135	82
jig	296	12577	90	239	1087
polka	339	207	2602	209	200
hornpipe	920	120	198	2537	106
slide	614	1058	115	188	408
overall (%)	88.3	87.5	71.8	76.7	21.7

Table 4: Aggregate confusion matrix at window-level for experiment B (column: reference, line: prediction)

Finally, the confusion matrix for tune-level prediction is given in Table 5. The overall score on slides is low, which is in line with the observation made on tune-level predictions for experiment A. Most of the slides are misclassified as jigs. Both are in duple compound metres, which suggests that the model did manage to capture relevant features, but could not make a good enough distinction between these two tune types. However all 18 hornpipes in the dataset have been correctly classified, despite sharing the $\frac{4}{4}$ time signature with reels. Our method manages to distinguish two tune types having distinct "rhythm signatures" but the same metre.

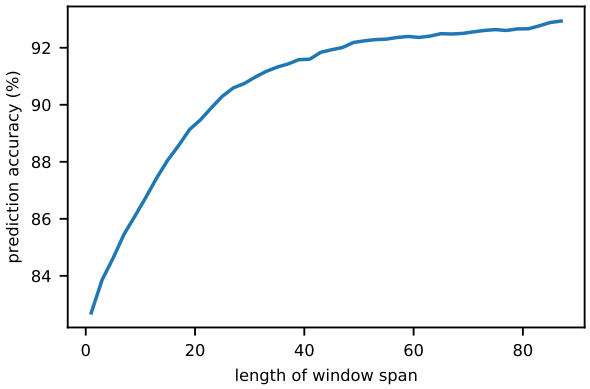


Figure 4: Prediction accuracy by span length for experiment B

	reel	jig	polka	hornpipe	slide
reel	137	0	3	0	0
jig	1	104	0	0	9
polka	0	0	25	0	2
hornpipe	1	0	0	18	0
slide	0	0	0	0	3
overall (%)	98.6	100	89.3	100	21.4

Table 5: Aggregate confusion matrix at tune-level for experiment B (column: reference, line: prediction)

6. CONCLUSION

We introduced a new method for inferring rhythm information from an audio recording, using low-level spectral features and logistic regression classifiers. The performance on the dataset was very good, or even perfect, for some types of tunes (jigs, hornpipes). Other tune types proved to be more challenging (slides), while others were too rare in our collection to be considered.

In future work, we hope to be able to predict more tune types. In order to do so, a larger collection of recordings will have to be used, so more examples can be used to train our models. Testing our models on solo recordings would be useful to further assess the robustness of our proposed approach. Indeed, although our onset detection function relies on spectral content and not on hard onsets from percussive instruments, drums or plucked string instruments (guitar, banjo) are present in most of the recordings in our dataset. Applying our method to flute or fiddle solo recordings could establish to what extent hard onsets help the rhythm inference.

7. REFERENCES

Bello, J., Daudet, L., Abdallah, S., Duxbury, C., Davies, M., & Sandler, M. B. (2005). A Tutorial on Onset Detection in Music Signals. *IEEE Transactions on Speech and Audio Processing*, *13*(5), 1035–1047.

Brown, J. C. (1993). Determination of the meter of musical scores by autocorrelation. *The Journal of the Acoustical Society of America*, *94*(4).

Coyle, E. & Gainza, M. (2007). Time Signature Detection by Using a Multi-Resolution Audio Similarity Matrix. In *Audio Engineering Society Convention 122*.

Duggan, B. (2009). *Machine annotation of traditional Irish dance music*. PhD thesis, Dublin Institute of Technology.

Fouloulis, T., Pikrakis, A., & Cambouropoulos, E. (2013). Traditional asymmetric rhythms: a refined model of meter induction based on asymmetric meter templates. In *Proceedings of the Third International Workshop on Folk Music Analysis*, (pp. 28–32).

Gainza, M. (2009). Automatic musical meter detection. In *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, (pp. 329–332). IEEE.

Gouyon, F. & Herrera, P. (2003). Determination of the Meter of Musical Audio Signals: Seeking Recurrences in Beat Segment Descriptors. In *Proceedings of the 114th Convention of the Audio Engineering Society*, Amsterdam, The Netherlands.

Kelleher, J. D., Mac Namee, B., & D'Arcy, A. (2015). *Fundamentals of machine learning for predictive data analytics: algorithms, worked examples, and case studies*. Cambridge, Massachusetts: The MIT Press.

Pikrakis, A., Antonopoulos, I., & Theodoridis, S. (2004). Music meter and tempo tracking from raw polyphonic audio. In *Proceedings of the 5th International Conference on Music Information Retrieval (ISMIR 2004)*, Barcelona (Spain).

Toiviainen, P. & Eerola, T. (2006). Autocorrelation in meter induction: The role of accent structure. *The Journal of the Acoustical Society of America*, *119*(2), 1164.

Vallely, F. (2011). *The Companion to Irish Traditional Music* (Second edition ed.). Cork: Cork University Press.

Varewyck, M., Martens, J.-P., & Leman, M. (2013). Musical Meter Classification with Beat Synchronous Acoustic Features, DFT-based Metrical Features and Support Vector Machines. *Journal of New Music Research*, *42*(3), 267–282.

AN INFORMATION ETHICS-CENTRED APPROACH TO MUSIC AS INTANGIBLE HERITAGE

Christian Benvenuti
Universidade Federal do Rio Grande do Sul
cbenvenuti@gmail.com

ABSTRACT

This paper discusses ethical principles in the preservation of our increasingly digital musical cultural heritage, particularly in the context of the impact of information and communication technologies. UNESCO's Information for All Programme has recently drawn attention to the digitisation of intangible cultural heritage as a primary safeguarding measure. This paper will focus on the unfolding ethical issues concerning digitisation policies, such as the likely excessive reliance on information and communication technologies, the ethics of the decision-making powers regarding the selection of what musical heritage is worth keeping, the vulnerability of digital depositories, and future ethics-oriented paths.

Keywords: intangible cultural heritage; information ethics; archiving; data storage

1. DIGITISATION POLICIES

Modes of music production, transmission, and preservation are in a constant state of reshaping in a world where music has been increasingly being mediated by technology. In particular, information and communication technologies (ICTs) play a crucial role in this mediation as far as our musical intangible heritage is concerned, encouraging new approaches that bring ethical issues to the foreground.

Information ethics and information preservation were listed as priorities in UNESCO's Information for All Programme (IFAP) (UNESCO, 2017) with great emphasis on:

1. fighting the 'digital divide' through universal access to information and to ICTs;
2. safeguarding measures usually involving digitisation of intangible cultural heritage (ICH);
3. making the internet a safe place for its users.[1]

Information ethics, as proposed by Luciano Floridi, is the radical assumption that *existence* has moral intrinsic worth (Floridi, 2015). This is a non-biocentric approach that replaces *life* with *being*; it is not only about our biosphere, but about our *infosphere* which is the whole ontological kit and caboodle. In this context, any form of negative agency by destroying, corrupting, polluting, and depleting informational objects can be treated as something more fundamental than suffering.[2] This means that anything capable of altering the integrity of information raises questions about its moral implications.

In spite of the fact that we are the moral agents of the infosphere, our understanding of the moral implications, as well as the nature, of something as intangible and fundamental as information is still in its infancy. Information society as a whole expands in a pace exceedingly faster than the growth of our ethical comprehension of it. Nevertheless, the evolving consideration given to our ICH puts the integrity of our cultural memory at the centre of moral concerns. Threats to ICH, as laid out by UNESCO's Convention for the Safeguarding of the Intangible Cultural Heritage (UNESCO, 2003), are generally – and perhaps deliberately – vague, but we can sum up the described threats as including the following:

- Turning traditional forms of art into commodities;
- Contextual and structural distortion due to tourism;
- Social or environmental factors (for example, an economic crisis might impact the amount of time a community devotes to its musical practices while an environmental issue might impact the availability of materials used to make certain musical instruments);
- Traditions fed back into a community as a simplified version due to cultural appropriation or limited notation/transcription procedures.

[1] Other 'information curation' initiatives besides UNESCO's IFAP include the National Digital Stewardship Alliance (NDSA) and the International Internet Preservation Consortium (IIPC).

[2] Floridi suggests the use of the word 'entropy' to describe such 'impoverishment of reality' (2010, p. 103), but with a completely different meaning than the one intended by Claude Shannon in his information theory, which is more akin to complexity, disorder, or unexpectedness.

Traditional music is a notable example of the fragility of cultural heritage. The occasional co-existence between traditional music and their commoditised counterpart might seemingly operate as a symbiotic relationship, in the sense that the actors of such traditions could benefit from an additional source of income. However, commercial pressures tend to promote radical changes of context or content to traditional manifestations, eventually leading them to miss an important piece of their identity. The apparent inevitability of such changes encourages safeguarding measures in the form of the digitisation of musical practices – digital audio/video recording, digitised symbolic representations (MIDI files, notation, transcriptions), and databases. I will refer to these measures under the umbrella name 'digitisation policies'.

1.1 From Intangible to Tangible (and Vice-Versa)

We are currently experiencing an ontological shift. In order to preserve our intangible musical heritage, digitisation policies resulting in very tangible storage media (ranging from vinyl records to hard drives) are considered to be a necessary measure to protect worthy information. Musical information becomes then accessible, portable, and replicable. The next shift is from a musically materialist reality where the tangibility of storage media is taken for granted, relying on physical objects to ensure storage and reproduction, to an informational one, which promotes the illusion that music is independent from a physical support in order to be archived and retrieved, in spite of the fact it is more technology-dependent[3] than ever. This de-physicalizing shift is a core element of current music listening and sharing habits (Benvenuti, 2017), which are characterised by streaming digital music files and their inherent ability to be faithfully copied and accessed from anywhere, provided that some infrastructure requirements are met.

Storage of data still depends on a physical support, whose most significant infrastructure is the one provided by data centres. The increasing need of physical space, energy, and maintenance for such infrastructure is of enormous importance.

2. ICT-DEPENDING SOCIETIES

Only information societies are at risk of informational threats; the more they depend on ICTs, the higher the stakes when it comes to the nourishment and proper preservation of information. Information ethics can be seen as a tripartite approach concerned with information as:

a *resource* which must be accurately accessible within moral limits. In terms of our ICH, this refers, for instance, to ensuring proper access to sound archives but restricting access to personal data when appropriate;

a *product*, which is to be ethically created in order not to be characterised as plagiarism or 'fake news', for instance;

a *target* subject to vandalism (by hacking, for example), social control, and claims to freedom of expression. In terms of our ICH, this can refer, for instance, to filtering content by using (immorally acquired) data about a person's browsing behaviour in order to promote certain music genres or artists. It can also refer to the deliberate destruction of data by acts of state policy or terrorism, for example.

Therefore, since the informational shift has radically transformed the moral context in question, our relationship with technological mediation must be considered critically.

2.1 Relying on ICTs: How Much Is Too Much?

Particularly (but not only) in the state members of the G7 – Canada, France, Germany, Italy, Japan, the United Kingdom, and the United States of America – the Gross Domestic Product (GDP) is mostly comprised of intangible goods – information. In 2016, global expenditure on entertainment and media got a 2.54% share of the global GDP (see Figure 1) ("Global Entertainment and Media Outlook 2017-2021," 2017), of which digital entertainment and media comprise 33.9% (Floridi, 2016).[4] The projected decrease in global expenditure, according to the report, might be due to the increase in digital activities not actually traceable by its methodology. At any rate, this is a higher share than military expenditure in the same year (2016), which was 2.2% (US$ 1.69 trillion) of the global GDP. In the first half of 2017, 184.3 billion streams of on-demand audio were logged in the United States alone (Nielsen Music, 2017).

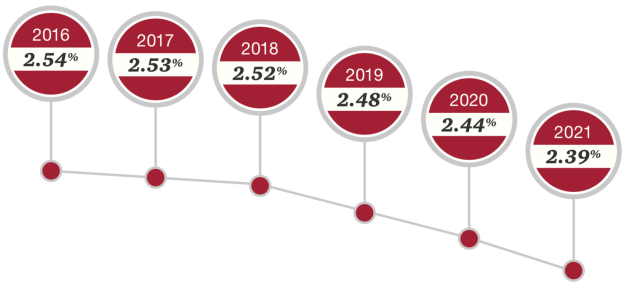


Figure 1 - Global Entertainment and Media Revenue as a Share of Global GDP. Source: Global entertainment and media outlook 2017−2021, PwC, Ovum

[3] By 'technology' I specifically mean ICTs, not just any technology such as the modern piano or the ballpoint pen.

[4] Based on 2010 data.

Streaming is now the dominant revenue source for recorded music. This impressive figure, which does not include music being played via other digital means (offline music files, CDs, DVDs), expresses the profound role of technology-mediated music in our daily lives. It is not surprising that traditional art forms are not immune to the pervasive influence of ICTs in these information societies.

As emphasized by UNESCO:

> *Performances may also be researched, recorded, documented, inventoried and archived. There are countless sound recordings in archives all around the world with many dating back over a century. These older recordings are threatened by deterioration and may be permanently lost unless digitized. The process of digitisation allows documents to be properly identified and inventoried* (UNESCO, n.d.)

This point is important because it assumes that digitisation is free from deterioration and 'entropy'. Digitisation is seen as a safe harbour for information worthy to be kept. Relying too much on digital storage for safeguarding our musical heritage might inadvertently promote some carelessness toward traditional music: once our musical heritage is deemed to be safe, we can (wrongly) come to the conclusion that the focus of our concerns could change to something else, neglecting the need to *continuously ensure the viability of traditional art forms*. However, as Floridi points out, '[t]he gradual informatization of artefacts and of whole (social) environments means that it is becoming difficult to understand what life was like in pre-digital times' (2016, p. 43). By progressively digitising our musical heritage, we are faced with an ethical concern often involving the ontological transformation from intangible (tradition) into tangible (storage): some traditional musical practices might eventually come to exist – albeit in a radically decontextualized, deconstructed way – solely on digital storage media. As we will see below, digital storage – notwithstanding its enormously convenient features – is an archiving method which is eminently unstable.

2.2 The Ethics of Decision-Making

As the infosphere grows, so grows the logical need to select between what to maintain and what to delete. This refers to the basic problem of information theory as a selection problem: one must know what to accept or reject during communication (Benvenuti, 2010, p. 40) and, as it has been pointed out, preservation always involves choices of some objects over others which might eventually disappear (Lundberg, 2015, p. 681). This calls for two questions with far-reaching ethical implications:

1. Who has the decision-making powers to decide what information, within the scope of our intangible musical heritage, is worth keeping and what is not?
2. Who owns – and under which commercial circumstances, if any – the depositories of such information?

Someone has to make decisions, and unlike the information curation initiatives discussed above might suggest, political decision-making is carried out not by states but by individuals invested with such power. The ownership of data centres should, in turn, be given the same ethical consideration. While their role in preserving information is bound to some contractual obligations, the fate of stored data in case of significant changes to the circumstances of their management – such as insolvency – is still uncharted territory. The 'cloud' in cloud computing, after all, is still simply somebody else's computer.

Research and historiography might be important motivators in the preservation of cultural heritage, but this in itself does not prevent ideological agendas from preferring and selecting musical manifestations that are deemed more valuable over others. For example, it is reasonable to suppose that considerable effort will have been spent in the preservation of *samba de roda*, since it has been recognised by UNESCO as Intangible Cultural Heritage of Humanity. On the other hand, it is also reasonable to suppose that Rio de Janeiro's *funk carioca*, an essentially urban Brazilian genre of electronic dance music not currently recognised by any information curation initiative, will not be given similar consideration, regardless of being a cultural manifestation enjoying phenomenal popularity.

2.3 Dependability of the Global ICT Infrastructure

There is evidence that a strong coronal mass ejection (CME) event – a solar superstorm – might be catastrophic for our power grids, basically rendering any electrical device useless (National Research Council, 2008). This would impact not only our computers, mobile phones, and other electronic devices. In most large cities, even the water supply is supported by pumps which depend on electricity (see Figure 2 for a projection of the impact of a strong CME event). A study by the National Academy of Sciences estimates an economic impact of such an event as 20 times greater than the costs inflicted by Hurricane Katrina; restoring the global power grids might take years (Phillips, 2014).

According to recent evaluations, such tragedy might not be as unlikely as one would wish. The probability of occurrence of a catastrophic CME event between 2012 and 2022 was estimated at around 12% (Riley, 2012). This is a chance roughly equivalent to getting the flu in the US sometime along the year (Molinari et al., 2007, p. 5092). A catastrophic CME event actually nearly missed the Earth by a relatively small margin in 2012 (Phillips, 2014).

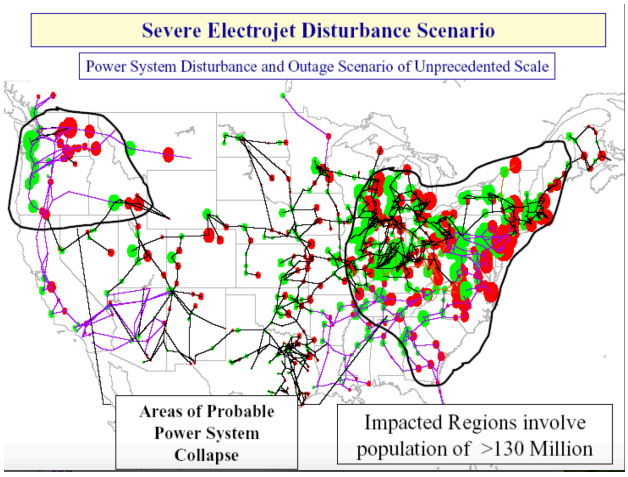


Figure 2 – Estimated Impact of a Solar Superstorm on USA Power Grids. Source: National Research Council, 2008

Seeing that most music in technology-dependent societies is transmitted over digital or electric devices, and that the subjacent infrastructure is sensitive to harmful interference – from unwarranted social control and acts of terrorism to solar storms and coronal mass ejection (National Research Council, 2008)–, the wellbeing of the infosphere does not seem to have been given satisfactory attention. The increasing reliance of our musical culture on the health of an essentially vulnerable infrastructure raises serious ethical issues, at the very least from the perspective of a negligent ecological management of information.

Furthermore, most information in the musical infosphere might be perceived as an intangible cloud asset but is not secured by an intangible cultural heritage status. Consequently, an eventual energy collapse might permanently prevent access to it, hence annihilating one of the most significant musical manifestations of our time. These are ethical implications of the mismanagement of music information which are not being adequately addressed. An information ethics-centred reflection on the imbalance between the expansion of the information society and its ethical roots must be a priority in the context of our musical heritage, at the very least.

3. ALTERNATIVES TO DIGITISATION POLICIES?

Organisations – governmental or otherwise – can play a crucial role in ensuring the viability of traditional art forms. The development of wider audiences and raising public awareness might be the single most important safeguarding measure. This is, in a way, a counter-argument to the notion that the contact of wider audiences with some musical practices might be harmful. It is the equivalent of increasing storage redundancy (where the 'storage' is carried out by individuals). By gaining familiarity with traditional music, audiences can promote its protection and popularity, attracting institutional interest and even a more active role of individuals in participating in the musical manifestations.

Digitising our intangible musical heritage might not be sufficient to ensure the proper transmission of our cultural memory on to future generations. These are unprecedented challenges as our digital environment calls for the development of a more active and empowered ethical framework. This is not to say that digitisation policies are superfluous or not necessary, but safeguarding information requires providing for the worst. As discussed above, the worst might effectively happen, as a catastrophic CME already did in fact happen in 1859 and seriously affected telegraph lines (Cliver & Svalgaard, 2004).

This paper is not advocating a return to a pre-electricity society, not even to a pre-digital society. As informational agents in the 21st century, we have ourselves been reshaped by the infosphere; to remove that aspect from our own selves would be an ontologically radical feat. Many of us might be 'digital immigrants', but many more others are bound to be 'digital natives' even more embedded ecologically in the infosphere. In a *post*-digital era, however, we are left with only our collective musical memory and whatever music is archived in non-electrical media (an archaic gramophone, as primitivist as this example may sound, could still play a record even if humanity were thrown back into the Stone Age).

What this paper advocates is thoughtful examination on the preservation of our intangible musical heritage, which requires ensuring the integrity of knowledge about modes of production, techniques, *lutherie*, and the whole of musical manifestations – not only what is able to be recorded – in an ethical, collectively responsible way.

ACKNOWLEDGEMENTS

This work was supported by the *Coordenação de Aperfeiçoamento de Pessoal de Nível Superior* (CAPES), an agency under the Ministry of Education of Brazil.

4. REFERENCES

Benvenuti, C. (2010). *Sound, noise and entropy : an essay on information theory and music creation*. University of Surrey. https://doi.org/10.13140/RG.2.1.2067.0480

Benvenuti, C. (2017). Craving Information, Drowning in Sound: Ethics and aesthetics of music information. *Submitted*.

Cliver, E. W., & Svalgaard, L. (2004). The 1859 Solar–Terrestrial Disturbance And The Current Limits Of Extreme Space Weather Activity. *Solar Physics*, (224), 407–422.

Floridi, L. (2010). *Information: A very short introduction*. New York: Oxford University Press.

Floridi, L. (2015). *The Ethics of Information*. Oxford University Press.

Floridi, L. (2016). *The Fourth Revolution: How the Infosphere is Reshaping Human Reality*. Oxford University Press.

Global Entertainment and Media Outlook 2017-2021. (2017).

Retrieved from https://www.pwc.com/gx/en/entertainment-media/pdf/outlook-2017-curtain-up.pdf

Lundberg, D. (2015). Archives and Applied Ethnomusicology. In S. Pettan & J. T. Titon (Eds.), *The Oxford Handbook of Applied Ethnomusicology*. New York: Oxford University Press.

Molinari, N.-A. M., Ortega-Sanchez, I. R., Messonnier, M. L., Thompson, W. W., Wortley, P. M., Weintraub, E., & Bridges, C. B. (2007). The annual impact of seasonal influenza in the US: Measuring disease burden and costs. *Vaccine*, *25*, 5086–5096. https://doi.org/10.1016/j.vaccine.2007.03.046

National Research Council. (2008). *Severe Space Weather Events: Understanding societal and economic impacts: A workshop report*. Washington, DC. https://doi.org/https://doi.org/10.17226/12507

Nielsen Music. (2017). *Nielsen Music US 2017 Mid-Year Report*.

Phillips, T. (2014). Near Miss: The solar superstorm of July 2012. Retrieved March 11, 2018, from https://science.nasa.gov/science-news/science-at-nasa/2014/23jul_superstorm/

Riley, P. (2012). On the probability of occurrence of extreme space weather events. *Space Weather*, *10*(2), n/a-n/a. https://doi.org/10.1029/2011SW000734

UNESCO. (n.d.). Performing arts (such as traditional music, dance and theatre).

UNESCO. (2003). Text of the Convention for the Safeguarding of the Intangible Cultural Heritage. Retrieved March 11, 2018, from https://ich.unesco.org/en/convention#art2

UNESCO. (2017). *Information for All Programme*. Paris. Retrieved from http://en.unesco.org/programme/ifap

AUTOMATIC MAKAM RECOGNITION USING CHROMA FEATURES

Emir Demirel
Music Technology Group
Universitate Pompeu Fabra
emir.demirel@upf.edu

Baris Bozkurt
Music Technology Group
Universitate Pompeu Fabra
baris.bozkurt@upf.edu

Xavier Serra
Music Technology Group
Universitate Pompeu Fabra
xavier.serra@upf.edu

ABSTRACT

This work focuses on the automatic makam recognition task for Turkish Makam Music using chroma features. Chroma features are widely used for music identification and tonal recognition tasks such as key estimation or chord recognition. Most of prior work on makam recognition largely rely on use of pitch distributions. Due to the imperfection of automatic pitch extraction for non-monophonic audio, use of chroma features is an alternative that has been showed to be effective in a previous study and we follow the same approach. Our work does not propose a new architecture but rather considers parameter optimization of chroma based recognition for makams. In our tests we use an open-content dataset and perform comparisons with previous studies. As a result of parameter optimization a better performance is achieved. All resources are shared for ensuring reproducibility of the presented results.

1. INTRODUCTION

This study is a continuation of automatic makam recognition studies carried in the CompMusic project (Serra, 2017) and targets improving the performance of chroma feature based automatic makam recognition.

The term '*makam*' mainly refers to a modality system in middle of a continuum defined by a particularized scale and generalized tune on its two poles (Powers & Wiering, 2001). Here we specifically consider the modality system of the Turkish makam music tradition where the following descriptors are considered to be most essential: scale description (involving micro-tonal intervals), overall melodic progression (*seyir*) describing a path from one emphasis note to another until the '*karar*' is reached, preference of specific tri-tetra-penta chords used to form melodies, typical phrases and dynamic range for the melodic contour. For an in-depth review of basic concepts of Turkish makam music and previous computational studies the readers are referred to (Bozkurt, et al., 2014).

Automatic makam recognition can be carried on symbolic or audio data. In (Ünal, et al., 2014), the authors use an n-gram approach for makam detection on symbolic data and report very high accuracies. Makam recognition from audio is a much difficult task due to various characteristics such as heterophony, high variability in interpretations by musicians. The most common used approach in literature (for detection from audio) is the use of pitch distributions (extracted from audio recordings) with a template-matching (or nearest neighbor) strategy (Gedik & Bozkurt, 2010; Karakurt et al., 2016). Pitch histograms have indeed been used as a feature in various automatic recognition tasks since early days of Music Information Retrieval (MIR) (Tzanetakis, et al., 2003). Karakurt, et al. (2016) also presents application of this approach on two other music traditions: Hindustani and Carnatic music (with accuracies 0.92 for 30 ragas and 0.73 for 40 ragas respectively).

Chroma features are frequently used for many tonality related MIR tasks such as chord recognition, tonality detection, audio classification (Dighe et al., 2013; Müller & Ewert, 2011; Jiang et al., 2011) and is a good alternative to pitch distribution features for non-monophonic audio. Chroma based makam recognition has been previously considered by Ioannidis et al. (2011), where the authors follow two distinct approaches for automatic makam classification. First, they apply a makam template-matching method where the templates are constructed from annotated data. Secondly, automatic classification is performed using support vector machines. In our study, we follow a similar approach to the second approach in the aforementioned paper since it shows considerably better performance than template matching approach. We focus on parameter factorization for improving the performance via use of larger window size which reduces noise in features, testing various dimensions for the chroma representation and hyperparameter optimization for the automatic classification stage. We conduct our experiments on the Ottoman-Turkish Makam Music Dataset (Karakurt, et al., 2016), which is the most comprehensive dataset available for computational research on Turkish Makam music. The proposed method is compared with all past approaches using the same set of makams. The performance of our methodology on the same nine makams show that our approach outperforms the prior work of Ioannidis et al. (2011), by more than %10 in overall accuracy and achieves slightly better accuracy scores compared to the state of the art over 20 makams (Karakurt, et al., 2016) which uses pitch histograms as feature.

To sum up, the work presented in this paper provides a bottom-up demonstration of a chroma-based supervised *mode* recognition architecture, and an evaluation method on an open-content dataset for future research on the topic.

2. DATASET

The *Ottoman-Turkish makam recognition dataset* (Karakurt et al.,2016) is the most comprehensive dataset for computational research on makam music, that is open content and is available for researchers. The entire set for the analysis in this study is composed of 997 audio tracks within the OTMM, which are distributed over 20 makams (Table 1). The tonic frequency of each track (available in the dataset) has been obtained and annotated by extracting pitch at the approximate mid-point of the last note in the performance (which has been annotated manually).

Makam Type	#_of_ Tracks	Makam Type	#_of_ Tracks	Makam Type	#_of_ Tracks
Acemaşiran	50	Huzzam	50	Rast	50
Acemkürdi	49	Karcigar	50	Saba	50
Bestenigar	50	Kurdilihicazkar	50	Segah	50
Beyati	49	Mahur	50	Sultaniyegah	50
Hicaz	50	Muhayyer	50	Suzinak	50
Hicazkar	50	Neva	50	Ussak	50
Huseyni	49	Nihavent	50	**Total**	**997**

Table 1. OTMM – Makam Set / number of tracks

Initially, experiments were performed on the entire OTMM dataset. Even though there exist hundreds of variations of makam types, the set of makams in OTMM is representative of this music tradition. In the previous works of Gedik & Bozkurt (2010) and Ioannidis et al. (2011), experiments contained data from only 9 commonly used makams, which are *Hicaz, Huseyni, Huzzam, Kurdilihicazkar, Nihavent, Rast, Saba, Segah, Ussak*. For the second stage of the experiments, the experimental procedures are performed on this makam set (449 tracks) to observe the effects of parameter factorization on HPCP (Harmonic Pitch Class Profiles) features and performance of supervised learning classifiers, in comparison with the work of the aforementioned study.

3. SYSTEM ARCHITECTURE

Our system uses chroma features for automatic classification. The main advantage of using chroma features for this task is that it discards the need of automatic melody extraction of polyphonic audio, which introduces many complexities.

There are several ways to extract chroma features. Most commonly used techniques include either applying spectral analysis on audio frames and quantizing the frame spectrum into frequency bins (Fujishima, 1999), or employing suitable filter banks for the pitch classes (Müller & Ewert, 2011). In our methodology, we use Harmonic Pitch Class Profiles (HPCP), extracted in a similar fashion as explained in the study of Gómez (2006). Figure 1 shows the general structure of the proposed system. The choices of parameters for each step are explained in detail in this section.

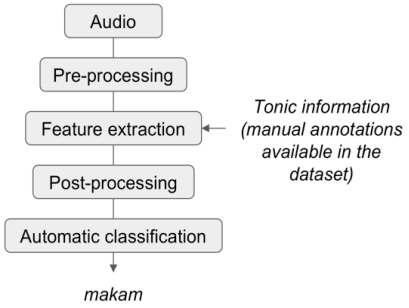


Figure 1. Schematic representation of the system architecture

3.1. Audio Signal Processing

3.1.1. Preprocessing

At the preprocessing stage, DC offset removal is applied on the audio signal using Infinite Impulse Response (IIR) filters. Then, to account for human perception non-linearity, the audio signal is filtered with inverted approximation of equal loudness curves. During our experiments, we have observed that application of these preprocessing steps show some improvement in the robustness of chroma features against transient noise.

3.1.2. Feature Extraction

After filtering out, spectral analysis based on Fourier transform is performed on a frame-based analysis strategy. The frame sizes are chosen as 200ms and hop size as 100ms, which outputs 10 frames per second. As mentioned in Jiang, et al. (2011), larger size windows are preferred over smaller window sizes for mid-level musical information recognition task like chord recognition. Also a smaller window sized window tends to capture more transient noise on the audio signal as opposed to using a larger size window. Besides obtaining chroma features more robust to noise, larger hop size also reduces the computation cost.

For the computation of the chromagram on the frame level, spectral peaks are detected from the local maxima in the frame spectra (Serra & Smith, 1990). The spectral peaks to be detected are limited within the frequency range 100 – 5000 Hz. The spectral peaks are then mapped to the finite number of frequency bins (12, 24, 36,..) with Equation 1 (Gomez, 2006), where n denotes the HPCP bin of to be considered and a_i and f_i denote the linear magnitude and frequency values of the peak i .

$$HPCP(n) = \sum_{i=1}^{nPeaks} w(n, f_i).a_i \qquad (1)$$

$$HPCP: Harmonic\ Pitch\ Class\ Profile$$
$$nPeaks: number\ of\ spectral\ peaks$$

The frequency mapping process require two important considerations to obtain features that are representative of

the musical signal. Initially, a reference frequency must be set for frequency mapping of the frame spectrum. This reference frequency can also be referred as the first bin of the HPCP vector. Moreover, the number of equally separated bins within one octave (or in other words the size of HPCP vectors) needs to be taken into account as a parameter for music traditions that exploit microtonal intervals to construct melodies. One of the main goals of this work is to shed light on the effect of varying sizes of HPCP vectors for the analysis of non-Western music traditions.

3.1.3. Reference Frequency

In chromagram computation, the center frequencies of each bin in HPCP vectors are determined with respect to a reference frequency. The general approach is to estimate the reference frequency by computing spectral peaks with respect to the standardized tuning frequency of 440Hz. Here, we use the manually annotated tonic frequencies of the tracks available in the data set. Frequency mapping of spectral peaks is performed directly with respect to the tonic as the reference frequency.

3.1.4. Normalization:

As explained in detail in (Müller & Ewert, 2011), normalization on a frame basis is necessary at the post-processing step in order to discard the effects of dynamic variations. In our study, we employ l^1 -norm (Equation 2) which corresponds to normalizing elements of the chroma vector x with respect to the sum of all elements of the vector. By doing so, we obtain the chroma histogram representation of each track in the dataset.

$$\|x\|_1 := \left(\sum_{i=1}^{N} |x(i)| \right) \quad (2)$$

3.2 Classification

3.2.1. Feature Set

The initial set of features are the bins of N-bin normalized and averaged HPCP histograms which are computed as the global mean chroma for each track. The normalized global HPCP mean histograms of a musical performance can also be referred as the normal distribution of averaged pitch-classes of a track. It is expected that this distribution would give an insight about the harmonic structure of the piece. In addition to the averaged HPCP vectors, we also include variance related information in our feature set, by simply computing the standard deviation of bins of HPCP vectors separately and globally for each track. As depicted in Figure 2, the standard deviation histogram shows a similar trend with the mean HPCP histogram. This implies that standard deviation also contains some information related to the makam scale and its emphasized tones, which can be used for automatic classification. In our experiments including standard deviation in the feature set, we have observed a considerable increase in the accuracy of automatic classification, which is explained in Section 4.

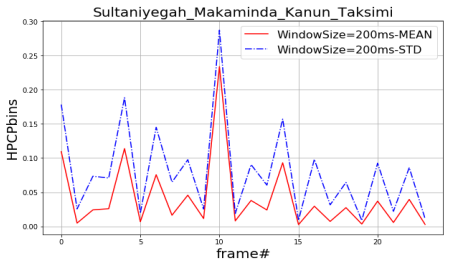


Figure 2. 24-bin - Mean vs. Standard Deviation HPCP histograms

An essential aspect of the makam concept is the fixed 'tonal-spatial' (or tonal-temporal) organization referred as *seyir*. As part of that aspect, the melodic organization and emphasis in the opening of a piece or improvisation plays a crucial role in forming a makam as defined in the theory. To account for more accurate makam estimation, characteristics of the melodic progression also needs to be taken account. It has been previously shown that the first part of the overall melodic contour (i.e. beginning of the performance) carries some discriminative characteristics in Turkish makam music (Bozkurt, 2012; Bayraktarkatal & Öztürk, 2015). To incorporate that, alongside with the global features set, statistical features are also computed locally from certain portions of the beginning of tracks. To determine a more suitable portion of the song, varying percentages of the full track are tested.

3.2.2. Supervised Learning

Automatic classification of the songs according to makam classes are performed using supervised learning. Different combinations of the statistical features set defined above were used to train a support vector machine with radial basis function kernel. The evaluation of each test feature subset are given in Section 4.

The training of support vectors is done with radial basis function kernels. For each test on feature subsets, the hyperparameters of the support vector classifier has to be optimized for each iteration. For training support vector machines with RDF kernel, there are two hyperparameters that need to be tuned to achieve good performance: penalty parameter (regularization constant) C and the kernel coefficient γ. We apply grid search method for hyperparameter optimization which is an exhaustive searching process over a set of defined parameters. For the regularization constant, iterations are done on the following set: C={0.001,0.01,0.1,1,10,100,1000 } for the penalty parameter and γ={0.001,0.01,0.1,1} for gamma. The grid search parameters are validated using 10-fold cross validation on the train set. The parameter combination that gives the best average cross-validated accuracy is used to build the overall model. Then, we test our classifier model, which is trained on the training set, and make pre-

dictions on the test set. To ensure that there is no over-fitting in the results and achieve a high generalization power, the experiments are repeated 10 times over randomized and stratified test & train data splits. In the results, we report the average accuracy and F measure of these experiments. Furthermore, to maintain reproducibility, the random seed is fixed and documented in our shared code. The overall result of predictions with the best performing feature set are shown via confusion matrix in Section 4.

4. RESULTS

The effects of HPCP parameterization on automatic classification with SVMs are tested using stratified 10-fold cross validation. The hyperparameters of the classifier are tuned on the training set using grid search and makam estimations are performed on the testing set which is randomly selected but stratified %10 of the whole set. In order to obtain statistically less biased results, the evaluation pipeline is performed on ten different and randomized train/test splits.

4.1. Experiments on 20 makams:

Automatic classification is performed over 997 songs in 20 makams. In our study we test the performance of using 12, 24, 36 and 48 bins in the HPCP vectors. In addition to the scale of vector size, the effect of different combinations of statistical features in the feature set are tested. Finally, F-measures and accuracy scores are reported. Since the dataset is balanced, weighted macro scores over the dataset are appropriate measures for evaluation.

Table 2 presents resulting F scores of cases with varying number of bins and feature sets which represent the global statistics of HPCP vectors. It is seen that standard deviations of HPCP vectors improve the classifier's performance. This improvement is more significant as the number of bins increase.

F-Measures	**12 - bins**	**24 - bins**	**36 - bins**	**48 - bins**
Mean	0.64	0.64	0.65	0.66
Stdev.	0.65	0.7	0.69	0.7
Mean+Std	0.65	0.7	0.7	0.7

Table 2. F Measure of varying number of bins and feature set combinations (Full track)

As explained in Section 3.2, the makam of the song is generally introduced with emphasis at the beginning of the track. In addition to the above experiments, we also provide a comparison of the global chroma features and chroma features obtained locally from the beginnings of the songs. To determine a good estimation of size of such a region for the beginning of the songs, an iteration over varying portions of the track (from %5 to %40) is performed. (Figure 3) In the figure only the combination of mean and standard deviation features are shown since they outperform the only-mean HPCP features. This analysis has a potential for revealing some future directions for automatic makam recognition task, including a further structural analysis for this tradition.

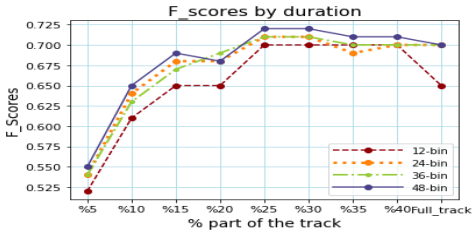


Figure 3: F_scores of classification with local statistical features (Mean + St.dev.)

The results in Table 2 and Figure 3 indicate that higher resolution in the HPCP vector increases the classification scores. Features with vector size of 48 shows slightly better performance than the rest, hence this resolution is set for the following experimental steps. Moreover, the classifier has a better performance when trained with the local chroma histograms instead of global. Regarding statistical features, further investigation is performed to highlight possible directions to improve classification performance. Table 3 shows the evaluation of the combinations of local and global statistical features. In this step, the local features are obtained from the first %30 of the whole track.

Feature_Set (HPCP)	**12-bins**		**48-bins**	
	F Measure	Accuracy	F Measure	Accuracy
Mean(full)	0.64	0.65	0.66	0.67
Std(full)	0.65	0.65	0.67	0.72
Mean(Full)+Std(Full)	0.65	0.66	0.67	0.7
Mean(Local)	0.65	0.65	0.67	0.67
Std(Local)	0.66	0.67	0.73	0.74
Mean(Local)+Std(Local)	0.7	0.71	0.72	0.72
Mean(Full)+Std(Local)	0.71	0.72	0.74	0.74
Mean(Local)+Std(Full)	0.71	0.72	0.74	0.75
Std(Local)+Std(Full)	0.72	0.73	**0.76**	**0.77**

Table 3. Evaluation scores of classifier models with varying feature set combinations, 12-bin vs. 48-bin (Analysis on 20 makams)

With the best performing feature set combination and HPCP parameters, our system is able to score **77%** overall accuracy. These results are comparably better that the current state of the art methodology on the task (Karakurt et al., 2016), where their best performing parameters result in 71.8% accuracy.

In Figure 4, we present the confusion matrix for our system using 48-bin HPCP vectors (mean HPCPs of the

first 30% together with standard deviation of HPCPs of the whole track). Here, the confusion matrix includes the classification instances in the tests over all of the ten randomized sets. We discuss our observations on the confusion matrices in Section 5.

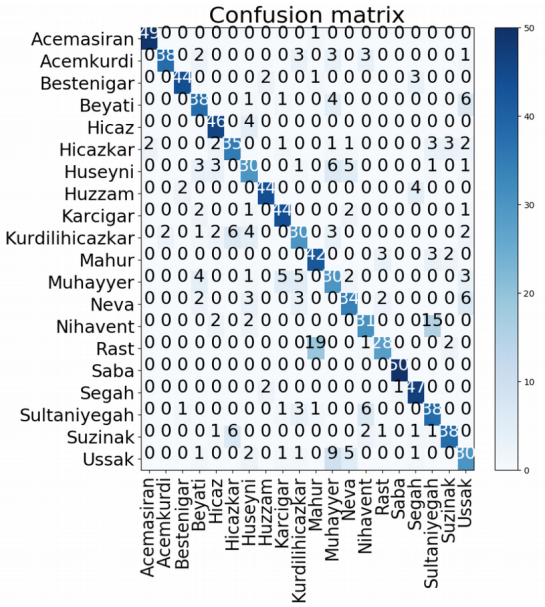


Figure 4. Confusion Matrix of 20 makams analysis

4.2 Experiments on 9 makams:

To have a clear comparison with prior work, we have performed the same experimental procedure over 449 songs in 9 makams. The makam set for this stage of experiments is chosen in consideration with previous research on the topic. (Ioannidis, et al., 2011) In their study, Ioannidis, et al. (2011) use 159-bin HPCP vectors as the feature set which is constructed in parallel with the theoretical knowledge. The experiments in this paper consider bin variations within [12,24,36,48].

Feature Set(HPCP)	**12-bins**		**48-bins**	
	F_Measure	Accuracy	F_Measure	Accuracy
Mean(Full)	0.76	0.77	0.8	0.8
Std(Full)	0.81	0.82	0.82	0.82
Mean(Full)+Std(Full)	0.81	0.81	0.85	0.85
Mean(Local)	0.77	0.77	0.82	0.82
Std(Local)	0.8	0.81	0.86	0.86
Mean(Local)+Std(Local)	0.82	0.82	0.86	0.87
Mean(Full)+Std(Local)	0.82	0.83	**0.89**	**0.89**
Mean(Local)+Std(Full)	0.84	0.84	0.87	0.87
Std(Local)+Std(Full)	0.86	0.86	**0.89**	**0.89**

Table 4. Evaluation scores of classifier models with varying feature set combinations (Analysis on 9 makams)

In order to provide a concise comparison, only the results of the best performing local features of 12-bin and 48-bin feature vectors, in combination with global statistics are shown. (Table 4) The classification scores of proposed methodology shows a robust performance with the classification accuracy of **%89**. This result outperforms the prior works of Ioannidis et al. (2011) where their best performing approach scores an F-measure of %73. Additionally, Table 4 shows the scores for the case where the HPCP vector resolution is 12-bins per octave, which is the standard resolution in MIR. Finally, confusion matrix of the best resulting model for the test with 9 makams is illustrated in Figure 5.

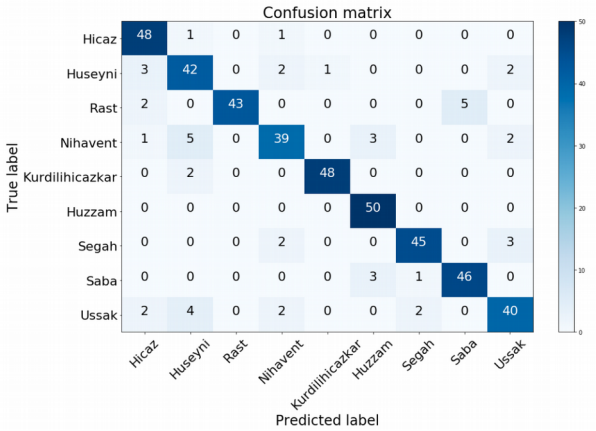


Figure 5. Confusion Matrix of 9 makams analysis

5. DISCUSSIONS

The research presented in this paper explores the significance of parameter selection for extracting chroma features and proposes the use of different statistical features for automatic makam classification. The outperforming results of second stage of the experiments (analysis over 9 makams) is highly due to parameter factorization. Larger size for windows serves better than a smaller size one for automatic modality detection tasks. This observation agrees with the previous work done in Western music traditions (Müller & Ewert, 2011; Jiang, al., 2011). Besides the window sizes, higher HPCP vector resolution has a positive effect on the classification performance. Another important aspect that may have an impact on the performance increase is the hyperparameter optimization for the automatic classifier.

Our study shows that adding various statistical features to the feature set shows a significant improvement for automatic classification as well as the other factorization explained above. Tests over the number of bins reveal that further research is necessary to study the size of chroma vectors when analyzing music from non-Western traditions. The results show better performance for sizes greater than 12. In Figure 3, Table 3 & Table 4, it is observed that the features obtained from the beginning of the tracks result in improvement of classification accuracy. Even though this does not contradict with the the-

ory, further research is necessary on the structure of songs in makam music tradition.

Studying the confusion matrices in Figure 4 and Figure 5, we observe that most of the mis-classifications occur for makams which have very similar or the same scales, like *Mahur* and *Rast*, or *Muhayyer* and *Huseyni*, or *Beyati* and *Ussak,* or *Nihavent* and *Sultanıyegah.* This implies that the classifier learns and extracts musically meaningful information from the chroma features, regarding makams. Moreover, the common confusions in similar scales indicate the necessity of expanding the analysis into more other dimensions, like expanding chroma feature vector into two octaves, since octave equivalency does not hold for certain makams. Further tests are done to observe the effect of smoothing the chroma frames in time, which did not show any improvement in the performance of our system. At the classification stage, we have tried to reduce the dimensionality of our feature space using principal component analysis, which neither showed an increase in the results. Thus detailed discussions related to smoothing and dimensionality reduction are not provided in this study.

6. CONCLUSION

Our approach of automatic makam classification with chroma features sets a baseline for further research on the topic. Moreover, for reproducibility purposes, we share a Jupyter Notebook demonstration of our work[1]. The list of MusicBrainzIDs of the songs in this study can be found in the same repository. The future directions of our research include testing various other chroma features (NLSS Chroma, Chroma Toolbox) on automatic makam classification task and applying structural analysis for segmentation of makams by detecting the harmonic changes in the performance.

8. REFERENCES

Bayraktarkatal, M. E., Öztürk, O. M. (2012). Ezgisel kodların belirlediği bir sistem olarak makam kavramı: Hüseyni makamı'nın incelenmesi. *Porte Akademik*, 3, 24.

Bozkurt B. (2008), An Automatic Pitch Analysis Method for Turkish Maqam Music", *Journal of New Music Research*, 1-13.

Bozkurt B. (2012). Features for analysis of makam music, In P*roceedings of the 2nd CompMusic Workshop*; 12-13, 61-65.

Bozkurt B., R. Ayangil & A. Holzapfel (2014). Computational Analysis of Turkish Makam Music: Review of State-of-the-Art and Challenges. *Journal of New Music Research*, 43(1) pp. 3-23.

Dighe, P., Agrawal, P., Karnick, H., Thota, S. & Raj, B. (2013). Scale independent raga identification using chromagram patterns and swara based features. In *Multimedia and Expo Workshops (ICMEW), IEEE International Conference* on, 1–4.

Gedik, A. C. & Bozkurt, B. (2010). Pitch-frequency histogram based music information retrieval for Turkish music. *Signal Processing*, 90(4), 1049-1063.

Ioannidis, L., Gómez, E. & Herrera, P. (2011). Tonal-based retrieval of Arabic and middle-east music by automatic makam description. In *Proceedings International Workshop on Content-Based Multimedia Indexing*.

Karakurt, A., Sentürk, S. & Serra, X. (2016). MORTY: A Toolbox for Mode Recognition and Tonic Identification. In *Proceedings of the 3rd International Workshop on Digital Libraries for Musicology*.

Fujishima, T. (1999). Realtime Chord Recognition of Musical Sound: A System Using Common Lisp Music. *International Computer Music Conference* (ICMC), 464-467

Gómez, E. (2006). Tonal Description of Music Audio Signals. *The Astronomical Journal*, *35*(5), 220.

Müller, M. & Ewert, S. (2011). Chroma Toolbox: Matlab Implementations for Extracting Variants of Chroma-Based Audio Features. *12th International Society for Music Information Retrieval Conference* (ISMIR).

Jiang, N., Grosche, P., Konz V. & Müller, M. (2011). Analyzing Chroma Feature Types for Automated Chord Recognition In *Proceedings of 42nd AES Conference*

Powers, H. S.& Wiering, F. (2001). et al. Mode. *Grove Music Online, Oxford Music On*line: http://www.oxfordmusiconline.com/subscriber/article/grove/music/43718pg5S

Serra, X. (2017). The computational study of a musical culture through its digital traces. *Acta Musicologica*; 89(1), 24-44.

Serra, X., and Smith, J. (1990). "Spectral Modeling Synthesis: A Sound Analysis/Synthesis Based on a Deterministic plus Stochastic Decomposition. *Computer Music Journal* 14 (4), 12-24.

Tzanetakis, G., Ermolinskyi, A., & Cook, P. (2003). Pitch histograms in audio and symbolic music information retrieval. Journal of New Music Research, 32, 43–152.

E. Ünal, B. Bozkurt, and M. K. Karaosmanoglu. (2014). A Hierarchical Approach to Makam Classification of Turkish Makam Music, Using Symbolic Data. *Journal of New Music Research*, 43(1) 132-146

1 *https://github.com/emirdemirel/Supervised_Mode_Recognition*

SUBIR LA LLAMADA: NEGOTIATING TEMPO AND DYNAMICS IN AFRO–URUGUAYAN CANDOMBE DRUMMING

Luis Jure, Martín Rocamora
Universidad de la República
lj@eumus.edu.uy, rocamora@fing.edu.uy

ABSTRACT

The leader–follower relationship among performers is an important aspect in the studies of interpersonal entrainment in the context of musical performance, specially when analysing the role of leadership in instances of changing tempo and/or dynamics. This research focuses on Uruguayan Candombe, a rich drumming tradition deeply rooted in the Afro–Atlantic culture. The purpose of this paper is to analyse the mechanisms by which Candombe drummers may coordinate and synchronize changes in tempo and dynamics during the performance, specifically at the process called *"subir la llamada"*. Of special interest is the analysis of the cues given by the drummer that leads the rest of the group in the process. Taking one particular recording by three expert Candombe drummers as case study, several computational tools were applied to extract features relevant to the analysis from the audio and video signals.

1. INTRODUCTION

The study of interpersonal entrainment in the context of musical performance is an area of research that has received increased attention in recent times. Its aim is to develop a better understanding of the ways in which groups of musicians coordinate their behaviour during performance (Clayton, 2012). An important aspect is the analysis of the leader–follower relationship among musicians, specially in instances of changing tempo and/or dynamics.

This research was carried out in the context of the Interpersonal Entrainment in Music Performance project, [1] and its purpose is to analyse the mechanisms by which Candombe drummers coordinate and synchronize changes in tempo and dynamics during the performance, specifically at the process called *"subir la llamada"*. Of special interest is the analysis of the cues given by the drummer that leads the rest of the group in the process.

The research is based on a case study, analysing a specific performance with the aid of a set of computational tools. The chosen performance was taken as a complete musical statement by three individual performers, underlining at the same time idiomatic features commonly found in the corpus. The tools used are described in Section 4, and were applied to extract features relevant to the analysis from the audio and video signals (Figure 4).

Thus, this study departs from the corpus analysis approach very common in computational musicology when dealing with non–Western traditional music, based on the statistical analysis of large amounts of data.

[1] https://musicscience.net/projects/iemp/

2. MUSICAL BACKGROUND

Deeply rooted in the Afro–Atlantic culture, Uruguayan Candombe drumming is internationally less known than other Latin American musics of African origin, such as Afro–Cuban or Afro–Brazilian. It possesses however a considerable rhythmic wealth and deserves wider recognition. Its most important and representative manifestation is the *llamada de tambores*, a drum–call parade of a group of drummers (typically between 20 and 60) marching on the street playing the characteristic Candombe rhythm, also called *ritmo de llamada*.

Like in other musics of the Afro–Atlantic tradition, the rhythm of Candombe is clave–based, with a cycle of four beats subdivided in sixteen pulses. The rhythm is the result of the interaction of the patterns of three drums of different size and pitch, called *chico*, *repique* and *piano*. The drum–head is hit with one hand bare and the other holding a stick that is also used to hit the shell when playing the *clave* or *madera* pattern (Figure 5). This timeline pattern is played by all the drums as an introduction to and preparation for the *llamada* rhythm; during the *llamada*, it may be played only by the *repiques* in between phrases (Jure, 2017).

Tempos in Candombe may vary from ca. 100 bpm for a slow *llamada* to around 150 bpm for very fast performances. The most characteristic tempos, however, are in the range of ca. 130 to 136 bpm. It is relatively common to begin the *llamada* at a slower tempo and then increase the speed to reach a typical tempo. After that, minor fluctuations are idiomatic (Figure 6). Essential to this practice is the concept of *"subir* ("raise") *la llamada"*, a term shared and understood by all the members of the community, although not formally defined. This process is primarily associated with an acceleration in tempo, but also involves an increase in dynamics and the use of certain patterns perceived as conveying more energy. The instance of one of the performers giving the cue to begin this process is referred to as *"llamar* ("call") *a subir"*.

2.1 The three drums and their rhythmic patterns

The three drums have different functions in the rhythm and specific patterns associated with their respective registers. The small, high–pitched *chico* drum is the timekeeper, establishing the pulse by repeating a simple one–beat pattern throughout the whole performance (*chico de dos* or *chico liso*). The only possible variant is playing an alternate pattern in sections with a slower tempo (*chico de tres* or *chico*

repicado) (Figure 1). [2]

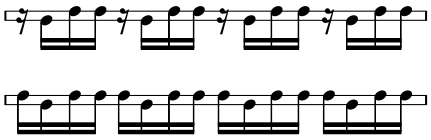


Figure 1: Standard *chico* pattern (top) and a common variant used in slower tempos.

The middle–sized *repique* drum, on the other hand, is regarded as a soloist and improviser, and has the greatest degree of variability among the three drums. During performance, a *repique* player typically interposes cycles of *madera* pattern in between *repique* phrases. These can be characterized by having a higher degree of syncopation and rhythmic and technical complexity. The *repique* has, however, a primary pattern (*repique básico* or *repique corrido*), that may constitute a significant portion of the performance of a *repique* during the *llamada* (Jure, 2013). The short excerpt transcribed in Figure 2 displays these three behaviours.

The piano drum, the largest and lowest sounding of the three drums, has two different functions. The primary one (*piano base*) is to delineate the timeline with characteristic one–cycle patterns. There are many variants, that depend on both the style of each neighbourhood and on the individual style of the performer. But the piano drum can occasionally interpose more ornamented *repique*–like patterns (*piano repicado*), typically one or sometimes two cycles long (Rocamora et al., 2014).

Figure 3 shows the two main *base* and *repicado* patterns found in this recording. They are notated in their basic configuration; [3] during actual performance several subtle variants are introduced by means of added strokes and ghost notes (see also Figure 8).

3. CASE STUDY

The recording taken as a case study in this work is part of the audio–visual dataset of Candombe performances presented in (Rocamora et al., 2015). It features three expert drummers of the same generation, members of families of long–standing tradition in the community of barrio Palermo (Ansina): Héctor Manuel Suárez (b. 1968), Luis Giménez (b. 1969), and Sergio Ortuño (b. 1966). The three are known as accomplished players of the three types of drum, but in this particular take they played *repique*, *chico* and *piano*, respectively (see Figure 5).

The performance was recorded using a multi–track audio system and filmed with a multiple–camera video set–up. The audio set–up provided a stereophonic recording of the ensemble and separate audio channels of each drum—yielding clean direct sound from a given drum, with almost no interference from the others. Therefore, the separate audio channels were used for automatically extracting information of each drum independently. As for the video, only the wide shot of the ensemble was used in this work.

[2] In all the examples, the lower line represents the hand and the upper line the stick, with an X representing the *madera* sound. Parenthesized notes are de–emphasized or ghosted.

[3] The technique of the *piano* drum is more complex and requires some additional symbols: a cross represents a muted note (the hand and/or stick rest on the drum–head after striking it), and a stem without note head means dampening the vibration with the palm without producing a sound. The triangular note head means palming the drum head with the fingers.

4. INFORMATION EXTRACTION

Some computational methods for information extraction are applied to the audio–visual record of the performance, oriented towards capturing and representing the evolution over time of the most relevant aspects noted above, namely tempo, dynamics and rhythmic patterns. For the analysis of dynamics only the *chico* drum is considered, since—given that it always repeats the same one–beat pattern—it allows for a consistent and comparable estimation throughout the whole performance. Two different kinds of information are extracted for this purpose: the root mean square (RMS) value of the audio waveform of the separate track, and the amplitude of the trajectory of the left hand of the performer obtained from the video. In addition, an onset–based asynchrony analysis is carried out, for providing information on interpersonal entrainment and leadership.

4.1 Tempo curve

The evolution of the tempo is computed as the inverse of the difference between two adjacent downbeats (first beat of the four–beat cycle) and expressed in beats per minute (bpm). The downbeats were automatically extracted by using *BayesBeat* (Krebs et al., 2013) trained with the dataset released in (Nunes et al., 2015). The extraction of downbeats was very reliable, yielding an F-measure of 100% when compared to manual annotations using the standard ± 70 ms tolerance, as in (Nunes et al., 2015). The resulting tempo curve is depicted in the second plot of Figure 6.

4.2 Dynamics

The root mean square (RMS) of the audio waveform of the *chico* separate track was computed for consecutive signal frames (using a frame length of 1 second and a hop size of 0.5 seconds) and expressed in decibels. The third plot of Figure 6 shows the RMS values obtained (solid line).

The left hand of the *chico* performer exhibits a cyclic up–down movement in relation to the drum head, that corresponds to the hand stroke at the second subdivision of each beat (in both patterns, see Figure 1). A measure of the extent of the trajectory of the left hand of the *chico* performer is considered as an indirect estimate of the dynamics of the performance. To do that, an existing computer vision system called *OpenPose* was applied, that detects human body, hand, and facial keypoints for multiple persons from single images (Cao et al., 2017). An example of the detections for one video frame can be seen in Figure 4.

The location of the left hand in each video frame is provided as an x–y point in pixels. Then, the locations are further processed using moving maximum and moving minimum filters to estimate the boundaries of the hand move-

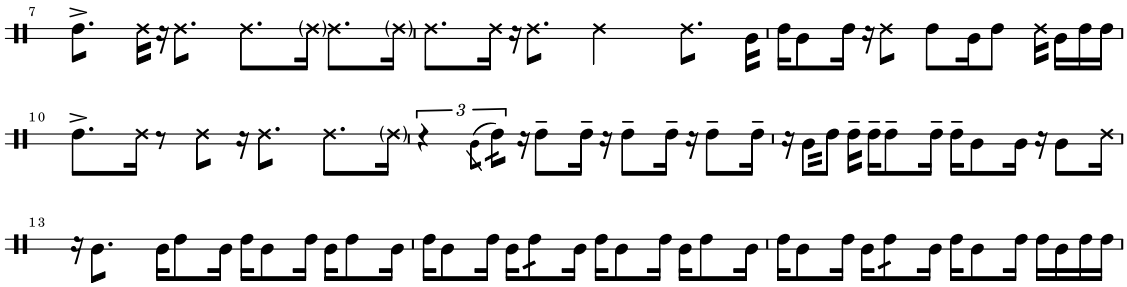


Figure 2: Transcription of mm. 7–15 of the *repique* solo in this performance, displaying the *madera* pattern, complex improvised phrases and the primary *repique* pattern at the end.

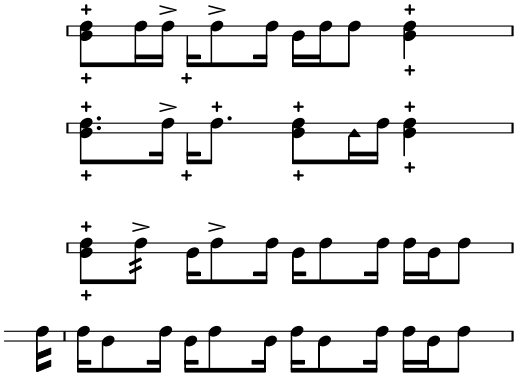


Figure 3: Rhythmic patterns of the *piano* drum in this performance. From top: *base 1*, *base 2*, *repicado 1* and *repicado 2*.

ment within a certain time interval. Finally, the difference between the estimated boundaries is considered as the extent of the trajectory over time. The result of this procedure is represented in the third plot of Figure 6 (dashed lined). It is worth noting that, not surprisingly, the RMS values and the hand motion signal show roughly a similar behaviour. [4]

4.3 Rhythmic pattern analysis

The analysis of rhythmic patterns is based on the spectral flux, a feature extracted from the audio signal that captures changes in the energy content in different frequency bands. The separate audio track of each drum is processed to conduct two different type of analysis: 1) the detection of *madera* rhythm cycles (Rocamora & Biscainho, 2015; Jure & Rocamora, 2017), and 2) the extraction of a feature map of rhythmic patterns (Rocamora et al., 2014).

The spectral flux feature is computed through the Short–Time Fourier Transform of the signal mapped to the MEL scale for sequential 40 ms duration windows in hops of 10 ms. The resulting sequences are time–differentiated and half–wave rectified. The spectral feature is summed across all the MEL bands for onset detection, whereas the first MEL bands (< 1500 Hz) are used for sound classification.

Onset detection is based on a combination of a fixed and an adaptive threshold, as in (Böck et al., 2012). A Support Vector Machine classifier trained on isolated sounds is used to detect *madera* sounds. The proportion of onsets classified as *madera* within a rhythm cycle is used to detect *clave* patterns (Rocamora & Biscainho, 2015).

Then, the spectral feature summed across all the MEL bands is amplitude–normalized and time–quantized to the 16–subdivisions grid using the manual beat/downbeat annotations. A representation in the form of a map of cycle–length rhythmic patterns is straightforwardly obtained by building a matrix whose columns are consecutive feature vectors. Figure 8 depicts the map obtained for the *piano* drum track, where the horizontal axis corresponds to the cycle index and the vertical axis is the subdivision index. The columns of the map virtually correspond to each of the cycle–length rhythmic patterns performed by the *piano* drum along the whole recording. To aid the analysis of their differences and similarities, the rhythmic patterns are clustered using the K–means algorithm and the Euclidean distance, the number of cluster specified as an input parameter. The clusters obtained for the *piano* drum—shown with different colors in Figure 8—match the characteristic rhythmic patterns actually performed. The centroids of the clusters are also depicted in Figure 8 for reference.

A simplified schematic representation of the rhythmic patterns obtained for each drum is provided in the top plot of Figure 6, and will be analysed in Section 5.

4.4 Asynchrony between ensemble parts

An analysis of the asynchrony between onsets by different ensemble parts in the same metric position was carried out, following (Polak et al., 2016; Rocamora et al., 2017). Given the tempo changes of the performance, the onsets timing data was normalized to the four–beat rhythm cycle. To do that, the annotated downbeats were used as an initial reference, which was further refined by estimating downbeats positions as the average of the onsets of the different drums at the beginning of each rhythm cycle. Then, an aggregated histogram of all the onsets was computed, heuristic boundaries were defined between metric positions, and each onset was assigned to its corresponding metric bin. A virtual reference for each subdivision was obtained as the mean of all onsets within each metric bin.

Signed asynchronies were computed for each onset of each drum relative to the virtual reference subdivision. The values of mean and standard deviation of the signed asynchronies, computed for windows of ten consecutive rhythm cycles, are schematically depicted at the bottom of Figure 6.

[4] Pearson correlation coefficient: $r(342) = 0.56$, $p < 0.001$.

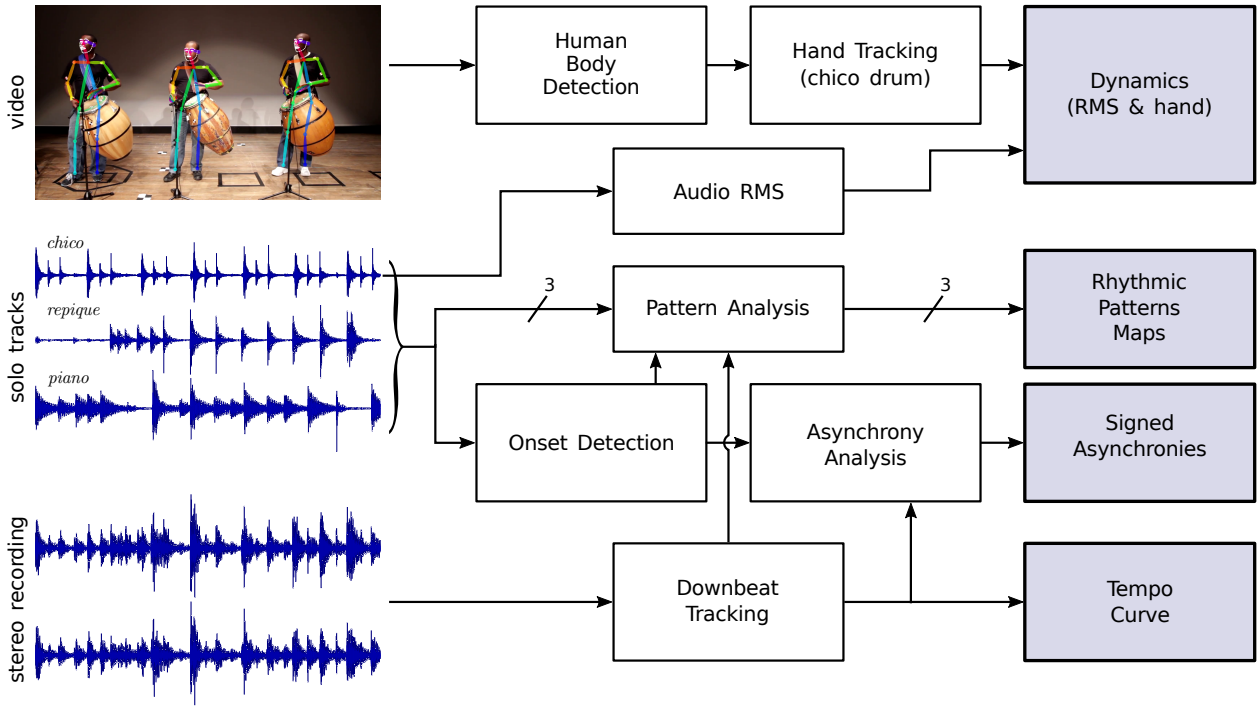


Figure 4: Block diagram of the computational tools applied and the information obtained.

Figure 5: Left to right: Héctor Manuel Suárez, *repique*, Luis Giménez, *chico*, Sergio Ortuño, *piano*.

Despite minor differences, the profile of the averaged asynchronies is similar along the whole performance, showing that the *repique* tends to be before the other two drums. The mean asynchronies obtained for each drum are below 2% of the normalized local beat duration, which corresponds to mean asynchronies in the range 8.2 and 12 ms, depending on the tempo value.

Figure 7 provides a detailed representation of the location of the onsets for the rhythm cycles corresponding to the first tempo increase (8 to 13). Note that the obtained grid of virtual reference subdivisions is not isochronous.

5. ANALYSIS

The recording has a duration of ca. 2:45 min. and comprises 86 complete cycles, ending in the downbeat of cycle 87. After a short introduction (a tremolo of the *repique* and two cycles of *madera* followed by a *repicado* in the *piano*), the *llamada* rhythm begins in the fourth cycle. The tempo curve shows an initial tempo that can be considered slow for Candombe (ca. 105–106), but—in a paradigmatic example of *subida*—there is a notorious increase in tempo beginning at around m. 9–10, reaching a more typical tempo of ca. 128–130 at m. 15, and another (minor) increase around m. 22. This first section of the performance (ca. 50 seconds) was analysed with some detail, revealing some relevant aspects:

a) there is a strong but not linear correspondence between the tempo curve and dynamics of the *chico* drum, with increments in sound level and trajectory related with the increases in tempo (Figure 6, mm. 10–20);

b) the first *subida* is led by the *piano* drum by means of microrhythmical displacements of the notes. While the average asynchrony between the three drums remains approximately constant throughout the whole performance (Figure 6, bottom), Ortuño plays systematically ahead of the *chico* certain groups of notes in mm. 9–11 (Figure 7, compare with mm. 8 and 12–13);

c) the rhythm patterns play a fundamental role: the *piano* calls to raise the rhythm by playing ahead the notes located in specific places in the rhythmic cycle, the *repique* reinforces the raise by playing *repicado corrido*, and the *chico* switches from *chico de tres* to *chico de dos* when the new tempo is reached;

d) the second *subida* is essentially pattern–based: it is led by the *repique* by playing *repicado corrido* in m. 21, and is immediately responded by the *piano* by switching from the ornamented first base pattern to the "straighter" second base in m. 22 (Figure 3). The *chico* drum also reacts with a local increment in dynamics. Although quantitatively small, this second raise is perceived as a significant

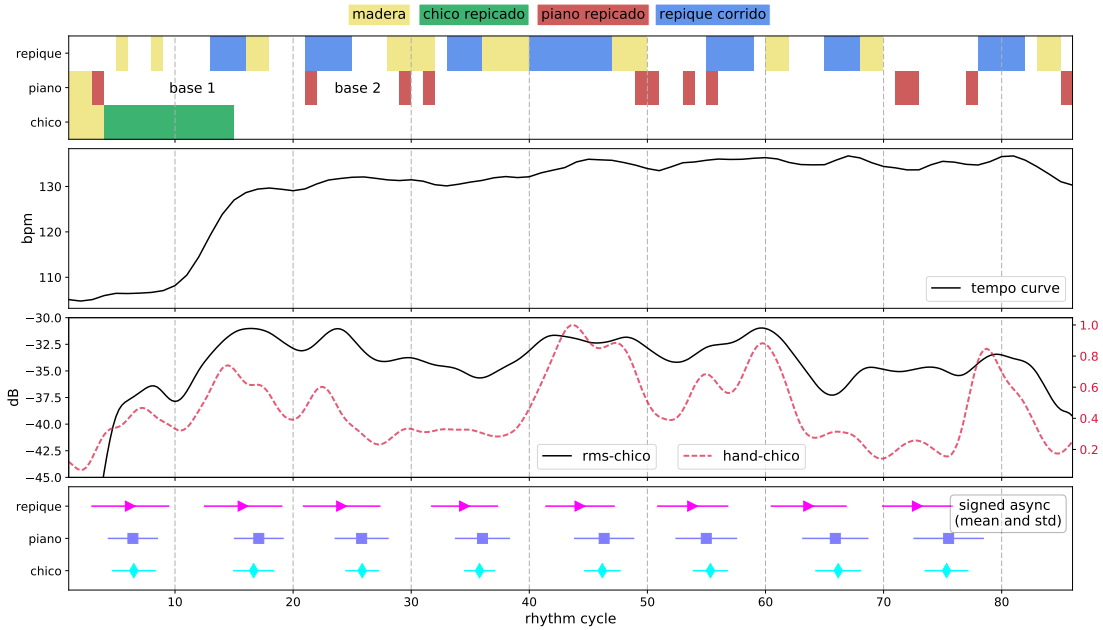


Figure 6: From top: main patterns of the three drums during the performance; tempo curve in bpm; dynamics curves of the *chico* (RMS and trajectory of the hand); asynchrony between the three drums in 10–cycle windows.

increase in energy, due to the patterns involved.

6. DISCUSSION AND FUTURE WORK

In this work, a particular Candombe performance was analysed with the aim to study the processes by which the players negotiate tempo and dynamics. Several computational tools were applied to the audio–visual record of the performance and succeeded in providing relevant information for the analysis.

Three phenomena were found to be involved in the process of *"subir la llamada"*: an increase in tempo (either moving from a slower initial tempo to a faster tempo, or a local *accelerando*); an increase in dynamics, measurable both in levels of sound energy and in the extent of the hand trajectory; the use of certain patterns considered with a more propulsive rhythm: the *repicado básico* (as opposed to more complex figurations), the *chico liso* (as opposed to *chico repicado*) and a straight *llamada piano base* (as opposed to more ornamented *base* patterns).

With respect to the means by which one player leads the process (*"llama a subir"*), also three types of cues were found. As was expected, microtiming played an important role (arguably the most important), with the drum leading the process playing "ahead" to "push" the rhythm. Dynamics was also a factor (playing louder to signal an increase in energy), as well as the use of specific rhythmic patterns recognized as "callers" (*llamadores*).

In future work, the analysis of the facial key points provided by the computer vision system could be addressed, in order to extract further information on the interaction between musicians. Another relevant research strand to develop is the automatic detection of the information governing the mechanisms of coordination and synchronization, such as small variations in onset asynchrony, so as to be able to predict the changes in tempo and dynamics.

7. ACKNOWLEDGEMENTS

The authors wish to acknowledge the performers. Thanks to Guillermo Carbajal for providing support on image processing and for running *OpenPose* on the analysed video.

This work was partially funded by Comisión Sectorial de Investigación Científica (CSIC) and Agencia Nacional de Investigación e Innovación (ANII), Uruguay. The IEMP project is funded by the Arts and Humanities Research Council (AHRC), United Kingdom.

Software tools were implemented in Python, using Scipy, Numpy, Matplotlib and Scikit-learn libraries. Music examples were typeset using LilyPond.

8. REFERENCES

Böck, S., Krebs, F., & Schedl, M. (2012). Evaluating the online capabilities of onset detection methods. In *Proc. of the 13th International Society for Music Information Retrieval Conference (ISMIR)*, (pp. 49–54).

Cao, Z., Simon, T., Wei, S.-E., & Sheikh, Y. (2017). Real-time multi-person 2d pose estimation using part affinity fields. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (pp. 1302–1310).

Clayton, M. (2012). What is entrainment? definition and applications in musical research. *Empirical Musicology Review*, *7*, 49–56.

Jure, L. (2013). Principios generativos del toque de repique del candombe. In C. Aharonián (Ed.), *La música entre África y América* (pp. 263–291). Montevideo, Uruguay: Centro Nacional de Documentación Musical Lauro Ayestarán.

Jure, L. (2017). Timeline patterns in Uruguayan Candombe drumming. In *44th International Council for Traditional Music World Conference (ICTM)*.

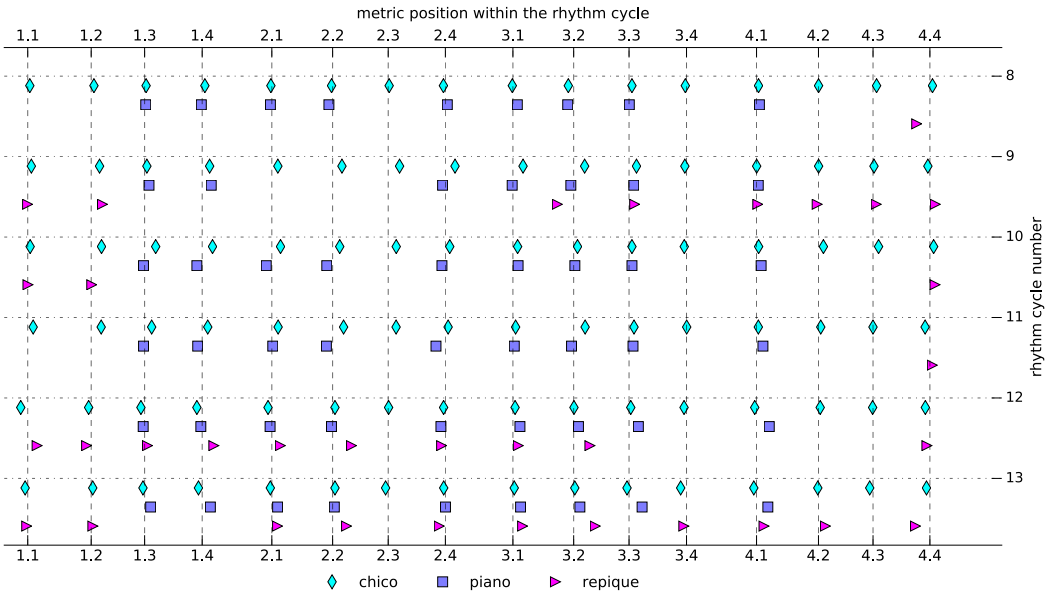


Figure 7: Normalized location of the onsets of the three drums for cycles 8 to 13.

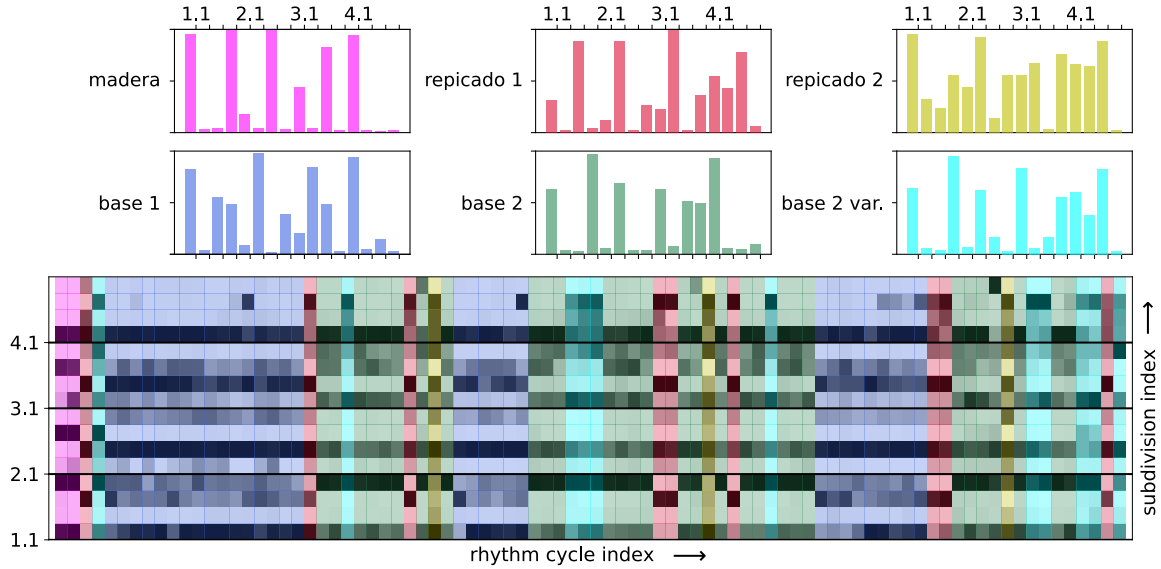


Figure 8: Analysis of the rhythmic patterns of the *piano* drum.

Jure, L. & Rocamora, M. (2017). Clave patterns in Uruguayan candombe drumming. In *16th Rhythm Production and Perception Workshop (RPPW)*.

Krebs, F., Böck, S., & Widmer, G. (2013). Rhythmic pattern modeling for beat and downbeat tracking in musical audio. In *Proc. of the 14th International Conference on Music Information Retrieval (ISMIR)*, (pp. 227–232).

Nunes, L., Rocamora, M., Jure, L., & Biscainho, L. W. P. (2015). Beat and downbeat tracking based on rhythmic patterns applied to the Uruguayan candombe drumming. In *Proc. of the 16th International Society for Music Information Retrieval Conference (ISMIR)*, (pp. 264–270).

Polak, R., London, J., & Jacoby, N. (2016). Both isochronous and non-isochronous metrical subdivision afford precise and stable ensemble entrainment: A corpus study of malian jembe drumming. *Frontiers in Neuroscience*, *10*, 285.

Rocamora, M. & Biscainho, L. W. P. (2015). Modeling onset spectral features for discrimination of drum sounds. In *Proc. of the 20th Iberoamerican Congress on Pattern Recognition (CIARP)*, (pp. 100–107).

Rocamora, M., Jacobi, N., Jure, L., & Polak, R. (2017). Interpersonal music entrainment in Afro-Uruguayan candombe drumming. In *44th International Council for Traditional Music World Conference (ICTM)*.

Rocamora, M., Jure, L., & Biscainho, L. W. P. (2014). Tools for detection and classification of piano drum patterns from candombe recordings. In *Proc. of the 9th Conference on Interdisciplinary Musicology (CIM 2014)*, (pp. 382–387).

Rocamora, M., Jure, L., Marenco, B., Fuentes, M., Lanzaro, F., & Gómez, A. (2015). An audio-visual database of candombe of performances for computational musicological studies. In *Proc. of the Congreso Internacional de Ciencia y Tecnología (CICTeM)*, (pp. 17–24).

TWO METHODS TO COMPUTE MELODIES FOR THE LOST CHANT OF THE MOZARABIC RITE

Geert Maessen
Gregoriana Amsterdam, Amsterdam, The Netherlands
gmaessen@xs4all.nl

Darrell Conklin
Department of Computer Science and Artificial Intelligence, University of the Basque Country UPV/EHU, San Sebastian, Spain
IKERBASQUE, Basque Foundation for Science Bilbao, Spain
darrell.conklin@ehu.eus

ABSTRACT

Several medieval chant traditions are preserved in precursors of modern music notation. Virtually all chants of the Mozarabic rite are only preserved in the earliest of these: pitch-unreadable neumatic notation. Melodic intervals are not available. This paper sketches two computational methods to produce melodies based on a comparison of transcriptions of the early notation with pitch-readable preserved traditions encoded in a data set.

1. NOTATION, ENCODING & AN EXAMPLE

In 1973 Don Randel published a detailed description of over 5,000 chants of the Mozarabic rite preserved in about forty manuscripts and fragments dating from the early eighth until the thirteenth centuries. Several genres were included, from very simple to most complex: Randel orders 28 genres in five manuscript groups with neumatic notations: León, Rioja, Silos, Toledo A and Toledo B. The most important manuscript is the León antiphoner (E-L 8) dating from the early tenth-century. Unlike Gregorian chants, many Mozarabic chants appear in only one or two manuscripts, often with much greater differences in musical detail than in Gregorian chant. Only a few dozen relatively simple melodies have been found in pitch-readable notation. Some scholars, however, have shown melodic relations with other chant traditions for some specific Mozarabic chants (Levy, 1998). Therefore we have compiled a data set with preserved medieval melodies, as a base for the construction of melodies for the lost Mozarabic chant (Van Kranenburg & Maessen, 2017).

Although melodic information is virtually absent, the neumatic notation of the Mozarabic rite sketches the contours of the melodies. From note to note we can mostly see if the melody goes up or down (Rojo & Prado, 1929). An elementary way to represent this contour information is using six letters: `h` a note higher than the previous note; `l` a note lower; `e` a note of equal pitch; `b` higher or equal; `p` lower or equal; `o` a note with unclear relative height. Figure 1 shows the beginning of the second part of one of Levy's chants; the sacrificium *Sanctificavit Moyses altare.* Shown at the top of the figure are three lines from the León antiphoner. Following that, three parallel lines show the transcription of the neumes to contour letters and two different melodies produced with our two methods for chant reconstruction. Encircled in the manuscript image is a repeated *intra-opus* neumatic pattern that should be instantiated by the same musical material. In the contour string and melodies the corresponding patterns are underlined. Capitals in the string indicate notes of the patterns.

It should be noted that the context of this particular pattern is interesting for the comparison of chant traditions. The chant text is a narrative from Exodus (Ex. 34:2-5): The Lord said to Mozes "come up to me unto mount Sinai" (ascende ad me in montem Syna). Then Mozes went up unto the mount (ascendit in montem) and the Lord descended towards him (descendit ad eum). The three words "ascende", "ascendit" and "descendit", share the pattern. This pattern is also part of two extended patterns; the first shared by "ascende" and "ascendit", the second by "ascendit" and "descendit". So here the music can be seen to express the "meeting" of the Lord and Mozes, something hardly imaginable in other chant traditions. A closer look may even reveal a metaphor for the idea that Mozes is in God's "hands": "His" words and deeds, "ascende", and "descendit". Misunderstanding of this kind of sophistication may well have been used in the repression of the Mozarabic rite in the late eleventh century when, in the ongoing power struggle between the advocates of the different rites, the supposed heretical character of the Mozarabic rite was repeatedly stressed by Pope Gregory VII (Vones, 2007). In the parallel chants of the surviving traditions these details are blurred, completely absent, or at best reduced to different up and down movements only.

2. METHOD 1

The first method to produce melodies for the lost chant has been described in detail in a previous paper (Maessen & Van Kranenburg, 2017) and is reviewed here briefly. In search for a melody of a specific lost chant, we first

transcribe the chant notation to a string of contour letters. Then we divide this string into segments, guided by the grammatical structure of the chant. Our method implemented a brute force string matching algorithm that searches all matches of all segments of the contour string in all chants of the data set, allowing for a variable number of n skips in all segments. The algorithm then lists the best matching melodies of the data set conforming to a computed evaluation score S. This score is defined by the positions of the matches of the segments in the data set melodies and the number n of allowed skips.

In order to produce a singable melody for the lost chant we may need three additional steps using this method. The first combines the matches of the segments in the best data set melody to one melody for the lost chant. For some of the segments we may need to increase the number of allowed skips n, or shorten the segments. In a second step we may give repeating patterns (*intra-opus* patterns, as appearing in the early notation of a single chant) the same pitch sequences. The third step consists in rehearsing the chant and correcting some of its “uncharacteristic” pitches. Some segments may be in need for transposition, and especially at the borders of segments melodic lines sometimes need to be smoothed.

With this method we produced over 100 chants, some of them still on the internet (Gregoriana Amsterdam, n.d.). Although some of these chants are considered beautiful, there are some problems (see Section 4).

3. METHOD 2

The basis for the second method is the construction of statistical models of a coherent corpus, and then “inverting” this model to generate new music having high probability according to the model (Conklin, 2003). A statistical model is trained on a data set of 137 Gregorian offertories, comprising a total of approximately 65,000 notes. Given the size of the corpus, it was possible to create a bigram model of pitches that does not have data sparsity problems. Following model training, a sequence of pitches can be generated based on the probabilities derived from the data set by performing statistical Gibbs sampling and settling on sequences at the high end of probability space.

To capture the specified positional constraints and also *intra-opus* repetition, Conklin (2016, 2017) introduced an important improvement to statistical generation that makes it appropriate for the production of melodies for the lost chant at hand. Template pieces are encoded using a *semiotic pattern*, which specifies constraints on individual notes and also equality relations between segments of notes. Sequences are sampled from the trained statistical model while ensuring that the semiotic pattern is maintained for each sample. Positional constraints, the most important facet of the pattern in the case of chant reconstruction, are given by the string of contour letters referred to in Section 2 above and illustrated in Figure 1. Repeating (*intra-opus*) contour patterns are generated as patterns of equal pitches, for example, see the three (overlapping) repeated patterns in capitals in Figure 1. These *intra-opus* patterns are annotated manually for each template.

Given a model and a semiotic pattern, a large space of sequences is sampled in two steps. In the first, an iterative random walk (Conklin, 2016) is used to create an initial solution compatible with the semiotic pattern. In this step the contour pattern, using the specified ambitus desired, is compiled into further positional constraints that make the search for an initial solution feasible. For example, if an `h` contour is at a position, it is clear that the previous position cannot be at the maximum height of the ambitus.

Following the successful production of an initial solution, Gibbs sampling is performed: positions are selected uniformly over the semiotic pattern variables, and all new possible notes are considered in that position. A new sequence is then sampled from the distribution of pieces with successfully substituted notes. Unlike standard Gibbs sampling, here some preference is given to sampling sequences that increase rather than decrease the current sequence probability. In our implementation, we have found that, for most templates, after approximately 100,000 iterations the highest probability solutions no longer change.

Until now we successfully generated several chants using positional and *intra-opus* patterns together. Some of these melodies we performed in videos with the early notation running along (Gregoriana Amsterdam, n.d.). We also successfully experimented with other constraints, such as the ambitus (different for different parts), and the first and last (and other) pitches of the chant.

4. COMPARISON

As Figure 1 and Gregoriana Amsterdam (n.d.) may show, both methods are able to generate singable melodies. Of importance here, however, are the differences:

1. Due to the segmentation and the working hypothesis of the existence of (nearly) identical matches of segments to parts of existing melodies, Method 1 seems infeasible without manual editing of the constructed melody. In Method 2 manual editing is not necessary because, unlike Method 1, there are no “problematic” borders between segments. Any sequence following the semiotic pattern will be a solution to the contour and *intra-opus* pattern specification.

2. The score S of Method 1 provides a good criterion for the relation with the lost melody. When S is higher than about 70 %, Method 1 also indicates serious candidates for historically related melodies in the data set. However, until today, Method 1 seldom produced a score above 40 %. Method 2 is only able to generate general characteristics about probabilistic space as it uses only a low order statistical model.

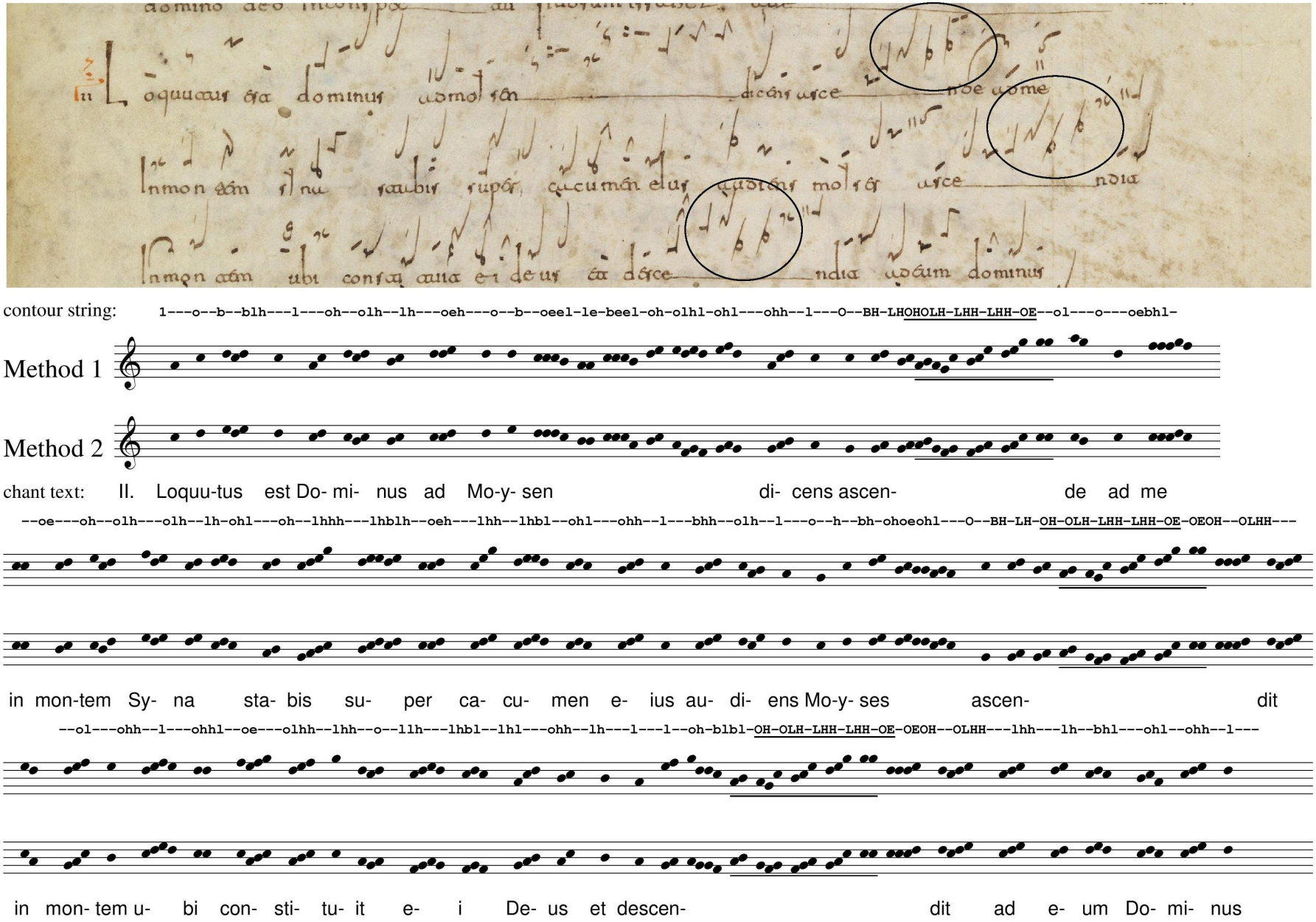


Figure 1. A passage from *Sanctificavit Moyses*: neumatic notation (E-L 8; 305r13), contour string and two melodies.

3. In Method 1 we did not yet fully use the information included in the contour string about syllables, words and sentences. In Method 2 this information is automatically processed as part of the statistical model.

4. The more constraints we introduce to Method 1 (patterns, ambitus, specific pitches), the more problematic it will be to construct a melody, unless the desired melody (or a very close variant) is included in the data set. In Method 2 the only things of importance are the sampling algorithm and the statistical model used.

5. In order to construct a considerable self-consistent corpus of chants it will be necessary to produce *inter-opus* patterns of equal pitches, i.e. to generate patterns occurring in different chants and having equal sequences of pitches. Given its reliance on segmentation and the facts relating to its general hypothesis this seems almost impossible in Method 1. However, we are already successfully experimenting with this in Method 2.

5. CONCLUSION AND FUTURE WORK

The overall impression is that both methods are able to generate singable melodies. The second, however, even in this stage, seems less laborious. No manual editing is needed and the generation of *intra-opus* repeating patterns is implicit in the method.

A fascinating point opened up by our research is the role of overfitting in statistical models. Usually this is viewed negatively as the inability of a model to generalize past the known data. However in the chant reconstruction problem there are cases where overfitting is desired, as for example when a template melody may contain *inter-opus* patterns (seen in another chant). These patterns should be used when available. Thus high on our agenda is the consideration of how to handle *inter-opus* recurring patterns. In summary Method 2 seems not only the best option to generate unique chants, but also to construct a self-consistent repertory agreeing with the early notation, something that seems hardly feasible with the first method. And, of course, this last option is high on our agenda.

Presently we are, therefore, working on the implementation of *inter-opus* patterns. We are also constructing some cross-validation cases: chants where neumes and corresponding pitches are known, e.g. the Gregorian chant offertory *Scapulis suis* (Gregoriana Amsterdam, n.d.). There are other items on our agenda. Pattern discovery algorithms might be used to find *intra-opus* patterns in the templates, thus automating the laborious step of hand annotation of a template for patterns. To create large collections of reconstructions for many templates this seems even necessary. In Method 2 it will be necessary to handle church modes and ambitus constraints: these may vary throughout the piece and will require some broad segmentation of the template. Also in Method 2 the reference of patterns to the conservation of exact pitch sequences seems unnecessarily strict and we plan to allow any feature (intervals, neume shapes) to be conserved between pattern instances. Until now Method 2 only made use of a single tradition. Since we know that several traditions were related to the lost chant (Levy, 1998), it will be necessary to handle the differences between these traditions in Method 2. We are working on ways to define the relations between the lost chant and these traditions. Finally, until now we only used six contour letters. However, the neumatic information in the León antiphoner is much richer. Since the meaning of Mozarabic neumes is similar to Gregorian neumes and we do know the Gregorian melodies, it will be wise to include still another data set in our algorithms: Gregorian chants in pitch-unreadable neumes.

6. REFERENCES

Conklin, D. (2003). Music generation from statistical models. In *Proceedings of the AISB Symposium on Artificial Intelligence and Creativity in the Arts and Sciences, Aberystwyth, Wales,* (pp. 30-35).

Conklin, D. (2016). Chord sequence generation with semiotic patterns. *Journal of Mathematics and Music*, 10(2):92-106.

Conklin, D. (2017). Modelling and sampling jazz chord sequences. In *Proceedings of the 10th International Workshop on Machine Learning and Music, 6.10.2017, Barcelona, Spain,* (pp. 13-18).

Gregoriana Amsterdam. (n.d.). YouTube Channel of Gregoriana Amsterdam, with some sung examples produced with both methods (video titles with numerals 1 and 2) and showing the original notation: https://youtube.com/lelalilu

Levy, K. (1998). Toledo, Rome and the Legacy of Gaul. In *Gregorian Chant and the Carolingians*, 31-81.

Maessen, G. & Van Kranenburg, P. (2017). A Semi-Automatic Method to Produce Melodies for the Lost Chant of the Mozarabic Rite. In *Proceedings of the 7th International Workshop on Folk Music Analysis, 14-16 June 2017, Málaga, Spain,* (pp. 60-65).

Randel, D. (1973). An Index to the Chant of the Mozarabic Rite. New Jersey: Princeton University press.

Rojo, C. & Prado, G. (1929). El Canto Mozárabe, Estudio histórico-critico de su antigüedad y estado actual. Barcelona: Diputación Provincial de Barcelona.

Van Kranenburg, P. & Maessen, G. (2017). Comparing Offertory Melodies of Five Medieval Christian Chant Traditions. In *Proceedings of the 18th International Society for Music Information Retrieval Conference, Suzhou, China, October 23-27,* (pp. 204-210).

Vones, L. (2007). The Substitution of the Hispanic Liturgy by the Roman Rite in the Kingdoms of the Iberian Peninsula. In *Hispania Vetus: Musical-Liturgical Manuscripts: from Visigothic Origins to the Franco-Roman transition,* 43-59.

Going Deep with Segmentation of Field Recordings

Matija Marolt
University of Ljubljana
matija.marolt@fri.uni-lj.si

ABSTRACT

In the paper, we explore the performance of deep residual convolutional networks for labelling ethnomusicological field recordings. Field recordings are integral documents of folk music performances captured in the field, and typically contain performances, intertwined with interviews and commentaries. As these are live recordings, captured in non-ideal conditions, they usually contain significant background noise. Labelling of field recordings is a typical step in segmentation of these recordings, where short sound excerpts are classified into one of a set of predefined classes. In the paper, we explore classification into four classes: speech, solo singing, choir singing (more than one voice) and instrumental performances. We describe the dataset gathered for the task and the labelling tools developed for gathering the reference annotations. We compare different input representations and convolutional network architectures based on residual modules for labelling short audio segments and compare them to the more standard feature based approaches, where an improvement in classification accuracy of over 5% was obtained.

1. INTRODUCTION

Field recordings are documents of folk song and music performances taken "in the field", usually in environments familiar to musicians. They aim to preserve entire recording sessions and the context in which they were recorded, and are thus a mix of performances and speech, which often consists of interviews with musicians. Since recordings are taken in everyday environments, they are often very noisy due to background noise (e.g. people talking, doors closing etc.), poor recording equipment or the recording environment itself. Segmentation of field recordings is one of the first tasks that ethnomusicologists perform when studying the recorded materials, as they separate the contents into different units, such as speech or individual performances. It is also a prerequisite for computational analysis of field recordings.

In the audio processing and music information retrieval research fields, automatic segmentation of recordings is a well-studied task. It is important for segmentation of broadcast news and radio broadcasts, where recordings are usually separated into speech and music units, as well as in other domains such as for removal of non-speech parts in speech recognition systems. Most approaches either first label short segments of the recording into a set of classes (e.g. speech, music) and then find segment boundaries (Lie, Stan, & Hong-Jiang, 2001; Williams & Ellis, 1999), or first find the segment boundaries and later apply classification into classes (Panagiotakis & Tziritas, 2005; Tzanetakis & Cook, 1999). Pikrakis et al. (2008) used a three step approach: first they identified regions in the signal which are very likely to contain speech or music with a region growing algorithm. Then, they segmented the remaining regions with a maximum likelihood model and finally, a boundary correction algorithm was applied to improve the found boundaries. Marolt (2009) also used a three step procedure where signal fragments were first labelled into five classes, then candidate boundaries established and finally the actual boundaries estimated with a maximum-likelihood criterion.

More recently, within the Mirex 2015 Music/Speech Classification and Detection task ("Mirex 2015 Results," 2015), 9 authors submitted their algorithms for classifying recordings into either speech or music, and for finding segment boundaries in a set recordings, which also included a number of field recordings. The algorithms were very successful for the first task, reaching 99.7% accuracy (Lidy, 2015), which might indicate that the task of music/speech classification is *solved*, however it is more likely that the evaluation dataset was too basic and did not include enough challenging cases for the algorithms. This is already obvious if we observe results of the same approaches for the second task, where frame-based F1 measure of the best system dropped to 89.4% (Marolt, 2009), while the F1 score of finding segment boundaries was only at 40.3%.

In the past years, deep learning had become the prevalent approach for classification problems in image and audio domains. It is therefore not surprising that it was also applied to segmentation of audio recordings. The aforementioned best music/speech classifier at Mirex 2015 by Lidy (2015) was based on convolutional neural networks. Similarly, Kruspe et al. (2017) use deep networks to discriminate between speech and music sections in broadcast signals and reports over 99% F1 measure for speech and 91% for music discrimination. Authors from Google (Hershey et al., 2017) compared a number of deep architectures for large-scale audio classification on tagged audio from the YouTube-100M dataset, as well as on a large scale dataset of labelled sound clips from YouTube videos – Audio Set (Gemmeke et al., 2017).

In this paper, we explore deep neural networks for labelling ethnomusicological field recordings. Unlike broadcast recordings, field recordings are more challenging to label and segment due to their noisy nature. In contrast to most speech/music discriminators, we aim to separate between four rather than two classes: speech, solo singing, choir singing (more than 1 voice) and instrumental recordings. We chose the four classes as they are very representative for a number of field recordings from different regions that we analyzed. Also, in contrast to most previous work, we do not aim to segment (clean) broadcast recordings, but field recordings, which may be of varying quality, as already described previously. We describe the architecture used for classification, the dataset used and our first results.

2. DATASET

Exploration of field recordings revealed four major classes of recordings that appear in a variety of cultures: solo singing, choir (more than one voice) singing, instrumental performances, and speech. Our goal was therefore to classify field recordings into the four classes, and not to limit ourselves to just speech and music. To train deep learning classifiers, large datasets are needed - the larger the better as recent deep learning experiences show. Apart from the Audio Set (Gemmeke et al., 2017), which is an excellent large-scale audio classification dataset, there are few suitable datasets available for the task. In the presented work, we decided not to begin with the Audio Set, as its categories are not ideal for our purpose; for example, there is no solo singing category, examples labeled with singing are mostly accompanied by music, while musical genres are mostly oriented towards popular music genres (pop, rock etc.).

We therefore gathered short excerpts from a variety of recordings from ethnomusicological (and related) archives that put their collections online in recent years. The sources include: the British Library world & traditional music collection[1], Alan Lomax recordings[2], sound archives of the CRNS[3] and a number of recordings from the Slovenian sound archive Ethnomuse and the Norwegian national library, which are not available online, but were made available to us by ethnomusicologists with the respective institutions. These field recordings were augmented by the well-known GTZAN music/speech collection and the Mirex 2015 music/speech detection public dataset.

Altogether 7,000 5 second long excerpts were extracted from these sources. To manually label them into the four target classes, we enhanced the web-based audio annotator tool (Cartwright et al., 2017), so that it can be controlled exclusively by the keyboard. This enabled fast multi-user annotation of audio excerpts into the four categories, augmented by three additional categories of "voice over instrumental", "noise" and "not clear". The latter was to be applied when the audio clip was either too noisy to be recognized or contained too many short fragments of different types of materials, so that it was difficult to select a single label. The annotator's goal was namely, to select a single label for the five second clip, where clips were randomly chosen from the set of unlabeled clips for each participating annotator. The user interface was kept very similar to the original audio annotator and is shown in Figure 1.

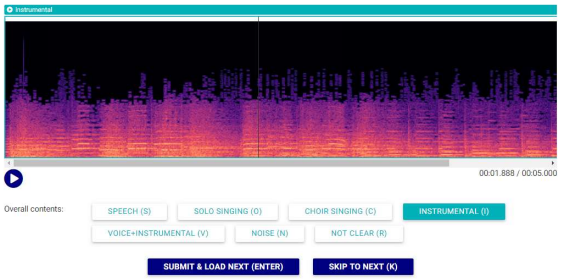


Figure 1. The annotation interface.

3. EXPERIMENT

Our goal was to evaluate the performance of deep networks for the classification task at hand. All the audio excerpts were first downsampled to 22050 Hz, mixed to a single channel and normalized.

We compared several input representations for the task: 46 ms FFT frames (252 bins between 50 and 7000 Hz) and 64 channel mel-scale spectrograms (50-8000 Hz) extracted from FFT frames of 23ms, 46 ms, and 92ms. We log-scaled all representations (adding 1e-5 before applying the logarithm) and used 1 or 2 second long feature blocks with 50% overlap as network inputs. Stacking of different resolution frames (23ms, 46ms, 96ms) was also tested.

We chose convolutional deep networks as our main classification tool and focused specifically on residual networks (K. He, Zhang, Ren, & Sun, 2015), which previously demonstrated good performance for a variety of image, as well as audio-based tasks. The main feature of residual networks are their shortcut connections that implement identity mappings and enable convolutional blocks to learn residuals between the underlying mapping of features and the input.

The overall network architecture is shown in Figure 2. The input layer is first processed by m nxn convolutions, optionally enhanced with m nxn dilated convolutions with rate 2, to expand the receptive field of filters. A max pooling layer was added to reduce the size of feature maps, followed by p resnet v2 blocks (Kaiming He, Zhang, Ren,

[1] https://sounds.bl.uk//World-and-traditional-music
[2] http://research.culturalequity.org/home-audio.jsp
[3] http://archives.crem-cnrs.fr/

& Sun, 2016), where the size of feature maps is halved (in each dimension) within each block and the number of filters doubled. The batch normalized output of resnet blocks is gathered by 1x1 convolutions into a 2D feature map. The map is finally processed by a small fully connected layer with four outputs, where the softmax activation yields final class probabilities. We tested different values for the described parameters, which we outline in section 4. Batch normalization, as well as l2 regularization were used for regularizing the network, to avoid overfitting. To introduce non-linearity, we compare the performance of standard ReLU activation functions with exponential linear units ELU (Clevert, Unterthiner, & Hochreiter, 2015).

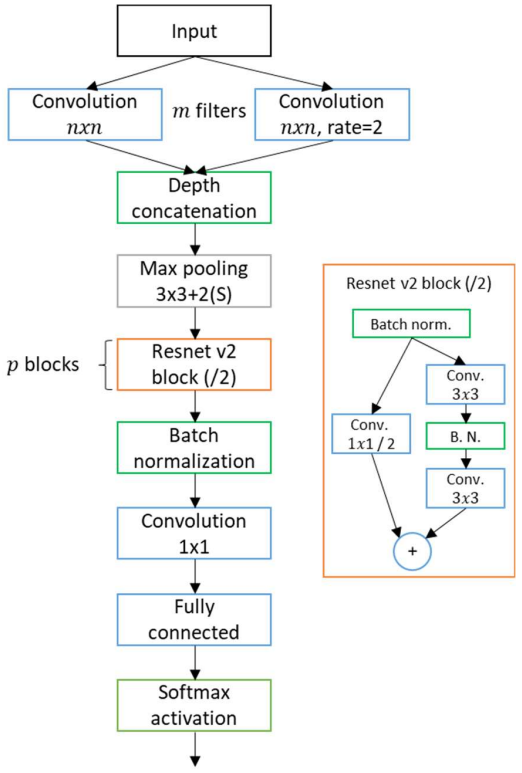


Figure 2. The network architecture.

Three-fold cross validation was used to assess the performance of each network, where 2/3 of the dataset was used for training, the remaining 1/3 for testing, and the procedure repeated three times. The networks were trained with minibatches of 128 examples. For each audio example, the block of input features was drawn from a random location within the audio, so that for each epoch, the feature blocks used to train the network differed in their location within training files. Such *time translation* diversifies the limited training data available and improves performance, as was also demonstrated elsewhere (Jansen et al., 2017). For testing, the entire test files were used.

Stochastic gradient descent was used for training over 500 epochs, and the learning rate set to decay from 0.1 by 0.75 each 500 steps. The experiments were implemented in Tensorflow.

4. RESULTS

4.1 Input representation

A comparison of different input representations is shown in Table 1. We report average accuracy over all classes over the three cross-validation splits in the last column. The same network architecture (described in 4.2) was used for all comparisons. We compare two different input representations: mel compressed spectrograms vs. FFT, two different block sizes (1.1 vs. 2.2 second long blocks of input features), three different window and two different step sizes for FFT calculation.

We see a significant difference only in the choice of block sizes: features covering 1.1 seconds of audio give around 2% lower accuracy as 2.2 second blocks, indicating that it is beneficial for the network to have more context in order to distinguish between the categories. Indeed, even when listening to, for example speech vs. solo singing, in many cases one second of audio cannot not reveal the correct category. This is even truer for field recordings, which are typically amateur performances, many times by older people, include strong dialects etc. There are no significant differences between different window and step sizes in 2.2 second blocks. Stacking of different window sizes also does not improve the performance significantly. We therefore decided to use 2.2 second blocks of 64 channel mel spectrograms calculated from FFT frames of 46 ms with 23ms step size (network input size 96x64) in our further experiments.

feature	*block (s)*	*step (ms)*	*window (ms)*	*input size*	*accuracy*
mel	1.1	12	12	96x64	0.861
			23	96x64	0.868
			46	96x64	0.868
			12,23,46	96x64x3	0.875
	2.2	12	12	192x64	0.882
			23	192x64	0.887
			46	192x64	0.887
			12,23,46	192x64x3	0.891
		23	23	96x64	0.886
			46	**96x64**	**0.890**
			92	96x64	0.891
			12,23,46	96x64x3	0.895
fft	2.2	23	46	96x252	0.892

Table 1. A comparison of different input representations.

4.2 Network architectures

The overall network architecture was described in section 3. We tested the influence of the following parameters on network performance: the number of filters in the first convolutional layer (2, 4, 6, 8), the sizes of these filters (4, 6, 8, with or without stacked dilated convolutions of the same size), the number of resnet blocks (3, 4, 5) and the activation function (ReLU vs ELU). Table 2 lists the key results.

The networks are not very sensitive to the size of input filters. When the number of layer one filters m increases

up to 6 filters, performance improves, while higher numbers do not have a large effect. Adding an additional set of dilated filters (rate=2) helps, although this also increases the number of network parameters. The optimal number of resnet blocks was determined to be 4, an additional block does not add much to accuracy, but increases the number of network parameters substantially. The ELU activation function seems to improve training (consistently higher accuracy by approx. 1%) over ReLU.

activation	*dilated*	*nxn L1 size*	*m L1 filters*	*p resnet blocks*	*accuracy*
elu	yes	4	6	5	0.893
				4	**0.890**
				3	0.882
			2	4	0.874
			4	4	0.881
	no	4	6	4	0.883
relu	yes	4	6	4	0.882

Table 2. Comparison of network architectures.

Based on the evaluation, our final network architecture uses ELU activations, 6 4x4 convolutions stacked with 6 dilated 4x4 convolutions (rate=2) on the first layer, followed by 4 resnet blocks. The final fully connected layer is small (24x4) and has no hidden layer, but directly maps into the four outputs. The entire network is not very deep, as we have a limited amount of training data, and contains 172,936 trainable parameters.

4.3 Comparison to other approaches

To put the obtained results into perspective, Table 3 lists the performance of three other approaches on the same dataset (also using 3-fold cross-validation):

- a *standard* deep convolutional network with two 3x3 convolutions (one with stride 2) in place of each resnet block (no shortcut links), trained on the same mel-spectrogram input data representation;
- a multilayer perceptron with one hidden layer of 16 neurons trained on VGGish (Hershey et al., 2017) features extracted from the data. VGGish are audio classification features extracted from a VGG-like deep model trained on a large YouTube dataset and made available by Google. Input to the MLP consisted of two consecutive 128-dimensional VGGish vectors, each summarizing 1 second of audio;
- a simple logistic regression model trained on handcrafted features, as described in (Marolt, 2009).

model	*number of parameters*	*accuracy*
proposed resnet	172,936	**0.890**
standard deep	166,556	0.862
MLP on VGGish	4,180	0.881
logistic regression	51	0.837

Table 3. A comparison to other approaches.

The proposed model outperforms the others. It has the highest number of trainable parameters, however care has been taken to avoid overfitting by including batch normalization and l2 regularization during training, as well as using 1/3 of the dataset for testing at each run, so it is safe to assume that its performance is realistic for a wide variety of materials. VGGish features come close second.

An analysis of errors showed many *logical* mistakes, which can be attributed to several factors. First, some of the recordings are very noisy and even a human listener can have some difficulty to discern the contents. Such recordings were often mistakenly classified as instrumentals, as the noise was considered part of the performance.

The confusion matrix in Table 4 shows that many mistakes are made between *neighboring* classes: solo singing is misclassified as choir singing or speech, choir mostly as solo, instrumentals as choir or speech as solo. Some confusions may be due to the particularity of the contents, e.g. some short excerpts of dialectal speech may sound very much like singing. Some mistakes are not really mistakes – an excerpt may be correctly classified, and wrongly labelled. Namely each five second audio clip in our dataset is only labelled with a single class, even though parts of it may contain another class. An example is a choir recording, where some parts are sung solo and then evolve into choirs. As the network only classifies short 2 second excerpts, it may correctly label the solo part as solo, however the entire example is labelled as choir, so this is considered a misclassification. Choir parts sung in unison are another case that is difficult to classify – they are labelled as choir singing in our dataset, but may sound very similar to solo singing.

The final trained network is integrated into the publicly available SeFiRe tool for segmentation of field recordings[1].

		predicted			
		solo	choir	instr.	speech
true	solo	0.87	0.07	0.01	0.05
	choir	0.06	0.89	0.02	0.03
	instr.	0.02	0.04	0.92	0.02
	speech	0.06	0.01	0.01	0.92

Table 4. The confusion matrix.

5. CONCLUSION

In the paper, we demonstrated the performance of a medium sized deep convolutional network applied to classification of field recordings into four classes. We also provide a comparison of different input representations and network architectures for the task. The database used and the final trained model will be made available to the community.

In our future work, we will aim to enhance the dataset with additional sources of field recordings. We will also

[1] http://lgm.fri.uni-lj.si/portfolio-view/sefire/

make use of the Audio Set, currently the largest annotated audio classification dataset, to enlarge our training data. Our second goal is to increase the number of target categories into typical instrument categories and introduce non-exclusive categories (e.g. singing over instrumental), which are currently labeled as instrumentals.

6. REFERENCES

Cartwright, M., Seals, A., Salamon, J., Williams, A., Mikloska, S., MacConnell, D., . . . Nov, O. (2017). Seeing Sound: Investigating the Effects of Visualizations and Complexity on Crowdsourced Audio Annotations. *Proc. ACM Hum.-Comput. Interact., 1*(CSCW), 1-21. doi:10.1145/3134664

Clevert, D.-A., Unterthiner, T., & Hochreiter, S. (2015). Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs). *CoRR, abs/1511.07289*.

Gemmeke, J. F., Ellis, D. P. W., Freedman, D., Jansen, A., Lawrence, W., Moore, R. C., . . . Ritter, M. (2017). *Audio Set: An ontology and human-labeled dataset for audio events*, New Orleans, LA.

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Deep Residual Learning for Image Recognition. *ArXiv e-prints*.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Identity Mappings in Deep Residual Networks. *CoRR, abs/1603.05027*.

Hershey, S., Chaudhuri, S., Ellis, D. P. W., Gemmeke, J. F., Jansen, A., Moore, C., . . . Wilson, K. (2017). CNN Architectures for Large-Scale Audio Classification. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Jansen, A., Plakal, M., Pandya, R., Ellis, D. P. W., Hershey, S., Liu, J., . . . Saurous, R. A. (2017). Unsupervised Learning of Semantic Audio Representations. *CoRR, abs/1711.02209*.

Kruspe, A., Zapf, D., & Lukashevich, H. (2017). *Automatic speech/music discrimination for broadcast signals*.

Lidy, T. (2015). *Spectral Convolutional Neural Network for Music Classification*. Paper presented at the Mirex 2015, Malaga, Spain.

Lie, L., Stan, Z. L., & Hong-Jiang, Z. (2001). *Content-based audio segmentation using support vector machines.* Paper presented at the IEEE International Conference on Multimedia and Expo.

Marolt, M. (2009). *Probabilistic Segmentation and Labeling of Ethnomusicological Field Recordings.* Paper presented at the ISMIR, 10th International Society for Music Information Retrieval Conference, Kobe, Japan.

Mirex 2015 Results. (2015). Retrieved from http://www.music-ir.org/mirex/wiki/2015:Music/Speech_Classification_and_Detection

Panagiotakis, C., & Tziritas, G. (2005). A speech/music discriminator based on RMS and zero-crossings. *Multimedia, IEEE Transactions on, 7*(1), 155-166.

Pikrakis, A., Giannakopoulos, T., & Theodoridis, S. (2008). A Speech/Music Discriminator of Radio Recordings Based on Dynamic Programming and Bayesian Networks. *Multimedia, IEEE Transactions on, 10*(5), 846-857.

Tzanetakis, G., & Cook, P. (1999). *Multifeature audio segmentation for browsing and annotation.* Paper presented at the Applications of Signal Processing to Audio and Acoustics, 1999 IEEE Workshop on.

Williams , G., & Ellis, D. P. W. (1999). *Speech/music Discrimination Based On Posterior Probability Features.* Paper presented at the Eurospeech'99, Budapest, Hungary.

TOWARDS THE STUDY OF EMBODIED METER IN SWEDISH FOLK DANCE

Olof Misgeld, Andre Holzapfel
Department of Media Technology and Interaction Design (MID)
Royal Institute of Technology (KTH)
{misgeld,holzap}@kth.se

ABSTRACT

The interrelation of playing and dancing is central for understanding performance practice in Swedish folk music, as it plays an important role for the metric and rhythmic qualities of *spelmansmusik*, and playing for dancing is considered a key competence for musicians in this tradition. As part of a research project into performance practice, sound, video and motion capture (MoCap) data were recorded from live performances of three musicians and two dancers in different combinations. In addition, dancing to two recordings by an influential musician and to live and prerecorded beat clapping was recorded.

This paper incorporates measurements and visualizations of performance data in combination with performer participation and interviews. As a starting point for our project, we focus on metric qualities in a historical recording, and on the dance movement patterns to a Swedish polska style with asymmetrical beat patterns. For this paper - as a preliminary investigation into the material - the recordings of one dancer dancing to an isochronous clapped beat, and to a recording by an influential player have been used for comparison of a central movement pattern in dancing. The findings show that asymmetric beat patterns contained in the recording cause wider variation among the movement patterns when compared to the patterns observed to isochronous clapping. Considering the performers reactions towards using MoCap as a tool for viewing and discussing their performances, we propose further investigations by combining scientific, ethnomusicological and artistic research methods into the research of performance practice in folk music.

1. INTRODUCTION

Triple-meter music forms like the Swedish polska and the Norwegian springar/pols/springleik are central in folk music and dance traditions in the Nordic countries. In Sweden, historical references to "polish dances" date back to the 16th century (Gustafsson, 2016) and these music and dance forms have been a main focus for collectors and researchers of folk music. The term *polska* encompass local substyles and variations, but can for the common variant of *rundpolska* generally be described as a couple dance in triple time including two parts: *försteg*, most commonly a couple walking forward side by side, stepping on beat one and three, and, second, *omdans* (turning), where "the couple is rotating clockwise around its own axis and at the same time anti-clockwise around the room" (Nilsson, 2017). In this, one turn of the clockwise rotation is completed over one measure.

Research on meter and rhythm in polska and related music - including some suggestions for typologies for music and dance types - took into account various aspects of music and dance: patterns of metrical markings, rhythmical variations, melodic rhythms, articulations and dance movements of the different styles. Styles that include asymmetric beat patterns, like some Swedish polska styles and Norwegian springar/pols/springleik styles have attracted special attention (Sandvik, 1943; Groven, 1971; Ahlbäck, 1989; Kvifte, 1999; Blom, 1993; Johansson, 2009; Haugen, 2017). Johansson (2017) discusses empirical research approaches on asymmetrical rhythm including suggestions for future research and points at the lack of measurement and analyses of different styles, different performers and situations (e. g. playing for dancing) and the lack of experimental studies including the performers view on dance interaction, timing, synchronization, learning practices etc.

Ahlbäck (1986, 1989) has formulated a Folk Music Theory approaching meter and rhythm in Swedish folk music, which has been a major contribution for the emerging folk music educations in Sweden (Ahlbäck et al., 2009). Ahlbäck (2003) discussed asymmetric beat in the polska, with reference to folk music collector Einar Övergaard's (1871-1936) ambiguity on where to place the downbeats in his notations of asymmetric polska tunes in Elverum, Norway (Övergaard & Ramsten, 1982). Ahlbäck exemplified how different metrical interpretations can be achieved by ways of articulation and foot-tapping and discussed asymmetric beat patterns in relation to melodic/rhythmical structures in some early Swedish and Norwegian polska recordings.

The Norwegian ethnomusicologist Blom's pioneering approach to draw curves depicting the typical patterned libration of the body´s center of gravity (*sviktkurva*) in different dances, is a well-established concept among folk dancers and folk dance researchers in the Nordic countries. Libration patterns for various Norwegian folk dances were illustrated in Blom (1993). These were obtained from Blom's observations - as an experienced fiddler and dancer - of step combinations and the oscillations of rising and falling movements (*thesis and arsis*) in the different dances.

Haugen (2017) used Motion Capture (MoCap) recordings of Norwegian *Telespringar* performances to explore the relation in time between vertical oscillation periods - libration curves - obtained from markers on the dancers hips with the foot-tapping and body movement of a *hardingfela* (fiddle) player. Haugen found a stable correlation between the foot-tapping of the player and the libration curves of the dancers which was taken as an indication of a

shared embodied asymmetric meter (long-medium-short). Haugen also showed that the local minima and maxima points of the librations were not in synchronization with the foot-tapping. From this Haugen suggested a modified libration curve for Telespringar, similar in shape but shifted in relation to the beats compared to the one suggested by Blom (1993) - which had the local minima and maxima points aligned with the beats. Another approach was used by Mårds (1999), where, as described by Haugen (2014), dancers were recorded with MoCap while at the same time moving on force plates used to measure the weight pressure of the steps. The player's foot tapping was also registered by a pressure mat but no sound was recorded. This showed similar results for the shape of libration patterns as Haugen (2017) and Blom (1993). Bakka (2014) suggested further comparisons of pressure measurements with libration patterns to determine how these correspond to the perception of musical beat in the dance. Naveda & Leman (2010) have suggested a topological analysis for representing spatiotemporal relations of dance and music gestures in some popular dance forms.

Haugen (2017) showed that MoCap recordings can offer a useful analytical approach to meter and body movement in Telespringar. However, MoCap studies of other styles, including *polska*, remain yet to be presented. As an additional difficulty, whereas Haugen assumed that the libration pattern remains constant throughout the performance, this is likely not to hold for polska styles with two-part structures and with a larger variance to the degree of beat asymmetry.

The present study focuses on a style of polska historically connected to players and dancers from the region of Orsa in Dalarna, Sweden. The fiddle player Gössa Anders Andersson (1878-1962) is a central influencer for this style, as indicated by the large number of published recordings (Musica Svecia, 1999; Andersson et al., 1998; Musica Svecia, 1995) and notations (Andersson, 1922; Forslund, 1921) with him as performer. Filmed recordings of the dance *Orsapolska* covering a time period of 1947-2000 have been presented in Norman et al. (2000) among which a silent film of Gössa Anders playing for dancing in 1947 is a key influence for the performers in this study. These films offer interesting examples on how performances of Orsapolska have varied, for instance, with the tempo being considerably lower in the later recordings. The beat patterns in Orsapolska are often asymmetric, but in a different way than the Telespringar: the beat patterns in Gössa Anders' playing have been shown (Ahlbäck, 1989) to fluctuate between two main beat patterns, one symmetric (three isochronous beats) and one asymmetric (short-long-medium). Ahlbäck (1989) has suggested the use of additive time signatures of 9/16 to express these two patterns in music notation: 3+3+3/16 and 2+4+3/16, respectively. These categorizations of beat patterns should not be regarded as a prescription of precise beat proportions over time, but rather as perceived categorical proportions between beats, where in the asymmetrical beat pattern the short first beat being roughly half the duration of the second beat. These conceptual categories are mirrored in music notation, as exemplified in Figure 1, and relate directly to the categorical successive note durations. Kvifte (1999) and Johansson (2017) questioned the use of the beat ratio of 2:4:3, arguing that it would cause an adaption to a oversimplification of the asymmetric beat structure among musicians in the tradition. This would not hold unless musicians used notation as their only source for interpreting the music, and not as Ahlbäck (1989) suggested, in combination with other approaches to musical meter. It could also be questioned whether such alleged changes in tradition should be ascribed to notation systems or for example the emergence of new cross-genre ensemble forms (introducing instruments new to the tradition, like percussion and plucked string instruments in accompaniment functions).

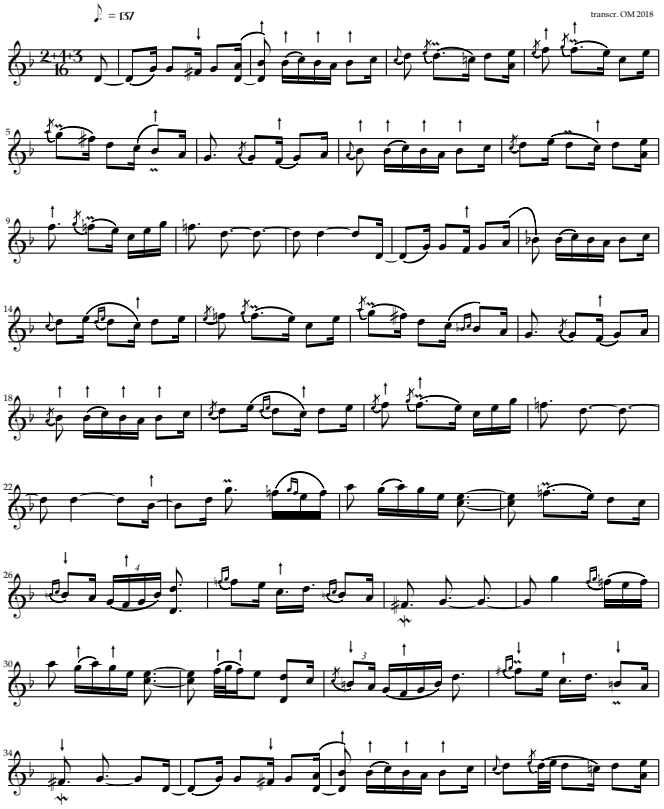


Figure 1: Lorikspolskan played by Gössa Anders, first round.

2. RESEARCH MOTIVATION

Strategies, tools and methods for learning the skills of dancing and playing for dancing have for a long time been developed in the teaching of performance practice [1] , however much of this remains tacit or oral knowledge. This paper is the beginning of a dissertation project on the interaction of musicians and dancers as one key element in performance practice of traditional Swedish folk dance tunes, (*spelmansmusik*). The project is part of ongoing research on performance practice in Swedish folk music where the performers' understanding, knowledge and skills are cen-

[1] One example in compulsory courses on dancing and playing in Folk Music Programs at Kungl. Musikhögskolan, Stockholm

tral for enabling a deeper understanding and formulation of knowledge. This is attempted by the researchers themselves being participants as performers and by performers contributing to the analysis of their own performance recordings. The aim is to formulate tools and concepts for performance practice that will be evaluated through situated performance practice: in teaching situations and by extension in performance situations involving music and dance interactions. Thus, the larger objective of gaining knowledge on performance practice is approached through a combination of scientific, folk music theoretical, ethnographic and artistic research methods.

In this project, experimental methods based on MoCap analysis are applied for measuring and analyzing performance and interaction parameters in combination with performer interviews. This paper presents only a small part of the collected body of experimental data, in a preliminary comparison of the dance movements of one dancer in two different settings, with the purpose to explore analysis methods for further use in the research project. The basic questions that we will address this way are:

1. How far do the libration patterns differ between the two phases (walking, turning) of the dance?

2. What are the timing characteristics of the non-isochronous beats in the recordings?

3. What insights can be obtained from libration patterns of a dancer, when dancing to a static recording with strongly varying beat patterns?

In the following, we describe the complete data collection process in Section 3, and provide subsequently preliminary results from an analysis of part of the obtained data in Section 4.

3. METHODOLOGY AND DATA

For the recording sessions, an experimental setup similar to Haugen (2017) was applied. Two dancers and three musicians were recorded in five different musical setups (described below). The recordings were made in the PMIL-studio at KTH, using an Optitrack Motion Capture recording system [2] of 16 infra-red cameras recording at a frame rate of 120 frames/second. The motion data were saved together with an aligned sound recording from the room. All sessions were in addition to this filmed with two cameras. The participating dancers were Ami Dregelid and Andreas Berchthold, both teachers at the School of Dance and Circus, Uniarts, Stockholm (DoCH) and the Royal College of Music in Stockholm (KMH). The musicians were Ellika Frisell, Sven Ahlbäck and Olof Misgeld - all teachers at the Department of Folk Music, KMH. All dancers and musicians are experienced performers and familiar to the style.

[2] http://optitrack.com

3.1 Setups

1. The first setup had each dancer dancing solo to a track of looped beat claps. The loop was constructed from three slightly different sounding claps, providing an isochronous three-beat cycle at 138 BPM, the mean tempo of the two recordings with Gössa Anders used in setup 3. The dancers were asked to dance in two different ways for the different takes - one time only walking (försteg) and the second time with turning (omdans).
2. The second setup had each dancer dancing solo to a clapped beat, this time performed live by a musician. The tempo was steady but affected by the interaction as the musician was watching the dancer during the take. The dancers were asked to do both walking and turning as they liked.
3. The third setup had each dancer dancing solo to two recordings of Gössa Anders: Lorikspolskan and Polska efter Pellar Anna [3] . Also here the dancers were asked to variate the dance movements between walking and turning.
4. The forth setup had each dancer dancing solo to each of the three musicians. The musicians were all playing the same two tunes as in the recordings with Gössa Anders. Here, the dancers and musicians were instructed to play and dance well together as they would normally do, which then included both walking and turning in the dance.
5. Finally, the two dancers were dancing as a couple to each musician, walking and turning together - the most typical way to dance polska.

The purpose of these different setups was to get a set of data for comparing the correlation of dance movements to music in takes that, (a): did or did not involve non-isochronous beat patterns, (b): did or did not involve musical context apart from beat markings (clapping), and (c): did or did-not involve the interaction between musician and dancer(s). In this paper, a preliminary study of a small part of this dataset is presented, focusing on the second and the third setup. This choice is motivated by the goal to document the typical libration patterns of the dancers in the second setup, and the comparison of these patterns with the ones that emerge from the dance to a recording.

3.2 Annotating the music

The two tunes by Gössa Anders were manually annotated with beat times using Sonic Visualizer [4] . Beats were placed by listening for the articulation, (bow turns and/or ornamentation) and moving the marker to the start of each note considered to correspond to a beat. The basis for these annotations were the transcriptions of the pieces using the notation software ScoreCloud [5] (Figure 1 shows one example). The beat patterns in the two polskas played by Gössa Anders are, as shown by Ahlbäck (2003) and Johansson (2009) non-isochronous, and the proportion of each beat within each bar varies. Using Ahlbäcks model of two main

[3] First recorded in 1950 (Andersson, 1950), re-issued on CD (Andersson et al., 1998)
[4] https://www.sonicvisualiser.org
[5] http://scorecloud.com

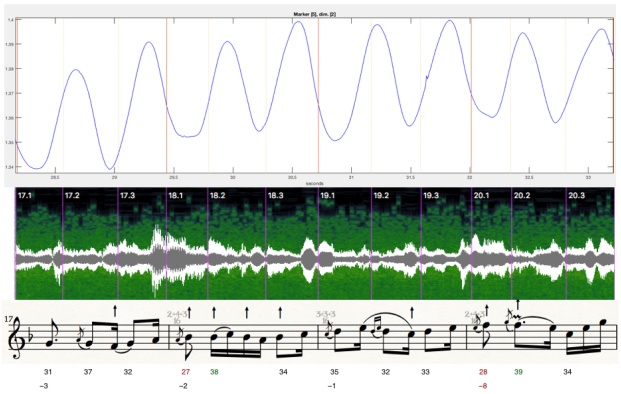


Figure 2: Libration pattern, sound envelope and music notation. Rows of numbers below show beat proportions and measure length deviations from the mean.

beat patterns all bars were classified into two categories, determined by the proportion of the first beat to the bar length, as described in Section 4.1.

3.3 Categorization of the dance

The dance varies between sections of walking and turning. Walking sections mostly include stepping on the first and third beat, with the left foot on the first beat, (in some parts stepping with the right foot on the first beat). Turning sections include turning with steps on the first (left foot) and third beat (right foot) and turning with stepping on the second (left) and third beat(right). The different sections were classified into the below categories, of which 1 and 2 are used in the following comparison in this paper. This is motivated by them being the most common movement patterns in the dance. [6]

1. Walking, beat 1 left foot, beat 3 right foot.
2. Turning, beat 1 left foot, beat 3 right foot.
3. Turning, beat 2 left foot, beat 3 right foot.
4. Other less frequent variations of the above.

The classifications are made to allow comparison of marker movements between different sections of the dance.

4. RESULTS

Figure 2 illustrates an example of the time series obtained from the data collection. In the upper part of the Figure, the vertical curve obtained from the marker of the dancer's upper back is plotted aligned with the beat annotations. The notation of the performed tune is provided in the lower part. Watching the libration in the upper part of the body of a dancer is a common approach among dance musicians, which motivated our particular choice regarding the selected marker. The depicted case, however, illustrates dance to a recording, and not a live music performance.

[6] A small sample of the libration pattern from category 3 is depicted in Figure 2 and differs by containing two oscillation periods per measure instead of - as in 1 and 2 - three.

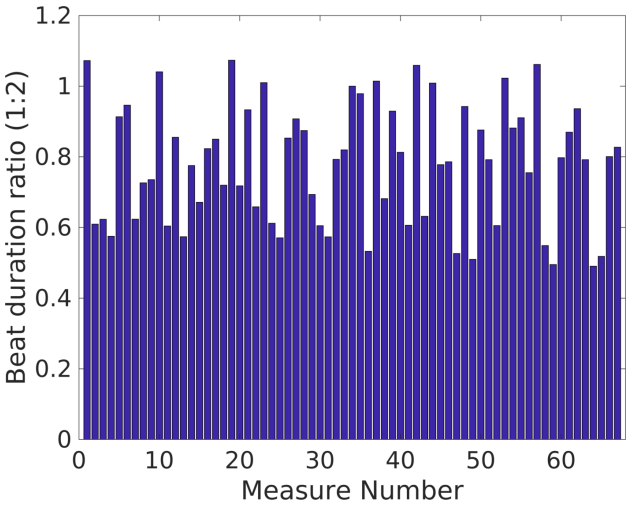


Figure 3: The ratio between the first and second beat duration in Lorikspolskan played by Gössa Anders.

In the example, measures 2 and 4 have a shorter first beat as reflected in the notation. Nevertheless, the libration patterns remain more or less constant throughout the measures of the depicted example. This implies that in this short example no musical interaction between the beat asymmetry in the recording and the dance could be observed. We will now further explore this relation between beat asymmetry and libration patterns.

4.1 Asymmetric beat in the playing of Gössa Anders

The plotted ratio between the beat duration of the first and second beat in Figure 3 confirms the variation in beat lengths of Gössa Anders' playing. As shown, a large proportion of the first beats align at around 0.6 of the lenght of the second beat, close to the 2+4+3 group suggested by Ahlbäck (1989). Other measures group around a larger ratio, however smaller than 1, and therefore not precisely isochronous. In Figure 4 the categorical classification of beats in two patterns is plotted in a boxchart. The threshold value to divide between 2:4:3 and 1:1:1 has been placed in the middle, at a bar proportion of 28 percent for the first beat. The chart in Figure 4 - which can be considered as a summary of Figure 3 - confirms a large frequency combination of a shorter first beat with a longer second beat, and the slightly shorter first beat in the supposedly isochronous class. Further recordings by Gössa Anders need to be examined to see if this is a general characteristics of his style.

4.2 Libration patterns

The attempt has been to show the libration pattern of the center of body gravity from the dancers, as an indication of the embodied metric patterns (Haugen, 2017; Blom, 1993). To this end, the y-direction of the marker placed on the upper back between the shoulders of the dancers was selected for plotting. The patterns of the libration in walking and turning are depicted in Figures 5 and 6, for setups 2 (claps) and 3 (recording), respectively. The subfigures depicting walking and turning differ regarding their shape in both

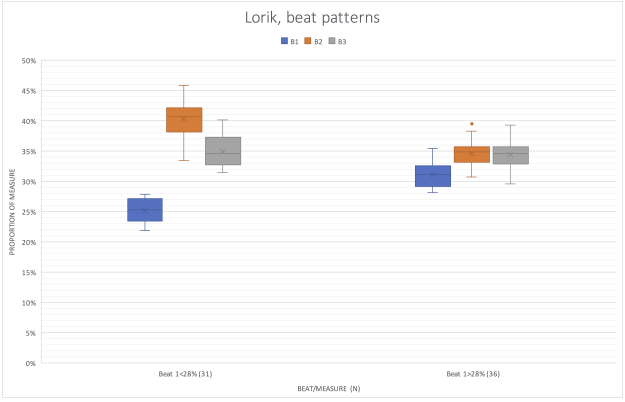


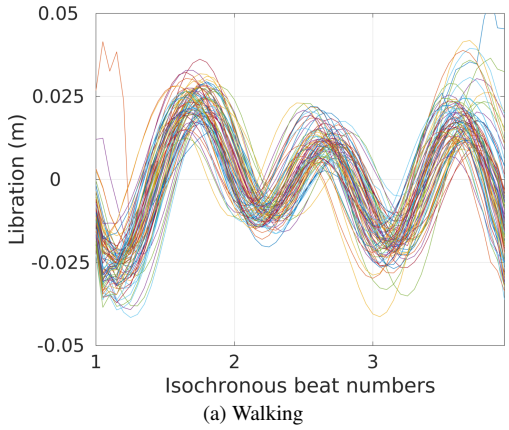
Figure 4: Beat proportions in Lorikspolskan, divided in accordance with the notation, in two beat patterns.

setups. Both figures from walking display a steady pattern of three oscillation periods within each measure, with the deepest points close to the first and third beat, where the dancer is stepping with her full body weight on one foot. There is also a vertical libration on the second beat in the turning (Figure 5b), however, smaller than in the walking in Figure 5a. A possible explanation of this could be the dancer - when turning - putting more emphasis to the horizontal rotation of the body than to the vertical libration on the second beat. The libration patterns to the clapping in Figure 5 are more coherent, while the patterns of dancing to the recording in Figure 6 are more varied and shifted in time. In Figure 5a, where the dancer is moving to the clapping, the second minima of the librations are slightly after the marked isochronous beat, whereas in the dancing to the recording Figure 6a the second minima are partly before the isochronous beat, which could indicate that the dancer relates to the variations in asymmetric beat patterns. To further address the question of the correlation between libration patterns and asymmetric beat variation, the recordings with live playing will be analysed in future work.

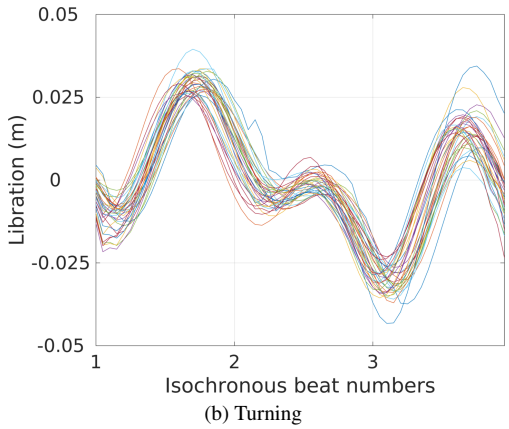
4.3 Interviews with performers

During the recording sessions and when watching their performances in the Motive software [7], dancers and musicians were asked to comment on their performance. The performers expressed that watching their performance, using the possibilities to view from different angles and follow marker movements by assigning tails to the markers, was highly interesting, exciting and informative. Furthermore, watching performers as skeletons in the graphical rendereding instead of watching a conventional film made it easier to focus on the movement. This triggered reflections and comparisons with concepts for describing and thinking about movements. In specific the observations of the variations in Figure 6a and Figure 6b correlate with the interview statement by Dregelid (2018) that "her experience was that she was not dancing with him (Gössa Anders)". At the time of the recording Dregelid said it might have been easier if Gössa Anders would have been

[7] http://optitrack.com/products/motive/



(a) Walking



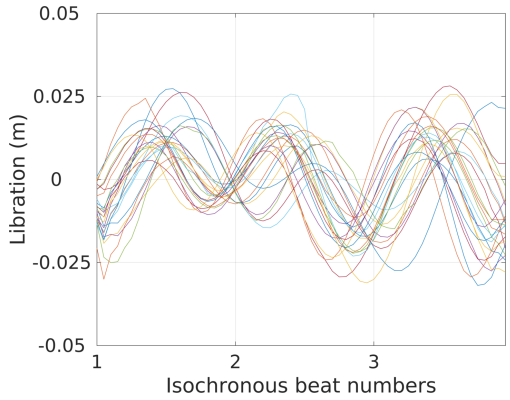
(b) Turning

Figure 5: Libration patterns from the dancing to a clapped beat.

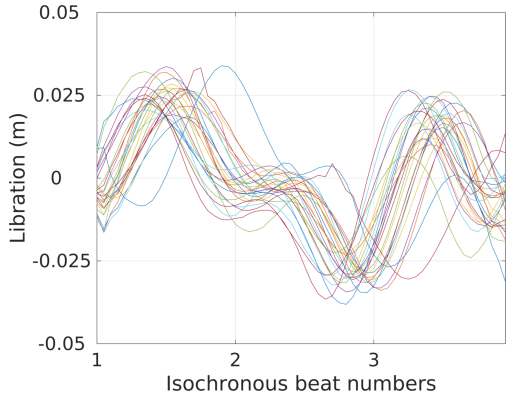
present in the room, so she could have watched his body movements while dancing. Further comparisons with the live recordings that contain musicians and dancers in interaction would be needed to address the question of how the dance movements correlates with non-isocronous beat patterns in realistic settings.

5. CONCLUSION

The accurate description of asymmetric beat patterns in polska music is a challenging task that may not be sufficiently addressed through the analysis of historical recordings. Recording performances with Motion Capture is assumed to allow for a more detailed analysis, taking into account bowing movements, bowing patterns and foot tapping. Using Motion Capture to plot libration patterns in the dancing is proposed as a method for examining how dancers relate to asymmetric beat patterns. The results indicate that (1) the libration patterns are consistent in shape inside the different sections of the dance by which they are assumed relevant for examining how dancers relate to musical beats. (2) The material confirms the variation in asymmetric beat patterns in this polska style. Including more recordings and the recordings with the musicians' interpretations of the same tunes is assumed to add to the presented findings on metrical characteristics. (3) The timing variations in the libration curves obtained from dancing to a recording with asymmetric beat suggest comparisons with the recordings of dancing to a live musician. A larger



(a) Walking



(b) Turning

Figure 6: Libration patterns from the dancing to a recording.

analysis of the correlation of musical events with dance movements will facilitate a more detailed understanding of the relation between dance and music in the particular style. Using MoCap recording in a performer-participating setting seems from this first attempt rewarding and worth exploring further as a method for studying performance practice. In conclusion, the presented study is suggested as a starting point for further explorations of music and dance interaction in Swedish folk music.

6. REFERENCES

Ahlbäck, S. (1986). *Tonspråket i äldre svensk folkmusik*. Stockholm: Udda Toner.

Ahlbäck, S. (1989). Metrisk analys - en metod att beskriva skillnaden mellan låttyper. *Norsk folkemusikklags skrifter*, (4).

Ahlbäck, S., Berndalen, P., Hjalmarsson, J., Marsden, B., Misgeld, O., Rosenberg, S., & Willman, P. (2009). *1976-2006, 30 år med folkmusik på KMH*. Stockholm: KMH.

Ahlbäck, S. (2003). About Asymmetrical Beat in the Polska. In M. Ramsten (Ed.), *The Polish Dance in Scandinavia and Poland* (pp. 165–80). Stockholm: Svenskt visarkiv.

Andersson, G. A. (1950). Lorichspolskan, Polska efter Pellar Anna. recorded by Matts Arnberg.

Andersson, G. A., Andersson, G. A., & Orsa Spelmanslag (1998). *Historical recordings of Swedish Folk Music: Fiddle tunes from Orsa: 1*. Malung: Hurv.

Andersson, N. (1922). *Svenska Låtar, Dalarna 1*. Stockholm: P A Nordstedts.

Bakka, E. (2014). Svikt, meter, force and weight. *downloaded from www.academia.edu/6649805/Svikt_meter_force_and_weight*.

Blom, J.-P. (1993). Rytme og frasering - forholdet til dansen. In B. Aksdal & S. Nyhus (Eds.), *Fanitullen. Innføring i norsk og samisk folkemusikk.* (pp. 161–184). Universitetsforlaget.

Dregelid, A. (2018). interview. conducted by Olof Misgeld.

Forslund, K. E. (1921). *Med Dalälven från källorna till havet, Öster-Dalälven Orsa och Våmhus (del1 bok 4)*. Stockholm: Åhlén & Åkerlunds förlag.

Groven, E. (1971). Musikkstudiar - ikkje utgjevne for 1. Rytmestudiar. In O. Fjalestad (Ed.), *Eivind Groven. Heiderskrift til 70-årsdagen* (pp. 93–102). Noregs Boklag.

Gustafsson, M. (2016). *Polskans historia : En studie i melodityper och motivformer med utgångspunkt i Petter Dufvas notbok*. PhD thesis, Lund.

Haugen, M. R. (2014). Studying rhytmical structures in norwegian folk music and dance using motion capture technology: A case study of norwegian telespringar. *Musikk og Tradition, Tidsskrift for forskning i folkemusikk og folkedans*, (28).

Haugen, M. R. (2017). Investigating periodic body motions as a tacit reference structure in norwegian telespringar performance. *Empirical Musicology Review*, *11*(3-4), 272–294.

Johansson, M. (2009). *Rhythm into style: studying asymmetrical grooves in Norwegian folk music.* Ph.d. thesis, University of Oslo.

Johansson, M. (2017). Empirical research on asymmetrical rhythms in scandinavian folk music: A critical review. *Studia Musicologica Norvegica*, *43*(01), 58–89.

Kvifte, T. (1999). Fenomenet ”asymetrisk” takt i norsk og svensk folkemusikk. *Studia musicologica norvegica*.

Musica Svecia (1995). *Folk music in Sweden: Folk tunes from Orsa and Älvdalen*. Stockholm: Caprice Records.

Musica Svecia (1999). *Folk Music in Sweden: Äldre svenska spelmän*. Stockholm: Caprice Records.

Mårds, T. (1999). *Svikt, kraft og tramp: en studie av bevegelse og kraft i folkelig dans.* Hovedfagsoppgave, Norges Idrettshøgskole, Oslo.

Naveda, L. & Leman, M. (2010). The spatiotemporal representation of dance and music gestures using topological gesture analysis (tga). *Music Perception: An Interdisciplinary Journal*, *28*(1), 93–111.

Nilsson, M. (2017). *The Swedish Polska*. Stockholm: Svenskt visarkiv/Musikverket.

Norman, I., Egler, L., Arnberg, M., & Sälgström, G. (2000). *Orsapolska 1947-2000*. Rättvik: Folkmusikens hus.

Övergaard, E. & Ramsten, M. (1982). *Einar Övergaards folkmusiksamling*. Sv. visarkiv i samarbete med Dialekt-och folkminnesarkivet i Uppsala.

Sandvik, O. M. (1943). *Østerdalsmusikken*. Oslo: Tanum.

MODELING SONG SIMILARITY WITH UNSUPERVISED LEARNING

Matevž Pesek, Manca Žerovnik, Aleš Leonardis, Matija Marolt
University of Ljubljana, Faculty of Computer and Information Science
{firstname.lastname}@fri.uni-lj.si

ABSTRACT

The SymCHM, which was developed for the pattern discovery, was applied to the music similarity task. It was used as a feature extractor, unsupervisedly learning repeated patterns on pitch-time representation, eliminating any additional high-level information. The output was used in the retrieval, where the model achieved 74.4 % classification accuracy on the Dutch folk dataset. By the unsupervised aspect of model's training and the ability to perform similarity using only the most basic song representation, we find the results sufficient to further explore the use of the model on datasets with a low number of additional features and basic music representations.

1. INTRODUCTION

The concept of similarity in music has been studied in different research areas. The similarity in cognition plays a significant role in psychological accounts of problem solving, memory, prediction, and categorization (Holyoak & Morrison, 2005). Many research topics in musicology are inherently related to similarity and categorization in music, e.g. the study of motivic-thematic relations, comparison of musical motifs, categorization of songs into tune families and many others (Volk & Van Kranenburg, 2012). Understanding of the basic processes underlying perception of musical similarity is necessary for acquiring a deeper comprehension of music perception in general (Toiviainen, 2007).

The categorization of folk songs into tune families, where a tune family represents "a set of folk songs which have a common origin in history" (Bayard, 1950), also relies on the similarity. Many current approaches for this task rely on alignment algorithms. Mongeau & Sankoff (1990) were one of the first to use alignment algorithms for music, establishing the basis for several future approaches, employing the alignment algorithms and profile modeling for classification and retrieval tasks, e.g. using the pop and rock songs datasets Bountouridis & Van Balen (2014). Walshaw (2017) investigated enhancements of the well-established local alignment algorithms to also classify Dutch folk songs into tune families.

Bountouridis et al. (2017) used biologically-inspired techniques for MIR tasks. They identified several shared concepts between music and bioinformatics, such as melody (DNA), oral transmission (evolution), variations (homologues), tune families (homology) etc. and showed that bioinformatics algorithms are suitable for MIR tasks. (Savage & Atkinson, 2015) also used an adapted alignment algorithm from the field of bioinformatics to classify songs into four diverse tune families (two English, two Japanese).

Several developed approaches were evaluated on the Dutch folk song dataset compiled by Van Kranenburg et al. (2013). Among the most recent, the alignment approach by Van Kranenburg et al. (2016) produces the best classification accuracy on the Dutch folk song dataset. The approach models various features of music as substitution scoring functions, which are incorporated into the Needleman--Wunsch-Gotoh Gotoh (1982) algorithm. The model employs several 'viewpoints', such as pitch, duration, score time, time in bar, onset, current bar number, current phrase number, upbeat, current meter, free meter, accented, inter-onset-interval ratio, normalized metrical weight and the time position within phrase. Van Kranenburg had analyzed combinations of these attributes and had discovered the best results were given by using the pitch and position within phrase attributes. Despite the high accuracy, it requires a considerable amount of time to produce such attributes for each dataset. Consequently, the majority of these attributes are usually not available in music collections. To eliminate the need for expert knowledge, Velarde et al. (2013) classified Dutch songs using Haar-wavelet filters. The results are not on par with Van Kranenburg et al. (2013), but the approach does not require any encoded expert knowledge.

In this paper, we explore how unsupervised learning can be used for modeling tune similarities and classification into tune families. Specifically, we study the compositional hierarchical model that has been previously applied to several audio-based tasks, such as chord estimation and polyphonic transcription (Pesek et al., 2017a) and pattern discovery (Pesek et al., 2017b), using a modified symbolic version of the model (SymCHM).

2. METHODOLOGY AND EXPERIMENT

The compositional hierarchical model has been previously applied to several tasks, including spectral-oriented tasks of multiple fundamental frequency estimation (Pesek et al., 2017a) and automated chord estimation (Pesek et al., 2014), and symbolic-oriented tasks of pattern discovery (Pesek et al., 2017b). A model able to perform on symbolic music representations, denoted SymCHM will be used to tackle the tune family classification problem in this paper.

2.1 Model

The idea behind SymCHM lies in the organizing of frequently co-occurring events into compositions. Starting at its input, the model observes the statistics of events' pres-

ence and the relation between them. For example, if two events frequently co-occur in a given time window on a specific interval, both events can be joined into a composition. The composition is relatively encoded, meaning, should two events co-occur at one pitch location and again at a different one, the same composition would be formed. This procedure is repeated on consecutive layers. In contrast to the first layer, instead of observing the input, the model observes co-occurrences of compositions and forms new relatively encoded compositions on the next layer, based on the previous layer. In the SymCHM, we name all compositions *parts*, similar to nodes in other models.

Since each part may occur at several locations in a single input, such occurrences must be defined by its location in time and pitch (a single part may occur at two different pitch heights at the same time). We denote such occurrences *activations* and define their position by their time, and pitch. The parts learned by the model can be observed as melodic patterns and their activations as pattern occurrences.

Once the model is built, it can be inferred over another (or the original input). The inference may be exact or approximate, where in the latter case biologically-inspired hallucination and inhibition mechanisms enable the model to find variants of part occurrences with deletions, changes or insertions, thus increasing its predictive power and robustness. The hallucination mechanism provides means to activate a part even when the input is incomplete or changed. In symbolic music representations, such changes often occur in melodic variations and ornamentation. The hallucination enables the model to robustly identify patterns with variations. The inhibition mechanism is also essential in the SymCHM for removal of redundant co-occurrences. As the model does not rely on any musicological rules, parts may produce a large number number of competing patterns. Inhibition may be used to reduce the number of activations and find the patterns that best correspond to the learned hierarchy.

The SymCHM therefore learns a hierarchical representation of patterns occurring in the input, where patterns encoded by parts on higher layers are compositions of patterns on lower layers. The inference produces part activations which expose the learned patterns (and their variations) in the input data. Shorter and more trivial patterns naturally occur more frequently, longer patterns less frequently. On the other hand, longer patterns may entirely subsume shorter patterns.

2.2 Experiment

We tackled the melody classification using the SymCHM. The MTC-ANN annotated dataset [1] was used. The dataset consists of 360 Dutch folk songs, accompanied by tune family annotations. Similar to the experiment presented in Van Kranenburg et al. (2013), we classified the folk tunes in to tune families using features. To gather the features, we employed the symbolic version of the compositional hierarchical model for pattern discovery for this task. The SymCHM, shown in Fig. 1, was presented by Pesek et al. (2017b) and was evaluated for the MIREX discovery of repeated patterns and sections task.

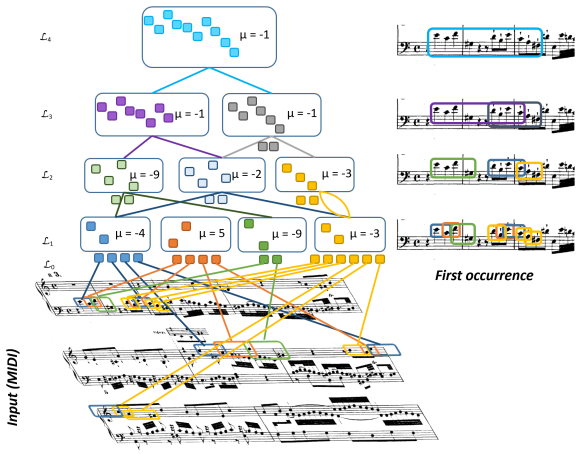


Figure 1: An abstraction of the SymHM's performance over a symbolic music representation. Each part has multiple activations (the number of activation is expressed under each part). Composing parts can partially overlap. While composing, the parts are joined into a composition by a relative offset, represented by μ. The activations for the first pattern for the selected example are shown on the right side per layer.

The SymCHM can be used for intra-opus task, such as the mentioned MIREX task, by building a model for each music piece individually. The model is first built and later inferred on a single symbolic representation. In the intra-opus task, the statistical drive behind the model's building procedure reflects the frequency of co-occurrences within a single music piece. It is therefore difficult to compare the learned patterns due to the difference in models' hierarchies, when comparing multiple music pieces. The music similarity task is, on the other hand, an inter-opus task. We have therefore built a single model on several songs. The compositions therefore still reflect the frequent proximity of events (i.e. pattern occurrences) within each single music piece, but is also regulated by the frequency of such occurrences across the given input dataset. The model accepts multiple inputs separately and analyzes them piece by piece. The statistical nature in the model is invariant to the length or the number of inputs—it calculates the co-occurrence of events within a single input and produces compositions of such events. If a similar co-occurrence of events occurs in another input, its statistic is added to the existing composition reflecting the occurrence.

Any symbolic music representation with the following two features is accepted as an input to the model: as a set of note onsets (e.g. in seconds) and note pitches (e.g. MIDI pitch). MIDI format may be used, extracting only these two attributes; all other attributes, which can be extracted from the MIDI format (e.g. meter, bar, phrase number

[1] Dataset accessible here: `http://www.liederenbank.nl/mtc/`

etc.), are discarded.

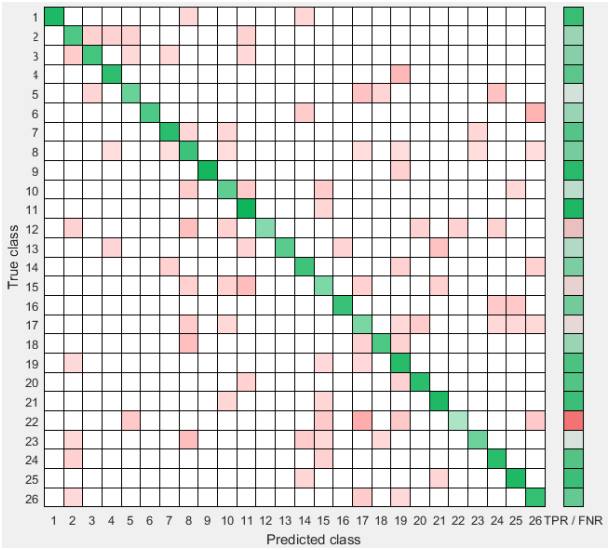


Figure 2: The confusion matrix of the tune family classification. The reference annotations are represented in rows (left) and the predicted classes in columns (bottom).

The model was built on the 360 songs of the MTC-ANN dataset. The songs are categorized into 26 tune families. No annotations were used during the training. The built model produced a list of discovered patterns across the input songs. The output was encoded into a feature vector where each part represented a vector element and its value represented the sum of activations for the represented part. The output was encoded into a feature vector where each part was mapped onto a vector element, and the value represented the sum of the part's activations, as described in the first experiment. The model generated 3750 parts across layers 3–7. The vector values were adjusted as described in (Van Kranenburg et al., 2013). For each element, the values were scaled to have zero mean and standard deviation of 1. As described by (Van Kranenburg et al., 2013), the cosine distance was used for comparison of vectors. The result of the vector comparison was a 74.4 % classification accuracy. The confusion matrix is depicted in Figure 2.

3. CONCLUSIONS

The results are about 20 percent lower when compared to Van Kranenburg et al. (2013). However, we believe the results are interesting, considering the fact that a pattern discovery model, relying only on onset-pitch notation, was used for this task. The model was not specifically trained or parameter-tuned for this task and was applied to the dataset without any dataset-specific adjustment. The model provided compositions of relatively-encoded melodic patterns learned in an unsupervised manner. In contrast to several approaches applied to this dataset, no know-how about the dataset or folk and western music in general was used in the procedure of patterns which were used for classification. Nevertheless, such incorporation could also be beneficial to the proposed model's results and will therefore be further explored in our future work.

4. REFERENCES

Bayard, S. P. (1950). Prolegomena to a study of the principal melodic families of british-american folk song. *The Journal of American Folklore*, *63*(247), 1–44.

Bountouridis, D., Brown, D. G., Wiering, F., & Veltkamp, R. C. (2017). Melodic similarity and applications using biologically-inspired techniques. *Applied Sciences*, *7*(12), 1242.

Bountouridis, D. & Van Balen, J. (2014). The cover song variation dataset.

Gotoh, O. (1982). An improved algorithm for matching biological sequences. *Journal of molecular biology*, *162*(3), 705–708.

Holyoak, K. J. & Morrison, R. G. (2005). *The Cambridge handbook of thinking and reasoning*. Cambridge University Press.

Mongeau, M. & Sankoff, D. (1990). Comparison of musical sequences. *Computers and the Humanities*, *24*(3), 161–175.

Pesek, M., Leonardis, A., & Marolt, M. (2014). A compositional hierarchical model for music information retrieval. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, (pp. 131–136)., Taipei.

Pesek, M., Leonardis, A., & Marolt, M. (2017a). Robust Real-Time Music Transcription with a Compositional Hierarchical Model. *PLoS ONE*, *12*(1).

Pesek, M., Leonardis, A., & Marolt, M. (2017b). SymCHM—An Unsupervised Approach for Pattern Discovery in Symbolic Music with a Compositional Hierarchical Model. *Applied Sciences*, *7*(11), 1135.

Savage, P. E. & Atkinson, Q. D. (2015). Automatic tune family identification by musical sequence alignment. In *Proceedings of the 16th ISMIR Conference*, volume 163.

Toiviainen, P. (2007). Discussion Forum 4A - Editorial. *Musicæ Scientiæ*, *1*(1), 3–6.

Van Kranenburg, P., Janssen, B., & Volk, A. (2016). The meertens tune collections: The annotated corpus (mtc-ann) versions 1.1 and 2.0. 1.

Van Kranenburg, P., Volk, A., & Wiering, F. (2013). A Comparison between Global and Local Features for Computational Classification of Folk Song Melodies. *Journal of New Music Research*, *42*(1), 1–18.

Velarde, G., Weyde, T., & Meredith, D. (2013). Wavelet-filtering of symbolic music representations for folk tune segmentation and classification. In *Proceedings of the Third International Workshop on Folk Music Analysis (FMA2013)*, (pp.5̃6).

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.

Walshaw, C. (2017). Tune classification using multilevel recursive local alignment algorithms.

PERCEPTION OF ASYMMETRIC RHYTHMS IN TRADITIONAL GREEK MUSIC

Marcelo Queiroz[†1], Katerina Peninta[2], Roberto Piassi Passos Bodo[1], Maximos Kaliakatsos-Papakostas[2], and Emilios Cambouropoulos[2]

[1] *Computer Science Department, University of São Paulo, Brazil,* {mqz,rppbodo}@ime.usp.br
[2] *School of Music Studies, Aristotle University of Thessaloniki, Greece,* {kpeninta,maxk,emilios}@mus.auth.gr

ABSTRACT

This paper explores aspects of rhythmic perception within the context of traditional Greek music, more specifically Demotika songs, which display a rich variety of asymmetric rhythmic patterns, *i.e.* patterns comprising beats of different durations. A listening experiment with volunteer university music students was conducted, in order to investigate basic questions regarding timing accuracy and meter structure as perceived by subjects. This study suggests that identifying accurately rhythmic meter patterns in traditional Greek music is not an easy task, even among Greek music students, although statistically significant differences may be observed depending on cultural background. Statistical analysis also reveals correlations between elements associated with the difficulty of the task, such as the degree of agreement between participants, the response times and the number of times each excerpt was heard, and musical aspects such as tempo, meter structures and symmetry/asymmetry of rhythms.

1. INTRODUCTION

Musical time is commonly organized around a hierarchic metrical structure, having as most salient metric level the beat level, also referred to as tactus (Lerdahl & Jackendoff, 1983). Most western musics assume an isochronous beat level, and divergences from isochrony are treated as exceptions or special cases. In traditional musics from the Balkan and Middle-East, on the other hand, rhythms commonly feature non-isochronous metric structures, referred to as *additive* or *aksak* or *asymmetric meters* (Fracile, 2003; Moelants, 2006). Such metric structures are based on asymmetric beat levels comprising repeating asymmetric patterns of long and short beats at a 3:2 temporal ratio, such as 5/8 (3+2), 7/8 (3+2+2), 8/8 (3+3+2) and so on; this asymmetric beat level stands between a lower isochronous sub-beat level and a higher isochronous metric level (Cambouropoulos, 1997). Enculturated listeners spontaneously use an asymmetric tactus to measure time (clapping hands, tapping feet), since this is presumably the most plausible and parsimonious way to organize given rhythmic stimuli from specific musical idioms. In a sense, asymmetric beat structures organize time similarly to how asymmetric pitch scales organize pitch/tonal spaces (Fouloulis et al., 2012, 2013).

[†]The first author acknowledges the support of FAPESP grant 2014/25686-5 and CNPq grant 309645/2016-6.

The main driving question behind this study is: do listeners with different backgrounds perceive asymmetric rhythms the same way in terms of beat structure and beat accents? This question entails follow-up questions such as: are there differences in perceptual timing accuracy and perceived beat ordering due to enculturation? The interest in these questions is not restricted to analytical or psychophysical considerations; empirical data may also aid in developing novel Music Information Retrieval methods accounting for both beat asymmetry and subjectively perceived accents (one such method is proposed in Fouloulis et al. (2012)).

The main goal of this paper is to present and discuss experimental data of a listening test dealing with actual examples of traditional Greek music and the rhythmic patterns used to represent their metric structure. This classification experiment was conducted with volunteer university-level music students in Greece, in Brazil and in several other countries. Through statistical analysis of the collected answers it is possible to have an idea of the difficulty of this classification task as a function of the types of patterns encountered in traditional Greek music, of musical aspects such as tempo, harmony or instrumentation, and of the cultural background of the participants.

Recent work dealing with Greek rhythmic patterns include Fracile's study of aksak structures in Balkan folklore (Fracile, 2003), mapping the occurrences of their most common forms, Moelants' discussion of the influence of tempo in the ratio between long and short beats during performance of aksak metres (Moelants, 2006), and Fouloulis, Pikrakis and Cambouropoulos' investigation on automatic beat-tracking systems when confronted with asymmetric repertoire (Fouloulis et al., 2012, 2013), where basic implementation premises, such as the existence of a steady pulse, fail. In an experimental study motivated by questions similar to ours, Tekman et al. (2003) compared listeners familiar and unfamiliar with musical idioms that frequently use asymmetric meters in a recognition task, in which musicians and non-musicians classified pairs of symmetric/asymmetric/irregular meter structures as *same* or *different*; their results did not provide support for the existence of schematic representations of asymmetric meters. Our work, on the other hand, proposes an identification task for musicians, classifying perceived

rhythms according to formal symbolic music notation, and investigates how this task is affected by meter structure, tempo and cultural background.

The next section brings a broad overview on the types of meter structure frequently found in traditional Greek music. Section 3 presents the experimental methodology and Section 4 brings experimental results and their discussion. Conclusions and open questions for future work are presented in Section 5.

2. ASYMMETRIC RHYTHMIC METERS IN TRADITIONAL GREEK MUSIC

Greek traditional music is a very important part of Greek education. Lullabies and cradle songs, chants of toddlers, nursery rhymes, carols for seasons greetings and even the sneering and satyric songs of carnival reflect popular customs throughout Greece, and offer important material for all stages of education. Having said that, it is not evident whether non-Greek listeners unfamiliar with Greek music would have a harder time (compared to Greek musicians) figuring out asymmetric meter structures found in ordinary Greek dances, such as *Kalamatianos* (a 7/8 measure of the form 𝄆 ♩. ♩ ♩ 𝄇 or 3+2+2) or *Karsilamas* (a 9/8 measure of the form 𝄆 ♩ ♩ ♩ ♩. 𝄇 or 2+2+2+3). It is also not evident whether any listener, Greek or non-Greek, would judge the longest beat of the former to be in the first position of the pattern (as **3**+2+2) whereas in the last position for the latter (2+2+2+**3**). Any asymmetric meter structure gives rise to rotated alternatives (*e.g.* 2+2+3 or 3+2+2+2) that might be perceived as more fitting for a particular music piece, according not only to beat durations, but other aspects such as musical dynamics (energy), positions of instrumental and vocal entries, articulations, etc.

Another important aspect of meter structures in this context is the fact that actual musical instances consist of several instruments playing varying rhythmic patterns, and not unusually alternating between patterns that fit several meter structures. As a simple example, a binary 4+4 (rhythmically equivalent to 2+2) would easily accommodate instrumental lines playing 4+2+2 and 3+3+2 and any of their rotations, in fact any other pattern that adds up to 8 eighth notes. The same applies to 6+4 (rhythmically equivalent to 3+2), which accommodates 2+4+2+2 and 3+3+2+2, among others. Saying that a piece adhere to the style of a certain dance or corresponds to a certain meter structure does not entail that all instruments will play homorhythmically, and thus listeners will judge the perceived meter structure according to subjective (and possibly unconscious) criteria.

3. EXPERIMENTAL METHODOLOGY

In a nutshell, the experimental session here proposed consists of a simple questionnaire with audio excerpts of Demotika songs, followed by a number of alternatives in music notation for representing them. Each experimental subject listens to each excerpt and chooses the rhythmic pattern that best matches the perceived meter structure.

3.1 Selection / Formatting of Audio Excerpts

Songs were included that reflect the diversity of rhythmic patterns found in traditional Greek music (and particularly Demotika songs). An attempt has been made to balance the selection across meter structures, with a varied palette within each rhythmic style (*e.g.* excerpts displaying different tempi, instrumentation, rhythmic ornaments, etc). In order to keep the difficulty of the task within a reasonable level, only excerpts with rhythmic patterns of up to four beats per measure were be included. Many other interesting and much more complicated rhythmic patterns surely exist in traditional Greek music, but exploring this repertoire is the subject of future work. We aimed at keeping the session at about 15 minutes, which allowed the inclusion of 30 excerpts of 30-seconds each. The selected songs are presented in Table 1 along with their corresponding rhythmic patterns [1] .

3.2 Selection of Rhythmic Patterns

In order to give the subject a wide range of options for the rhythmic representation of the meter structure of each song, a dictionary of meter structures was built by taking all the ground-truth formulae in Table 1, including all possible rotations of each asymmetric pattern, and also all patterns obtained from the above by replacing each long beat (a dotted quarter note) with a half note (*e.g.* 2+2+3 would give rise to a 2+2+4 alternative). This produced a set of 27 possible responses for each excerpt, classified according to the number of beats (one to four beats per measure).

3.3 Selection of Subjects / Volunteers

Due to the technical nature both of the listening task as well as the notation used to represent meter structure, participation in this experiment was restricted to trained musicians. This is not supposed to mean that such an experiment is impossible to conduct with non-musicians but it is understood that the training effort that would be necessary to allow non-musicians to express their rhythmic perceptions using a formal and quantitatively accurate music notation would jeopardize the feasibility of the experiment.

In order to reduce the heterogeneity of the population, the experiment was addressed at university-level (undergraduate and graduate) music students. On the one hand this is an audience that is used to

[1] For each of the songs, a 30-second excerpt was produced by cutting an arbitrary portion of the audio signal (between 1:00 and 1:30 from the beginning) and using 100ms fade-in and fade-out ramps, in order to ensure that each excerpt would start in an apparently random position (relative to the beginning of a measure).

Table 1: Dataset used in the experiment. The first column corresponds to the annotated ground-truth, and the second column displays the track information for each excerpt included. Excerpt #22 alternates between 2+2 and 3+3+2, having both as ground-truths.

Pattern	Style	Song/Artist
2+2, ‖: ♩ ♩ :‖	Syrtos/Sta Tria	#1=Perdika (Anastasios Xalkias), #2=Ta Matia ta Glika (Georgia Mittaki), #3=Kotsarin (Tsimaxidis), #22=Vlaxa Paei Gia Th Stani (Georgia Mittaki)*
2+3, ‖: ♩ ♩. :‖	Baiduska	#4=Vasilarxontissa, #5=At Xavasi, #6=Mia Kali Geitonopoula (Elina Papanikolaou)
3+2, ‖: ♩. ♩ :‖	Zagorisios/Tik	#7=Sou'pa Mana Pantrepse Me
3+3, ‖: ♩. ♩. :‖	Zonaradikos	#8=Aleksis Andriomenos (Elina Papanikolaou), #9=Vasil'kouda (Xronis Aidonidis)
2+2+2, ‖: ♩ ♩ ♩ :‖	Tsamiko	#10=Enas Aetos Kathotane (Giorgos Papasideris), #11=Pame sti Roumeli (Maikantis), #12=Davelis (Georgia Mittaki)
2+2+3, ‖: ♩ ♩ ♩. :‖	Mantilatos	#13=Hicaz Mantilatos, #14=Siko Koukounouda M', #15=Serenitsa
3+2+2, ‖: ♩. ♩ ♩ :‖	Kalamatianos	#16=To Papaki (Giorgos Papasideris), #17=M'Ekapses Geitonissa (Elena Maggel), #18=Pou Eisai Lenio Den Fainesai (Maikantis), #19=Tromaxton Laziko Xorontikon (Kemetze/Gkogkotsis)
2+3+3, ‖: ♩ ♩. ♩. :‖	Berati	#20=Berati
3+3+2, ‖: ♩. ♩. ♩ :‖	Nisiotikos Syrtos / Ballos	#21=Ta Ksila, #22=Vlaxa Paei Gia Th Stani (Georgia Mittaki)*, #23=Na'xa Ena Milo Na'rixna (Miltos Stanos), #24=To Paneigiri (Giorgos Papasideris)
2+2+2+3, ‖: ♩ ♩ ♩ ♩. :‖	Karsilamas / Zeibekikos	#25=Vasilepsen Avgerinos (Xronis Aidonidis), #26=Katsivelikos, #27=To Enteka, #28=Sfarlis
2+3+2+2, ‖: ♩ ♩. ♩ ♩ :‖	Argilamas	#29=Manio (Xaris Aleksiou), #30=Dimitroula Mou

take rhythmic perception tests, so that this experiment would not seem so out-of-the-ordinary (apart from the repertoire which might be unfamiliar); on the other hand, this would reduce the number of senior musicians and/or novices, which would have to be analysed as separate groups.

3.4 Interface and Experimental Session

The interface for the experiment (available in English, Greek and Portuguese) was implemented in PHP using SQL for accessing the database, and was made available on the Internet using a dedicated server. The entry page contained an explanatory text about the nature of the experiment and a consent term, along with a small form for personal data (name, email, institution and nationality, only to be used anonymously). The experimental session proceeded through 30 pages, one for each excerpt, as illustrated in Figure 1.

For each user the presented excerpts would follow a randomized order defined when the experiment begins. This is done to minimize the fatigue effect as well as other order-related effects. In a drop-down menu there was also an option labeled "I cannot decide", to let the user skip an excerpt and continue the experiment.

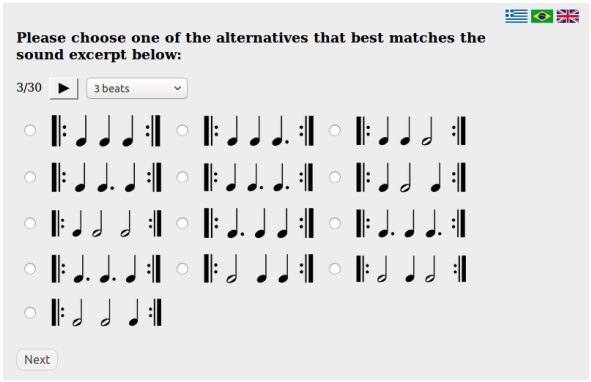


Figure 1: Interface of the experiment.

4. RESULTS AND DISCUSSION

The experiment was available between January 9th and February 9th, 2018, and had 56 participants (36 Greeks and 20 from other nationalities). The raw results of the experiment consist of a database that relates participants, nationalities, excerpts and meter structures. One of the first issues that must be addressed when dealing with these results is to define what is considered a correct answer. As discussed in the concluding paragraph of Section 2, many possible variations of a given notated pattern are musically meaningful and may be considered differences of interpretation rather than errors. Therefore, for a

given ground truth, all answers that add up to the same amount of eighth notes [2] are considered equally valid/correct. These include all rotations of a given pattern (*e.g.* 3+2+2, 2+3+2 and 2+2+3), but also different patterns that could be easily superimposed, such as 2+2+2, 4+2 and 3+3. In the sequel, a series of analyses of different aspects of the experiment are considered.

When we consider the number of valid answers each participant produced (the participant's *score*), several relevant statistics may be extracted. The average of the score distribution is $\mu = 16.7$ valid answers per participant (out of 30), and the standard deviation is $\sigma = 7.4$; this sample, however, does not pass the D'Agostino and Pearson's normality test (the null hypothesis corresponding to the sample coming from a normal distribution is rejected with $p = 0.0002$), which renders such a Gaussian description a very poor statistical model for this data. Figure 2 displays the same data separated in two groups according to nationality (Greeks and non-Greeks). These two subpopulations pass the aforementioned normality test ($p = 0.0938$ for Greeks and $p = 0.2656$ for non-Greeks) and may be reasonably modeled as Gaussian distributions. Valid answers per participant amount to 18.6 ± 6.4 for Greeks and 13.4 ± 7.8 for non-Greeks; moreover, these averages may be considered significantly different ($p = 0.0104$ for a T-test, $p = 0.0230$ for a Kolmogorov-Smirnov or KS test). This difference may be explained due to enculturation of Greek participants, *i.e.* their lifelong immersion in a culture where such music is commonly heard. Yet it cannot pass unnoticed that, based on these numbers, the task does not appear to be easy even for Greek music students, as might be otherwise presumed.

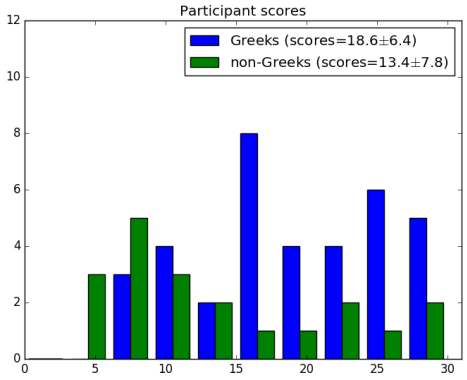


Figure 2: Number of valid answers given by each participant, grouped according to nationality.

The number of valid answers per excerpt may also be studied according to the groups defined above. Figure 3 displays these values separated for Greeks and non-Greeks. The two samples formed by these percentages may be compared through their Pearson's correlation coefficient $r = 0.5374$ ($p = 0.0022$), indicating that they do behave with a certain degree of correspondence (when the blue bar rises the green bar also tends to rise, and vice-versa). Both T and KS tests indicate that the averages of these two samples (62% of valid answers for Greeks and 44.5% for non-Greeks) are significantly different ($p = 0.0012$ for T test and $p = 0.0046$ for KS), reconfirming the observation that enculturation may render the task somewhat easier for Greeks, but not so much, as several counterexamples may be found (excerpts #1, #5, #7, #14, #20 and #22).

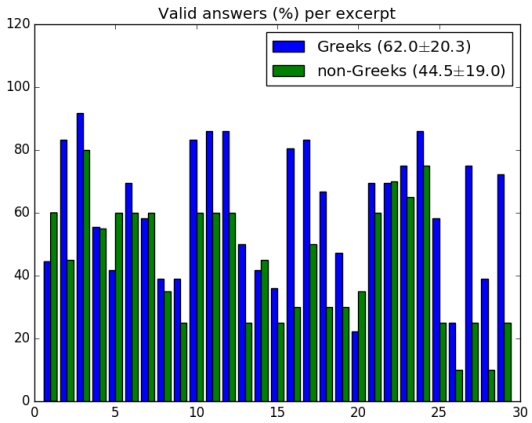


Figure 3: % of valid answers for each group / excerpt.

Considering specific excerpts with exceptionally low values of valid answers, a few patterns emerge. Excerpts #8 and #9 are symmetric compound binary pieces (3+3), both of which received more answers of type 2+2 than 3+3; these are binary meter structures that differ on the sub-beat representation level. It might be the case that some students were assuming the sub-beat level would use tuplets (which theoreticaly represent exception rather than rule), or else these are differing opinions on which were the metric and the tactus levels (phrases for these fast excerpts did comprise two measures). Excerpts #13, #14 and #15 corresponded to the structure 2+2+3, a rotated form of the popular Kalamatianos 3+2+2 that appeared in excerpts #16-#19, a well-known structure for Greek participants; invalid answers for these excerpts included 2+2, 4+2 and 2+2+4, all of which would imply an identification of the long-to-short-beat ratio of the form 2:1 rather than 3:2. Excerpt #20 (annotated 2+3+3, a rotated form of the popular Ballos 3+3+2) received the largest number of "I cannot decide" answers (35.7%), possibly due to its very slow tempo (more on tempo below). Finally, the quaternary asymmetric structures 2+2+2+3 and 2+3+2+2 (Excerpts #25–#30) were harder for non-Greek participants (but #26 reached only 25% of valid answers among Greeks).

Figures 4 and 5 present the average response times (in seconds, per excerpt) and average play counts (how

[2] For this characterization to be formally well-defined one needs to consider a rhythmic sort of "octave equivalence", allowing 2+2 for instance to be considered equivalent to 4+4, or 2+3 and 3+2 to be equivalent to 4+6 and 6+4, respectively.

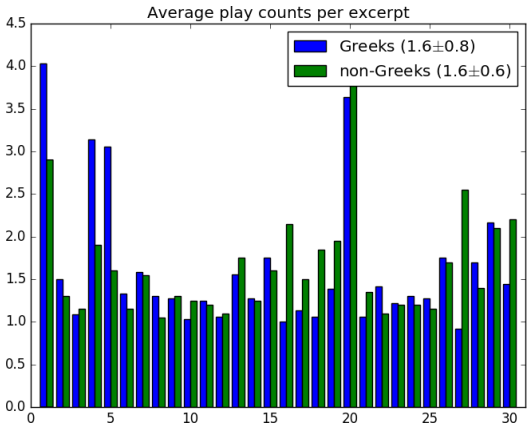


Figure 4: Average play counts for each excerpt.

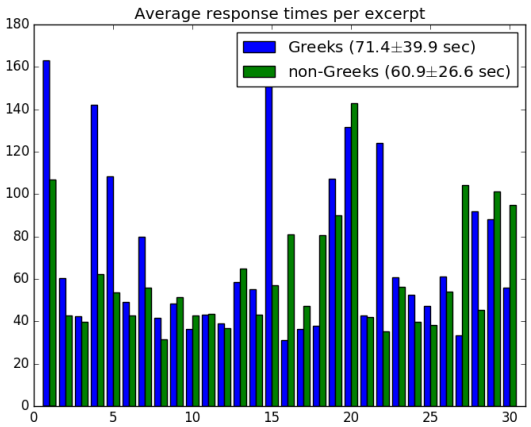


Figure 5: Average response times (s) for each excerpt.

many times in average each participant pressed 'play' for each excerpt), grouped according to nationality. Play counts as a function of the excerpt are highly correlated among Greeks and non-Greeks ($r = 0.6439$, $p = 0.0001$), and response times follow a similar trend ($r = 0.3600$, $p = 0.05071$). No significant differences between groups have been observed for these statistics ($p >> 0.1$ for T and KS tests in both cases). As should be expected, play counts and response times are highly correlated ($r = 0.8363$, $p < 0.0001$) and no relevant conclusions should be drawn from this correlation (it simply takes more time to hear an excerpt repeatedly).

Another indication of the perceptual difficulty of assigning a meter structure to each excerpt is the measure of *spread* of the answers obtained by each particular excerpt. The measure of spread adopted, also known as Gini-Simpson diversity index, is displayed in Figure 6 for each one of the 30 excerpts. This graph suggests that for a few excerpts a relatively high level of agreement (small spread) was obtained, such as excerpts #3, #16, #17 and #18. We have seen that play counts and response times are highly correlated, but less obvious are the correlations between play counts and spread ($r = 0.3915$, $p = 0.0324$), and between response times and spread ($r = 0.5216$, $p = 0.0031$) These observations might support the interpretation that play counts, response times and spread of answers have some underlying relationship with the difficulty of the task.

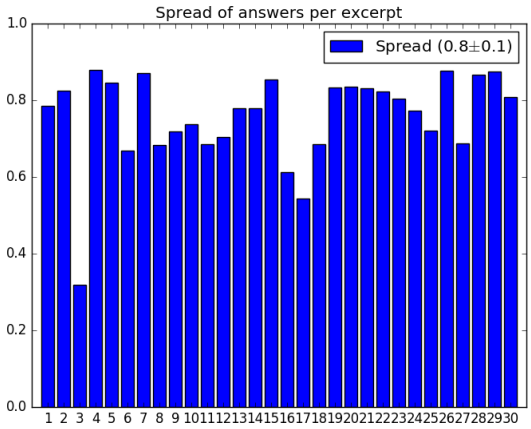


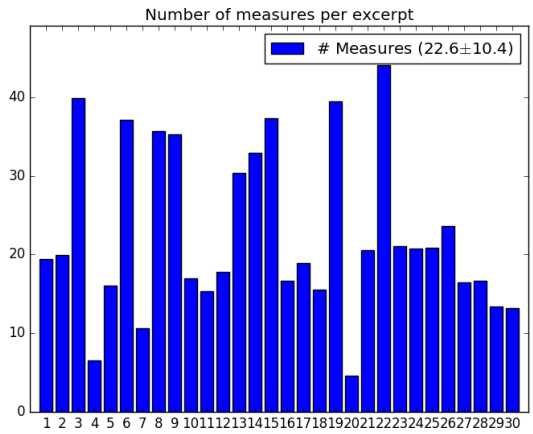
Figure 6: Measures of spread for each excerpt.

Participants observed an apparent relationship between tempo of the excerpts and the difficulty in assigning a meter structure: the intuition was that for slower tempi we would pay more attention to the lower rhythmic levels (*e.g.* beats and beat subdivisions) than to the higher metrical levels (*e.g.* measures and beats). Figure 7 displays the number of measures for each excerpt, and also the estimated tempo (based on the number of measures and the ground truths for meter structure). Since beats are asymmetric in many examples, we have adopted a uniform tempo measurement in units of eighth notes per minute (an eighth note is the common beat subdivision in formulae such as 2+3 or 3+2+2).

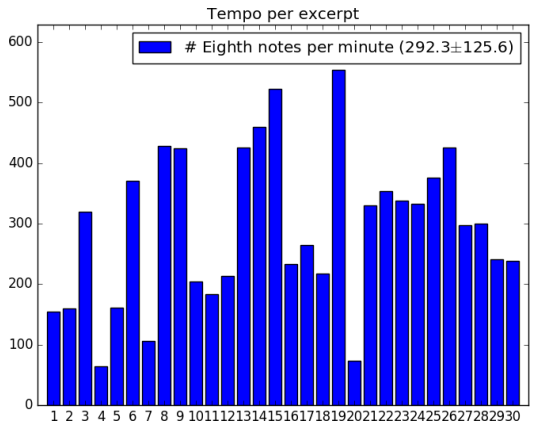
Negative correlations between number of measures and play counts ($r = -0.4298$, $p = 0.0177$) and between tempo and play counts ($r = -0.4801$, $p = 0.0073$) appear to corroborate the observed point: the lower the tempo the higher the number of times participants heard the excerpt. On the other hand, no relevant correlations have been observed between number of measures or tempo on one hand and response times or spread on the other ($p > 0.1$ for all such correlations).

5. CONCLUSIONS

In this paper we approached an important and seldom studied issue in rhythmic perception, namely that of recognition of asymmetric rhythms in Greek traditional music by musically literate subjects based on an accurate symbolic rhythmic notation. An experimental task consisting of assigning rhythmic patterns in common music notation to audio excerpts was designed and applied to university-level music students in Greece and other countries. Statistical analysis of the responses led to preliminary results that answer some of the questions that motivated this study, while leaving other questions requiring further investigation.



(a) Number of measures.



(b) Estimated tempo.

Figure 7: number of measures per excerpt (a) and estimated tempo in eighth notes per minute (b).

One of the questions that was partially answered by empirical data is the relationship between enculturation and accuracy of rhythmic perception. Greek participants had a slight advantage in producing valid answers for the selected excerpts. On the other hand, relatively average-to-low scores even for Greek participants indicate that the proposed task is not as easy as one might assume. Some excerpts (such as #20) proved to be exceptionally hard for most participants, a fact associated to a high number of "I cannot decide" answers.

We also investigated other metrics that might relate to the difficulty of the proposed experimental task. Considering the time each participant required to arrive at a decision as a plausible indication of task difficulty, a significant correlation was observed between the spread of answers (representing the degree of disagreement between participants responses) and the temporal metrics of play counts and response times. This may be interpreted as evidence that all those factors are somehow interrelated, and maybe represent converging aspects of difficulty in the perception of complex rhythmic structures.

Another question for which a tentative answer was achieved refers to the relationship between task difficulty and tempo of the excerpts. Lower tempi meant a smaller number of measures covered by the fixed-length audio excerpts, which could undermine the ability of participants to match their perceived rhythms with the given written alternatives. Negative correlations between spread of answers and both tempi and number of measures in each excerpt may corroborate these intuitions.

This study barely scratched the surface of the problem of characterizing the underlying processes in the perception of asymmetric rhythms in traditional Greek music, and there are several questions that remain open. One such question is the relationship between symmetry / asymmetry of meter structures and the difficulty of the task. Another such question refers to temporal accuracy and the alternative interpretations of the ratios 3:2 and 2:1 for long:short beats. In order to address these questions it would be important to investigate the sources of individual variations in responses to specific meter structures and individual examples, by proposing musicological explanations for both valid and invalid rhythmic alternatives, and by trying to endorse or refute them based on further statistical analyses and possibly new experimental methodologies.

6. REFERENCES

Cambouropoulos, E. (1997). Musical rhythm: A formal model for determining local boundaries, accents and metre in a melodic surface. In *Music, gestalt, and computing* (pp. 277–293). Springer.

Fouloulis, A., Pikrakis, A., & Cambouropoulos, E. (2012). Asymmetric beat/tactus: Investigating the performance of beat-tracking systems on traditional asymmetric rhythms. In *Proceedings of the Joint Conference ICMPC-ESCOM 2012*.

Fouloulis, T., Pikrakis, A., & Cambouropoulos, E. (2013). Traditional asymmetric rhythms: a refined model of meter induction based on asymmetric meter templates. In *Proceedings of the Third International Workshop on Folk Music Analysis*, (pp. 28–32).

Fracile, N. (2003). The "aksak" rhythm, a distinctive feature of the balkan folklore. *Studia Musicologica Academiae Scientiarum Hungaricae*, *44*(1-2), 191–204.

Lerdahl, F. & Jackendoff, R. (1983). *A Generative Theory of Tonal Music (Cambridge, MA)*. MIT Press.

Moelants, D. (2006). Perception and performance of aksak metres. *Musicae Scientiae*, *10*(2), 147–172.

Tekman, H. G., Kurt, S., & Peynircioğlu, Z. (2003). Perception of symmetric and asymmetric meters by listeners familiar and unfamiliar with asymmetric meters. In *Proceedings of the 5th Triennial ESCOM Conference*.

AUTOMATIC TRANSCRIPTION OF FLAMENCO GUITAR FALSETAS

Sonia Rodríguez, Emilia Gómez, Helena Cuesta
Music Technology Group, Universitat Pompeu Fabra
{sonia.rodriguez,emilia.gomez,helena.cuesta}@upf.edu

ABSTRACT

This work deals with the automatic transcription and characterization of flamenco guitar, with a focus on short melodic interludes improvised between sung verses. These are called *falsetas* in the flamenco argot and are very challenging for manual and automatic transcription due to their fast and highly ornamented nature. However, they are a key resource for guitar players to practice. We adapted a state of the art singing transcription algorithm to process an audio signal containing one or several guitar *falsetas* and extract their symbolic representation. The algorithms first perform a segmentation to locate the guitar fragments and then a symbolic transcription of these segments into symbolic representation. In order to evaluate it, we collected the first (to our knowledge) annotated *falseta* datasets. Our results confirm the difficulty of the task, and a detailed study of two transcriptions revealed that combining the algorithm with specific musical knowledge about the scale used by the song, improves the performance of the system. Our approach follows the principles of research reproducibility, and the system is integrated in a computer-assisted paradigm, where the user complements the automatic annotation with a priory knowledge to generate a final transcription.

1. INTRODUCTION

Flamenco is a musical genre that includes three basic elements: *cante* (singing), *toque* (guitar), *baile* (dancing), and has its own rules and traditions. This sociocultural movement has extended internationally beyond its geographical origin, becoming an Intangible Cultural Heritage of Humanity by UNESCO [1] in 2010.

Unlike other musical genres, flamenco guitar performance is orally transmitted; both songs and terminology have passed down across generations without a standard writing system. Flamenco *falsetas* are defined as short improvised melodies played between sung verses.

1.1 Related work

The COFLA [2] project deals with how computational models can support the analysis and synthesis of flamenco music to provide an adaptation of the general Music Information Retrieval (MIR) methodologies. Some of these aspects are linked to standard MIR tasks such as melodic similarity (Kroher et al., 2014) or genre classification (Salamon, Rocha & Gómez, Salamon et al.) which have been evaluated and adapted to flamenco music. Previous research has addressed the automatic transcription of flamenco singing, which has revealed to be challenging compared to other singing styles (Gómez & Bonada, 2013) (Kroher & Gómez, 2016). The present study focuses on the flamenco guitar.

[1] http://www.unesco.org/culture/ich/en/RL/flamenco-00363

[2] COFLA: COmputational analysis of FLAmenco music. www.cofla-project.com.

1.2 The flamenco guitar

When analyzing pieces of traditional flamenco music, we observe a dialog between instruments that appears throughout most of the songs. In particular, the most important dialogue is found between singing (*cante*) and guitar (*toque*), as they alternate on the roles of soloist and accompanying instrument. This is a key factor in our research because after a singing section, the lead is taken by the guitar player during the *falseta*. The detection and segmentation of the *falsetas* are the first steps of the proposed system, which are then followed by the transcription into a symbolic representation using the MIDI format as depicted in Figure 1. Regarding the transcription stage, we study relevant sound characteristics of flamenco guitar and typical playing techniques in order to build an optimal transcription method.

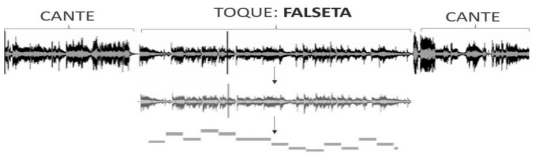


Figure 1: *Falseta transcription given the traditional dialogue between cante and toque*

1.3 Goals and contributions

We aim at providing a computer-aided system as a first step to the manual transcription, which requires advanced music and flamenco knowledge. This project is motivated by the lack of scores of flamenco guitar pieces and the high difficulty of creating them. These manual transcriptions provide complementary information to the note's representation such as fingering or dynamics. However, this information is very hard to find in an automatic way with existing techniques.

Our system provides a first step to reduce the cost of a complete transcription. The main purpose of this work is to develop an algorithm capable of automatically segmenting and transcribing flamenco guitar sections, also called *falsetas*, which would be a useful tool for learning and studying guitar. To evaluate the system, we created two manually annotated datasets: one for the segmentation and one for the MIDI transcription.

2. DATASET BUILDING

It is hard to find flamenco guitar datasets due to the absence of related research, mainly in the transcription field. Since we want to evaluate two different parts of our system, we built two datasets: one for the segmentation stage and another one for the transcription.

2.1 Segmentation dataset

Falsetas are musical segments delimited in time and in the segmentation stage we look for their boundaries, i.e. start and end times. We come upon a big controversy regarding the required length of a guitar section to be considered as a *falseta*, as well as identifying start and end points, because there is not a clear agreement on this topic. For this reason we define the minimum length as an input parameter of our system, which is set to 15 seconds by default.

The segmentation dataset contains twenty songs, including 43 falsetas (19.5 minutes of audio) from Camarón de la Isla and Paco de Lucia (1969-1977). For each of them we manually annotated the start and end times as ground truth for the segmentation step. All of them are recognized as an exemplary repertoire for classical flamenco dialog between *cante* and *toque*.

2.2 Transcription dataset

The dataset created for automatic transcription, called *ToqueFlamenco* contains the manual annotations of ten *falsetas* including onset, offset and pitch for each note. To create this data, we obtained [3] and edited the score of each piece and then converted them into MIDI files. Finally, we manually aligned the MIDI and the original audio to increase the accuracy. The dataset and details of the annotation procedure are provided in our web page [4] .

3. PROPOSED METHOD

In this section, we detail the steps of the algorithm: given an audio file, the system automatically finds and transcribes the guitar sections taking into account some classical flamenco features. Our algorithm is based on a state of the art method for flamenco singing transcription (Kroher & Gómez, 2016) which is published as a python library *PyCante* [5] . We adapt it to flamenco guitar in a similar python library *PyToque* [6] . The method consists of three main stages: *falsetas* segmentation, transcription and post-processing, each of them explained in the following subsections.

3.1 Falsetas segmentation

The aim of this part is to detect and segment *falsetas* from a song that also includes vocal parts (*cante*). *PyToque*, as a user's choice, allows to skip this step if the input contains only guitar sections.

[3] www.canteytoque.es, www.tabsflamenco.com
[4] https://doi.org/10.5281/zenodo.804050
[5] https://github.com/NadineKroher/PyCante
[6] https://github.com/SoniaLuque/PyToque

3.1.1 Channel selection

In flamenco stereo recordings, the vocals are usually more predominant in one of the channels: in order to create an artificial panorama that simulates a live performance, the guitar is strongly separated from the vocals. To make the *falsetas* segmentation easier, one of the channels is automatically selected to reduce noise and irrelevant information. To carry out this task, both channels are analyzed in terms of the distribution of their spectral energy. As showed in (Kroher & Gómez, 2016), the energy density increases between 500 Hz and 6 kHz when vocals are present. The channel selection strategy proposed in *PyToque* is based on picking out the one with the lowest average density in that range. Alongside the guitar section, the singer commonly says short sentences, also known as *jaleos*. Because of this, by choosing the channel where the vocals are not predominant, the delimitation is more precise since it avoids having a *falseta* divided by a *jaleo* by mistake.

To analyze the spectral content we compute the Short Time Fourier Transform (STFT) using a Hanning window of size N=2048 samples. According to the spectral vocal features that we mentioned above, we define the suitable frequency range both for vocals and guitar. Then, the spectral band ratio (SBR) is computed frame-wise dividing the sum of normalized magnitudes spectrum $|\dot{X}|$ of the vocal frequency band by the one corresponding to the guitar (see Eq.1), being k the bin corresponding to frequency f. After computing the average along the entire signal for each channel, the system selects the one with lowest vocal presence i.e. with lowest SBR average (unlike *PyCante*)

$$SBR[n] = 20 \cdot log10 \left(\frac{\sum_{k(500Hz)<k<k(6KHz)} \left| \dot{X}[k,n] \right|}{\sum_{k(80Hz)<k<k(400Hz)} \left| \dot{X}[k,n] \right|} \right) \quad (1)$$

3.1.2 Melody extraction

As we mentioned before, flamenco songs contain a dialogue between the voice (*cante*) and the guitar (*toque*). The proposed method exploits this feature to locate the *falsetas*. We use the melody extraction algorithm MELODIA (Salamon & Gomez, 2012) to extract the predominant pitch of the whole piece, which will correspond to the singing voice part, as the fundamental frequency range is adapted to singing, as well as rules for selecting pitch contours with fluctuations which are characteristic of singing. The result is an array, $f_0[n]$, that contains a pitch value per frame. As shown in Figure 2, we assume the segments detected as unvoiced by MELODIA to be *falsetas*.

As a parameter of the algorithm, we set the frequency range between 120 Hz and 720 Hz to track the vocals by using an analysis window of 4096 samples, as suggested by the paper authors (Gómez et al., 2012).

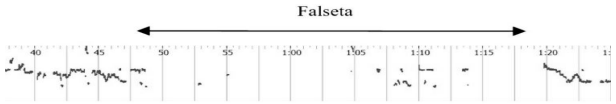


Figure 2: *Visualization of the vocal melodic line extracted using MELODIA. The segment with no melody is likely to be a falseta.*

3.1.3 Contour Classification

In this section, we classify each frame as voiced/unvoiced if it corresponds to vocal content or not. We will focus on unvoiced segments as candidates for guitar *falsetas*. This process adapts the one proposed in (Kroher & Gómez, 2016) to get rid of the guitar sections based on spectral features differences as exemplified in Figure 3. We first extract the energy in the lower twelve bark bands (see Eq. 2) computed frame-wise to carry out a preliminary discrimination.

$$B[n,m] = \sum_{k(f_{1,m})<k<k(f_{2,m})} |X[k,n]|^2 \qquad (2)$$

The result for each band m, delimited by f_1 (lower frequency limit) and f_2 (upper frequency limit) and where $k(f)$ is the frequency bin corresponding to the frequency f, is stored in a a 12-length vector $\vec{x}$ for each analyzed frame n. By using predominant melody information (i.e. the output of MELODIA), an initial label is assigned to each vector $\vec{x}$: the melody frames are marked as voiced and the non-melody frames as unvoiced. We then compute the mean and the covariance for both groups and fit a single multivariate Gaussian distribution to both sets separately. The fitting process is done for each recording and no training is needed beforehand. We obtain a probability p for each element to be voiced or unvoiced and we perform a binary classification taking into account the highest probability.

In order to avoid fast fluctuations in this prediction, a binary moving average filter of length 1 second is applied to make the vocal detection smoother. Finally, we search for segments of consecutive melody frames in f_0 and evaluate the result of the prediction for each of them. Those segments where all the prediction values are equal to zero, i.e. non-melodic according to MELODIA, are removed from the f_0 list because following our hypothesis, they will correspond to *falsetas*, and their boundaries are then used for segmentation. Instead of directly removing the vocal sections, we observed that if we first eliminate the guitar parts, as in *PyCante*, the *falsetas* delimitation is more accurate.

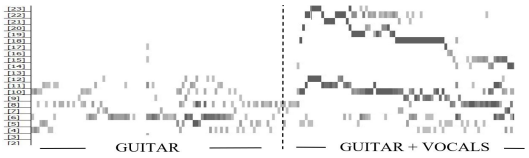


Figure 3: *Bark coefficients representation for vocal and guitar sections*

3.1.4 Falsetas identification and segmentation

In the last step of the segmentation, we locate the null segments in f_0 (see Figure 4) and extract their boundaries. These segments correspond to the non-vocal parts of the recording and thus we assume they are guitar sections. Finally, we compute the duration of each segment: if it is longer than the minimum duration specified by the user (15 seconds by default), it is classified as a *falseta*. Otherwise, we eliminate the segment.

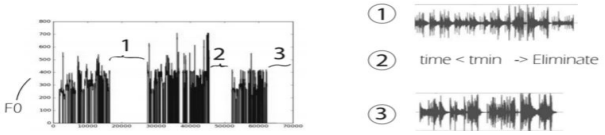


Figure 4: *Identification and evaluation of the falsetas candidates*

The output of the segmentation stage consists of an audio wave file that contains a concatenation of all the detected *falsetas*. Since we can recover the original audio file, if it is a stereo recording we obtain one audio file per channel. In addition, the system also generates a text file with the boundaries (i.e. start and end points) of each **falseta** in seconds. This information is further used for evaluation.

3.2 Transcription

We use the output of the previous stage to obtain a symbolic representation (i.e. a MIDI file) of the *falsetas*. First, we analyze the guitar melody with a pitch tracking algorithm and then we find the onsets and offsets of the notes to define their boundaries. We finally label each note with its corresponding pitch value. All the steps are detailed in the following subsections.

3.2.1 Guitar melody extraction

If the input is a stereo audio file, we compute the mean of both channels because although the guitar is more predominant in one of them, there is still some guitar in the other one that we also need for a complete analysis. At this point, we extract the guitar melody using an algorithm proposed by (Klapuri, 2006) and implemented in the *Essentia* library (Bogdanov et al., 2013). This method estimates multiple pitch values per frame, which correspond to the melodic lines present in a polyphonic music signal. By default, the transcription is monophonic but it can also be polyphonic as a user's choice:

- For monophonic *falsetas*, the algorithm only selects the first frequency value for each frame. We restrict the frequency range to 80-750 Hz, which corresponds to the guitar range; however, this parameter that can be modified.

- If we have a polyphonic guitar line, as a limitation of our algorithm, we take a maximum of two values within the mentioned frequency guitar range for each sample. Due to restrictions of the multi-pitch tracking algorithm, in the polyphonic case we do not

limit the frequency range; instead, we remove the values which are out of the range afterwards.

3.2.2 Onset and offset detection

In this step we present the methodology used for note segmentation. We consider typical flamenco guitar techniques such as contiguous notes that can be tied together, as well as very fast *staccato* notes, called *picado*, played with the index or middle finger, or *alzapúa* played with the thumb.

For onset detection we use an algorithm based on diverse novelty functions (Dixon, 2006), which is implemented in *Essentia*. In our case, even though the guitar strings have a percussive component, we consider spectral features, specifically the spectral flux. This novelty function is obtained by computing the euclidean distance between two consecutive and normalized spectra. This method provides the best results for instruments like guitar, defined as pitched and percussive by (Dixon, 2006).

To determine the offsets, we computed the average duration of all notes within the transcription dataset which was found to be $d_{avg} = 0.16s$. If we consider subsegments as sections between onsets:

- If the subsegment is shorter than d_{avg}, the offset is set as the previous sample of the next onset. In this case, we consider that the subsegment is too short and does not need to be analyzed in depth because the note is muted by immediately playing another one on the same string.

- Otherwise, we analyze the energy in each subsegment as showed in Figure 5. We compute the RMS (root mean square) using a window of size $N = 256$ and define E_{max} as the maximum energy value within the subsegment. We also define $thr = 0.1 \cdot E_{max}$. We define the offset as the first value that fulfills the condition: $E_m < thr$.

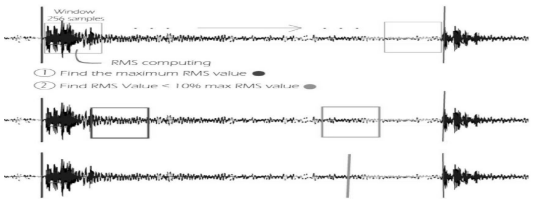


Figure 5: *Offset detection process*

3.2.3 Pitch estimation and labeling

After delimiting each note in time, we analyze its pitch content as depicted in Figure 6. We first convert pitch values from Hz to cents, using $f_T = 440$ Hz as the reference tone. Then, we compute a local pitch histogram for each note, $H[f_{cent}]$, and choose the most common pitch value - as long as it belongs to the guitar frequency range defined in previous sections. For polyphonic cases, we repeat this process for the second melodic line. Finally, we convert the obtained pitch values in cents into MIDI notes using:

$$MIDI_{note} = \left\lceil 12 \cdot log_2 \left(\frac{f}{f_T} \right) + MIDI_{ref} \right\rceil \quad (3)$$

Given the output of the onset detection function, we label each note by aligning each onset value (in frames) with the original signal to obtain the actual points in time (seconds) using Eq. (4). Afterwards, we use the computed offsets to find the duration of each note.

$$oTime = onset \cdot \frac{HopSize}{f_s} \quad (4)$$

Together with the MIDI note, the onset and the duration, we also add an energy value which is closely related to the volume for each note to provide better perceptual results when listening to the MIDI file. To this end, we use the energy function included in the *Essentia* library. Instead of the regular MIDI range (0-127), the output is bounded between 40 and 100 in order to avoid abrupt volume changes between notes.

With this information we create the resulting MIDI file using the MIDIUtil Python library [7], for which we need to set a *tempo* value in *bpm* (beats per minute). We use the algorithm proposed in (Percival & Tzanetakis, 2014) to estimate the tempo of the input recording and use it as a default value. This *tempo* can be later modified by the user using any sequencer or MIDI editor. Finally, each note is defined by an onset time, duration, pitch and energy value, and all this information is stored in a MIDI file as well as in a CSV (comma-separated values) file.

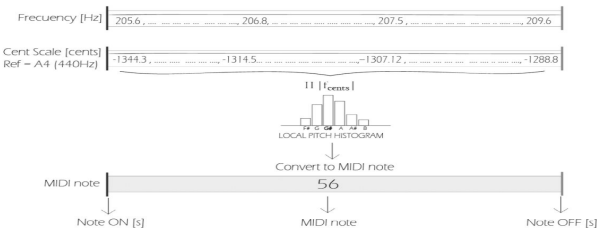


Figure 6: *Pitch labeling process*

3.3 Post-processing

In the last stage of the proposed system we aim to adjust some potential errors using pitch re-scaling: this method can only be applied in cases where the user knows the scale of the piece. It consists of scanning the whole set of notes and re-scaling the pitch values using tonal features and typical scales of flamenco music (Fernandez, 2004). Our system includes the following scales:

- From modern Western modes, the *Phrygian* mode on E and *Phrygian* dominant scale produced by raising the third scale degree when ascending as we displayed in Figure 6.

- A *Flamenco* mode which arises from the previous scale but using a transposition of two tones and a half as shown in Figure 8

[7] https://pypi.python.org/pypi/MIDIUtil/

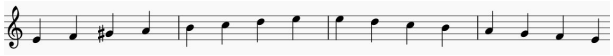


Figure 7: *Phrygian dominant scale also called Spanish Gypsy Scale*

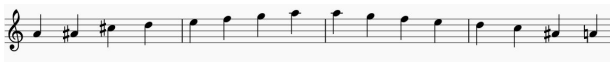


Figure 8: *Phrygian mode transposed to A*

- Major scale based on E shown in Figure 9

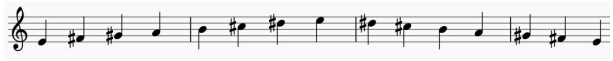


Figure 9: *E Major scale*

In flamenco, it is common to adapt the pitch range of the *falseta* to the pitch range of the singer by means of a *capo*. In order to allow for that we implement a transposition function where the user can specify a number of semitones.

4. EVALUATION METHODOLOGY

We evaluate the segmentation and transcription stages separately using the two datasets (see Section 2) and two different methods, both included in the *mir_eval* library (Raffel et al., 2014).

Regarding the segmentation stage, we evaluate two different aspects: first, the number of *falsetas* longer than fifteen seconds found by the algorithm, and second, the precision of their boundaries. In flamenco pieces, we usually hear short guitar resources used to open and close *falsetas* called *llamadas*, *remates* or *cierres*. Since it is not clear whether to consider them as part of the *falseta* or not, we use a tolerance window of size $N = 4$ seconds to evaluate the boundaries of each *falseta*.

For the transcription stage, we evaluate the onset, offset, and pitch value for each note in the *falseta* using the *transcription* method in *mir_eval*. According to MIREX 2015 [8], this method assumes an estimated note to be correct if its pitch value is within $\pm$ quarter tone of the corresponding reference note. Regarding the onset and offset rules, we increase the tolerance from $\pm$50 ms to $\pm$100 ms because of the relative inaccuracy of the manual annotations in the transcription dataset.

4.1 Standard Metrics

To evaluate the performance of our system we use the three standard information retrieval evaluation metrics: Precision (P), Recall (R), F-measure (F):

$$P = \frac{c}{c+f^+}, R = \frac{c}{c+f^-}, F = \frac{2 \cdot P \cdot R}{P+R}, \quad (5)$$

[8] http://www.music-ir.org/mirex/wiki/2015:Main_Page

Where c is the number of correct detections, and f^+ and f^- represent the number of false positives and false negatives respectively. For the transcription stage, we also obtain the average overlap ratio (AOR) as the mean overlap ratio computed over all matching reference and estimated notes.

5. RESULTS

We first examine the number of *falsetas* identified for each song without taking into account the segmentation accuracy, i.e. we count the number of *falsetas* that the algorithm finds even though their boundaries are not exact, and find that our approach is capable of locating as many *falsetas* as there are in the ground truth. This accuracy confirms that this method is able to discern the guitar sections in a flamenco piece.

The results of the evaluation of the second part of the segmentation stage (the delimitation of the *falsetas*) and transcription stage are summarized in Table 1, averaged for all the *falsetas*. Notice that even though we obtained an accuracy of 100% in the first part of the segmentation stage (i.e. the number of *falsetas*), the average precision of their boundaries falls to 75%. We observe that our system is better at segmenting *falsetas* than at transcribing them, which suggests that there is room for improvement, especially in this second step.

Stage	P	R	F	AOR
Segmentation (boundaries)	0.75	0.77	0.76	-
Transcription	0.61	0.62	0.615	0.618

Table 1: Results for both stages (independently) according to the methodology detailed in Section 4.

To understand the limitations of our system, we measure how pitch re-scaling (PRS), detailed in Section 3.3, affects the performance of the algorithm. We evaluate two excerpts of our dataset using the PRS procedure: the first one (*Soleá* 1) is a monophonic *falseta* which has a duration of 6 seconds and is re-scaled using the *Phrygian* dominant scale. The second one (*Alegrias*) corresponds to a 30 seconds polyphonic piece and is re-scaled using the E major scale. Table 2 shows these results, and illustrates an increase in precision and recall. This suggests that adding specific musical knowledge, such as the scale of the song, to the algorihm has a positive impact on the performance.

Data	P	R	F	AOR
Soleá 1	0.79	0.81	0.80	0.71
Soleá 1 (PRS)	0.823	0.848	0.83	0.718
Alegrias	0.735	0.59	0.654	0.412
Alegrias (PRS)	0.756	0.61	0.674	0.415

Table 2: Impact of the PRS applied on two specific cases

Since we are not able to compare the transcription results with any existing system, we manually inspect the

MIDI files together with their corresponding audio input. We observe that the onset detection is quite unstable and provides significantly different results throughout the dataset because of the wide range of different techniques used in flamenco guitar. To measure the performance of onset detection both perceptually and using standard metrics, we evaluate the system using two methods based on different novelty functions as shown in Table 3: the complex domain spectral difference function (Complex) and the high frequency content (HFC) detection. We compare these results with the spectral flux function used by default. Since the guitar is a pitched percussive instrument, we also include the evaluation removing the offset rule.

Method	P	R	F	AOR
Spectral Flux	0.61	0.62	0.615	0.618
Complex	0.608	0.607	0.60	0.61
HFC	0.618	0.496	0.54	0.59
Spectral Flux (no offset)	0.65	0.661	0.65	0.629
Complex (no offset)	0.658	0.656	0.652	0.614
HFC (no offset)	0.68	0.553	0.607	0.563

Table 3: Results for both stages according to the methodology detailed in Section 4

By manual inspection of the results, we observe that the HFC method is useful for regions with transients, but can be misleading when hand-clapping appears. The *complex* method works properly for pitch-changing notes (*legato* or *glissando*) but not for fast notes. As mentioned in Section 3.2.2, the spectral flux method performs well for pitched and percussive notes, although it still provides unstable results for techniques such as *glissando*, flamenco *tremolo* [9] or *alzapúa*.

6. CONCLUSIONS AND FUTURE WORK

We have addressed the problem of *falseta* detection and transcription and we consider that the obtained results are satisfactory. The dataset collection has been one of the most challenging tasks in our project, given the lack of scores and related research. Due to this fact, our dataset still contains few samples and needs to be expanded to obtain more representative results. In spite of that, we think that this work provides a good starting point for further research in this problem.

We consider that the segmentation provides reliable results but the system would sometimes need to disambiguate what is considered as a *falseta* or not. Our segmentation method is limited to pieces that contain dialogs between *cante* and *toque*. If a new instrument is present and has its own sections, the system will probably classify it as a guitar *falseta*. As a future work, we suggest to use spectral features to create timbre spaces allowing the discrimination between a varied set of instruments. Regarding onset detection, a multimodal fusion technique could be used to stabilize the results by improving the precision also for those techniques in which weak transients can be included.

[9] https://www.atrafana.com/flamenco-guitar-techniques–tremolo.html

7. ACKNOWLEDGMENTS

This work is partially supported by the Spanish Ministry of Economy and Competitiveness under the CASAS project (TIN2015-70816-R).

8. REFERENCES

Bogdanov, D., Wack, N., Gómez, E., Gulati, S., Herrera, P., Mayor, O., Roma, G., Salamon, J., Zapata, J. R., & Serra, X. (2013). Essentia: an audio analysis library for music information retrieval. In *ISMIR'13*, (pp. 493–498).

Dixon, S. (2006). Onset Detection Revisited. *Proc. of the 9th Int. Conference on Digital Audio Effects (DAFx-06)*, 1–6.

Fernandez, L. (2004). *Teoria musical del flamenco*. Acordes Concert.

Gómez, E. & Bonada, J. (2013). Towards computer-assisted flamenco transcription: An experimental comparison of automatic transcription algorithms as applied to a cappella singing. *Computer Music Journal*, *37*, 73–90.

Gómez, E., Cañadas, F., Salamon, J., Bonada, J., Vera, P., & Cabañas, P. (2012). Predominant fundamental frequency estimation vs singing voice separation for the automatic transcription of accompanied flamenco singing. In *13th International Society for Music Information Retrieval Conference (ISMIR 2012)*, Porto.

Klapuri, A. (2006). Multiple Fundamental Frequency Estimation by Summing Harmonic Amplitudes. *Proceedings of the International Symposium/Conference on Music Information Retrieval (ISMIR)*, 216–221.

Kroher, N. & Gómez, E. (2016). Automatic transcription of flamenco singing from polyphonic music recordings. In *IEEE/ACM Transactions on Audio Speech and Language Processing*, volume 24, (pp. 901–913).

Kroher, N., Gómez, E., Guastavino, C., Gómez-Martín, F., & Bonada, J. (2014). Computational models for perceived melodic similarity in a cappella flamenco cantes. In *15th International Society for Music Information Retrieval Conference*, Taipei, Taiwan.

Percival, G. & Tzanetakis, G. (2014). Streamlined Tempo Estimation Based on Autocorrelation and Cross-correlation With Pulses. *IEEE/ACM TRANSACTIONS ON AUDIO*, *22*(12).

Raffel, C., Mcfee, B., Humphrey, E. J., Salamon, J., Nieto, O., Liang, D., & Ellis, D. P. W. (2014). mir_eval: A TRANSPARENT IMPLEMENTATION OF COMMON MIR METRICS. In *15th International Society for Music Information Retrieval Conference*.

Salamon, J. & Gomez, E. (2012). Melody extraction from polyphonic music signals using pitch contour characteristics. *IEEE Transactions on Audio, Speech and Language Processing*, *20*(6), 1759–1770.

Salamon, J., Rocha, B., & Gómez, E. Musical Genre Classification using Melody Features Extracted from Polyphonic Music Signals.

QUANTITATIVE EVALUATION OF MUSIC COPYRIGHT INFRINGEMENT

Patrick E. Savage
Keio University Shonan Fujisawa Campus
psavage@sfc.keio.ac.jp

Charles Cronin
USC Gould School of Law
ccronin@law.usc.edu

Daniel Müllensiefen
Goldsmiths, University of London
d.mullensiefen@gold.ac.uk

Quentin D. Atkinson
University of Auckland
q.atkinson@auckland.ac.nz

ABSTRACT

Unfounded music copyright lawsuits inhibit musical creativity and waste millions of taxpayer dollars annually. Our aim was to develop and test simple quantitative methods in order to supplement traditional qualitative musicological analyses and improve the efficiency and transparency of music copyright lawsuits. We adapted automatic sequence alignment algorithms from computational biology to create a "percent melodic identity" (PMI) method that was initially developed to measure the cultural evolution of folk music from different cultures. This method automatically quantifies and visualizes the percentage of identical pitch classes shared between two melodic sequences. We applied the PMI method to a corpus of 20 pairs of melodies that had been the subject of legal decisions and that had previously been analyzed using automatic methods. We found that the PMI method was able to accurately predict 80% (16/20) of previous decisions, with PMIs below 50% usually resulting in decisions of no infringement (11/13 cases), and PMIs above 50% usually resulting in decisions of infringement (5/7 cases). Importantly, each of the four outlying cases could be explained by contextual factors not related to melodic similarity (e.g., lyrics, access). Our results provide promise for improving music copyright evaluation by supplementing traditional qualitative components with quantitative methods and visualization tools that are simple enough to be useful to juries, judges, and other non-musicologists.

1. INTRODUCTION

Copyright serves the public good by encouraging the creation of innovative expression by granting authors a limited term during which they alone have the right to capitalize on their works. In music, unauthorized copying of melodies, lyrics, or other attributes of music has been legally prohibited since the 18th century. Initially, copyright law was designed to protect simply against wholesale copying of entire musical works (e.g., J. C. Bach's sonatas, the first case in which music was recognized as protected by copyright). Gradually, however, copyright case law broadened the scope of impermissible copying, such that even subconscious copying of parts of a melody could constitute infringement if such copying was substantially similar to protectable expression in the earlier work. Exactly how much and what types of musical copying qualify as "substantial" is a multimillion dollar question that is being actively and intensely debated (Cason & Müllensiefen, 2012; Cronin, 2015; Fishman, Forthcoming; Fruehwald, 1992). These debates have important practical consequences for all, as inappropriate music copyright lawsuits not only inhibit musical creativity, but also waste millions of taxpayer dollars annually that cover the adjudication of these disputes, not to mention the financial and temporal losses of individual defendants. One reason for this waste is that judicial evaluation of claims of musical similarity, on which these disputes are grounded, typically involves expert testimony by musicologists, who tend to use subjective, idiosyncratic, and time-consuming methods, tailored to the interests of the party that has retained them.

Unlike other arts (e.g., visual arts) where no single dimension has been given priority in copyright claims, music is unique in that one dimension – melody – has traditionally been the focus of copyright debates. For example, the court in King vs. Northern Music (1952) declared: "It is in the melody of the composition—or the arrangement of notes or tones that originality must be found. It is the arrangement or succession of musical notes, which are the finger prints of the composition, and establish its identity."

The rise of fields such as music information retrieval and music cognition have resulted in the development of automated melodic similarity algorithms and their application to musical copyright cases (Cason & Müllensiefen, 2012; Cronin, 1998; Mongeau & Sankoff, 1990; Müllensiefen & Frieler, 2004; Müllensiefen & Pendzich, 2009; Robine, Hanna, Ferraro, & Allali, 2007; Selfridge-Field, Forthcoming). For instance, Müllensiefen and Pendzich (2009) developed an algorithm that compares the profile of successive pitch intervals in two disputed songs against each other, while weighting them against a database of comparable profiles from 14,063 pop songs using a weighting formula for estimating perceptual salience. They found that optimizing this algorithm to a cut-off similarity value of 0.24 allowed them to accurately predict 90% (18/20 cases) of court decisions centered on questions of melodic similarity between 1976-2006.

While such algorithms have been somewhat successful, they have also been hard to translate into terms that are meaningful for non-scientists. Not only is it hard for jury members to interpret a salience function value of 0.24, but even this value is dependent on the makeup of the 14,000-pop song sample, and thus redoing these analyses using a different reference sample would result in different cutoff values. Juries, judges, and other interested parties would benefit from an intuitive measure of melodic similarity that depends only on the two melodies in question and can be easily visualized through simple notation.

The goal of this article is to propose and test a simple quantitative measure of musical similarity against a series of influential past decisions. Supplementing subjective qualitative interpretations with clear and intuitive quantitative guidelines by which to compare new cases with past

cases should increase transparency and efficiency, reducing the chance of costly mistakes in the legal process and stemming the recent explosion of meritless claims that aim to force a quick payout from artists unwilling or unable to accept the risks of the current system.

2. THE PERCENT MELODIC IDENTITY (PMI) METHOD

We propose to adapt a "percent melodic identity" (PMI) method to musical copyright cases. This method is based on the automated sequence alignment and percent identity calculations used in molecular genetics to compare DNA and protein sequences (May, 2004). It was originally adapted to music in order to quantify the cultural evolution of English and Japanese folk song melodies in ways that could be meaningfully compared both with each other and with the evolution of other types of music from around the world (Savage & Atkinson, 2015). However, musical copyright represents an ideal application for this method, since copying of melodies with modification is simply another form of musical evolution. The PMI method is a general one that can also be applied to other types of folk and art music around the world (e.g., *gagaku*, Child ballads; Savage, 2017), justifying its inclusion in the *Folk Music Analysis* workshop despite its application to popular music.

The PMI method and other melodic sequence alignment algorithms are similar in principle to Judge Learned Hand's "comparative method" (Fishman, Forthcoming) for evaluating musical similarity. Like Hand, the PMI method begins by transposing two melodies transcribed in staff notation to a shared tonic, eliminating rhythmic information by giving all notes equal values[1], and then aligning and counting corresponding notes. However, while Hand's alignments were performed manually, the PMI method can take advantage of automated sequence alignment algorithms (Needleman & Wunsch, 1970) to eliminate subjectivity in alignment (although alignments can still be performed manually either from scratch or to correct errors in the automated alignment, as is also done in molecular genetics).

Automatic alignment requires penalties to be specified for opening or extending gaps in the alignment (represented by dashes in Fig. 1). Previously, we found that gap opening penalties (GOP) of 12 and a gap extension penalty (GEP) of 6 were the most successful in distinguishing whether two folk melodies shared ancestry (Savage & Atkinson, 2015; although further testing may be warranted in future to see whether these parameters are optimal for music copyright cases). Once the melodies are aligned, the number of identical notes (*ID*) are counted and divided by the average length of the two melodies (L_1 and L_2)[2] to calculate percent melodic identity (*PMI*, previously termed *PID* or "percent identity") as the percentage of identical notes shared between the two melodies, as follows:

$$PMI = 100\left(\frac{ID}{\frac{L_1+L_2}{2}}\right) \quad (1)$$

The PMI method can also be used to determine whether a given PMI value is statistically significant beyond what might be expected by two stylistically similar melodies that share similar scales. To do this, the PMI value for a given pair of sequence is compared against the distribution of 100 random PMI values given the same sequence lengths and compositions, as calculated by randomly reordering one of the sequences (Savage & Atkinson, 2015). Thus, an observed PMI value greater than 95% of randomly reshuffled values corresponds to a significant *P*-value of <.05.

Figure 1 shows an example of the PMI method using the famous case of Bright Tunes vs. Harrisongs. In this case, Judge Owen concluded that George Harrison had subconsciously plagiarized The Chiffons' *He's So Fine* because the melody of his song *My Sweet Lord* was "virtually identical". The PMI method is able to quantify this statement more precisely. For the opening three phrases shown in Figure 1, there are nine identical notes, while the average length of both melodies is 14, giving a PMI of 64%. When automatically aligning the full melodies of both songs, the PMI value drops slightly to 56% (27 identical notes, average length = 48 notes).

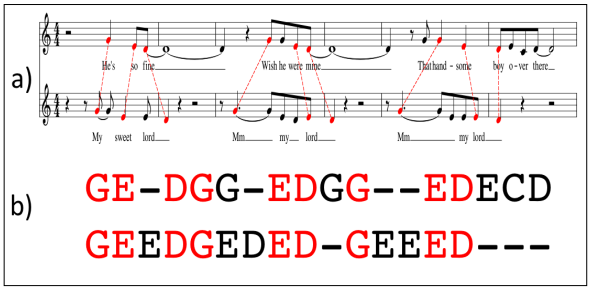


Figure 1. Comparison of the opening melodies of The Chiffons' *He's So Fine* (top) and George Harrison's *My Sweet Lord* (bottom) using ***a)*** standard staff notation, and ***b)*** the PMI sequence alignment method. In both cases, red represents aligned notes sharing identical scale degrees (joined with dashed lines in *a*). Dashes in *b* represent gaps inserted during the alignment process. PMI = 64% for these three phrases (56% for the full melodies). See Savage & Atkinson (2015) for details of how staff notation is converted into sequences of letters (including transposition to share a common tonic of C).

[1] Incorporating rhythmic information along with pitch information greatly increases computational complexity and does not appear to contribute substantial additional information (Cason & Müllensiefen, 2012; Cronin, 1998).

[2] This has been recommended as the most consistent denominator (May, 2008), but future investigation could explore whether other denominators may be more appropriate for music copyright. For instance, dividing by the length of the prior (plaintiff's) work may be more consistent with the principle that the shared content needs to constitute a substantial part of the original work.

3. MUSICAL COPYRIGHT INFRINGEMENT DATASET

For a ground-truth dataset to test the PMI method, we chose the set of 20 court decisions regarding melodic copyright infringement previously analyzed by Müllensiefen and Pendzich (2009). These decisions are a subset of the decisions available at the Music Copyright Infringement Resource (Cronin, 2018) selected by Müllensiefen and Pendzich because they contained clear rulings that were specifically focused on questions of melodic similarity (i.e., excluding cases focused on plagiarism of lyrics, unauthorized sampling of sound recordings, technicalities about the copyright registration process, etc.). This dataset seemed like an ideal starting point to test the PMI method against because the melodies had already been pre-selected and because automated similarity algorithms had already been used against them, providing a benchmark to compare the value of the PMI method. The list of cases is shown in Table 1, with the cases arranged by order of increasing PMI value.

4. RESULTS

4.1 Classification accuracy

Receiver operating curve (ROC) analysis using the area under the curve (AUC) measure confirms the intuitive impression from Table 1 that the optimal cutoff PMI value is 50% (AUC = 0.69). Using this cutoff, the PMI method was able to accurately classify 16 out of the 20 cases to match the court's decisions. For each of the four "failures", however, the following brief analyses show important non-melodic contextual factors that suggest that these exceptions were not due primarily to a failure of the melodic similarity algorithm but rather to the complex nature of musical copyright law (see Cronin, 2018 for further details on these and the other cases analyzed):

4.1.1 Grand Upright vs. Warner

There was no significant similarity between the melodies of Gilbert O'Sullivan's *Alone Again (Naturally)* and Biz Markie's *Alone Again* (PMI = 27%). However, there is obvious similarity in the lyrics, particularly in the title phrase "Alone again, naturally" used in both works. More importantly, Biz Markie uses an unauthorized sample of Gilbert O'Sullivan's piano accompaniment, and it appears that this was in fact the deciding factor in the case. Thus, this case may not have been appropriate for Müllensiefen and Pendzich to include (we have included it here for comparability).

4.1.2 Three Boys Music vs. Michael Bolton

This case is interesting because, although there is no significant melodic similarity between The Isley Brothers' *Love Is A Wonderful Thing* and Michael Bolton's song of the same name when taking the chorus as a whole (PMI = 36%), the opening phrase of each chorus uses not only the identical title lyrics but is also almost identical melodically (PMI = 86%; 5 out of 6 identical notes; Fig. 2). Thus, it appears that not only may similarities in the title/lyrics have influenced the jury's decision, but there may also be legitimate room for debate regarding how much of the melody should be included for purposes of melodic comparison and how long/complex a melody needs to be before it qualifies as original copyrightable expression.

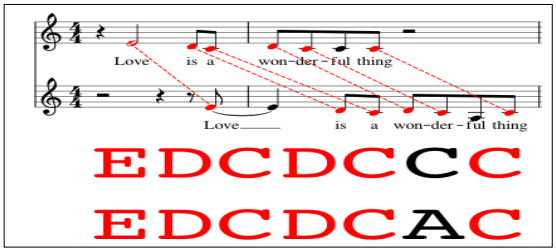


Figure 2. Comparison of the first phrase of the chorus of The Isley Brothers' *Love Is A Wonderful Thing* (top) and Michael Bolton's *Love Is A Wonderful Thing* (bottom). PMI = 86% for this phrase (but only 36% for the full choruses).

4.1.3 Selle vs. Gibb

The jury's verdict that the Bee Gees' *How Deep Is Your Love* infringed on Ronald Selle's *Let It End* was in fact consistent with the significant PMI value of 61% (Fig. 3). However, in this case the jury's verdict was overruled by the judge on appeal based on the fact that the Selle had not offered evidence to demonstrate that the Bee Gees had access to his work that would have allowed them to copy it. Such evidence is a legal requirement in addition to evidence of substantial similarity.

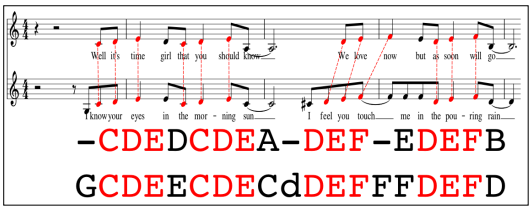


Figure 3. Comparison of the opening melodies of Ronald Selle's *Let It End* (top) and the Bee Gee's *How Deep Is Your Love* (bottom). PMI = 73% for these phrases (61% for the full chorus).

4.1.4 Fantasy vs. Fogerty

There was significant melodic similarity between John Fogerty's *Run Through The Jungle* and his *The Old Man Down The Road* (PMI = 67%) but a jury judged that the

No.	Case	Complaining work	Defending work	Decision	PMI
1	Suzane McKinley vs. Collin Raye	"I Think About You"	"I Think About You"	0	11%
2	Ferguson vs. N.B.C.	"Jeannie Michele"	"Theme 'A Time To Love'"	0	24%
3	*Grand Upright vs. Warner*	*"Alone Again (Naturally)"*	*"Alone Again"*	*1*	*27%*
4	Jean *et al.* vs. Bug Music	"Hand Clapping Song"	"My Love Is Your Love"	0	35%
5	*Three Boys Music vs. Michael Bolton*	*"Love Is A Wonderful Thing"*	*"Love Is A Wonderful Thing"*	*1*	*36%*
6	Cottrill vs. Spears	"What You See is What You Get"	"What U See is What U Get"	0	38%
7	Baxter vs. MCA	"Joy"	"Theme from 'E.T.'"	0	40%*
8	Intersong-USA vs. CBS	"Es"	"Hey"	0	40%**
9	Ellis vs. Diffie	"Lay Me Out By The Jukebox When I Die"	"Prop Me Up Beside The Jukebox (If I Die)"	0	40%
10	Granite Music vs. United Artists	"Tiny Bubbles"	"Hiding The Wine"	0	41%**
11	Repp vs. Lloyd-Webber	"Till You"	"Phantom Song"	0	45%**
12	McDonald vs. Multimedia Entertainment	"Proposed Theme Music 'Sally Jesse Raphael Show'"	"Theme Music 'Sally Jesse Raphael Show'"	0	46%
13	Benson vs. Coca- Cola	"Don't Cha Know"	"I'd Like To Buy The World A Coke"	0	49%*
14	Swirsky vs. Carey	"One of Those Love Songs"	"Thank God I Found You"	1	50%**
15	Bright Tunes Music vs. Harrisongs Music	"He's So Fine"	"My Sweet Lord"	1	56%**
16	Herald Square Music vs. Living Music	"Day By Day"	"Theme N.B.C.'s 'Today Show'"	1	56%**
17	*Selle vs. Gibb*	*"Let It End"*	*"How Deep Is Your Love"*	*0*	*61%***
18	*Fantasy vs. Fogerty*	*"Run Through The Jungle"*	*"The Old Man Down The Road"*	*0*	*67%**
19	Louis Gaste vs. Morris Kaiserman	"Pour Toi"	"Feelings"	1	73%**
20	Levine vs. McDonald's	"Life Is A Rock (But The Radio Rolled Me)"	"McDonald's Menu Song"	1	80%**

Table 1. The 20 music copyright infringement cases analyzed, ordered by increasing PMI (Percent Melodic Identity). See text for discussion of italicized exceptional cases. "0"=No infringement, "1"=Infringement. *P<.05, **P<.01.

two works were not musically substantially similar. The curious aspect of this case was that it involved a composer being accused by a recording company of plagiarizing his *own* work, the copyright to which he had assigned to the recording company. Given the substantial stylistic similarities expected among compositions by the same composer, it seems possible that the jury may have interpreted the judge's instructions regarding substantial similarity differently than they might for a case involving disputes between different composers. Furthermore, there was limited original expression in both melodies to begin with. Both are based predominantly on only two notes, so chance alone would already give a PMI of approximately 50%. In theory, this limited palette should be accommodated by the significance testing aspect of the PMI method, but – as discussed further below – this significance testing is complicated by other factors and cannot always be relied on.

4.2 Comparison with other algorithms

The best-performing algorithm tested by Müllensiefen and Pendzich (Müllensiefen & Pendzich, 2009) accurately predicted 90% (18/20) of these decisions. Their results were similar to our results using the PMI method, with the exception that Müllensiefen and Pendzich's algorithm resulted in *Three Boys Music vs. Michael Bolton* falling above their 0.24 optimal cutoff threshold, while *Fantasy vs. Fogerty* fell below this threshold.

Müllensiefen and Pendzich also tested other algorithms, including a "raw edit distance" algorithm that was more similar to the PMI method in that it was based purely on comparisons between disputed melodies without calibration against a database. The raw edit distance algorithm performed similarly to the PMI method, except that it failed to classify *Swirsky vs. Carey* as infringement[1].

However, as discussed above, it is not clear whether such differences in predicting court decisions truly imply that one melodic similarity algorithm is better. Indeed, the reverse may be true: Müllensiefen and Pendzich's algorithm may have overfit melodic similarity measures to match decisions that were affected by non-melodic factors such as lyrics or the identity of the composer. Future testing on a broader sample of cases should help determine whether there are substantive differences in the performance of these algorithms.

4.3 Statistical significance

The PMI method produced non-significant *P*-values in all cases where the PMI value was below 40%, and produced significant *P*-values in all cases where the PMI value was above 50%. PMI values between 40-50% gave mixed results, but generally produced significant *P*-values even though no infringement was found (5/7 cases). This suggests that the statistical significance measure is returning an inflated false positive rate. This is likely due to the fact that the comparison with completely random sequence is not a fair comparison, as even melodies that are completely unrelated will tend to share more pitch sequences than expected by chance alone due to universal regularities in melodic structure (e.g., tendencies for small, stepwise intervals and descending/arched melodic contours; Savage, Brown, Sakai, & Currie, 2015). We thus recommend caution in interpreting statistical significance of PMI values.

5. DISCUSSION AND FUTURE DIRECTIONS

We applied a simple and intuitive PMI (Percent Melodic Identity) method for measuring and visualizing melodic similarity to a ground-truth dataset of 20 court decisions on musical copyright. The PMI method performed similarly to existing, more complicated methods, accurately predicting 80% (16/20) of the decisions.

The major limitation of the current study is the limited sample of 20 cases and the fact that these cases include some complicating extra-musical factors. This makes it difficult to accurately evaluate automated melodic similarity algorithms against one another or conclusively determine whether they can usefully supplement future cases. In the decade since Müllensiefen and Pendzich compiled their sample, there have already been dozens of new decisions added to the Music Copyright Infringement Resource (including from countries such as China with different copyright regimes) and the number of active cases is increasing more rapidly than ever. In the future we plan to continue to expand and test the database to include these and many other cases from around the world. This broader sample will also allow us to address various technical issues such as the relative strengths of the similarity algorithms used, the effects of including rhythmic parameters, weighting different degrees of melodic similarity beyond simply identical or non-identical, etc. (Mongeau & Sankoff, 1990; Savage & Atkinson, 2015; Urbano, Lloréns, Morato, & Sánchez-Cuadrado, 2011; van Kranenburg, Volk, & Wiering, 2013). In particular, the cases discussed above highlighted the way similarity measurements can vary depending on the lengh of disputed melodic sections, and future studies may benefit by comparing different melodic lengths using both global and local alignment algorithms (van Kranenburg et al., 2013).

A broader issue is that the traditional reliance on melody as the key dimension by which to evaluate musical infringement may be changing along with the technology for making music (Cronin, 2015; Fishman, Fothcoming). This issue has been particularly actively debated recently following the jury verdict in *Pharrell Williams vs. Bridgeport Music* (currently under appeal) finding Williams and Robin Thicke liable for damages of over $5 million for infringing on Marvin Gaye's *Got To Give It Up* with their number-one hit *Blurred Lines* despite minimal melodic similarities. Specifically, the short "signature phrase" cited by expert musicologist Judith Finell as the primary melodic similarity, only gives a non-significant PMI value of 45% (5 identical notes out of 11; Fig. 4), while the PMI reduces to 19% when the full melodies are considered.

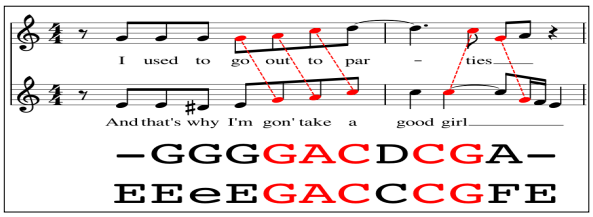


Figure 4. Comparison of the "signature phrase" from Marvin Gaye's *Got To Give It Up* (top) and Robin Thicke and Pharrell Williams' *Blurred Lines* (bottom). PMI = 45% for this phrase (19% for the full melody).

Hundreds of musicians, musicologists, and lawyers have weighed in on this decision, with some supporting the removal of arguably Eurocentric melodic notation from its dominant role in copyright law, while others fear the potentially stifling effect on creativity that vaguer and looser standards may cause (Cullins, 2018). Some have argued that we are already seeing a "Blurred Lines effect" (Fishman, Forthcoming) by which more dubious lawsuits ultimately settled out of court without a final legal decision as to the question of infringement.

[1] In fact, *Swirsky vs. Carey* is the other case that may not have been appropriate to include in the database, as it was

are being settled out of court due to fears that the old rules no longer apply. Given the changing norms in evaluating musical copyright infringement claims, and uncertainty about the relative weights given to the various factors going into past decisions, another important area for future work is to isolate the perceptual effects of melodic and extra-melodic similarities through controlled laboratory experiments (Lund, 2011; Müllensiefen & Frieler, 2004).

As even this small sample of cases shows, determinations of musical copyright infringement are too complex for it ever to be possible to predict outcomes perfectly through automated algorithms alone. Trial by algorithm will never replace trial by jury, nor should it. However, the automated, quantitative PMI method that we have presented is relatively accurate and easy for non-experts to understand and visualize. As such, we anticipate that it will help complement traditional qualitative analyses in future cases to create a more efficient, transparent, and just system for evaluating musical copyright infringement.

Acknowledgments: Funding support for this work was provided by a Japanese Ministry of Education, Culture, Sports, Science and Technology (MEXT) scholarship to P.E.S. and a Rutherford Discovery Fellowship to Q.D.A. We thank the anonymous reviewers for comments on a previous draft.

6. REFERENCES

Cason, R. J. S., & Müllensiefen, D. (2012). Singing from the same sheet: Computational melodic similarity measurement and copyright law. *International Review of Law, Computers & Technology*, *26*(1), 25–36.

Cronin, C. (1998). Concepts of melodic similarity in music-copyright infringement suits in melodic similarity: Concepts, procedures, and applications. In W. B. Hewlett & E. Selfridge-Field (Eds.), *Computing and musicology 11* (pp. 187–209). Cambridge, MA: MIT Press.

Cronin, C. (2015). I hear America suing: Music copyright infringement in the era of electronic sound. *Hastings Law Journal*, *66*(5), 1187–1254.

Cronin, C. (2018). Music copyright infringement resource. Retrieved February 20, 2018, from http://mcir.usc.edu/

Cullins, A. (2018, December 28). Marvin Gaye's family gets boost from songwriters, musicologists in 'Blurred Lines' appeal. *Hollywood Reporter*. Retrieved from https://www.hollywoodreporter.com/thr-esq/marvin-gayes-family-gets-boost-songwriters-musicologists-blurred-lines-appeal-959465

Fishman, J. P. (Forthcoming). Music as a matter of law. *Harvard Law Review*. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2931091

Fruehwald, E. S. (1992). Copyright infringement of musical compositions: A systematic approach. *Akron Law Review*, *26*(1), 15–44.

Lund, J. (2011). An empirical examination of the lay listener test in music composition copyright infringement. *Virginia Sports and Entertainment Law Journal*, *11*(1), 137–177.

May, A. C. W. (2004). Percent sequence identity: The need to be explicit. *Structure*, *12*, 737–738.

Mongeau, M., & Sankoff, D. (1990). Comparison of musical sequences. *Computers and the Humanities*, *24*, 161–175.

Müllensiefen, D., & Frieler, K. (2004). Cognitive Adequacy in the Measurement of Melodic Similarity: Algorithmic vs. Human Judgments. *Computing in Musicology*, *13*(2004), 147–176.

Müllensiefen, D., & Pendzich, M. (2009). Court decisions on music plagiarism and the predictive value of similarity algorithms. *Musicae Scientiae*, *13*(1 Suppl), 257–295.

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*, 443–453.

Robine, M., Hanna, P., Ferraro, P., & Allali, J. (2007). Adaptation of string matching algorithms for identificaton of near-duplicate music documents. *Workshop on Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection (PAN07)*, 37–43.

Savage, P. E. (2017). 音楽の文化的進化を測る ——ブリティッシュ・アメリカンと日本の民謡・ポップス・古典音楽の事例を通して [Measuring the cultural evolution of music: With case studies of British-American and Japanese folk, art, and popular music]. PhD dissertation, Tokyo University of the Arts.

Savage, P. E., & Atkinson, Q. D. (2015). Automatic tune family identification by musical sequence alignment. In *Proceedings of the 16th International Society for Music Information Retrieval Conference (ISMIR 2015)* (pp. 162–168).

Savage, P. E., Brown, S., Sakai, E., & Currie, T. E. (2015). Statistical universals reveal the structures and functions of human music. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(29), 8987–8992.

Selfridge-Field, E. (Forthcoming). Substantial musical similarity in sound and notation: Perspectives from digital musicology.

Urbano, J., Lloréns, J., Morato, J., & Sánchez-Cuadrado, S. (2011). Melodic similarity through shape similarity. In *Exploring Music Contents: 7th International Symposium, CMMR 2010* (pp. 338–355).

van Kranenburg, P., Volk, A., & Wiering, F. (2013). A comparison between global and local features for computational classification of folk song melodies. *Journal of New Music Research*, *42*(1), 1–18.

Tension perception in Greek traditional folk music: examining the role of timbral semantics.

Asterios Zacharakis
Aristotle University of Thessaloniki
aszachar@mus.auth.gr

Konstantinos Pastiadis
Aristotle University of Thessaloniki
pastiadi@mus.auth.gr

Athena Katsanevaki
Aristotle University of Thessaloniki
atekatsa@mus.auth.gr

ABSTRACT

This paper presents an empirical experiment aiming to investigate the potential influence of timbral semantics on tension induction in Greek traditional folk music. To this end, a group of seventeen listeners rated the evolution of auditory *luminance*, *texture* and *mass* together with the felt *tension* over sixteen musical excerpts in real-time. Correlation and regression analyses between these four quality profiles for each particular stimulus showed that all three examined timbral qualities had instances of very strong association with tension. Although auditory mass featured the greatest number of such instances, no safe conclusion can be reached based on current findings regarding the most influential timbral semantic dimension for tension induction. Instead, it seems that a combination of conditions (i.e., musical parameters) can either maximise or minimise the influence of each timbral dimension.

1. INTRODUCTION

Music is believed to draw a significant amount of its appeal from its ability to stimulate emotional responses that alternate between tension and relaxation (e.g., Huron, 2006; Lehne & Koelsch, 2015). Krumhansl (2015) sums up the cognitive view of musical tension by stating that it is created when an expected event is delayed or when the context is ambiguous. In general, the tension-relaxation phenomena in Western music have mostly been studied with a focus on harmony, melody, dynamics and rhythm, while timbre –being a domain-general psychological feature as Farbood (2012) puts it- has been largely underrepresented with the notable exception of Pressnitzer et al. (2000). However, Farbood & Price (2017) have recently investigated some timbral attributes with respect to tension and reported that higher degrees of roughness, inharmonicity and spectral flatness of musical tones were associated with higher tension ratings. The authors of this work suggest that the effect of timbre on tension should be further investigated in more ecologically valid settings.

In Greek traditional folk music extreme manipulations of expectations are not the norm. With the exception of improvisational parts, melodic and rhythmic structures tend to be repetitive, the orchestrations are generally fixed throughout a given song and climactic moments - that according to Huron (2006) constitute the epitome of a tension-relaxation schema in music making- are rare. Does this mean that tension build up is not intended by the folk music creators and in turn not experienced by the listeners? Or could it be that in more predictable musical creations, the timbral characteristics of a piece may have an important role to play with respect to conveying tension? If it so, which of the timbral qualities are most influential?

This study will aim to address the above questions based on the previously developed luminance-texture-mass (LTM) framework for musical timbre semantics (Zacharakis, Pastiadis & Reiss, 2014; 2015; Zacharakis & Pastiadis, 2016). The LTM framework suggests that the most salient semantic dimensions of musical timbre are luminance (i.e., bright vs. dull), texture (i.e., rough vs. smooth) and mass (i.e., full vs. empty). Therefore, this work will seek to identify for possible associations between the three dimensions of the LTM framework and felt tension in Greek folk music.

2. METHOD

A selection of 16 instrumental excerpts from various types of Greek traditional folk music (e.g., dances, dirges, Akritika, from Epirus, Asia Minor, Aegean islands, etc.) also including a variety of lead instruments was presented to seventeen listeners. Most excerpts constitute introductory or improvisational parts and range from 24 seconds to 68 seconds long. Vocal parts were avoided partly due the increased timbral heterogeneity (Sandell, 1995) that is introduced by a human voice and partly due to the semantic charge of the lyrics whose impact could not be controlled. The stimuli can be accessed online at: http://ccm.web.auth.gr/timbreandtension.html. Stimuli were equalised in loudness through informal listening within the research team and their RMS playback level was measured to be between 65 and 75 dB SPL (A-weighted, slow response). The presentation of the stimuli was made using a pair of PreSonus HD7 headphones. All listeners reported an equal loudness for all stimuli.

The participants (17) were students at the School of Music Studies of the Aristotle University of Thessaloniki (6 male, mean age: 21, average years of musical practice: 12.4) and they received course credit as compensation.

Participants rated continuously and in real-time the change in the timbral qualities *luminance*, *texture* and *mass* according to the LTM model, plus the felt *musical tension*[1] of each excerpt. The timbral qualities were orally

[1] Felt musical tension refers to the amount of tension that is actually felt by the participant as opposed to the amount of tension that he/she thinks that the stimulus is supposed to express.

elaborated very briefly by defining their two extremes to ensure that listeners have a common understanding of the concept. Positive luminance was defined as auditory brightness and negative luminance as auditory dullness, positive texture as auditory roughness and negative texture as auditory smoothness, finally positive mass as auditory fullness and negative mass as auditory emptiness. At this point, it has to be noted that the LTM framework for timbral semantics has been developed by empirical experiments on monophonic timbres. Therefore, the assessment of polyphonic music based on the LTM components is a novel element of this study. The concept of tension was elaborated as *inner tension (εσωτερική ένταση)* in order to avoid confusion that may have arisen due to the fact that the word for tension in Greek coincides with the term for sound volume. Inner tension was defined similarly to Farbood & Price (2017) as: *less tension corresponds to a feeling of relaxation or resolution, while more tension corresponds to the opposite direction*.

The rating device used was a small (16.9 x 21 cm) Wacom Intuos draw pen tablet set up like a mouse. Movement of the pen on the tablet on the right indicated increase of the quality under judgement while movement on the left indicated decrease of the quality. The values obtained were not limited by the physical dimensions of the tablet since the participant could simply reposition the pen on the tablet to get more available space just like he/she could do with a normal mouse. The data acquisition interface was custom designed in LabVIEW. It sampled the pen's horizontal axis coordinate every 5 milliseconds and offered participants a real-time visualisation of the profiles they were creating.

Rating on each of the four components was made in blocks of random order for each participant. In addition, the presentation order of the stimuli within each block was also randomised. As a result, all listeners eventually listened to each stimulus four times. The duration of the experiment was a little over an hour for most of the participants not including breaks (which they were advised to take whenever necessary in order to keep their concentration levels high).

3. ANALYSIS & RESULTS

Raw responses were subsampled by calculating the mean value over adjacent non-overlapping rectangular time windows (.5 secs = 10 samples). The resulting time series were subjected to first-order differentiation and replacement of positive/negative values with 1 and -1, respectively. The time series were next integrated and each participant's data were normalised within each quality by his/her maximum rating on this particular quality. Finally, the profiles were smoothed using a cubic spline interpolant and linear trends were removed from each individual participant's time series to ensure 'stationarity' (Dean & Bailes, 2010).

The inter-participant reliability analysis showed a good agreement for all qualities under study with luminance, texture, mass and tension featuring a Cronbach's Alpha of .81 and .86, .88 and .86 respectively.

Therefore, the processed time series were averaged over every stimulus and each of the four qualities under study. Figure 1 presents the mean profiles along with their 95% confidence intervals. Correlation analysis between the averaged time series revealed the relationships between timbral semantics and tension. The Pearson's correlation coefficient between tension and the timbral qualities luminance, texture and mass are presented in table 1. Overall, tension variation seems to be associated with variation in all three timbral semantic components to different extends depending on the particular stimulus.

Stimulus	Lead instrument	Luminance	Texture	Mass	Tension MIRToolbox
In a foreign land since a little boy	Qanun	.50**	.71**	.85**	.30*
The Rasti	Bagpipes	.94**	.89**	.89**	.41**
Zonaradikos Dance	Bagpipes	-	.50**	-.34*	.30*
The King throws a party	Violin	.80**	.91**	.92**	-
Dirge and Stroto Pogonisio	Clarinet	.23*	.75**	.40**	.35**
Lament	Lute	.69**	.95**	.87**	-
Karsilamas	Politiki Lyra	.79**	.24*	.69**	.41**
Tik dance	Pontiaki Lyra	.89**	.76**	.96**	.27*
Dirge from Epirus	Ney	.87**	.25*	.92**	.63**
Servikos dance	Ney	.71**	.96**	.31*	.52**
Yannis and the dragon	Violin	-	.78**	.80**	.77**
If you are going to foreign lands	Ney	.90**	.97**	.74**	-
Trygona	Santouri	.65**	-.48**	.72**	.39**
Sousta dance of Patmos	Bagpipes	.68**	.56**	.48**	.32*
The little Vlach boy	Clarinet	.95**	.90**	.97**	-.31*
Sebastian dance	Tabouras	.92**	.34*	.27*	-.60**

Table 1. Pearson's correlation coefficients between tension profiles and the profiles of the timbral qualities luminance, texture, mass together with tension calculated by the MIR Toolbox's miremotion. (**: p<.001, *:p<.05)

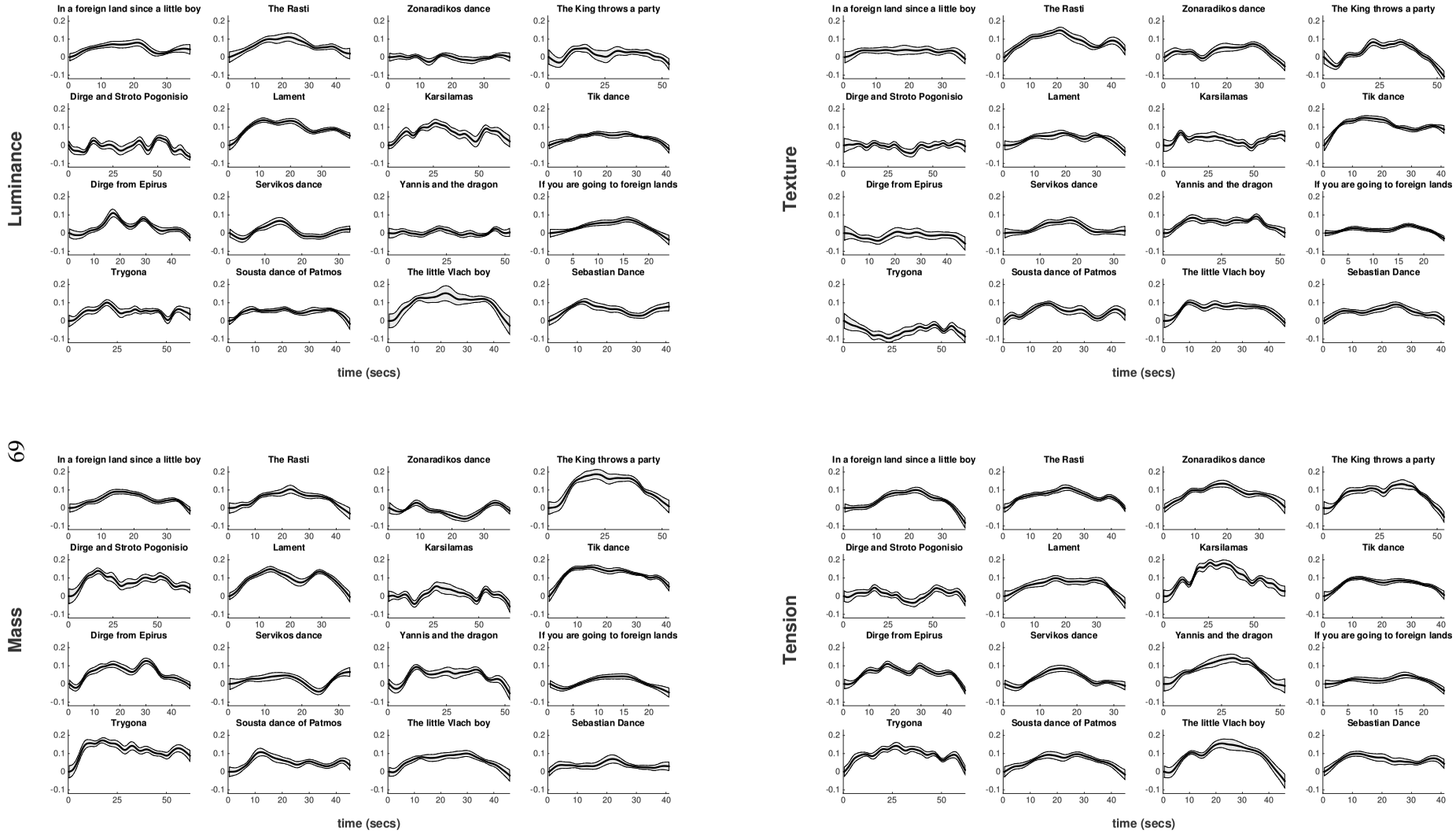


Figure 1. Mean temporal profiles corresponding to luminance, texture, mass and tension (in black) for each stimulus along with the 95% confidence intervals area in light grey. The values on the y-axis result from the normalisation process described in section 3.

In addition, a computational tension prediction calculated from the audio signal according to the *miremotion* function (Eerola et al., 2009) included in the MIR Toolbox (Lartillot & Toiviainen, 2007) was also compared to the empirically acquired tension profiles. The MIR tension estimation is based, among others, on calculation of dynamic, tonal and rhythmic variations. The window used for calculating tension through miremotion was 4 seconds long with 50% overlap. Subsequently the time series of the tension calculation were linearly extrapolated to exactly match the number of samples corresponding to the empirical tension profile for each stimulus. Finally, the time series of the miremotion tension were also smoothed using a cubic spline interpolant. Figure 2 shows the sixteen tension profiles that resulted from the above procedure and the last column of table 1 presents the Pearson's correlation coefficients between the empirical and the calculated tension profiles for each stimulus. With the exception of 'Yannis and the Dragon' where the relatonships between the MIR-tension and the

Stimulus	**Standardised beta**				**R-squared**
	Luminance	Texture	Mass	Tension MIRToolbox	
In a foreign land since a little boy			.85		.72
The Rasti	.90			.15	.91
Zonaradikos Dance		.51		.34	.36
The King throws a party			.92		.84
Dirge and Stroto Pogonisio		.71		.21	.62
Lament			.87		.76
Karsilamas	.75			.29	.72
Tik dance			.96		.93
Dirge from Epirus			.80	.29	.92
Servikos dance		.96			.93
Yannis and the dragon			.52	.41	.73
If you are going to foreign lands		.98			.95
Trygona			.66	.17	.55
Sousta dance of Patmos	.66			.24	.52
The little Vlach boy			.97		.94
Sebastian dance	.92				.84

Table 2. Multiple regression models for each stimulus using tension as dependent variable and the three timbral qualities plus tension calculated by the MIR Toolbox as predictors.

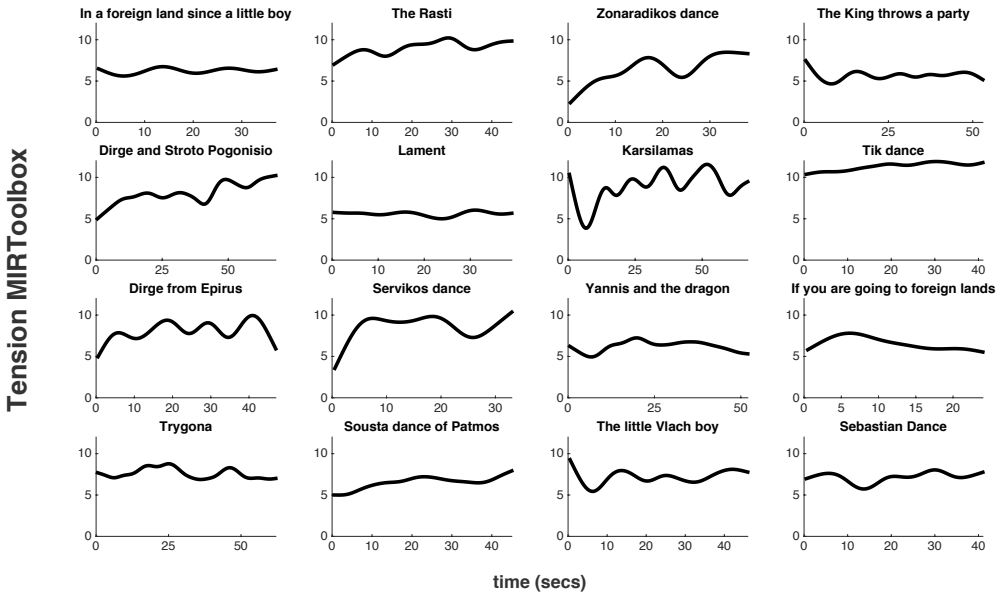


Figure 2. Temporal tension profiles calculated from the miremotion function of the MIR Toolbox.

LTM components with tension is comparable, in all other cases the MIR tension is more weakly correlated to tension compared to the LTM components, albeit in most cases the correlations are statistically significant.

Wanting to examine the influence of the timbral semantics along with other musical structures represented by tension-MIR we performed a two-step linear multiple regression analysis with one of the LTM components and MIR-tension as predictors and tension as the dependent variable. In the first step we picked the best predictor out of the LTM components and in the second step we examined whether additional inclusion of the MIR-tension contributed towards a better prediction of tension. The models were evaluated based on a combined maximisation of the explained variance (i.e., R-squared) and minimisation of the Akaike Information Criterion (AICc; Hurvich & Tsai, 1989). Table 2 succinctly presents the models that were favoured through this process.

Mass features eight appearances as the best predictor while *luminance* and *texture* appear only four times each, thus implying that *mass* may be more influential than the other two semantic components for tension perception. Also, in half of the stimuli, pairing MIR-tension with one of the LTM predictors contributes to a better model. Overall, the amount of tension variance explained (reflected by the R-squared values) is in many cases very high, as expected by the generally high correlations between tension and the LTM components.

4. DISCUSSION

This study constitutes a preliminary attempt to investigate a potential relationship between timbral semantics and tension perception using traditional folk Greek music as a vehicle. It should be viewed as a stimulation for further discussion on this thorny topic rather than a study that provides definitive answers having in mind that this approach is novel not only in the field of folk music but also in music perception literature in general. Using real-world polyphonic music as stimuli is particularly challenging for various reasons. First of all, in such a scenario all musical parameters (e.g., melodic contour, rhythmical patterns, dynamics, expressivity, timbre, etc.) vary concurrently and cannot be easily isolated. Secondly, the timbral heterogeneity (Sandell, 1995) inherent in some of our polyphonic stimuli makes the rating of timbral semantics of a whole excerpt of music a non-trivial task. Considering this fact, the agreement exhibited for the timbral ratings in particular is impressive.

All tension profiles except for 'Dirge and Stroto Pogonisio' feature one or two tension peaks at some point during the stimulus that are statistically significant (i.e., the lowest limit of their 95% confidence intervals is higher than the highest limit of the 95% interval of the 1st second of the stimulus). Such tension profiles have been typically identified in music perception literature (Krumhansl, 2015). The 'Dirge and Stroto Pogonisio' features a statistically significant minimum at the second half of the stimulus length. Even this sole appearance of decreasing tension profile shows that our participants were not biased towards reporting a bell-type tension profile at all instances. The LTM profiles feature a higher variability in shape, including increasing, decreasing, relatively stable and fluctuating profiles.

At this point, it has to be noted that the design of this experiment does not allow making any judgement with respect to the absolute value of the qualities under study for each stimulus. This is because participants judged only the variation of the qualities in time but did not have a way to inform us of the initial value of each quality. That is to say, a quality profile that does not vary much (e.g., the luminance profile for the 'Zonaradikos dance') does not necessarily imply a low overall judgement of this quality. The information on the initial absolute values of the four qualities for all our stimuli is going to be obtained and exploited in future work.

Such information will help to better examine the influence of timbre on perceived tension and to properly validate the perspective suggested by Huron (2006) according to which tension can be viewed as the dynamic subcategory of a generalised feeling of uneasiness that he calls *dissonance*. The other branch of this general dissonance is the static *sensory dissonance* in which timbral information belongs to. In this sense, the stress induced by static sensory means, such as timbre, should not be strictly viewed as tension but rather as a feeling of uneasiness. However, participants' agreement on tension ratings –reported in the present as well as in other studies- suggests that the terms tension and uneasiness can probably be used interchangeably.

Despite the above described caveats, our results demonstrate that timbral semantics are in many cases very strongly correlated with tension perception and that tension as calculated by the miremotion from the MIR Toolbox can -in half of the stimuli- be used in combination with LTM components to better account for felt tension variation. That is, of course, not to say that timbre is the most important attribute for tension perception. Since correlations do not imply causal relationships, it could well be the case that an underlying musical parameter (such as melodic contour, rhythmical density or dynamics) is a latent variable, affecting both timbre and tension perception. This hypothesis will also be investigated in future work.

In general, probably the most important finding of this study is that all three components of the LTM timbral semantics framework can potentially influence tension perception. This, however, does not always happen in a consistent manner as demonstrated by table 1. As an example, the two stimuli featuring the highest tension peaks ('Karsilamas' and 'The little Vlach boy') also feature the

highest ('The little Vlach boy') and third highest luminance peaks ('Karsilamas'). At the same time, however, the lowest ranked stimuli in terms of maximum luminance ('Yannis and the Dragon' and 'Zonaradikos dance') are ranked third ('Zonaradikos dance') and fifth ('Yannis and the Dragon') in terms of maximum tension. The same also stands for texture and mass. These results may pose a partial challenge to past findings supporting that sonorities of higher roughness and mass are tension-provoking means in music (Pressnitzer et al., 2000; Huron, 2006; Farbood & Price, 2017) by suggesting that this may only be conditionally true.

A similar type of inconsistency emerges from a preliminary musicological analysis of our excerpts. For 'Karsilamas' (featuring one of the highest maxima in both luminance and tension profiles) the peak in the profiles seems to coincide with a rise in melodic pitch. The same stands for 'Servikos dance', 'Rasti', 'The King throws a party' and 'Yannis and the dragon'. A notable exception, however, is 'The little Vlach boy' where while pitch decreases, tension rises to reach the maximum peak out of all tension profiles. In this case, the violation of the norm could be attributed to the existence of strong melodic attractions and existence of chromatisism. The above examples, demonstrate that tension perception is a multifaceted phenomenon that is not influenced by one attribute alone.

Some other types of idiosyncratic elements that affect the profiles of our qualities can also be reported. The small fluctuations evident in the profiles of 'Trygona' could probably be due to the impulsive character of Santouri, while local rises in tension, luminance or mass could be attributed to glissandi or melodic embellishments (e.g., in 'Rasti', 'Dirge and Stroto Pogonisio', 'Dirge from Epirus' and 'Servikos dance'). A detailed musicological analysis in respect with the acquired LTM and tension profiles will be another scope of future work.

Finally, another issue that warrants further investigation is a fine-tuning of the window size for the MIR-tension calculation. The selection of a four-second window was made in this work as a reasonable choice that would successfully simulate human reaction time and working memory. However, other possible lengths and/or temporal adjustments may yield better associations with the empirical data.

5. ACKNOWLEDGEMENTS

This study was supported by a post-doctoral scholarship issued by the Greek State Scholarships Foundation (grant title: "Subsidy for post-doctoral researchers", contract number: 2016-050-050-3-8116), which was co-funded by the European Social Fund and the Greek State.

The authors would like to thank the participants of this study.

6. REFERENCES

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

Eerola, T., Lartillot, O., & Toiviainen, P. (2009). Prediction of Multidimensional Emotional Ratings in Music from Audio Using Multivariate Regression Models. In *Proceedings of the International Society of Music Information Retrieval Conference (ISMIR 10)* (pp. 621-626).

Farbood, M. M. (2012). A parametric, temporal model of musical tension. *Music Perception*, *29(4)*, 387-428.

Farbood, M. M., & Price, K. C. (2017). The contribution of timbre attributes to musical tension. *The Journal of the Acoustical Society of America*, *141*(1), 419-427.

Huron, D. B. (2006). *Creating tension* in Sweet anticipation: Music and the psychology of expectation. MIT press: London, UK.

Hurvich, C. M., & Tsai, C. L. (1989). Regression and time series model selection in small samples. *Biometrika*, *76*(2), 297-307.

Krumhansl, C. L. (2015). Statistics, structure, and style in music. *Music Perception, 33(1)*, 20-31.

Lartillot, O., & Toiviainen, P. (2007). A Matlab toolbox for musical feature extraction from audio. In *Proceedings of the International Conference on Digital Audio Effects (DAFx)* (pp. 237-244).

Lehne, M., & Koelsch, S. (2015). Towards a general psychological model of tension and suspense. *Frontiers in Psychology*, 6, 79.

Pressnitzer, D., McAdams, S., Winsberg, S., & Fineberg, J. (2000). Perception of musical tension for nontonal orchestral timbres and its relation to psychoacoustic roughness. *Perception & psychophysics*, *62*(1), 66-80.

Sandell, G. J. (1995). Roles for spectral centroid and other factors in determining" blended" instrument pairings in orchestration. *Music Perception, 13(2)*, 209-246.

Zacharakis, A., & Pastiadis, K. (2016). Revisiting the luminance-texture-mass model for musical timbre semantics: A confirmatory approach and perspectives of extension. *Journal of the Audio Engineering Society*, *64(9)*, 636-645.

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

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

Abstracts

The Hidden Modes: A computer-assisted approach to tonality analysis of Swedish Folk Music

Sven Ahlbäck
Kungliga Musikhögskolan,
Stockholm, Sweden
Sven.ahlback@kmh.se

1. INTRODUCTION

The term tonality is used in numerous ways with different significations in ethnomusicological, musicological and music cognition literature, and in its most general sense it refers to a cognitively and/or culturally significant hierarchical system of relationship between pitches in a piece of music, in a repertoire or music culture (Bengtsson 1977). Tonality in this general sense corresponds to the concept of tonal hierarchy, representing relative prominence, stability and structural significance of musical tones in a musical context, a concept which has been frequently addressed in empirical studies within the field of music cognition over the past 40 years (Krumhansl & Cuddy 2010). The empirical research in tonal hierarchy, involving both psychological studies and computer modelling, show that listeners cognition of tonal hierarchy can be related to statistical distribution of tones in musical contexts, a correlation which is also shown to be cross-culturally valid (e.g. Eerola 2004, Krumhansl & Cuddy 2010, Stevens 2004).
In the present study computer-assisted modelling of tonal hierarchy by means of analysis of statistical distribution of tones, is applied on traditional Swedish folk music in order to examine idiomatic features of tonality in this music corpus. The aim is to explore whether this approach can provide new insights regarding stylistic unity and diversity within this musical repertoire. More specifically, the concept of key-profile is applied, representing the structural prominence of chromatic tones within a mode and key.
This is not trivial for a number of reasons, such as the frequent use of micro-tonal variation of intonation within the style, observed by scholars as early as in the beginning of the 19th century as a significant trait of Swedish folk music (Boström, Lundberg & Ramsten 2010, Ahlbäck 2010). Some previous studies of tonal hierarchy are problematic with regards to this issue, assuming a chromatic pitch category set as a fundament, an underlying, invariant scale structure and not taking musical context into consideration in the estimation of structural prominence of tones.
Moreover, studying stylistic features of traditional folk music in Sweden we are faced with the problem of a diverse source material, encompassing contemporary commercial recordings, contemporary & historical field recordings as well as transcriptions and collections of music notations made under a period of over 200 years, under very different conditions, for different purposes and of varying quality regarding e.g. detail.
A purpose of the present study is to develop methodology in order to be able to make comparisons between source material with different level of musical detail.
In order to compare different source material and obtain comparable musical representations, a model for automatic musical structure analysis developed by the author is used

2. THE PRESENT STUDY

The source material used in the study is a combination of recordings and notated material of Swedish traditional folk music digitized into the same digital music representation (Ahlbäck 2004, ScoreCloud 2013). This system automatically transcribes sound or MIDI input into standard western staff notation, including quantization, metrical analysis, pitch categorization, ornamentation analysis and segmentation of melodies, which makes it possible to compare sounding and symbolic input on different level of detail.

The source material includes vocal and instrumental herding music (vallåtar), related to the traditional herding culture in Scandinavia and Fiddle music, mainly from the same geographic area, consisting of notations/transcriptions from the 20th century and field recordings. Handwritten manuscripts, so called "fiddlers books" with instrumental popular fiddle tunes from the 18th century and 19th century from the same geographic area as the recordings and collected notations was used as a reference material. The historical handwritten manuscripts shared a number of tunes (melodic themes) with the recorded material. A total of 2100 melodies were used in the study.

The method used for obtaining the key profiles was developed from methods used in previous studies, based on mainly relative duration and frequency of appearance of chromatic pitch categories, assuming octave-equivalence by Krumhansl and others (Krumhansl 1990). In the present study, other statistical features, mentioned in ethnomusicological literature (Nettle 1964), were also included in the measure of structural prominence, such as metrical prominence and melody structural prominence, as well as features motivated by psycho-acoustical research (Parncutt 1994). Moreover, melodic context was also included in the measure of structural prominence in terms of connectivity between pitch categories within a musical context. This was motivated by the significance of melodic structure in relation to intonation of scale degrees typical for many modal systems, including western major-minor tonality.

From the total dataset an initial categorization was made based on the correspondence with major or minor mode profile respectively. It was only possible to obtain a balanced subset of the data for the three different data sets (herding calls, fiddle music transcriptions and historical manuscripts) for match with the minor mode profile (Krumhansl 1990). This particular subset consists of 450 melodies, 150 melodies from each source corpus and constitutes the comparative material for this study. The analysis was performed automatically by the system.

In order to make comparison possible micro-tonal alterations of pitch were automatically assigned to closest chromatic pitch category.

3. RESULTS

The results show that the obtained mode-profiles (key-profiles transposed to the same "key") differ between the different repertoires in the study, showing interesting features for different repertoires. These differences were most significant when using the more elaborate feature set for measuring structural prominence of tones.

However, also when using simple statistical measures such as relative duration and frequency of appearance of tones, differences and connections between repertoires show.

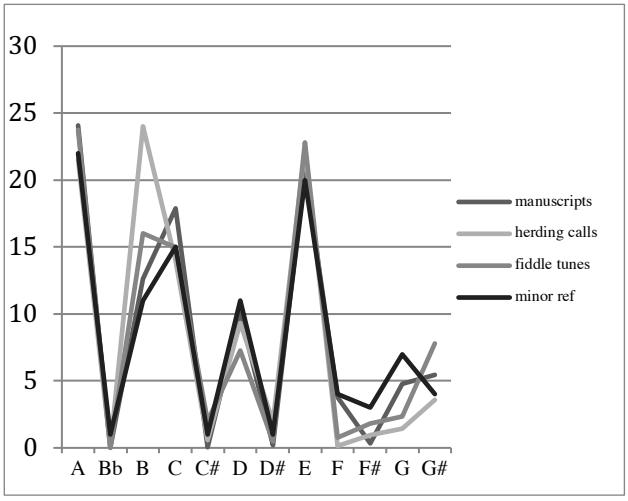


Figure 1. Relative prominence of chromatic pitch categories (percentage) for different repertoires in the study within category "minor mode" key profiles, in comparison with minor key profile (Krumhansl 1990). The neg. correlation between herding call and manuscript and herding call - minor reference are significant $p > 0.05$ (Pearson linear), as well as the pos. correlation between manuscript and minor

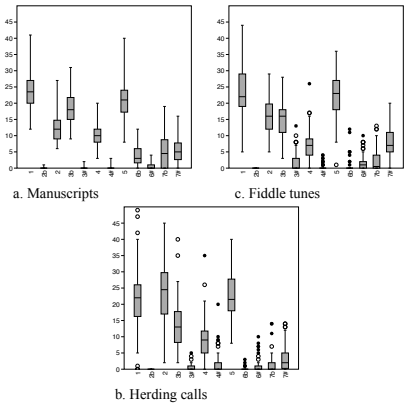


Figure 2. Data distribution for each chromatic category within the three data sets. Mean standard error for a 0.40, for b 0.58 and for c 0.52. N=150 for each set.

As can be seen in figure 1 and 2, the herding music and the fiddle repertoire share certain structural features such as the relatively higher prominences of the second scale degree in relation to the third scale degree in comparison with the manuscript and minor mode profiles. It might be interpreted as an influence of the herding call music in the fiddle music repertoire. The study indicates that sub-modes can be found within what is traditionally in a Swedish context categorized in terms of major and minor mode, but actually challenges this categorization, even when compensating for micro-tonal variation of intonation.

Furthermore, the study indicates that taking melodic contextual factors into account in statistical measurement of mode profile, can reveal structural stylistically significant features of tonal hierarchy.

4. REFERENCES

Ahlbäck, S. (2004) Melody Beyond Notes. A study in melody cognition. (Doctoral thesis) Publications from the department of Musicology, Göteborg University

Ahlbäck, S. 2010, Svenska Låtar som musikalisk källa, In Boström, Mathias, Lundberg, Dan & Ramsten, Märta (ed.) 2010, *Det stora uppdraget*, Nordiska museets förlag, Stockholm

Bengtsson, I., (1979). Rytm, *Sohlmans Musiklexikon*, volume 5:248-250, Stockholm: Sohlmans förlag.

Boström, M., Lundberg, D. & Ramsten, M. (ed.) 2010, *Det stora uppdraget*, Nordiska museets förlag, Stockholm

Eerola, T. (2004) "Data-driven influences on melodic expectancy: Continuations in North Sami yoiks rated by South African traditional healers", In S. Lipscomb, R. Ashley, R. O. Gjerdingen and P. Webster, (ed) *Proc. 8th ICMPC.*

Krumhansl, C.L. (1990) *Cognitive Foundations of Musical Pitch*. New York: Oxford University Press

Krumhansl, C.L. & Cuddy, L. (2010) A Theory of Tonal Hierarchies in Music. In M.R. Jones et al. (eds.), *Music Perception*, Springer Handbook of Auditory Research 36, (3), 51-87
Nettl, B. (1964). *Theory and Method in Ethnomusicology*. London, Macmillan Ltd
Parncutt, R. (1989) Harmony: A Psychoacoustical Approach. Berlin: Springer-Verlag.
ScoreCloud http://www.scorecloud.com. DoReMIR Music Research AB 2011-2016.
Stevens, C. (2004) Cross-cultural studies of musical pitch and time. In Acoust. Sci. & Tech. 25, 6, 433-438

IMITATIONS-TRANSFORMATIONS: BIRDS OF PARADISE IN PERFORMANCE FROM THE CENTRAL PROVINCES OF PAPUA NEW GUINEA

George Athanasopoulos
Aristotle University of Thessaloniki
georgenathanasopoulos@gmail.com

1. EXTENDED ABSTRACT

Could singing and courtship displays of birds be considered music and dance? If so, could they provide the basis for ritualized human behaviour aiming at social cohesion and bonding? Birdsong has inspired western composers in the past, and was noted down using WSN up until the middle of the 20th century, implying inspired by the perception of inherent music-like qualities (Mundy, 2009), such as rhythm, melody, repetition and variation. Few in the western world today would call birdsong music, however, as it lacks human creativity and meaning (Titon, 2015). The fact that the imitation of birdsong and bird courtship display acquires meaning as coordinated musico-ritual activity in performance for many tribes inhabiting the mountainous grasslands of central Papua New Guinea is unfortunately often ignored. These activities often aim at strengthening social organization and group cohesion (Patel, 2010), and on occasion is thought to possess metaphysical properties (Feld, 1982). Furthermore, movement to music, and the activity of making music in collaboration with others, is considered a key component of activities which are central to ritual, courtship, identity, and human expression across the majority of human cultures.

The aim of this paper is to present a sampling of ritual "bird-song" performances from the Huli, Melpa, Enga, Bena-Bena and Abau tribes of Papua New Guinea, and to compare them to the actual birdsongs and courtship displays (where applicable) from which they yield their inspiration. This juxtaposition involves comparison between birdsong and bird movements with human music activity and dance.

As this ethnomusicological analysis of songs and dances imitating local bird fauna runs parallel with sonic information recorded in nature, this juxtaposition is carried out through sonogram and frequency analysis of song performances using Audacity and through audio data captured during fieldwork in 2010 in Goroka and Mt. Hagen. Audacity is a free open source digital audio editor and recording computer software application which was considered as an appropriate tool for its easiness of use in the field by a non-expert in sound editing and processing. Additionally, the Raven Interactive Sound Analysis Software is also deployed in further analysis of the sound data as it specializes in birdsong.

Synchronizing movements in performance, a common element in Papua New Guinean "bird imitation" dances, is thought to "merge" an individual's self with others, via neural pathways that code for both action and perception (Overy and Molnar-Szakacs, 2009). Though the short sample size does not permit broader assumptions through observation, it is possible to yield interesting results regarding this organized display of social behaviour in ritual song/dance performance in Papua New Guinea. It has to be stressed however, that, as this correlational exploratory research study focused on participant performance activities occurring in natural context, it was only through interviews and bibliographic research that the actual causes of such behaviours were determined.

The results of this study forces us to reconsider the nature of "bird imitation" dances not as mere mimicry of nature, as Allen and Dawe rightly observe (2015), but as a form of a collective group activity, pervading through the history of music as a social interaction among our species (Nettl, 2000; 2010).

Keywords: *ecomusicology; ethnomusicology; Papua New Guinea; bird-of-paradise; performance ritual.*

2. REFERENCES

Allen, A. S., & Dawe, K. (Eds.). (2015). *Current Directions in Ecomusicology: Music, Culture, Nature*. Routledge.

Feld, S. (2012/1982). *Sound and Sentiment: Birds, Weeping, Poetics, and Song in Kaluli Expression, with a new introduction by the author*. Duke University Press.

Mundy, R. (2009). Birdsong and the Image of Evolution. *Society & Animals*, *17*(3), 206-223.

Nettl B. (2000). "An ethnomusicologist contemplates universals in musical sound and musical culture," in *The Origins of Music*, eds Wallin N. L., Merker B., Brown S., editors. (Cambridge, MA: MIT Press;) 463–472

Nettl, B. (2010). *The study of ethnomusicology: Thirty-one issues and concepts*. Urbana, IL: University of Illinois Press.

Overy, K., & Molnar-Szakacs, I. (2009). Being together in time: musical experience and the mirror neuron system. *Music Percept.* 26, 489–504. doi: 10.1525/mp.2009.26.5.489

Patel, A. D. (2010). *Music, language, and the brain*. Oxford university press.

Tarr B, Launay J & Dunbar RIM (2014) Music and social bonding: "self-other" merging and neurohormonal mechanisms. *Front. Psychol.* **5**:1096. doi: 10.3389/fpsyg.2014.01096

Titon, J. T. (2015). *Worlds of Music: An Introduction to the Music of the World s Peoples*. Nelson Education.

A NON-MELODIC CHARACTERISTIC TO COMPARE THE MUSIC OF MEDIEVAL CHANT TRADITIONS

Geert Maessen
Gregoriana Amsterdam, Amsterdam, The Netherlands
gmaessen@xs4all.nl

Peter van Kranenburg
Meertens Intstituut, Amsterdam, The Netherlands
Utrecht University, Utrecht The Netherlands
peter.van.kranenburg@meertens.knaw.nl

1. INTRODUCTION

In the scholarly discourse on the origins of Gregorian chant (GRE) several medieval chant traditions play key roles, notably Old Roman (ROM), Milanese (MIL) and Beneventan chant (BEN). Although hardly anything is known with certainty about Gallican chant, which was in existence in Gaul before GRE, this chant also played an important role. The chant of the Mozarabic rite has long been considered of major importance as well. However, due to the lack of pitch-readable sources it was virtually absent in the discussion. (Levy, 1998)

At the end of the eleventh century the Mozarabic rite and its chant were officially replaced by the Roman rite with its chant (GRE). Over 5,000 chants, however, are preserved in neumatic notation dating from the ninth to thirteenth centuries. In this notation, precise intervals (apart from incidental primes) cannot be read (Randel, 1973). The most complete manuscript is the early tenth-century León antiphoner (LEO) with over 3,000 notated chants (for an example see Figure 1).

In Toledo, six parishes were allowed to continue the tradition. Oral descendants of this tradition, heavily mixed with a newly invented tradition, were finally written on the staff in the early sixteenth century (MOZ). Only a few dozen melodies agreeing with the early neumatic notation of the Mozarabic rite were ever found in pitch-readable notation. We know, however, that there must have been melodic relations between the lost tradition and traditions preserved in pitch-readable notations. Based on the chant texts and the number of notes per syllable Kenneth Levy (1998) has shown in detail that some LEO sacrificia must have been musically related to offertories on the same texts in GRE, ROM and MIL.

In a previous paper, we have shown that it is possible to produce melodies agreeing in all detail with our knowledge of the early neumes (Maessen & Van Kranenburg, 2017). In order to produce melodies with higher historical probability we compiled a data set of all GRE, ROM, MIL, BEN and MOZ offertories (Van Kranenburg & Maessen, 2017). For the current study, we additionally have encoded 25 out of all 102 LEO sacrificia (2 to 5 parts each); some from beginning, end and middle of the manuscript, and several pieces of specific interest, including Levy's sacrificia (20,000 notes in total). Based on the intervals between consecutive notes in the traditions of this data set we have shown that parts of the chants can be classified with very high accuracy. In order to extend this classification to include LEO we needed a pitch-independent feature that is shared by all six traditions. The number of notes on a syllable of chant text is such a feature. We defined 15 categories for the number of notes per syllable. All melisma lengths up to 10 we consider separate categories, and we added categories for 11-15, 16-20, 21-25, 26-50, and 50 or more notes. In this study, we perform a (zeroth-order) dimension reduction analysis, and we train (first-order) bigram language models.

2. CLASSIFICATION

For each tradition, we trained a bigram language model on the representation of the chants as sequences of melisma categories. We followed the same procedure as in Van Kranenburg & Maessen (2017): we computed for each chant part the perplexity for each tradition, and we assigned the chant part to the tradition with the lowest perplexity. In all cases the query chant was excluded from the language model of its own tradition. Although classification appeared to be less precise compared to

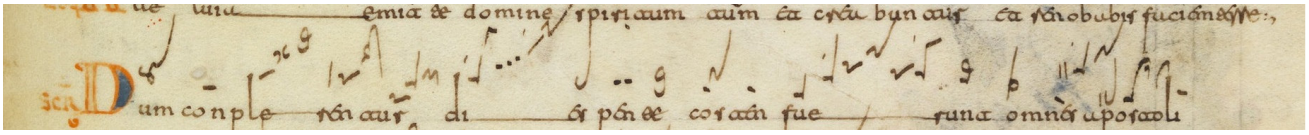


Figure 1. The first line of the sacrificium *Dum complerentur* in the León antiphoner (E-L 8; 210r14).

pitch-based models, most chant parts in each tradition, could be classified correctly. The fourth row in Table 1 shows that LEO chant parts are classified 65 % as LEO and never as MOZ. ROM has highest classification score (80 %), BEN lowest (54 %). When we drop LEO as a target class, 70 % of LEO chant parts are classified as GRE and again none as MOZ (see Table 2).

3. DIMENSION REDUCTION

For a better understanding of the relations between the six traditions we also performed a dimension reduction using the t-SNE algorithm (Van der Maaten & Hinton, 2008). We used the occurrence rates of the melisma categories as features. Here, the result was somewhat dependent on the way we categorized the number of notes per syllable. Nevertheless, there are clear trends observable which are consistent across different configurations and different runs of the algorithm. Figure 2 shows a typical 2D embedding of the chant parts. Here, again, it is observable that LEO is most close to GRE, and that ROM and MOZ are most alien to each other.

	MOZ	MIL	GRE	LEO	BEN	ROM	parts
MOZ	74,1	8,6	10,8	4,3	1,4	0,7	139
MIL	3,4	63,3	22,4	4,8	2,7	3,4	147
GRE	1,7	13,1	61,3	8,1	4,7	11,0	394
LEO	0,0	6,3	19,0	65,1	6,3	3,2	70
BEN	2,4	12,2	14,6	4,9	53,7	12,2	41
ROM	1,8	3,5	9,5	2,1	2,5	80,7	285

Table 1. Classification (in %) of parts in six traditions.

	MOZ	MIL	GRE	BEN	ROM	parts
MOZ	77,7	8,6	12,2	0,7	0,7	139
MIL	3,4	64,6	22,4	6,1	3,4	147
GRE	1,8	13,5	67,5	5,3	11,9	394
LEO	0,0	12,9	70,0	5,7	11,4	70
BEN	2,4	22,0	22,0	46,3	7,3	41
ROM	1,8	4,6	8,1	1,8	83,9	285

Table 2. Classification without LEO as a target class.

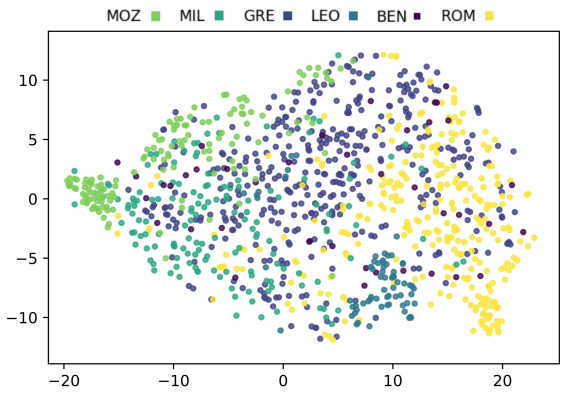


Figure 2. Dimension reduction for six traditions.

4. CONCLUSION & FUTURE WORK

Some of the misclassifications and outliers of LEO found in Sections 2 and 3 are striking. LEO 053, *Erit hic vobis,* and 075, *Oravi Deum meum,* e.g. are two of only three extreme outliers (the other is 027, *Sicut cedrus*). These two chants are also two of only three Levy chants that are clearly related to their GRE counterparts (the third is 098, *Sanctificavit Moyses*). The GRE counterpart of LEO 075 also was the most extreme outlier in our interval-based analysis. Since this indicates that *Oravi* is alien to both LEO and GRE, Levy may have been right in stating "the most plausible origin" for *Oravi* and *Erit* as of "Gallican or mixed Mozarabic-Gallican usage" (Levy, 1998). Other traditions he excluded on other grounds. Apart from *Oravi* and *Erit,* now also *Sicut cedrus* and 099, *Congregavit David,* are qualified as candidates for Gallican chant. Another observation by Levy, the possible Gallican origin of five cognate GRE-MIL pairs, is confirmed for two of them by our analysis based on numbers of notes only: *Angelus Domini* and *Oratio mea.*

Our main conclusion: it is possible to make claims about relations of chant melodies in different traditions without reference to their texts and pitch content. This means that "a systematic exploration of the preserved Mozarabic repertory with a view to identifying any Gallican residue" (Levy, 1998) has become much easier. For the lost melodies of the Mozarabic rite it also means that, using our classification results, we are able to relate each chant to a specific tradition and take this tradition as the basis for the production of a melody.

In order to make our claims more precise we will continue our investigation in the way how to handle the different numbers of notes on chant syllables and in encoding the sacrificia of the León antiphoner.

5. REFERENCES

Levy, K. (1998). Toledo, Rome and the Legacy of Gaul. In *Gregorian Chant and the Carolingians*, 31-81.

Maessen, G. & Van Kranenburg, P. (2017). A Semi-Automatic Method to Produce Singable Melodies for the Lost Chant of the Mozarabic Rite. In *Proceedings of the 7th International Workshop on Folk Music Analysis 14-16 June 2017, Málaga, Spain,* (pp. 60-65).

Randel, D. (1973). An Index to the Chant of the Mozarabic Rite. New Jersey: Princeton University press.

Van der Maaten, L.J.P. & Hinton, G.E. (2008). Visualizing High-Dimensional Data Using t-SNE. *Journal of Machine Learning Research* 9:2579-2605.

Van Kranenburg, P. & Maessen, G. (2017). Comparing Offertory Melodies of Five Medieval Christian Chant Traditions. In *Proceedings of the 18th International Society for Music Information Retrieval Conference, Suzhou, China, October 23-27,* (pp. 204-210).

EFFORT-VOICE RELATIONSHIPS IN INTERACTIONS WITH IMAGINARY OBJECTS IN HINDUSTANI VOCAL MUSIC

Stella Paschalidou
Dept. of Music Technology & Acoustics Eng.,
TEI of Crete, Greece
pashalidou@staff.teicrete.gr

Martin Clayton, Tuomas Eerola
Dept. of Music, Durham University, UK
martin.clayton@durham.ac.uk,
tuomas.eerola@durham.ac.uk

ABSTRACT

In Hindustani (dhrupad) vocal improvisation singers often engage with melodic ideas by manipulating intangible, imaginary objects with their hands while singing, such as through stretching, pulling and pushing. Such engagements ('MIIO' for Manual Interactions with Imaginary Objects) suggest that some patterns of change in the acoustic features relate to rudimentary interactions and the levels of effort that the respective objects may afford due to their physical properties. Through this work we seek to gain a deeper understanding of performance practice in the dhrupad music tradition in the specific cases where the singer seems to interact with imaginary objects, by examining whether effort and gesture types appear in an arbitrary fashion along with the voice or if they are related to the sound in a consistent way. The results suggest that a good part of the variance in both physical effort and gesture type can be explained through a small set of audio and motion features.

1. INTRODUCTION

In recent years we have seen a shift towards more embodied approaches in the study of music performance. However, while sound-producing gestures have drawn strong attention, studies on sound accompanying gestures (as in singing) have been extremely rare, even more so in terms of computational approaches (Luck & Toiviainen, 2008) and in the case of non-western 'oral' music traditions (Clayton & Leante, 2013) like the one portrayed here. Additionally, although physical effort has been stressed as one of the important aspects in music performance, systematic approaches in its role still remain limited.

Here effort is understood as a concept which reflects the active or passive attitude of the person in fighting against or giving in to the physical conditions that influence the movement while trying to achieve an intentional task (Hackney, 1998). During MIIIOs Hindustani singers seem to act as if they encounter an increased resistance upon their hands, presumably imitating the effort that would have been induced in handling real objects in our natural environment. Although MIIOs can be considered as founded on rudimentary knowledge of interacting with the environment, whether and how these are related to their melodic counterpart is a non-trivial question. Examining whether, how and to what extent effort and gesture types appear associated with the voice in a consistent way is where this work aims to offer a contribution.

2. METHOD

The work uses a mixed methodological approach, combining qualitative (ethnographic) and quantitative methods based on original material that was recorded in domestic spaces in India. This consists of interviews (8 vocalists), audio-visual material (4 vocalists) and motion (10-camera passive-marker Optitrack system) capture data (2 vocalists) of vocal improvisations by different dhrupad vocalists of the same music lineage. We argue that the sequential approach of combining the rich outcomes of qualitative methods with the compact results of quantitative methods can offer a more rigorous and comprehensive picture of the phenomena under study.

Real performances rather than designed experiments were used for the purposes of ecological validity. Singers were asked to improvise without any further instructions and were recorded only during the alap improvisation (the initial slow non-metered section, sung to a repertoire of non-lexical syllables), in order to concentrate on melodic factors rather than the metrical structure or lyrical content in the later stages of the rāga performance.

In the qualitative part of the analysis, we first applied a thematic analysis to the interview material in order to identify action-based metaphors, which informed the annotation process of the video material that followed. For the video analysis we relied on third-person observations that aimed at identifying, labeling and later classifying the audio-visual material in terms of recurrent types of MIIOs (categorical), as well as perceived effort levels (numerical) that appear to be exerted by the performer in a range between 0 (lowest) and 10 (highest). Such annotations were cross-validated by two choreographers.

In the quantitative part of the study, we used the cross-validated annotations as response values of linear models that were fit to measured movement and sound features in order to (a) estimate effort levels and (b) classify gestures as interactions with either elastic (through elasticity) or rigid (through weight/friction) objects. Two vocalists were used here, namely Afzal Hussain (rāga Jaunpurī) and Lakhan Lal Sahu (rāga Mālkauns). The features that were used for estimating the responses were computed by first extracting time-varying movement and audio features from the raw data, and then computing representative statistical global measures (such as mean, SD, min, max). We started by using the features reported in (Nymoen et al, 2013), but then a number of alternative features were also explored that were meant to raise the explained variance of the estimated responses.

3. RESULTS

Two variations of linear models were developed for each task (effort estimation and gesture classification):
(1) a model that best fits each individual performer, thus better reflecting the idiosyncratic aspect of each singer;
(2) a model that can better describe shared, more generic cross-performer behaviours.

3.1 Idiosyncratic schemes

3.1.1 Effort levels

Different idiosyncratic schemes of associating the perceived physical effort with acoustic and movement features were identified, that are based on the pitch space organisation of the rāga and the mechanical strain of voice production.
Hussain: Higher effort levels are required when the hands move slower and wider apart and with a larger speed variation. They are accompanied by melodic glides that start from lower degrees and ascend to higher degrees of the rāga scale within the boundaries of each individual octave, thus they are associated with characteristic qualities of the specific rāga. The use of 5 non-collinear audio and movement features in the linear models that were developed yielded a good fit (R^{2adj}) of about 60%.
Sahu: Higher bodily effort is required for hand movements that exhibit a larger variation of hand divergence (speed in moving the hands further apart), with a strong onset acceleration. They are accompanied by larger melodic glides that reach up to higher maximum pitches, reflecting the increased mechanical strain of voice production. As the alap is organised based on a gradual ascent towards the pitch climax, pitch is here also representative of the alap macro-structure. The use of 4 non-collinear audio and movement features in the linear model yielded an adequately good fit (R^{2adj}) of about 44%.

3.1.2 Gesture classification

Different modes of gesture class association with acoustic and movement features were identified, that are based on regions of particular interest in the rāga pitch space organization and analogous cross-domain morphologies.
Hussain: It is more likely that interactions with elastic objects (rather than rigid) are performed by hand gestures that exhibit a low absolute mean acceleration and a large variation in hands' divergence. They are associated with slower and larger melodic movements that ascend to a higher degree of the scale. Interestingly, the highest degree happens to be the most unstable degree of the scale (in rāga Jaunpurī), which imposes a subsequent pitch descent (i.e. a double pitch glide), similar to the change of direction observed by the hands when interacting with an elastic object. Thus, it could be suggested that MIIO types are associated with the grammatical rules of the rāga. The use of 5 non-collinear audio and movement features in the logistic models that were developed yielded a high classification rate (AUC) of about 95%.
Sahu: Interactions with elastic objects are more likely performed with pitch movements of a larger interval and larger duration and with the hands moving faster and remaining bound to each other. The use of 4 non-collinear audio and movement features in the logistic models yielded a high classification rate (AUC) of about 80%.

3.2 Generic scheme

3.2.1 Effort levels

Higher bodily effort levels are required by both singers for melodic movements that start from a lower and reach up to a higher pitch, reflecting the mechanical requirements of voice production. They are accompanied by movements which are slow on average but exhibit a large variation of speed, and in the specific case of Hussain when the hands move further apart. Two almost identical linear models were developed, yielding a good fit (R^2adj) of about 53% (with 5 features) for Hussain and 42% (with 4 features) for Sahu respectively.

3.2.2 Gesture classification

Interactions with elastic objects are more likely to be performed at lower pitches for larger melodic movements, and with the hands moving further apart for Hussain and less apart but faster in the case of Sahu. Two almost identical general logistic models were developed, yielding a good fit (AUC) of about 86% (with 3 features) for Hussain and 78% (with 4 features) for Sahu respectively.

4. CONCLUSIONS

MIIOs offer a special case where motor imagery is "materialised" through physical actions directed towards an imagined object. Despite the flexible character of music-movement correspondences, there is ample evidence of more generic associations that are not necessarily performer-specific or stylistic. I suggest that the vocalists' capacity of imagining musical sound is facilitated through the retrieval of motor programs and image schemata from well-known real interactions with real objects and that this may be exactly the reason for which imaginary objects are employed.
As much as bringing the advantages of ecological validity in combining ethnographic data with exact measurements of real performances, the approach that was followed has also posed important challenges and limitations, such as the limited dataset. Larger datasets of multiple performers, performances and rāgas for each performer would be beneficial for enabling a more systematic comparison between performers, performances and rāgas.

5. REFERENCES

Hackney, P. (2003). *Making connections: Total body integration through Bartenieff fundamentals*. Routledge.

Luck, G., & Toiviainen, P. (2008). Exploring Relationships between the Kinematics of a Singer's Body Movement and the Quality of Their Voice. *Journal of interdisciplinary music studies*, 2.

Nymoen, K., Godøy, R. I., Jensenius, A. R., & Torresen, J. (2013). Analyzing correspondence between sound objects and body motion. *ACM Transactions on Applied Perception (TAP), 10(2), 9*.

Clayton M., & Leante L. (2013). *Experience and Meaning in Music Performance* (Martin Clayton, Byron Dueck & Laura Leante (eds.)). Oxford University Press, USA, 188-207.

CREATIVE HARMONISATION OF FOLK MELODIES

Costas Tsougras **Maximos Kaliakatsos-Papakostas** **Emilios Cambouropoulos**
School of Music Studies, Aristotle University of Thessaloniki
tsougras@mus.auth.gr maximoskalpap@gmail.com emilios@mus.auth.gr

1. INTRODUCTION

Since the 19th century, many composers attempted to blend local national musical elements (such as traditional rhythms, modal thematic materials) with aspects of established western musical idioms (such as classical tonality, post-tonal harmony, atonality); this way, novel musical styles were created that have a characteristic local flavor. This paper focuses on issues of creativity involved in the interaction between traditional folk melodies and diverse harmonic idioms. Traditional melodies often embody characteristics outside the 'standard' western major-minor framework, posing a challenge for a composer that wants to reconcile partially incompatible music systems. Can a creative computer system assist such a task? This study employs a system that harmonises folk melodies in diverse harmonic styles and presents some results regarding its usage. This system is a rather rare instance of the application of creative technologies in the domain of traditional music.

2. THE *CHAMELEON* HARMONISER

The CHAMELEON melodic harmonisation assistant has been developed in the context of the COINVENT project framework[1] and is capable of harmonising a given melody in different harmonic styles (Kaliakatsos-Papakostas et al., 2016), and also of blending different harmonic idioms (Kaliakatsos-Papakostas et al., 2017). The proposed melodic harmonisation assistant is adaptive (learns from data), general (can cope with any tonal or non-tonal harmonic idiom) and modular (learns different aspects of harmonic structure such as chord types, chord transitions, cadences and voice-leading). This harmonisation system can be used to generate novel harmonisations for diverse melodies via the exploration of the harmonic possibilities provided by the implied harmonies of input melodies.

Harmonic blending, as performed by CHAMELEON, involves two different processes. The first process is *melody-harmony* blending whereby a melody originating from a given musical idiom (with certain implied harmonic qualities) is harmonised based on a harmonic space (chord types, chord transitions, cadences, basic voice-leading) derived via machine learning from a different harmonic idiom. The second, is *harmony-harmony* blending whereby the harmonic space that is used to harmonise a given melody is, itself, the product of blending between two different harmonic idioms. In this paper we use both processes, whereby an annotated melody is presented to the system to be harmonised in different single or blended harmonic styles (from medieval to contemporary harmonic styles). Several examples of creative harmonisations of different traditional melodies (Scottish, Greek, Russian, etc) have been produced (see Kaliakatsos-Papakostas et al., 2016 & 2017).

[1] http://coinvent.uni-osnabrueck.de/

3. CREATIVITY & ACTIVE EVALUATION

A number of *passive* listening empirical studies have been conducted aiming to evaluate the creativity of CHAMELEON (Zacharakis et al., 2018). In the context of an *active* evaluation of the system, a compositional 'assignment' was given to seven composers or composition students in Thessaloniki (2 graduates of the School of Music Studies of the Aristotle University of Thessaloniki and 5 students enrolled in Costas Tsougras's "Stylistic Composition" course during the 2016 spring semester). The composers were provided with three different Greek folk melodies to choose from: "Είχα μιαν αγάπη" (*Eicha mian agapē*, "I had a love", in D Aeolian mode), "Απόψε τα μεσάνυχτα" (*Apopse ta mesanychta*, "Tonight at midnight", in A Aeolian mode) and "Μωρή κοντούλα λεμονιά" (*Mōrē kontoula lemonia*, "Oh short lemon tree", in D minor pentatonic mode). For each melody 40 different harmonisations were provided, created by CHAMELEON from 16 diverse harmonic idioms and their blendings ('BC' - Bach chorales, 'BTL' - Beatles, 'JA' - Jazz, 'CN' - Constantinidis, 'HM' - Hindemith, 'KP' - Kostka Payne, 'PL' - Palestrina Stabat Mater, 'FB' - Fauxboudon, 'BN' - 'Bossa nova', 'CAN' - Chords added notes, 'GG' - Grieg, 'MDC' - modal chorales, 'PAC' - Parallel chromatic, 'POC' - polychords, 'WT' Whitacre major or minor, 'WT' - whole tone). The participants were asked to select the melody of their preference and the harmonisation/s that they considered more interesting/inspiring and compose a miniature for piano employing any compositional elaboration/variation techniques they deemed appropriate. The aim of the experiment was the creative use of the produced harmonisations as a structural harmonic framework for the building of rich musical textures and original compositional thought.

An excerpt from one of the seven short compositions, Lazaros Tsavdaridis's "Mōrē kontoula lemonia" is presented, in order to demonstrate the process. The song's harmonizations were produced from the annotated score of Fig. 1 (the annotation defines the mode, the phrasing and the harmonic rhythm). The composer chose a single CHAMELEON harmonisation, a blend between Jazz minor and Parallel Chromatic Harmony (see Fig. 2) and produced two piano variations, one simple and one more complex. In the first variation, presented in Fig. 3, he develops pianistic textures and voice-leading from the proposed harmony, choosing to deviate from it when necessary (e.g. in bars 15-16, to achieve a better cadence).

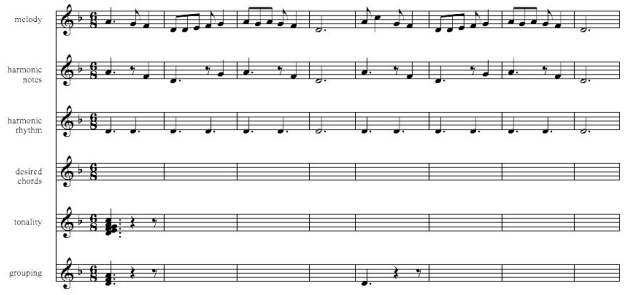


Figure 1. Annotated first phrase of "Mōrē kontoula lemonia".

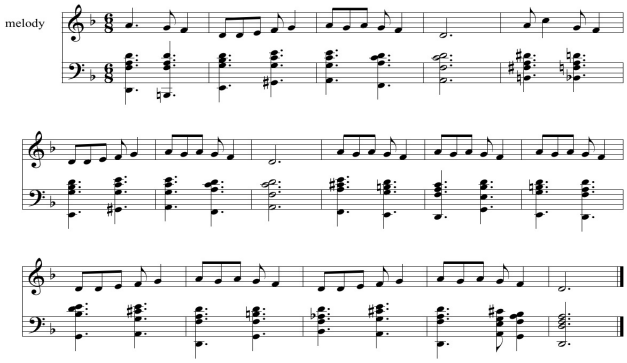


Figure 2. Harmonisation by CHAMELEON produced from the blending of the idioms 'JA minor' and 'PAC'.

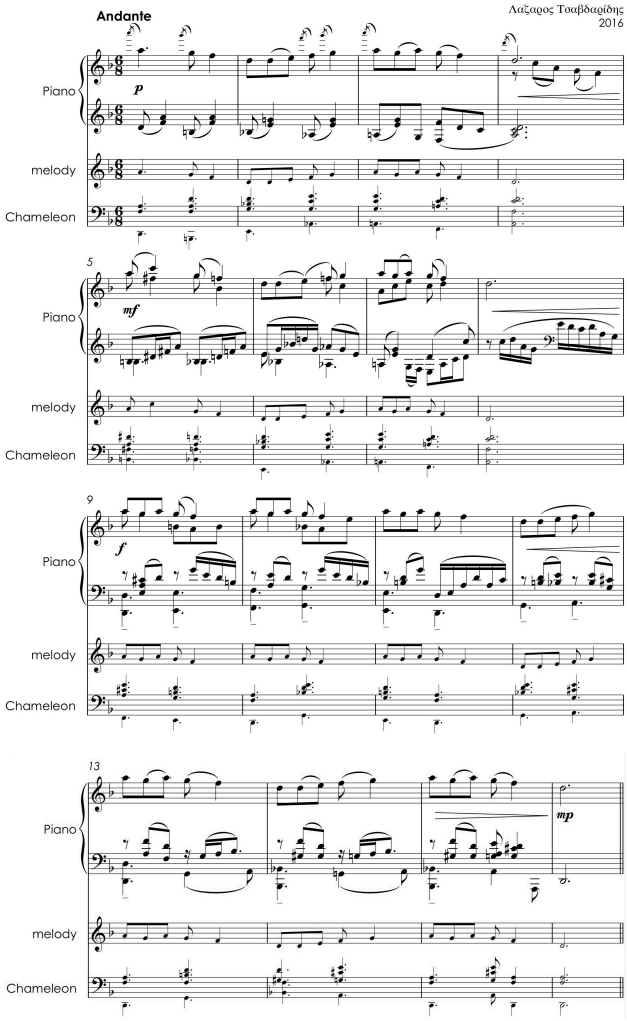


Figure 3. Excerpt (1st variation) of composed piano miniature.

The seven composers created very diverse piano miniatures by applying different elaboration techniques on the selected harmonic backgrounds, by creating various types of textures and layers and by developing original forms that optimally accommodated their material (see and listen to all the pieces at CHAMELEON's website[1]). The seven miniature pieces were performed on 19 October 2016 by pianist Fani Karagianni during a live concert at the Museum of Contemporary Art in Thessaloniki.

Most composers reported that they found the whole project very stimulating and that they considered some of the harmonisations particularly inspiring; some stated that they would have never come up with one or more of the harmonisations they used. Overall, there was a positive response regarding at least some of the creative products of this system. More extended creative evaluation studies are expected to be conducted in future research.

So, CHAMELEON aids composers at one of the most important stages of composition, the choice of pitch material and the creation of the harmonic (structural) background, by providing an easily-controlled computational environment which very rapidly produces a multitude of diverse creative options, thus by broadening their choices at the minimum of time and effort. Of course, what defines the quality and originality of the musical result is ultimately each composer's personal touch and inventiveness, as the creative kick provided by CHAMELEON is only the first stage of evolution of the pieces' character, form and design, parameters between which the seven compositions differentiated substantially.

4. CONCLUSIONS

This study proposes a way to use creatively artefacts of music heritage, namely, folk melodies. Creative systems such as CHAMELEON may enhance users' appreciation for and engagement with traditional music, enabling them not only to access such music in digital repositories but also to re-use it creatively in novel compositions.

5. REFERENCES

Kaliakatsos-Papakostas M., Makris D., Tsougras C., Cambouropoulos E. (2016). Learning and creating novel harmonies in diverse musical idioms: An adaptive modular melodic harmonisation system. *Journal of Creative Music Systems,* 1(1).

Kaliakatsos-Papakostas, M., Queiroz, M., Tsougras, C., & Cambouropoulos, E. (2017). Conceptual blending of harmonic spaces for creating melodic harmonisation. *Journal of New Music Research,* 46(5), 305-328.

Zacharakis, A., Kaliakatsos-Papakostas, M., Tsougras, C., & Cambouropoulos, E. (2018). Musical Blending and Creativity: An Empirical Evaluation of the CHAMELEON Melodic Harmonisation Assistant. *Musicae Scientiae,* 22(1), 119-144.

[1] http://ccm.web.auth.gr/creativeusecomposers.html

Extended Abstract: Visualising Melodic Similarities in Folk Music

Chris Walshaw
Computing & Information Systems
University of Greenwich, London, UK
c.walshaw@gre.ac.uk

1. INTRODUCTION

The aim of this extended abstract is to outline a technique for visually exploring melodic relationships within folk tune collections. It stems from related work known as TuneGraph, [1], which allows users of abcnotation.com to explore melodic similarity. TuneGraph uses a similarity measure to derive a proximity graph representing similarities within the abc notation corpus. From this a local graph is extracted for each vertex, aimed at indicating close variants of the underlying tune. Finally an interactive user interface displays each local graph on that tune's webpage, allowing the user to explore melodic similarities.

As it stands TuneGraph only gives a localised view of the melodic relationships: this paper aims to look at exploring those relationships at a global (corpus-based) level.

2. METHODOLOGY

The essential idea is that, given a collection of tunes and a melodic similarity measure which can measure pairwise similarity between tunes (e.g. [2]), it is possible to construct a complete proximity graph of the corpus. Here the melodic similarity measure used is multilevel recursive sub-sequence alignment discussed in detail in [1] with some additional enhancements, tested in [3], also applied. However, the ideas are generic.

In the proximity graph each vertex represents a tune and edge weights represent similarities between tunes: the greater the similarity the larger the edge weight. If a similarity threshold, T, is applied so that an edge is only included if the two tunes it connects are sufficiently similar (if they match across at least some proportion T of their length, [3]) then a sparse proximity graph can be induced (the higher the threshold, the more sparse the graph). 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.

Following [1], the value for T used here is 1/6, a good compromise between restricting the number of edges (in order to make the visualisation tractable) but including enough to make the graph sufficiently rich. Further testing with other of values including 1/4 and 1/8 will be shown.

However, most reasonable values of the threshold typically generate a corpus graph with several *disconnected components* (subgraphs that are not connected by any edges) and often many *isolated vertices* (vertices with no incident edges – i.e. tunes that are not sufficiently similar to any other tune in the corpus to generate an edge).

This presents a problem for the investigation discussed here. An option is simply to visualise the largest connected component: however, this may only represent a small portion of the dataset. Accordingly, a straightforward scheme has been devised for connecting up the graph with a minimal number of zero weighted edges to help with the layout.

The setting of the edge weight to zero is important for the visualisation: since edge weights influence vertex placement, a zero weight edge will have minimal impact on the graph layout but the edge will mean that the two insufficiently similar vertices that it connects are positioned as close together as possible.

Once the corpus graphs are constructed they can be visualised using MultiLevel Force-Directed Placement (MLFDP) algorithms, e.g. [4], a standard technique for visualising large unstructured graphs.

3. RESULTS AND DISCUSSSION

3.1 Annotated datasets

The initial investigation explores two small datasets known to contain many related tunes and which have been annotated manually to indicate similar melodies, specifically those belonging to the same tune family.

The first of these datasets is the Annotated Corpus of the Meertens Tune Collection, version 2.0.1, [5]. This contains 360 Dutch folk melodies, each identified by experts as belonging to one of 26 tune families.

The second dataset contains 368 English morris dance tunes taken from the Morris Ring website[1]. Since morris music has several (approx. 35) traditions, each typically associated with a village, there are many tunes found in more than one tradition, but each tradition typically has a different variant of the tune. This dataset therefore contains 368 tunes which have been manually identified as belonging one of 113 tune families.

Because the datasets are annotated, the tune families induce a partition on the graph and different families can be visualised with different colours. Fig. 1 shows the results of the corpus graphs, visualised using MLFDP and overlaid with the partition induced by the tune families (zero-weight edges used to connect the graph are not displayed).

As can be seen, for both datasets the graph construction and visualisation is very complementary to the attribution of tunes to tune families: most edges go between tunes that are in the same family and even where they are not connected, most isolated vertices and small components (e.g. pairs) are close to other tunes in the same family.

Together these suggest that this kind visualisation can help to disambiguate tune families.

[1] https://themorrisring.org/music

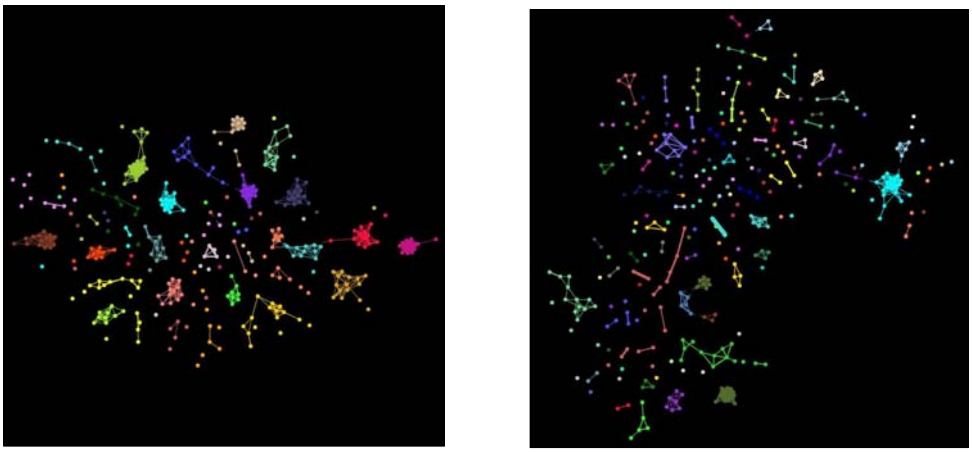


Figure 1. Visualisations of the Meertens (left) and Morris Ring (right) corpus graphs.

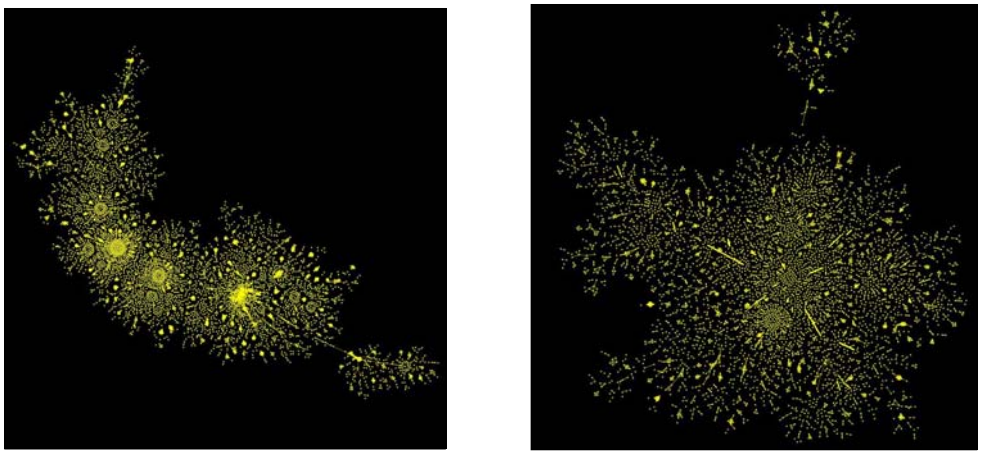


Figure 2. Visualisations of the Village Music Project (left) and TheSession (right) corpus graphs.

3.2 Collection-based datasets

The second set of results come from two tune collections which are not annotated (so not partition data is available). These are the Village Music Project[1] with around 5,600 English tunes transcribed from 18th & 19th century manuscripts and a subset of 5,000 of the ~30,000 tunes found at The Session[2], a community site which hosts a large collection of Irish traditional music.

Fig. 2 shows visualisations of the resulting corpus graphs. The structures are similar to their smaller counterparts in Fig. 1 although with many more disconnected vertices (these are much larger and much more disparate datasets). Nonetheless, tightly bound clusters of similar tunes are clearly visible, together with other structures, such as super-connectors (isolated vertices surrounded by sunflower like structures of other isolated vertices) and weak linkage (long, lightly-weighted edges which indicate loose connections between different subgraphs).

Although it is not easy to draw any conclusions from these two visualisations, it is encouraging to imagine an interactive exploration tool with user-driven zoom features and including score rendering and MIDI playing facilities.

4. REFERENCES

[1] C. Walshaw, "Constructing Proximity Graphs To Explore Similarities in Large-Scale Melodic Datasets," in *6th Intl Workshop on Folk Music Analysis*, 2016.

[2] B. Janssen, P. van Kranenburg, and A. Volk, "Finding occurrences of melodic segments in folk songs employing symbolic similarity measures," *J. New Music Res.*, p. (to appear), 2017.

[3] C. Walshaw, "Tune Classification using Multilevel Recursive Local Alignment Algorithms," in *7th Intl Workshop on Folk Music Analysis*, 2017.

[4] C. Walshaw, "A multilevel algorithm for force-directed graph-drawing," *J. Graph Algorithms Appl.*, vol. 7, no. 3, 2003.

[5] P. van Kranenburg, B. Janssen, and A. Volk, "The Meertens Tune Collections : The Annotated Corpus (MTC-ANN) Ver. 1.1 and 2.0.1," 2016.

[1] http://www.village-music-project.org.uk/

[2] https://thesession.org/

FEATURE ANALYSIS OF REPEATED PATTERNS IN DUTCH FOLK SONGS USING PRINCIPAL COMPONENT ANALYSIS

Iris Yuping Ren[1], Hendrik Vincent Koops[1], Dimitrios Bountouridis[1], Anja Volk[1], Wouter Swierstra[1], and Remco C. Veltkamp[1]

{*y.ren, h.v.koops, d.bountouridis, a.volk, w.s.swierstra, r.c.veltkamp*}*@uu.nl*
[1]Utrecht University

1. INTRODUCTION

Oral transmission plays a significant role in folk music. Through this often imperfect communication process, certain parts of melodies remain stable, variations are created, repeated patterns emerge [1]. Formulated in ethnomusicological studies, the concept of tune family describes the structures in this stream of transformations: folk songs that are supposed to have a common ancestor in the process of oral transmission are grouped into a tune family. Local structures within the melodies, namely characteristic motifs, or prominent, nonliterally repeated patterns, are detected to be useful in determining music similarity and classifying tune families [2]. Subsequently, in an annotation study on the influence of different musical dimensions on human similarity judgements of melodies belonging to the same tune family, repeated patterns between melodies turned out to play the most important role for similarity among all considered musical dimensions [3]. Therefore, algorithms which can extract these repeated patterns automatically would be useful for tune family classification.

Different pattern discovery methods have been introduced, such as sequence-based approaches [4, 5, 6, 7], geometric approaches [8]. Unfortunately, patterns extracted by state-of-the-art algorithms are not yet capable of replacing human annotations when we attempt to apply the patterns to classification and discovery tasks [9, 10].

This paper uses Principal Component Analysis (PCA) to better understand characteristics of musical patterns and to further use this information for designing and evaluating future pattern discovery algorithms. We show what features can summerise the data variance in musical patterns and propose using feature selection and extraction methods to improve pattern discovery algorithms.

There exists research that uses patterns in analysing tune families, modelling similarity, improving compression and retrieval tasks [10]. In this setting, it is common to either take the features of the whole song or the raw data of pitch and duration pairs of the patterns. We do not know of existing studies that focus on investigating the features only within patterns in music.

2. DATA AND SETUP

Dataset The corpus data we use is the Dutch folk song dataset MTC-ANN [11]. Three experts have annotated the prominent patterns in each song which could best position the song into one of the 26 tune families. The dataset consists of 360 Dutch folk songs with 1657 annotated patterns.

Feature calculation We calculate features from the patterns by using a common feature extraction tool: the jSymbolic2 toolbox in the jMIR toolset [12]. jSymbolic2 takes MIDI files as input and computes 155 musically meaningful features in six categories: texture, rhythm, dynamics, pitch, melody and chords.

Feature selection We perform a feature selection step and retain 64 features as follows: (1) Eliminating the features which are constant across all patterns; (2) Eliminating the features which are irrelevant to the music content of time and pitch, such as the dynamics features and artefacts introduced by MIDI conversion.

PCA After feature selection, we further combine and transform features to make new combined features, which is known as the feature extraction step. PCA is a well-known feature extraction and dimension reduction method. PCA gives new combinations of features which form orthogonal principal components. The principal components are in the same directions as the directions of the largest variances of the dataset. By examining the resulting principal components, we gain insights as to which features are of more significance in explaining the spread of the data points. PCA has been employed and shown to be effective in many MIR tasks [13]. We take a similar approach in the PCA analysis as [13] in which the author investigated audio features in popular music.

3. RESULTS

In Table 1, we report the prominent features and the weights in the first three PCA components. We make the following observations: (1) The most significant feature of the first component is *the number of strong rhythmic pulses*. Since rhythmic pulses are derived from beat histogram, it shows the importance of metric structures in the patterns. [1] More specifically, although there are both pitch and rhythmic features in the first principal component, we have three rhythmic features and two pitch features. In the second component, although there are more pitch features, the repeated notes feature is relevant both to pitch and duration. In the third component, we only have rhythmic features.

Furthermore, to give a fuller picture than the first five features in each component, we calculate the total weight sums of rhythmic and pitch related features. In the first component, the pitch related features have a total weight sum of 48.89% and the rhythmic features have a total weight

[1] For the details of other features, please refer to [12]. Given that we have 40 pitch related features, 20 rhythmic features and 4 features related to both pitch and duration, it is non-trivial that we have rhythmic features top-ranked in the first three principal components.

PC (Percentage of variance explained)	Features	Weight (Percentage)
PC1 (22.51)	Number of Strong Rhythmic Pulses	5.18
	Pitch Variety	5.15
	Number of Relatively Strong Rhythmic Pulses	5.07
	Number of Common Pitches	5.07
	Number of Moderate Rhythmic Pulses	5.07
	Other Features	74.46
PC2 (12.42)	Repeated Notes	8.24
	Relative Prevalence of Top Pitches	8.06
	Relative Prevalence of Top Pitch Classes	7.58
	Prevalence of Most Common Pitch	6.32
	Prevalence of Most Common Pitch Class	5.98
	Other Features	63.82
PC3 (8)	Combined Strength of Two Strongest Rhythmic Pulses	10.58
	Polyrhythms	9.98
	Rhythmic Variability	9.27
	Strongest Rhythmic Pulse	7.26
	Strength of Strongest Rhythmic Pulse	7.14
	Other Features	55.77

Table 1: The first three principal components of PCA and the weights of features. We omit the rest of $64 - 3 = 61$ components since they do not contribute significantly ($< 7.5\%$) to the variance and, for visualisation purposes, it is common practice that only the first three dimensions of PCA are considered.

sum of 46.38%. In the second component, pitch and rhythmic features have 64.45% and 27.89% weight sums respectively. The weight sums are 25.2% and 68.0% for the third component. In summary, looking at the first three dimensions of PCA, we see a balanced contribution from both the pitch and rhythmic features.

4. CONCLUSION AND DISCUSSIONS

Using PCA, we show the prominent features of MTC-ANN patterns. The pitch related and rhythmic features contribute together to the first PCA component; the second and third component is consist mainly of pitch related features and rhythmic features respectively. Despite the fact that we have less rhythmic features computed using the jSymbolic2 toolbox, the rhythmic features do not contribute less in the first three principal components. One might argue it is obvious that both pitch and rhythmic features are important, but it is remarkable that the two together contribute to each of the first few PCA dimensions.

The prominent features also give hints on potential improvements to current existing pattern discovery algorithms. Although metric structures have been considered in musical pattern research [14, 15, 16], many pattern discovery algorithms do not explicitly consider metric structures imposed by musical punctuations such as bar lines and measures. According to what PCA shows, in designing and evaluating pattern discovery algorithms, we should take metric structures into consideration as well as the repetitions and pitch related features in the patterns.

This investigation is a starting point for future work on using extracted pattern features for pattern classification and discovery. More concretely, we can further use the features to cluster and classify the patterns into tune families; using other metadata in the annotations, we can also correlate the features to the descriptions of annotators and motif classes; the features after PCA transformation can be used to explore, evaluate and compare algorithmically extracted patterns with human annotations.

5. REFERENCES

[1] Berit Janssen. *Retained or Lost in Transmission?* PhD thesis, University of Amsterdam, 2018.

[2] James R. Cowdery. A fresh look at the concept of tune family. *Ethnomusicology*, 28(3):495–504, 1984.

[3] Anja Volk and Peter Van Kranenburg. Melodic similarity among folk songs: An annotation study on similarity-based categorization in music. *MusicæScientiæ*, 16(3):317–339, 2012.

[4] Olivier Lartillot. Multi-dimensional motivic pattern extraction founded on adaptive redundancy filtering. *Journal of New Music Research*, 34(4):375–393, 2005.

[5] Darrell Conklin. Discovery of distinctive patterns in music. *Intelligent Data Analysis*, 14(5):547–554, 2010.

[6] Iris Yuping Ren. Closed patterns in folk music and other genres. *Proceedings of the 6th International Workshop on Folk Music Analysis*, pages 56–58, 2016.

[7] Matevž Pesek, Aleš Leonardis, and Matija Marolt. Symchman unsupervised approach for pattern discovery in symbolic music with a compositional hierarchical model. *Applied Sciences*, 7(11):1135, 2017.

[8] David Meredith, Kjell Lemström, and Geraint A. Wiggins. Algorithms for discovering repeated patterns in multidimensional representations of polyphonic music. *Journal of New Music Research*, 31(4):321–345, 2002.

[9] Iris Yuping Ren, Hendrik Vincent Koops, Anja Volk, and Wouter Swierstra. In search of the consensus among musical pattern discovery algorithms. *Proceedings of the International Society for Music Information Retrieval*, pages 671–680, 2017.

[10] Peter Boot, Anja Volk, and W. Bas de Haas. Evaluating the role of repeated patterns in folk song classification and compression. *Journal of New Music Research*, 45(3):223–238, 2016.

[11] Peter van Kranenburg, Berit Janssen, and Anja Volk. The Meertens Tune Collections: The Annotated Corpus (MTC-ANN) versions 1.1 and 2.0.1. *Meertens Online Reports*, 2016(1), 2016.

[12] Cory McKay. *Automatic Music Classification with jMIR*. PhD thesis, McGill University, 2010.

[13] Jan Van Balen. *Audio description and corpus analysis of popular music*. PhD thesis, University Utrecht, 2016.

[14] Peter Van Kranenburg and Darrell Conklin. A pattern mining approach to study a collection of dutch folk-songs. *Proceedings of the 6th International Workshop on Folk Music Analysis*, pages 71–73, 2016.

[15] Darrell Conklin and Christina Anagnostopoulou. Representation and discovery of multiple viewpoint patterns. In *Proceedings of the 26th International Computer Music Conference*, pages 1–7. Citeseer, 2001.

[16] Darrell Conklin and Mathieu Bergeron. Feature set patterns in music. *Computer Music Journal*, 32(1):60–70, 2008.