Proceedings of the 9th International Workshop on Folk Music Analysis (FMA2019)

Editors: Islah Ali-MacLachlan and Jason Hockman
Chair: Islah Ali-MacLachlan

2nd-4th July 2019
Birmingham City University
Birmingham, UK

Contents

Programme Committee

Chairs

- Jason Hockman (Birmingham City University, UK)
- Islah Ali-MacLachlan (Birmingham City University, UK)

Members

- Pierre Beauguitte (DIT, Dublin, Ireland)
- Emmanouil Benetos (Queen Mary University, London, UK)
- Baris Bozkurt (Universitat Pompeu Fabra, Spain)
- Emilios Cambouropoulos (Aristotle University of Thessaloniki, Greece)
- David Carroll (DIT, Dublin, Ireland)
- Matthew Cheshire (Birmingham City University, UK)
- Darrell Conklin (University of the Basque Country UPV/EHU, Donostia - San Sebastián, Spain)
- Jake Drysdale (Birmingham City University, UK)
- Carl Southall (Birmingham City University, UK)
- Maciek Tomczak (Birmingham City University, UK)
- Costas Tsougras (Aristotle University of Thessaloniki, Greece)
- Peter Van Kranenburg (Meertens Institute, Amsterdam, The Netherlands)
- Chris Walshaw (Old Royal Naval College, London, UK)

MULTI-MEDIA RECORDINGS OF TRADITIONAL GEORGIAN VOCAL MUSIC FOR COMPUTATIONAL ANALYSIS

Frank Scherbaum
University of Potsdam
frank.scherbaum@uni-potsdam.de

Nana Mzhavanadze
University of Potsdam
nana.mzhavanadze@uni-potsdam.de

Sebastian Rosenzweig
University of Erlangen
sebastian.rosenzweig@audiolabs-erlangen.de

Meinard Müller
University of Erlangen
meinard.mueller@audiolabs-erlangen.de

ABSTRACT

Traditional multipart-singing is an essential component of the national identity of Georgia. It has been an active field of ethnomusicological research since more than 100 years, with a whole series of thematically very diverse research questions. Here, we report on the generation of a new research corpus of traditional Georgian vocal music collected during a three-month field expedition in 2016. It employs new and partially unconventional field recording techniques and is intended particularly for the application of modern computational analysis methods. To circumvent the source separation problem for multiple singing voices in field expeditions, we used larynx microphones to record skin vibrations close to the larynx (additionally to using conventional audio and video equipment). The resulting multi-media recordings comprise audio and video material for more than two hundred performances, including more than fourteen hundred audio tracks based on different types of microphones (headset, larynx, ambient, directional), video tracks, as well as written documents of interviews with the performers. We demonstrate that the systematic use of larynx microphones, which to our knowledge has never been used before on a larger scale in ethno-musicological field expeditions, opens up new avenues for subsequent computational analysis regarding a multitude of aspects including pitch tracking, harmonic and melodic analysis, as well as for documentation and archiving purposes.

1. INTRODUCTION

There have been many efforts in the past to document and record the rich musical heritage of traditional Georgian polyphonic music, pioneered by phonograph recordings more than a century ago. However, many of them have been lost over the years and the available historic audio recordings are often of insufficient quality for the application of modern, quantitative analysis techniques. A notable exception, regarding the applicability of computerized analysis tools, is the collection of historical recordings by master chanter Artem Erkomaishvili. This dataset was processed and enriched by features informatically extracted from the recordings, see Müller et al. (2017) (accompanied by a website [1]).

In 2015, a pilot project was carried out to test the usefulness of body vibration recordings for ethnomusicological purposes, see Scherbaum (2016). Based on this experience, we started during the summer of 2016 to collect a new set of field recordings of traditional Georgian vocal music. The regional focus of our field work was on Upper Svaneti, which is one of the rare regions in the crossroads of Europe and Asia where very old (presumably pre-Christian) traditions are still cultivated as part of daily life. Svan songs as parts of these rituals occupy a special place within the Georgian music and are still maintained in a comparatively original form due to the remote geographic location. The style of Svan multi-part singing has been described in different terms as chordal unit polyphony Aslanishvili (2010), or drone dissonant polyphony Jordania (2010), and the judgments regarding the importance of the (moving) drone and/or of the role of dissonances differ between authors, e. g. between Dirr (1914) and Jordania (2006). Consensus, however, exists on the hypothesis that Svan music represents the oldest still living form of Georgian vocal polyphony.

There have been considerable efforts in the past to record traditional Svan music, first with phonographs, later with tape recorders. Already more than 100 years ago, Dirr (1914) discussed the musical characteristics of phonograph recordings from Svaneti (North-Western Georgia), which had been collected by Paliashvili (1909). Unfortunately, most recordings of Svan songs from the early days of the last century have not survived the ravages of time. The few audio files obtained are mostly in very poor quality. On the other hand, the Tbilisi Conservatory has also made recordings since the 1950s. These recordings, however, were lost during construction work in the 1990s. It should also be noted that during the 1980s, a set of recordings with the Mestia Regional Folk Song and Dance Ensemble Riho were made in the voice recording studio Melodia. These recordings, however, were only released very recently, see Khardziani (2017). A small number of more recent audio recordings were made by ethnomusicologists Malkhaz Erkvanidze and Keti Matiashvili in 2004 (recordings of approximately 25 songs in Lower Svaneti), and between 2007 and 2010 by the State Center of Folklore under the supervision of ethnomusicologist Nato Zumbadze, both in Lower and Upper Svaneti, partly with several microphones and in a mobile recording studio. In addition, within the crowd-funded Svan Recording Project performed by American singer Carl Linnich in 2010 with members of the Riho Ensemble in Lengeri, 32 songs were supposedly recorded. However, it is not known to us if this project was completed. In conclusion, before our own field expedition described in this paper, the publicly available audio material from traditional Svan songs known to us was very limited in number and quality.

[1] https://www.audiolabs-erlangen.de/resources/MIR/2017-GeorgianMusic-Erkomaishvili

2. NEW RECORDINGS OF SVAN MUSIC

During the summer of 2016, Frank Scherbaum and Nana Mzhavanadze performed a field expedition to record Georgian vocal music in Svaneti and other regions in Georgia. The recordings cover a wide range of examples of traditional Georgian singing, praying, and rare examples of funeral lament (roughly 120 pieces in total). The technical quality of the recordings is good to excellent. All the recordings were done as multi-media recordings in which a high resolution (4K) video stream is combined with a stream of 3-channel headset microphone recordings (one for each voice group), a stream of 3-channel larynx microphone recordings (one for each voice group as well), and a conventional stereo recording. The systematic use of larynx microphones, which to our knowledge has never been done before in ethnomusicological field expeditions, was motivated by the results of Scherbaum et al. (2015). Larynx microphone recordings allow the undistorted documentation of the contribution of each singer while all of them are singing together in their natural context. Secondly, larynx microphone recordings contain essential information of a singer's voice regarding pitch, intonation, timbre and voice intensity, which allows for using computer-based ways to document and analyze oral-tradition vocal music in new ways, e. g to perform computerized pitch analysis to document the pitch tracks (including the microtonal structure), to study the pitch inventory and scales used and the interaction between singers, see Scherbaum (2016). Each recording session was accompanied by extensive interviews of the performers conducted by Georgian ethnomusicologist Nana Mzhavanadze as described in Scherbaum & Mzhavanadze (2017).

All the initiatives to systematically record traditional Svan music before our field expedition have in common that the recordings were purely acoustic. Even with recordings using separate microphones (as in some of the more recent projects), the separability of the individual voices is problematic. In the context of our own work on the generation and propagation of body vibrations during singing described by Scherbaum et al. (2015), we have tested the acoustic separability of individual voices with directional microphones under studio conditions and found that this becomes problematic even under idealized conditions when singers sing with differing intensity (which they definitely do in Svaneti). In conclusion, the acquired research corpus seems optimally suited to address a large number of diverse research questions.

2.1 Recording Locations

The field recordings were done in 25 recording sessions spread over the summer months (July – September) of 2016. The recording locations are shown in Fig. 1.

Since our emphasis was on (Upper) Svaneti, the vast majority of the recording sessions involved Svan singers or people performing Svan prayers, either in Svaneti or in Svan settlements outside Svaneti (Didgori, Tsalka, and Udabno). In Sessions 3 and 4 we recorded two groups of Gurian singers (Shalva Chemo and Amaghleba in Ozurgeti and Bukitsikhe, respectively), while in Session 22 in Tbilisi we used the opportunity to record singers from a women's group of ethnomusicologists (Mzetamze) to perform songs from various regions. In addition, we recorded singers in the villages of Glola and Ghebi in the upper part of the Rioni river valley (which now belongs to Racha). In former days, this region used to be part of Svaneti as well.

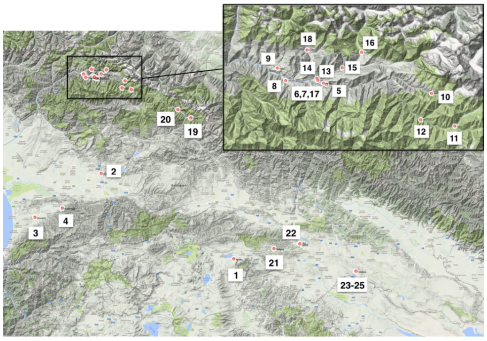


Figure 1: Location of the recording sessions.

2.2 Recording Equipment

Our standard recording setup consisted of three DPA 4066-F headset microphones and three (modified) Albrecht AE 38 S2 larynx microphones (one set of headset/larynx for each singer), which were recorded using an eight-track Zoom F8 field recorder. In addition, a stereo recorder (Olympus LS5) was used to record each group of singers from an approximately 2-m distance. Each session was documented by video in 4 K resolution using a Sony AX 100 video camera. The corresponding audio signal was either recorded by the internal stereo microphone (in cases of small rooms) or with a Sony XLR-K2M directional microphone (in cases where the camera was placed at larger distances from the singers). Finally, still photos and occasionally short videos were taken with a Sony HX90 camera and an iPhone 6. A Zoom Q4 video camera was occasionally used for interviews.

2.3 Pre-Processing

For each performance, audio, larynx microphone, and video tracks of similar length were manually cut and saved to disc. Subsequently, all tracks were aligned in time using the Plural Eyes software (Red Giant Inc.) and underwent a first visual quality control. In total, the collection contains 1444 files (tracks) of different media types (video, audio, and larynx microphone recordings) belonging to 216 different performances. Among them are 37 performances of prayers and 11 performances of funeral songs (Zari). The rest is referred to as songs (in a very general sense). Some of the songs were recorded several times by different ensembles, e. g. Kriste Aghsdga and Jgragish, each five times, Vitsbil-Matsbil, Tsmindao Ghmerto (the funeral version) and Dale Kojas, each four times. Overall, the dis-

tribution of song types of recorded songs is quite diverse. Hymns and ballads alone already make up one half of the song inventory. One quarter consists of dance songs, table songs, and mourning songs. Eleven recordings of funeral songs (Zari) were made in different contexts. Four of them were recorded during actual funerals (Session numbers 12 and 13 in Kala and Latali, respectively), while the rest were performed during conventional recording sessions.

3. REUSABILITY OF THE CORPUS

In the context of creating a research corpus for computational analysis, Serra (2014) discusses five criteria (purpose, coverage, completeness, quality, and reusability) relevant to its quality. The first four of them were already discussed in the previous chapter. Regarding the fifth criterion (reusability), we are following two strategies.

First, a curated version of the pre-processed tracks, together with the original audio and video material as well as descriptive material related to the individual recording sessions is permanently stored within the long-term archive of regional scientific research data (LaZAR), hosted at the University of Jena/Germany, see Scherbaum et al. (2018). It is accessible for research and other non-commercial purposes through a searchable web-interface [2] . The main purpose of this archive is the long-term preservation of the collected material.

Second, due to the systematic use of larynx microphones, the collected material allows for new ways to employ computational methods from audio signal processing and music information retrieval (MIR). It forms an important basis of the research project "Computational Analysis of Traditional Georgian Vocal Music" (funded by the German Research Foundation for the period 2018-2021), hereafter referred to as GVM project. In order to facilitate the access to the collected material within the framework of this project, e. g. to visually study a particular recording session or to perform conventional analysis such as transcription of the lyrics, we have developed a web-based interface with search, navigation, visualization, and playback functionalities. Based on the trackswitch.js architecture (see Werner et al. (2017)), this interface allows a user to play back all the media channels (audio, larynx microphone recordings, video) and seamlessly switch between different audio tracks. A preliminary version was demonstrated by Scherbaum et al. (2018) and can be accessed at the accompanying website [3] , see also Fig. 2.

4. ANALYSIS EXAMPLES

In the framework of the GVM project, we aim to improve the understanding of traditional Georgian vocal music by using computational tools. In the following, we will give a few examples how the collected multi-media recordings can be used for this purpose. The first aspect which we will discuss is how the larynx microphone tracks of a performance can be used to determine the "tuning system" employed by the singers. For traditional Georgian vocal music, the topic of "tuning system(s)" or –in more general terms –"tonal organization", has been a matter of intense debate for a long time (cf. Tsereteli & Veshapidze (2014) or Erkvanidze (2016)). The second aspect which we will address is how a combination of audio tracks and larynx microphone tracks can be used to explore new ways to digitally represent polyphonic non-Western orally transmitted music. As musical example we have chosen the Gurian song "Chven Mshvidoba" recorded with the group "Shalva Chemo" in Ozurgeti. Because of its musical complexity, this song is one of the more challenging examples to analyze.

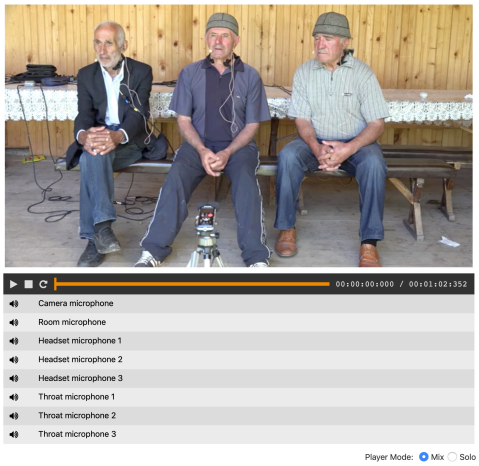


Figure 2: Web-repository interface. Shown is a screendump of the multi-track player.

4.1 Tonal Organization

As was demonstrated by Scherbaum et al. (2015), there is essentially no cross-talk between larynx microphone recordings for different singers. Therefore, the F0-trajectories for the individual voices can be obtained by using monophonic pitch trackers such as the pYIN algorithm (see Mauch & Dixon (2014)). Such a trajectory is shown in Fig. 3 for the top voice of the song "Chven Mshvidoba".

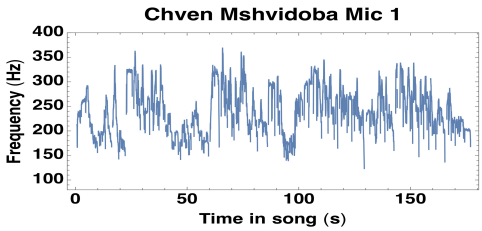


Figure 3: F0-trajectory of the top voice of the song "Chven Mshvidoba".

[2] https://lazardb.gbv.de/search

[3] https://www.audiolabs-erlangen.de/resources/MIR/2018-ISMIR-LBD-ThroatMics

For the analysis of the set of pitches used, we are only interested in stable segments of the F0-trajectories (where one can perceive a stable pitch) and not in utterances corresponding to very short transient signals. A computationally very efficient way to remove them is by morphological filtering as described by Vavra et al. (2004). In practice, the process consists of calculating the dilated and eroded F0-trajectory and considering only those time windows where the difference between the two stays below a chosen threshold (black solid lines in Fig. 4).

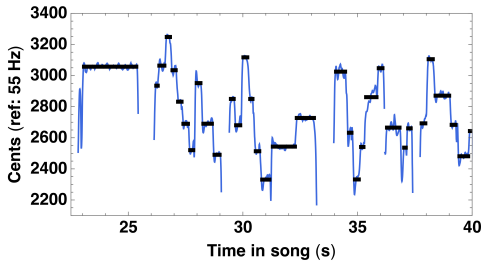
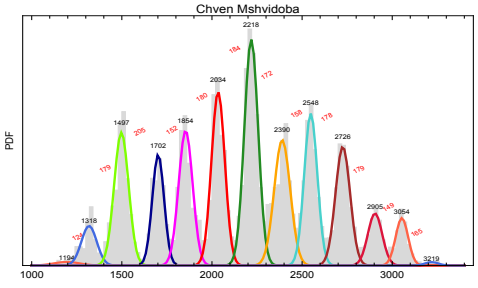


Figure 4: Determination of stable pitch elements (black solid lines) from F0-trajectories by morphological filtering.

Fitting a Gaussian mixture distribution to the set of F0 samples coming from the stable pitch segments from all three voices results in the probability density function (PDF) shown in the top panel of Fig. 5.

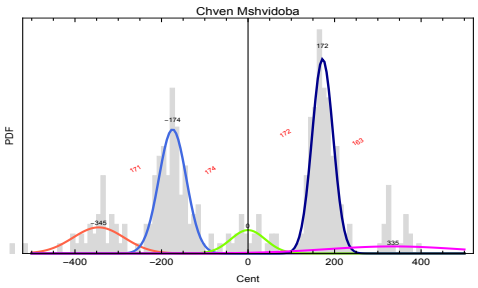


	μ	σ
0.0050502	1194.32	72.2934
0.0330392	1318.24	42.5507
0.10476	1497.46	40.1996
0.0709743	1701.58	32.6678
0.102658	1853.95	38.815
0.126277	2034.26	37.0818
0.169965	2217.89	38.3523
0.109547	2389.71	44.5427
0.107995	2548.13	36.2487
0.0946024	2725.8	40.5899
0.0405386	2904.53	39.6784
0.0320591	3053.52	34.474
0.0025343	3218.55	37.6776

Figure 5: Pitch distribution of the song "Chven Mshvidoba". The unit of pitch is in cents, relative to 55 Hz.

Each mixture component in Fig. 5 is displayed in a different color. The weights, mean values (μ), and standard deviations (σ), for the individual components are given in the first, second, and third column of the table in the bottom panel of Fig. 5, respectively. The mean values are also shown as black numbers in the top panel of Fig. 5. The tilted red numbers in the top panel show the differences between the mean values of neighboring individual mixture components. Fig. 5 describes the melodic pitch inventory which is used by the singers.

The distribution of the melodic steps used, calculated as the pitch difference between subsequent stable pitches, is given in Fig. 6. It shows that the melodic progression in this song is primarily stepwise and rarely contains melodic jumps. The most prominent melodic step sizes are approximately 173 cents with the upwards steps showing less variability than the downward ones.



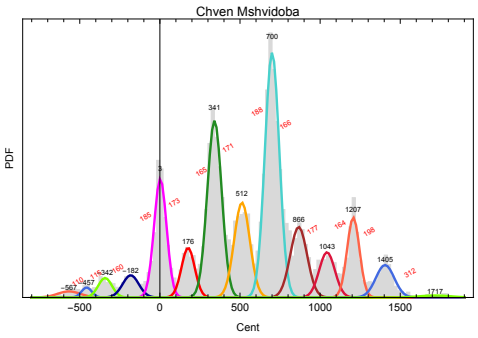
	μ	σ
0.126423	−345.391	59.2326
0.321274	−174.405	31.6349
0.0762331	−0.140965	39.8197
0.387171	172.391	24.3321
0.088899	335.266	154.438

Figure 6: Distribution of melodic step sizes in cents (relative to 55 Hz) of the song "Chven Mshvidoba". The table in the bottom panel shows the parameters of the Gaussian mixture model used to quantify the distribution. For further details see explanations for Fig. 5.

To complete the discussion of the tonal organization, Fig. 7 shows the distribution of harmonic intervals which was calculated from all concomitant intervals in all three voices. Figs. 5 to 7 quantitatively describe the complete tuning system as it was used by the singers. Representing the distributions in Figs. 5 to 7 as Gaussian mixture distributions is motivated by the fact that this facilitates the subsequent analysis, e. g. the comparizon of different performances, enormously. It is interesting to note in Fig. 7 that the mixture component representing the harmonic fifth (turquoise color) has a mean value of 700 cents, which approximately corresponds to just tuning (702 cents). This value differs by nearly 20 cents from a melodic fifth which would be obtained by three subsequent melodic steps of approximately 173 cents (cf. Fig. 6). This is believed to be achieved by intonation adjustments of the individual voices (cf. Nadel (1933)), a phenomenon which is also reflected in the variability of the melodic step size distribution shown in Fig. 6 and which we intend to study in detail within the GVM project.

4.2 Digital Representation of Polyphonic Non-Western Oral-Tradition Vocal Music

In the final example, we will illustrate how a combination of audio and larynx microphone tracks can be used



	μ	σ
0.00983433	−567.17	71.1854
0.00721699	−457.271	31.6286
0.0189798	−342.017	42.6342
0.0253974	−182.018	50.1485
0.105064	2.60931	38.8967
0.046553	176.039	41.1946
0.180325	340.77	44.7445
0.105827	512.304	48.9605
0.249322	700.38	44.6899
0.0842518	866.137	52.4137
0.0502955	1042.93	48.8734
0.0701926	1206.92	38.542
0.0422472	1405.08	57.1401
0.0044934	1716.91	98.6081

Figure 7: Distribution of harmonic interval sizes of the song "Chven Mshvidoba". The table in the bottom panel shows the parameters of the Gaussian mixture model used to quantify the distribution. For further details see explanations for Fig. 5.

to digitally represent the song "Chven Mshvidoba" in a form which overcomes some of the problems related to classical transcriptions. From the spacing of the mean values of the individual mixture components in Fig. 5 it can be seen that the tuning which is used for the song "Chven Mshvidoba" is significantly different from the 12 tone-equal-temperament (12-TET) system, where the pitch spacing would be expected to be either 100 or 200 cents. It would therefore be inappropriate to transcribe this song in western 5-line staff notation (which is based on the 12-TET system).

The representation which we have chosen here is in form of a movie ("pitch track movie")[4] in which the viewer listens to an audio mix from the three headset microphones while watching a vertical cursor (which represents the actual audio playback position) horizontally moving over a display window of a chosen duration (here 25 sec) (see Fig. 8). Shown in the display window are the stable pitch elements (horizontal bars) determined from the larynx microphone tracks together with the corresponding F0-trajectories segments (wiggly lines). The horizontal axis is time in seconds (with respect to the start of the song) while the vertical axis is in cents (with respect to a chosen reference frequency in Hz). In the right part of the display window the pitch distribution shown in Fig. 5 (rotated by 90 degrees) is plotted as a reference for the tuning system used. The used tuning system is also visualized by a set of horizontal lines grid lines, corresponding to the mean values of the individual mixture components (cf. μ values in the table in the bottom panel of Fig. 5).

Whenever the playback cursor falls within a stable pitch segment in any voice, the corresponding pitch element is highlighted and a horizontal bar is superimposed on the rotated pitch distribution. At the same time, the pitch of the lowest stable pitch element is shown within a green ellipse, subscripted by the interval of the second lowest stable pitch element, and superscripted by the interval of the highest stable pitch element (always with respect to the lowest one).

One of the advantages of the type of representation shown in Fig. 8 is that it is automatically adapting to any tuning system used. The display of the "active" stable pitch elements by vertically moving bars (one for each voice) in the right part of the display window is closely related to chironomic choir singing which is a very intuitive teaching practice, see D'Alessandro et al. (2014). "Understanding" the structure of a song from such a representation does not require the ability to read sheet music, which is yet another advantage.

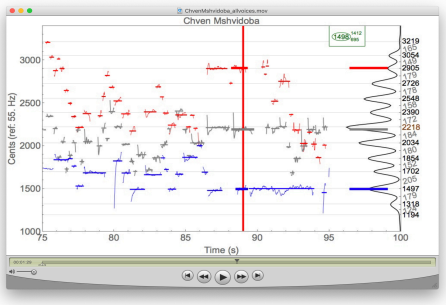


Figure 8: Screen dump of the "pitch track movie" for the song "Chven Mshvidoba". The red vertical cursor marks the audio playback position.

5. DISCUSSION AND OUTLOOK

In recent years, we have seen a revolution in how computer technology changes the way we live and interact with the world around us. Not surprisingly, these changes have also started to influence ethnomusicology and have led to the emergence of the new research field of Computational Ethnomusicology. The success of computational analysis, however, strongly depends on the availability and quality of data. With the new collection of recordings presented here, a high-quality multi-media data set is now available for the first time with which computational analysis of traditional Georgian vocal music can systematically be performed on a larger scale.

The multi-media nature of the new corpus, the unconventional recording strategy (using larynx microphones in addition to conventional audio recordings), together with the unique web-based interface, now enables researchers to

[4] https://www.uni-potsdam.de/fileadmin01/projects/soundscapelab/Videos/ChvenMshvidoba_allvoices.m4v

address a multitude of research questions related to problems such as pitch tracking, harmonic and melodic analysis, and the analysis of the interaction of singers, in completely new ways. In addition, classical ethnomusicological analysis can benefit as well from these corpora by the easy access to the multi-media recordings and the collected metadata.

6. ACKNOWLEDGMENTS

First and foremost, our gratitude goes to all the people who shared their cultural treasures with us and allowed us to be part of their rituals. We are immensely indebted to Levan (Leo) Khijakaze without whom the field expedition would not have been possible in its present form. The project has also been helped by many people who supported us in different ways in particular helping us with establishing contacts with some of the singers. In alphabetical order these are the Chamgeliani family in Lakhushdi, Malkhaz Erkvanidze, and Tornike Skhiereli. Didi madloba to all of you.

We also thank Daniel Vollmer for the continuing technical support and El Mehdi Lemnaouar for the support regarding the Web representation. Finally, we gratefully acknowledge the funding for the GVM project Computational Analysis of Traditional Georgian Vocal Music [MU 2686/13-1, SCHE 280/20-1] (2018 - 2021) through the German Research Foundation (DFG). The International Audio Laboratories Erlangen are a joint institution of the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) and Fraunhofer Institute for Integrated Circuits IIS

7. REFERENCES

Aslanishvili, S. (2010). Forms of multipart singing in Georgian folk songs. In R. Tsurtsumia & J. Jordania (Eds.), *Echoes from Georgia: Seventeen arguments on Georgian polyphony* chapter 4, (pp. 57–81). Nova Science Publishers, Inc.

D'Alessandro, C., Feugère, L., Le Beux, S., Perrotin, O., & Rilliard, A. (2014). Drawing melodies: evaluation of chironomic singing synthesis. *The Journal of the Acoustical Society of America*, *135*(6), 3601–12.

Dirr, A. (1914). Neunzehn Swanische Lieder (Statt eines Referates.). *Anthropos*, *9*(3/4), 597–621.

Erkvanidze, M. (2016). The Georgian Musical System. In *6th International Workshop on Folk Music Analysis, Dublin, 15-17 June, 2016.*, (pp. 74–79).

Jordania, J. (2006). *Who Asked the First Question? The Origins of Human Choral Singings, Intelligence, Language and Speech.* Tbilisi: Tbilisi State University Press.

Jordania, J. (2010). Georgian Traditional Polyphony in comparative studies: History and Perspectives. In R. Tsurtsumia & J. Jordania (Eds.), *Echoes from Georgia: Seventeen arguments on Georgian polyphony* chapter 15, (pp. 229–248). Nova Science Publishers, Inc.

Khardziani, M. (2017). *Svan Folk Songs. Collection of sheet music with two CDs for self-study.* Tbilisi: Tbilisi State Conservatory.

Mauch, M. & Dixon, S. (2014). pyin: A fundamental frequency estimator using probabilistic threshold distributions. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2014)*, (pp. 659–663).

Müller, M., Rosenzweig, S., Driedger, J., & Scherbaum, F. (2017). Interactive fundamental frequency estimation with applications to ethnomusicological research. In *Submitted to the AES Conference on Semantic Audio*, Erlangen, Germany.

Nadel, S. F. (1933). *Georgische Gesänge*. Lautabt., Leipzig: Harrassowitz in Komm.

Paliashvili, Z. (1909). Kartuli khalkhuri simgherebis krebuli (Collection of Georgian Folk songs). Technical report, Kartuli filarmoniuli sazoadoeba (Georgian Philharmonic Society, Tbilisi.

Scherbaum, F. (2016). On the benefit of larynx-microphone field recordings for the documentation and analysis of polyphonic vocal music. *Proceedings of the 6th International Workshop Folk Music Analysis,15 - 17 June, Dublin/Ireland*, 80–87.

Scherbaum, F., Loos, W., Kane, F., & Vollmer, D. (2015). Body vibrations as source of information for the analysis of polyphonic vocal music. In *Proceedings of the 5th International Workshop on Folk Music Analysis, June 10-12, 2015, University Pierre and Marie Curie, Paris, France*, volume 5, (pp. 89–93).

Scherbaum, F. & Mzhavanadze, N. (2017). A new archive of multichannel-multimedia field recordings of traditional Georgian singing, praying, and lamenting with special emphasis on Svaneti. Technical report.

Scherbaum, F., Mzhavanadze, N., & Dadunashvili, E. (2018). A web-based, long-term archive of audio, video, and larynx-microphone field recordings of traditional Georgian singing, praying, and lamenting with special emphasis on Svaneti. *9. International Symposium on Traditional Polyphony,Oct 30 - Nov 3*.

Scherbaum, F., Rosenzweig, S., Müller, M., Vollmer, D., & Mzhavanadze, N. (2018). Throat microphones for vocal music analysis. In *Demos and Late Breaking News of the International Society for Music Information Retrieval Conference (ISMIR)*, Paris, France.

Serra, X. (2014). Creating Research Corpora for the Computational Study of Music: the case of the CompMusic Project. In *AES Semantic Audio Conference*, (pp. 1–9).

Tsereteli, Z. & Veshapidze, L. (2014). On the Georgian traditional scale. In *The Seventh International Symposium on Traditional Polyphony: 22-26 September, 2014, Tbilisi, Georgia*, (pp. 288–295).

Vavra, F., Novy., P., Maskova, H., Kotlikova, M., & Netralova, A. (2004). Morphological filtration for time series. In *APLIMAT 2004*, (pp. 983–990).

Werner, N., Balke, S., Stöter, F.-R., Müller, M., & Edler, B. (2017). trackswitch.js: A versatile web-based audio player for presenting scientific results. In *Proceedings of the Web Audio Conference (WAC)*.

Per Åsmund Omholt
Associate professor
University of South-Eastern Norway
per.omholt@usn.no

Painting blue
-on measuring intonation in Hardanger fiddle tunes

This presentation is an attempt to demonstrate, by using a function in the software *Melodyne* as a polyphonic pitch detection tool, how patterns of variable intonations in Norwegian Hardanger fiddle music can be explored, analyzed and visually presented. The paper is a modest proposal for a reasonably accessible method of measuring and analyzing music, in this case traditional fiddling.

The main source for my research is the Hardanger fiddler Johannes Dahle from Telemark (1890-1980), who is regarded as an excellent performer with an authentic playing style, and who is well-known among other performers and insiders regarding his intonation. Nevertheless, I sense that his manner of tonal "coloring" is perceived as demanding among younger generations.

From the point of view of conventional music theory, the most striking and surprising detail in Dahle's tonal language is the raising of the expected tonic, often in the upper part of the pitch range. In several tunes in different tunings, a tonic, or better, a tonal center/frame, is established through melody and drone strings working together. In the soundscape, this appears as polyphonic structures, and is presumably recognized as a major triad in most cases. Intonations in the upper range – mainly on the E-string – challenge this basic frame when the fiddler fingers the expected tonic "much too high": 20, 50 or even 70 cents above the expected pitch (meaning the diatonic step). These intonations are definitely not accidental; rather, they are intended by the performer. Scales of diatonic intervals with octaves as a frame can hardly be described as basic concepts in this music. Non-diatonic intonations seem to be used as a conscious, expressive tool – the performer is "painting in blue".

My main purpose is to demonstrate how the visual outputs from Melodyne provide a functional point of entry for discussing possible patterns and systematics in an intonation practice that largely challenges conventional music theory. I consider Melodyne, in which pitch/intonations are displayed in relatively clear and legible graphs, to be an adequate tool for my purposes, despite, of course, several reservations concerning the technical measurements.

Figure 1. A screenshot from *Melodyne* showing intonations on the E-string; open string, 1st, 2nd and 3rd finger in first position performed by Johannes Dahle

Furthermore, the findings might bring new knowledge to a debate (in Norway apparently never-ending) on pitch and intonation among scholars in general, and, more specifically, to the case of traditional fiddling in Scandinavia. The performer's irregular intonation, sometimes referred to today as "hovering" or "floating" intervals, or, more frequently, as "blue notes", has been in focus in the Norwegian debate for a century (Omholt 2015 and 2008; Kvifte 2012; Sevåg 1993). A visual approach to older recordings may also serve as a useful guideline for younger generations of fiddlers concerning intonation, scales, timbres and harmony.

Empirical examples will consist of sound recordings of traditional fiddling, live performance, relevant visual examples from Melodyne and simple models and statistics based on a quantitative approach to the Melodyne outputs. Observations will be discussed in relation to existing research on pitch and intonation practice in Scandinavian fiddle music.

REFERENCES:

Recordings:

Dahle Johannes 1954, reel-to-reel recordings, Telemark folkemusikkarkiv, TFATr0701-0713

Literature:

Kvifte, Tellef 2012: "Svevende intervall – og svevende begrep". In Kolltveit (ed.) *Musikk og tradisjon*. Oslo: Novus, 93-112

Omholt, Per Åsmund 2015: "Mælefjøllvisa". In Søyland Moen (ed.) *Musikk og tradisjon*. Oslo: Novus, 29-57

Omholt, Per Åsmund 2008: "På jakt etter folkemusikkskalaen". In Ressem (ed.) *Norsk folkemusikklags skrift* nr.22, 27-59

Sevåg, Reidar 1993: "Toneartsspørmålet i norsk folkemusikk". In Aksdal & Nyhus (ed.) *Fanitullen*. Oslo: Universitetsforlaget, 342-376

Thedens, Hans-Hinrich 2001: "Durifisering eller hva?». In *Norsk folkemusikklags skrift* 15, 28-49

TIMING-SOUND INTERACTIONS IN TRADITIONAL SCANDINAVIAN FIDDLE MUSIC: PRELIMINARY FINDINGS AND IMPLICATIONS

Mats Johansson
University of South-Eastern Norway
mats.s.johansson@usn.no

ABSTRACT

This paper reports from a study of concepts and practices of microrhythm among skilled performers of traditional Scandinavian fiddle music, particularly the so-called *springar* tradition which features non-isochronous and variable beats and subdivisions within a triple meter framework. In this context, microrhythm refers to the overall shaping of musical events at the micro level, encompassing both timing (temporal placement and duration) and sound (shape/envelope, timbre and intensity). A particular focus is to explore how these musical features interact and how timing-sound interactions in turn are understood in terms of groove-forming elements. The referred study consisted of semi-structured interviews with five expert musicians, focusing on the defining features of a good groove, and how aspects of sound are envisaged to affect aspects of rhythm and timing, and vice versa. It was found that groove is largely conceptualized in terms of movement and embodiment; that musical features (timing, accentuation, ornamentation, tone production) are seen to interact and overlap, suggesting a multiparametrical and dynamic concept of groove; and that variation in the overall melodic-rhythmic crafting of the tunes is an important groove-forming element. To further highlight some of these findings, the paper also presents an analyzed sample of a springar tune.

1. INTRODUCTION

This study is part of the larger project TIME: Timing and Sound in Musical Microrhythm.[1] The project conducts comparative investigations of four different rhythmic genres – jazz, electronic dance music, R&B/hip-hop and Scandinavian fiddle music – in order to gain new insights into the relationship between temporal and sound-related aspects of musical perception and performance. The present study focuses on traditional Scandinavian fiddle music, particularly the Norwegian *springar* tradition. The springar (also called *springleik*, *pols*, *polsdans* and *rundom*), which is largely equivalent to the Swedish *polska*, is a traditional partner dance in triple meter with numerous variants across the country. When it comes to the music, differences in rhythm are by far the most important markers of stylistic distinction (Thedens, 2000). Overt differences are found in characteristic patterns of beat accentuation (where and how accents fall), beat duration (the measure may be symmetrically or asymmetrically divided) and beat subdivision (even/duple, uneven/triple, and shifting or ambiguous) (Ahlbäck, 1995). Within these distinctions, practices and concepts of groove, timing and sound are highly specialized. On the one hand, from a bird's eye view the Norwegian fiddle tradition as a whole may seem like a relatively coherent formation of musical practices. On the other hand, among performers it is considered difficult to master more than one of these apparently similar styles of playing, subtle nuances of phrasing and melodic-rhythmic articulation determining the stylistic identity and quality of a performance (Blom, 1981). At the same time, and somewhat paradoxically, stylistic categories are highly flexible in the sense that the same rhythmic style, or even the same tune, may materialize in a variety of ways (Kvifte, 1994; Omholt, 2012). This variability includes a number of musical features related to groove, timing and sound: beat duration patterns (varying from isochronous to highly non-isochronous), rhythmic subdivisions, dynamics (the distribution of accentual energy between and within notes and phrases), phrasing (which and how many notes that are tied together), ornamentation, intonation (the pitching of notes and intervals), harmonization (the use of double stops), onset quality (sharp, soft, gliding), and sound coloring.[2]

From these descriptions, questions arise as to how groove and groove-forming mechanisms are conceptualized among skilled performers in light of the project's overall focus on the relationship between temporal and sound-related aspects of musical perception and performance. To produce data that can shed light on these questions, interviews have been conducted with five expert musicians (see Methods). The paper presents some of the general findings from the interviews, as well as certain specific manifestations of timing-sound interactions. To further highlight some of these findings, the paper also presents an analyzed sample of a springar tune.

In this context, microrhythm refers to the overall shaping of musical events at the micro level, encompassing both timing (temporal placement and duration) and

[1] See info at: https://www.uio.no/ritmo/english/projects/flagship-projects/time/

[2] See https://youtu.be/Iw8Iae5YRdI for examples. The video features the interviewed fiddlers performing in different styles of the springar tradition.

sound (shape/envelope, timbre and intensity). Notably, this notion of microrhythm as a compound concept represents a rethinking that moves beyond existing scholarship's traditional focus on microrhythm as timing (Johansson et al., 2020). As Danielsen (2015) has noted, there is a severe shortage of research on the relationship between sound and timing (as defined above) beyond experimental studies using manufactured sounds devoid of musical context. Moreover, with regard to Scandinavian fiddle music existing rhythm research is largely devoid of ethnographic insight on how performers make sense of performance timing and its associated concepts (groove, flow, phrasing, accentuation and timing–sound relationships) (Johansson, 2017a). The present study thus complements existing research by taking on the complexity of performed music and including ethnographic data from practitioners of a particular style of rhythmic performance.

2. METHODS

To explore discourses on groove and sound–timing interactions among traditional fiddlers, I conducted in-depth semi-structured interviews with five expert performers. The interviews were conducted in 2017 and 2018, based on a semi-structured interview guide that opened with general considerations about what a good groove is, then moved on to more specific questions about the informants' reflections on the importance of timing, sound, and timing–sound interactions, respectively. While all topics and questions mentioned in the interview guide were touched upon, the interviews ended up being relatively open conversations. Moreover, all five informants had their instruments in hand, actively using them to demonstrate particular features of playing technique and associated modes of melodic-rhythmic articulation.

The sample of a springar tune featured in fig. 1 has been analyzed for beat and measure durations (inter-onset intervals), bowing patterns, ornamentation and double stops. Inter-onset intervals were analyzed manually, using the Adobe Audition software to mark the points in the sound graph that correspond to the start/end of the unit concerned and then measuring the distance between the points. It needs to be acknowledged that the fiddle produces sound images which are challenging to account for in terms of rhythmic onsets and that identifying the attack points between which to measure beat durations is a matter of interpretation rather than mere observation. While these sources of uncertainty are compensated for by means of a consistent measurement procedure – making the same decision of placement in all comparable occurrences – it remains that the relationship between physical onset, measurements and experienced rhythm cannot be analytically determined (Kvifte, 2004: 61). A similar objection can be directed against the level of precision with which my analysis operates: the temporal resolution of measurement data is in milliseconds, which is far beyond the threshold for listeners' perception of timing differences (Clarke, 1989). My response to these quandaries is twofold: 1) The precision with which rhythm is produced may be considerably higher than that of a listener attempting to detect such details (Johansson, 2010: 119). 2) It is not assumed that measured timing data corresponds to how temporal relationships are experienced by performers and listeners. For instance, onsets may be ambiguous, or substantial fluctuations in beat durations may go undetected (cf. below). Such discrepancies between measured beat positions and experienced rhythm are themselves interesting observations considering the focus on microrhythm as a multidimensional phenomenon. In this perspective, finding the precise beat onsets in an analysis of microtiming and feeling "the beat" (note the change in meaning) are not necessarily the same thing, and the latter is assumed to be dependent on other aspects of the music than timing alone.

3. RESULTS

Below, some of the main findings from the interviews are reported. The informants' responses are largely paraphrased and synthesized due to space restrictions, while a more comprehensive account will be available is a forthcoming publication (Johansson et al., 2020).

On an overarching level, it can be noted – in line with many other scholars – that groove is conceptualized in terms of movement and embodiment (Roholt, 2014): 1) The music is created with the aim of making listeners move their bodies. 2) The music is crafted in a way that corresponds to particular ways of moving the body, i.e. different styles of dancing. Accordingly, 3) movement is also represented in a rather direct sense: partly by references to the dancers' movements, and partly by references to the physical movements of the musician, particularly bowing patterns, foot-stomping and ornamental fingerings. 4) Rhythmic qualities of the sounding music are identified using movement metaphors, such as "lift," "drive," "flow," "breathing," and "forward thrust." Regarding the first three points, it is indeed striking that dance remains such an important reference point, even as few of the informants regularly play for dancers.

In the musicians' discourse, groove emerges as a multiparametrical phenomenon in that the duration of notes only partially accounts for the music being groovy. Equally important are the dynamics of tone production, how the melody is articulated and how the sound of the instrument is utilized to aid the continuous forward thrust in the music. Moreover, there are many examples of musical parameters overlapping or converging, e.g. timing (short/long) being conflated with accentuation (light/heavy) (cf. Clarke, 1989). Another, more specific example from the interviews is the notion that certain sound effects produced on the fiddle may allude early, late or ambiguous onsets partly independent of actual temporal placement.

Groove is also a dynamic phenomenon in the sense that a proper sense of weight, flow and drive is created through the interaction between all aspects of the music, which mutually influence one another during the course of performance. One implication of this concept is that there is no particular timing and/or accentuation pattern (a generalized groove template) that exists independently of the particularities of the individual performance and that can be translated between different tunes with a sufficient degree of accuracy. It is rather the unique combination of musical features that determines the span of musically viable timings and accentuations.

The above points, in turn, relate to the notion of *variation* as an important groove-forming element, which applies to a number of aspects of rhythmic performance: which notes are accentuated; how the music is phrased by means of varying bowing patterns, rhythmic subdivisions and ornamentations; the alternation between sharper and softer onsets, and the different sound colors of the instrument; and melodic and intonational variation. Notably, these qualities are as much features of sound as of what is commonly referred to as rhythm or microrhythm.

4. MUSIC EXAMPLE

Fig. 1 shows three versions of a two-measure motif from the springar tune "Fra morgon til kveld" performed by the Hardanger fiddler Bjarne Herrefoss.[1] This example can be used as an illustration of several of the points made above. Notably, instead of just repeating the motif, the fiddler uses a range of different variational techniques to breath rhythmic life into the phrase: bowing patterns and beat subdivisions are changing throughout; bow attacks vary along the sharp/soft axis; grace notes and ornaments both blur and accentuate beat positions; long notes exhibit internal dynamic development with a swelling in intensity, creating a surging rhythmic effect; notes are weighted differently between beats and measures; and certain notes are "prolonged" and "shortened" respectively (the quotation marks are justified as explained below). Referring to springar playing in general, the latter two points were talked about by three of the informants in terms of a tension-and-release strategy in which there is an alternation between "holding back" and "letting go" within the phrases. The mentioned strategies also allude that the music is structured in long phrases or sentences of varying length, rather than short repetitive chunks (1-2-3, 1-2-3, etc.).

In terms of beat timing, "Fra morgon til kveld" is a so-called *Tele-springar* in which the beat level is categorically non-isochronous with sequential beat durations of long-average-short in the cycle. However, as seen from the timing data this pattern is not consistent: the duration of the second beat alternates in every other measure, being shorter in the first measure, longer in the second measure, shorter in the third measure and so on. The second beat of the fourth measure has been additionally extended, being 231 ms longer than the shortest second beat. To assess the significance of these timing variations, I created an informal follow-up study in which my informants, together with six other highly knowledgeable springar performers, were asked to listen to the recording (fig. 1) and point out which of the beats that were prolonged and shortened respectively. Interestingly, no one observed the rather substantial durational fluctuation of the second beat and only one of the experts picked out the second beat of the fourth measure as being subjected to expressive variation.

5. DISCUSSION

Although the follow-up study lacks the required control and rigor to meet the standards of an experimental study, it supports the idea that the fluctuations in beat duration are so seamlessly embedded into the overall melodic-rhythmic flow of the performance that they remain largely undetectable. For this to make sense, it seems inevitable to return to the notion of a formative interaction between timing (*when* musical events occur) and sound (*what* is occurring). For instance, as I have suggested elsewhere (Johansson, 2010; 2017b), the melody "generates" durational patterns that could hardly be considered intentional as a particular distribution of time points. As shown in the line chart in fig. 1, while the individual measures have different timing profiles, the timing profile of the motif as a whole is largely consistent throughout the three repetitions with the exception of the second beat of the fourth measure. This, together with a number of similar observations in other springar performances (ibid.), supports the notion that beat timing is intrinsic to the overall melodic-rhythmic articulation of the motif. In this perspective, the difference between the two measures (1-2-3 vs. 4-5-6) is not to be considered a variation in timing as long as the difference is not produced and perceived with reference to the timing profile of the individual measures (or some neutral grid). Instead, it might be suggested that the motif as a whole references itself: when performed differently (cf. the extension of the second beat of the fourth measure), discrepancies will potentially be noticed and assigned an expressive function. As one of the informants cautioned: "When I speak of deviations, I'm simply referring to deviations from the last time I played the same motif."

The above discussion also highlights the multidimensional and emergent nature of the springar groove as expressed in the interviews. Concretely, melodic lines, ornaments, intonations, phrasings, timings and accentuations are not seen as occurring on top of or in relation to a groove. Rather, grooves are formed through the emergent interaction between these musical features. In line with this reasoning, the associated concept of performance timing can be defined in terms of the dynamic integration of all musi-

[1] Audio: https://vimeo.com/340747123

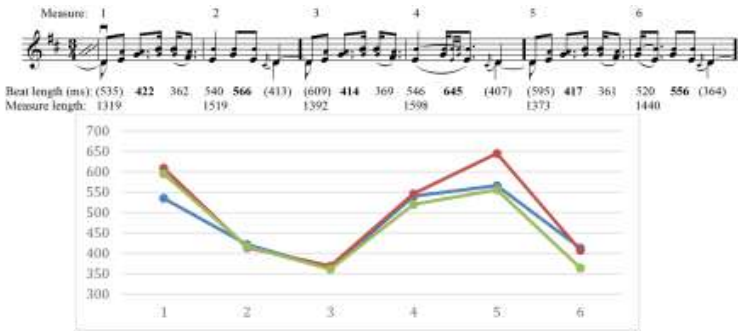


Figure 1. "Fra morgon til kveld." Three versions of a two-measure motif with timing data for beat and measure durations. The line chart shows the beat timing profile for the motif as a whole and its consistency across repetitions.

cal elements into a coherent or well-formed whole, as opposed to in terms of the direct control of time/durations per se. A promising focus for future research, then, is to map the interrelationships between expressive parameters in more detail.

6. REFERENCES

Ahlbäck, S. (1995). Karaktäristiska Egenskaper för Låttyper i Svensk Folkmusiktradition. Ett Försök till Beskrivning. Stockholm: Kungliga Musikhögskolan.

Blom, J. P. (1981). The dancing fiddle. In Blom, J. P., Nyhus, Sven, & Sevåg, Reidar (eds.): Slåttar for the Harding Fiddle. Norwegian Folk Music, vol. 7, 305-312. Oslo: Universitetsforlaget.

Clarke, E. F. (1989). The perception of expressive timing in music. *Psychological Research*, 51(1), 2-9.

Danielsen, A. (2015). Metrical ambiguity or microrhythmic flexibility? In Appen, R., Doehring, D. H., & Moore, A. F. (eds.): Song Interpretation in 21st-Century Pop Music, 53-72. Farnham: Ashgate.

Johansson, M. (2010). Rhythm into Style: Studying Asymmetrical Grooves in Norwegian Folk Music. PhD thesis, University of Oslo.

Johansson, M. (2017a). Empirical research on asymmetrical rhythms in Scandinavian folk music: A critical review. *Studia Musicologica Norvegica* 43.

Johansson, M. (2017b). Non-isochronous musical meters: towards a multidimensional model. *Ethnomusicology*, 61(1), 31-51.

Johansson, M., Danielsen, A., Brøvig-Hanssen, R., & Sandvik, B. (2020). Shaping rhythm: timing and sound in five rhythmic genres. [Forthcoming.]

Kvifte, T. (1994). On Variability in the Performance of Hardingfele Tunes and Paradigms in Ethnomusicological Research. Oslo: Taragot Sounds.

Kvifte, T. (2004). Description of grooves and syntax/process dialectics. *Studia Musicologica Norvegica*, 30, 54-77.

Omholt, P. Å. (2012). 48.600 måter å spille en slått på. *Musikk og tradisjon*, 26, 68-92.

Roholt, T. C. (2014). Groove: A Phenomenology of Rhythmic Nuance. London: Bloomsbury Academic.

Thedens, H.-H. (2000). It's the individual who is the dialect. In Lundberg, D., & Ternhag, G. (eds.): The Musician in Focus. Individual Perspectives in Nordic Ethnomusicology, 75-101. Stockholm: The Royal Swedish Academy of Music.

PHRASING PRACTICES IN NORWEGIAN SLÅTTE MUSIC:
PRELIMINARY RESULTS AND METHODOLOGICAL CONSIDERATIONS

Anders E. Røine
Phd Student
University of South-East Norway
Department of Traditional Art and Folk Music
anders.e.roine@usn.no

Introduction

This paper presents some findings from a broader investigation that aims at a comprehensive mapping of phrasing practices in the older layer of Norwegian traditional music referred to as *slåtter* (literally "tunes"). It also proposes a method for the visualization and subsequent comparison of different styles of phrasing Phrasing, in this context, refers to how musicians use different tools and techniques to combine individual notes into rhythmic patterns. The sounding result of this practice is a significant stylistic feature of traditional tunes. At the same time, phrasing is largely an area of tacit knowledge among traditional musicians, meaning that no explicit vocabulary is developed for its description and dissemination.

Related work

There is some existing research concerning slåtte music on the relationship between meter, phrasing patterns and the logic of dance movements (Blom & Kvifte, 1986). Some issues concerning phrasing patterns are also discussed in general terms by Eivind Groven (Groven & Fjalestad, 1971), Reidar Sevåg (Sevåg, Blom, Nyhus, Gurvin, & Norsk folkemusikksamling, 1981), Morton Levy (Levy, 1989) and Tellef Kvifte (Kvifte, 1987) (Kvifte, 2007). However, these writings do not rely on extensive empirical materials. They focus on the fiddle bow as the producer of the phrasings and do not address phrasing practices on other relevant instruments.

The present study

On this background, an important focus of the present study is to develop and explore a straightforward visualization of the rhythmic imprint produced by phrasings. The end goal is a simple tool for comparison, without the complexity of traditional scores and independent of the instrument at hand.

The empirical material is extensive and consists of archive recordings of Jew´s Harp, the Norwegian Dulcimer called Langeleik, Hardanger fiddle (Figure 1) and Lilting or Hulling in Norwegian.

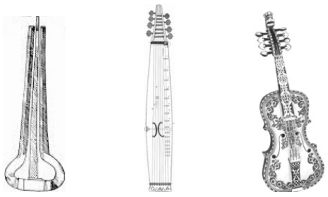


Figure 1. Left to wright: Jew´s Harp from Aust-Agder, Langeleik from Oppland, Hardangerfiddle from Hordaland.

These instruments have coexisted in Norwegian culture for approximately 400 years The Jew´s Harp and vocal traditions are supposedly much older. Today, they are all vital parts of the slåtte traditions. For practical reasons, the proposed presentation will be limited to analyses of so-called 3/8 gangar tunes performed vocally or on the Jew´s harp.

The figures below represent a sequence from 3/8 Jew's harp tune.[1] Figure 2 shows a traditional score notation while figure 3 shows solely the rhythmic imprint produced by the phrasing technique (any number of notes placed in between strokes on the lamella).

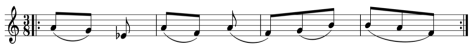


Figure 2. A sequence from a tune played on Jew´s Harp by Andres K. Rysstad (1893-1984)

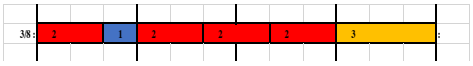


Figure 3. Rhytmic imprint of the phrasing pattern.

[1] Andres K. Rysstad (1893-1984)
https://www.youtube.com/watch?v=LLEBV-AWv4o

The length of the phrase segments (notes tied together) and their placement in time is identified by digits and colours (figure 3). The meter is represented by the vertical lines quantified into three subdivisions per measure.

Visualizations of this kind forms the basis of comparison regardless of how the phrasings are produced or what produces them, cf. strokes on a Jew´s Harp lamella, plectrum movements on the a langeleik, bowings on a fiddle and the usage of vocal chords and lungs in Hulling. The system also allows the empirical data to be quantified, which in turn provides the opportunity for comparative analysis of large amounts of data.

Preliminary results

There are important idiomatic differences between the instruments. This furnishes the musician with unique options and constraints. For instance, while the human voice is limited by the capacity of the lung, the striking of the lamella on a traditional Norwegian Jew's Harp produces a tone that is practically manipulable for approximately one second. In this perspective one could think that such differences would have led to the development of unique phrasing practices tied to the idiomatic constraints of each instrument. However, the preliminary analysis suggests that phrasing practices cannot be interpreted solely on the basis of idiomatic properties and that additional musical logics need to be taken in to account. Several interesting questions arise from this observation.

A: How do instrument specific characteristics influence playing technique and phrasing practices on the slåtte instruments? B: To what extent does the data support the notion of a fundamental shared phrasing practice governed by principles that are not directly linked to the instruments? C: If the phrasings to a certain extent are shared, is there a dominating influencer behind these practices, such as the fiddle? Or, would it be more fruitful to discuss whether a common phrasing practice is a result of a long and ongoing negotiation between instruments and humans that extends beyond the historic limitation of specific instruments.

A comprehensive discussion of these issues is too extensive for the present format. Instead, the presentation offers insights into an ongoing investigation and presents some preliminary results and methodological considerations that purportedly also might be of more general relevance to the study of traditional musics. The musical examples presented will consist of traditional tunes for Jew´s Harp and Hulling (Lilting/Diddling) played by the author and from archive recordings, along with visual analysis of phrasing patterns.

References

Blom, J.-P., & Kvifte, T. (1986). On the Problem of Inferential Ambivalence in Musical Meter. *Ethnomusicology*, *30*(3).

Groven, E., & Fjalestad, O. (1971). *Eivind Groven: heiderskrift til 70-årsdagen 8. oktober 1971*.

Kvifte, T. (1987). *Strøkfigurer — en side av bueteknikken i den norske hardingfele- og felemusikken*. 14.

Kvifte, T. (2007). *On Variability in the Performance of Hardingfele Tunes - and Paradigms in Ethnomusicological Research*. Taragot Sounds.

Levy, M. (1989). *The world of the gorrlaus slåtts : a morphological investigation of a branch of Norwegian fiddle music tradition. 3 - Nasjonalbiblioteket.*

Sevåg, R., Blom, J.-P., Nyhus, S., Gurvin, O., & Norsk folkemusikksamling. (1981). *Norsk folkemusikk: Serie 1 7 : Hardingfeleslåttar Springar i 3/4 takt*.

Analysis of mutual influence of music and text in Svan songs

Nana Mzhavanadze
University of Potsdam
mzhavanadze@uni.potsdam.de

Madona Chamgeliani
Lidbashi Foundation
madona_chamgeliani@yahoo.com

ABSTRACT

The present paper discusses the influence of musicological, linguistic, and ethnological aspects on the music-text relation in Svan songs. In a lot of cases, there is a deep bond between verbal texts and their musical counterparts (Mzhavanadze & Chamgeliani, 2015). Despite the fact that the lyrics of songs are of critical importance to the rituals in which they are performed, many of them are difficult (at times impossible) to transcribe. The present analysis shows that the reservoir of Svan melos is relatively modest in comparison with the verbal texts. In songs in which musical patterns repeat and texts alter, the latter get modified and often distorted to the degree to make them incomprehensible. This is also true for texts of pre-Christian origin, which has interesting ethnological consequences. The results of the current study challenge the isolated linguistic interpretation of verbal texts of Svan songs and emphasize the need for a joint analysis of lyrics, musical, and ethnological context.

1. INTRODUCTION

Svaneti is a high mountainous region in the West of the Republic of Georgia with its own pronounced sub-culture. The Svan language is one of the four Kartvelian languages – the others being Georgian, Megrelian, Laz. For historical and geopolitical reasons, Svans maintained their unique identity through the transmission of their traditions and customs. Therefore, often the Svan singing repertoire can be distinguished from other Georgian musical dialects through its own distinct musical qualities.

During the summer of 2016, we performed an ethnomusicological field expedition in Svaneti to record a new corpus of traditional Georgian vocal music, praying and lamenting (Scherbaum et al., 2019).[1] During the fieldwork, but also during the creation of the meta-data for the archive of the recordings made, we stumbled on various language-related issues which have the potential to strongly affect any subsequent ethnomusicological analysis. Examples for those issues are that:

- Ethnophores often argue among themselves about which words were the "right" ones to sing although, when asked about the meaning of those words, they often would not be able to translate them.
- Transcription of the text of the songs is a big challenge firstly because the words often are difficult to understand and, secondly, due to the phonetic reservoir (especially vowels in upper Svaneti) of Svans being rich and often having grammatical function. This raises a question about phonetic events being of musical or linguistic origin.
- In contradiction to the hypothesis of some authors that Svan songs are believed to be "song-poems" (Shanidze et al., 1939), the texts often do not show high integrity with the music as they are sung.

This suggests that in order to understand Svans' musical grammar, the interrelationship between music and verbal texts cannot be ignored. The presented paper touches on the issues what role either play in the forming process of a musical (artistic) image.

Since the 1930s, many studies have been devoted to the music-language interface in a wider sense. Regarding Georgian traditional music, the topic has been studied by philologists (Imedashvili, 1959), linguists (Zhghenti, 1963), and ethnomusicologists (Kalandadze, 1992, 1993, 2003; Bolle-Zemp, 2001).[2] In 1965, B. Asafyev astutely described the interaction between the language and music pair as "the pressing of melodic juice from living speech" (Асафьев, 1965, p.7). Concerning the ways of music-language interaction, as various studies show, different types of relationships have been revealed. As W. Bright notes "In many instances it is impossible to say which structure has influenced the other.… There are other examples, however, where it seems clear that music has influenced language, or vice versa" (Bright, 1963, p.26). Some studies show a deep link between musical and verbal texts to the extent that the rules of stress of a word and musical accent often coincide with one another (Palmer & Kelly, 1992) and speech rhythm is reflected in not only vocal but even instrumental music (Patel, 2008). But it is not only spoken language features which may have impact on music. Poetic forms of language also need to be examined since language prosody may not coincide

[1] https://lazardb.gbv.de/search

[2] Some studies show, that the neglect of musical context while analyzing texts can provide misleading results. For example, some symbols (signs) in manuscripts of medieval hymns were understood as the means of division of poetic lines (Ingoroq'va, 1954), however, these symbols turned out to be in service of musical rather than versification demands (Nakudashvili, 1996)

with the poetic rhythm, because the latter often obeys particular versification rules.

The issue of origins of poetry in Georgian traditional singing repertoire is the subject of an ongoing scholarly dispute. Some argue that a poem has never been an organic part of a song and it could not be a song which later broke up in parts (music and poem) but vice versa – a musical accompaniment was designed to fit in the versification model (Kurdiani, 1998, p.9). Others see a folk poem to be a syncretic phenomenon, arguing that the genesis of a verse as such was triggered by oldest forms of dancing song-poems (Bardavelidze, 1960, p.26) and that songs even reflect measures of poetic lines (Beradze, 1948).

Shanidze et al. (1939) believed that Svan songs can be seen as sung poetry (poetic texts in the preface of the Svan poetry collections are called "simgheraleksebi" meaning song-poems). However, the results of the analysis of the songs do not always support this thought. A Philologist and poet D. Tserediani, who translated Svan poetry into Georgian (Tserediani, 1968), was the first to detect that "a line of Svan poem and musical phrase do not cover each other in choral songs. Apart from refrains, the syllables inserted between words completely change the rhythm of poetic lines" (Barbakadze, 2011, p.246).[1]

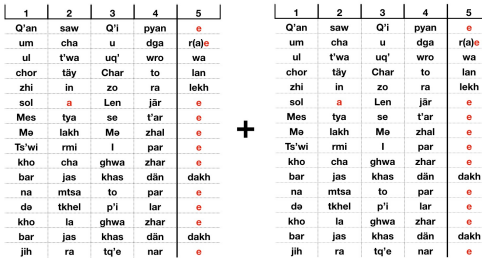
In the present study, two core aspects which affect the mutual relation of language and music in Svan songs are examined:

- The degree and type of interrelation between speech prosody and music.
- The relation of the versification model of sung texts with their musical pairs (rhythmic-melodic models).

Below we discuss two selected examples which illustrate how the joint analysis of verbal and musical texture of the songs can help to shed some light on the types and degree of the mutual influence of music and lyrics.

2. EXAMPLES

Svan songs often are the only original sources of information on the history and/or culture of this part of Georgia. In addition, the study of the texts of songs can shed light on some linguistics questions. A. Shanidze, in the preface of the collection of Svan poetry (Shanidze et al., 1939) suggests that Svan poems have preserved "root vowels which have been reduced and extinguished in three dialects of Svan language … and… also in the oldest forms of nouns and verbs, which must have been part of earlier Svan spoken speech" (ibid). In the following, it will be discussed if for the analysis of the lyrics of the "song-poems", the musical structure/context can actually be ignored. By comparing the rhythmic structure of texts with



1	2	3	4	5
Q'an	saw	Q'i	pyan	e
um	cha	u	dga	r(a)e
ul	t'wa	uq'	wro	wa
chor	täy	Char	to	lan
zhi	in	zo	ra	lekh
sol	a	Len	jär	e
Mes	tya	se	t'ar	e
Mə	lakh	Mə	zhal	e
Ts'wi	rmi	I	par	e
kho	cha	ghwa	zhar	e
bar	jas	khas	dän	dakh
na	mtsa	to	par	e
də	tkhel	p'i	lar	e
kho	la	ghwa	zhar	e
bar	jas	khas	dän	dakh
jih	ra	tq'e	nar	e

+

1	2	3	4	5
Q'an	saw	Q'i	pyan	e
um	cha	u	dga	r(a)e
ul	t'wa	uq'	wro	wa
chor	täy	Char	to	lan
zhi	in	zo	ra	lekh
sol	a	Len	jär	e
Mes	tya	se	t'ar	e
Mə	lakh	Mə	zhal	e
Ts'wi	rmi	I	par	e
kho	cha	ghwa	zhar	e
bar	jas	khas	dän	dakh
na	mtsa	to	par	e
də	tkhel	p'i	lar	e
kho	la	ghwa	zhar	e
bar	jas	khas	dän	dakh
jih	ra	tq'e	nar	e

Fig. 1 Selected verses of the text of *Q'ansaw Q'ipyane* in its poetic form**.** For further explanations see text.

the rhythmic-melodic structure of the corresponding music it will be demonstrated that the structural characteristics of texts can become strongly conditioned by musical demands. Besides the linguistic aspects, the concept of "song-poems" also gets challenged by considering the effect of the process they undergo while musicalized.

2.1 Q'ansaw Q'ipyane[2]

"Q'ansaw Q'ipyane" is a round dance song with a clear meter and rhythm.[3] It is isorhythmic and the articulation of the text does not change (although the tempo gradually grows faster). As a consequence, the text syllables are equally distributed within the given musical metric-rhythmic frame.[4]

Based on the study of the text of this song, which contains the word *p'ilare*, A. Shanidze hypothesised that this word, which is equivalent to the word *p'irebi* in Georgian (or *persons* in English), must be in its original form and the ending vowel 'e' must have been reduced only in modern day Svan language. If we look at the text systematically, however, we can observe that apart from *pilare* there are other words with the same ending: *tq'enar*(e), *ghwazhar*(e), *lenjar*(e), *məzhal*(e), etc.

In the following we will discuss possible reasons why these words with the ending vowel 'e' could appear in such forms. The first observation of the text is that the lyrics of the song represent a poetic form. The commonly assumed versification model of Svan poetic texts is the so called "maghali shairi". This means that the poetic lines are organized on a metric model of 4 + 4. However, the poetic lines of "Q'ansaw Q'ipyane" are based on a model of 5+5 (**la** la la la + **la** la la la la) (Fig. 1). A comparison of the metric/rhythmic accents of the text and that of the music shows pretty close alignment. This enables a listener to follow and understand the text of a song despite the additional vowels which are used to fill in the gaps in the sung text. Such an application of extra syllables seems to be conditioned by musical and versification demands. So, apart

[1] We are quoting from the book by T. Barbakadze since we failed to find the original source of the given quote. However, during the phone conversation, D. Tserediani confirmed that the given observation belongs to him.

[2] To ensure that the transcription of Svan texts is close to the original and reflects the phonetic peculiarities of Svan language, we have combined two transcription systems: for consonants – romanization of Georgian via using Latin script (national system, 2002; https://en.wikipedia.org/wiki/Romanization_of_Georgian); for vowels and some Svan-specific consonants – TITUS (http://titus.fkidg1.uni-frankfurt.de/didact/caucasus/kaukvok.htm).

[3] https://lazardb.gbv.de/ Title: QansavQipiane_TsalkaVillage_TsalkaPeople_20160909_VSOAX4.mov

[4] See one of the variants of the text and its translation at: http://titus.uni-frankfurt.de/texte/etca/cauc/svan/svapo/svapolex.ht

from the prosodic ‘o’, we have other prosodic vowels such as, in this case, ‘e’.

There are several factors supporting the argument for the conclusion that the ending vowel ‘e’ has a euphonic (poetic and mainly musical) function and that it is not a primarily linguistic phenomenon. The beginning lines of the pure text, as mentioned above, when freed from all extra syllables and vowels, are constructed within a five-syllable temporal frame. This renders the content of the song, via such words as: *umcha-udgara* (ageless and immortal), *uwltwa-uq’wrowa* (un-castrated), *chortay Chartolan* (Chartolan with weird walking legs), etc. These words and expressions are often used to characterize personages, which is a typical feature of Svan ballads. When telling stories, Svans always describe the characters in order to assist visualization of them, to build up an image, to render a storyline and to trigger the relevant emotions. Therefore, commonly, the very beginning lines already, as they start to tell the story, set a specific metric-rhythmic model, which on its own, in turn, sets a measuring example for the remaining lines. In the given song, “Q’ansaw Q’ipyane”, the versification model is based on a 5+5 syllable principle instead of 4+4 (which is typical for the Svan poetry) (Fig. 1). Other lines replicate this model. Despite the importance of the actual content of the text, the influence and dominant role of the musical and poetic forms and aesthetics over the information content of the words is clearly manifest. Syllabic vocalization of the text within the given musical model is impossible for a two-syllable foot cannot be equally distributed over a fourfold and a three-step motif. Thus, it naturally requires the filling of the tune with additional phonetic material, which happens at the expense of vowels and peripheral vocabulary. The text therefore becomes “ornamented” in order to fill in the gaps within the musical motif.

As a result, instead of the linguistically correct *Q’an-saw* (2 syllables), in the sung version (Fig. 2) we have *Q’a-nə-sa-wə* (4 syll.), *um-cha* (2 syllables) – *u-mə-cha-i* (4 syll.), etc. In the case of the next “feet” the same pattern applies. Instead of *se-t’ar* (2 syllables) we get *se-t’a-re* (3 syll.), *ghwa-zhär – ghwa-zhä-re* (3 syll.), and *pi-lar – pi-la-re* (3 syll.) (Fig. 2). In Fig. 1 the sung text is shown in its poetic form. The vowels highlighted in red describe poetic and musical events which appear to fill in the 5-syllable poetic model on the one hand, and on the other, to help to fit in with metric and rhythmic model of the tune and complete the musical phrase. That is why these extra vowels appear only in the end of a musical phrase/poetic line and not in the middle of it.

Fig. 2 shows the text coupled with the tune (melodic contour) with its own metric and rhythmic framework. The phonetic units, which are of musical and poetic origin are highlighted in red. Green ones denote the extra vowels/syllables which appear for purely musical reasons. The red lines on the top of the figure show temporal flow of the music within which the appearance of additional vowels is very regular and coincides with either mid-phrase rests or final endings of the phrase. Furthermore, precisely because of the influence of the ending ‘e’, which at the same time plays as a rhyme-making role in some (including the old archive) recordings, we hear *udgar-e*, although its correct grammatical form is *udgar-a* (cf. Fig. 1. At the end of the second line, the correct ending of the word ‘a’ is given in parenthesis which is replaced by ‘e’ highlighted in red).

It seems that musical and versification influences affect a word to the degree that the semantic meaning changes. For example: the correct form of *ghwazh-är* is *ghwazhmäre*: *ghwazh* (a male) and *märe* (a man) while *zuralmäre* means a woman; *ghwazhmärāl* is its plural form – *men*; however, for musical reasons the ending - *zhmärāl* or - *mär* is added the vowel ‘e’; Thus, if we make a literary translation of *ghwazhar(e)*, it will mean *one man* which is nonsense when telling about coward (or brave) men as a group, unless we consider it as mainly a musical event. It is worth noting that in the archives, some of the given words are

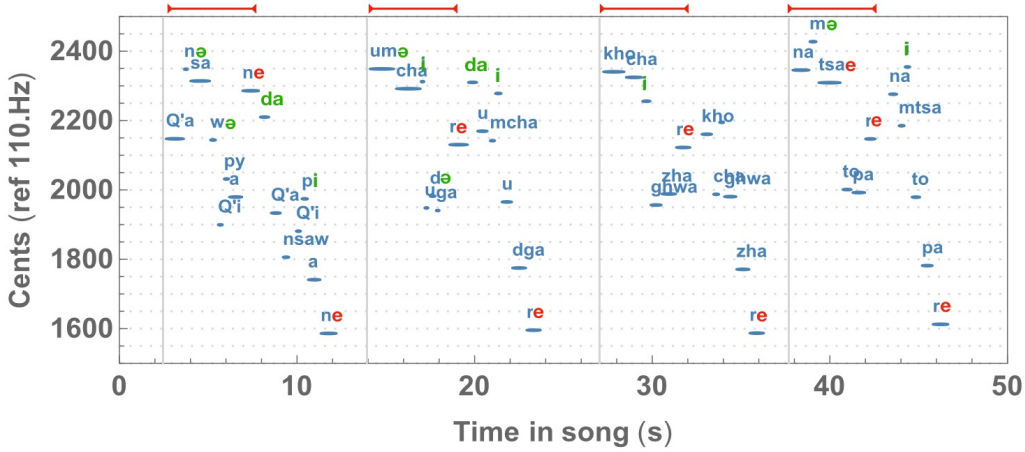


Fig. 2 Text of the sung version of *Q’ansaw Q’ipyane* coupled with the melodic contour.

sometimes presented in their correct form, i.e. *Ghwazhar,*[1] *lenjār*.[2]

2.2 Ushgulas makheghwazhare

Often in Svan songs, a part of the sung text, including the title is semantically so obscure that it is dismissed as nonsense. Deeper study of such examples, however, may reveal the empathic adoption of certain word fragments during the singing process, which takes place due to musical, mainly rhythmical requirements. A good illustration of this phenomenon is found in the round dance song "Ushgulas makheghwazhare". This song tells the story of young men from Ushguli who went hunting and were killed by avalanche (Akhobadze, 1957, p.31).
Observation of all available variants of the song made at different times reveals that the song often appears under a different title. They are listed below in chronological order:

- *Ushgulas makheghwazhare* (or *Ushgul lasma*), recorded by M. Gujejiani from Beka Gasviani (70) in Khalde village (Shanidze et al., 1939:200-204).
- *Ushgulas makheghwazhare*, recorded by V. Topuria in Ipari village from a teacher Gerasime Gulbani in 1926 (Shanidze et al. 1939:204-206).
- *Ushgulas makhe ghwazhare*, recorded by V. Akhobadze from Giorgi, Ivane and Grisha Nizharadzes in the vill. Ushguli in 1950 (Akhobadze, 1950).
- *Ushgwla lasma*, notated version preserved in the Archive of folklore State Centre. The song was transcribed by E. Sarishvili as it was sung by I. Pilpani in 1959 (Sarishvili, 1959).
- *Zhareda*, recorded by Frank Scherbaum and myself from Ruben Chark'viani in Kutaisi in 2016. (https://lazardb.gbv.de/detail/8079).

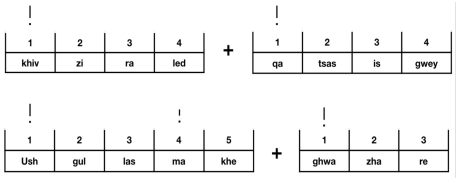


Fig. 3 Versification models used in the poetic form of *Ushgulas makheghwazhare*.

A comparative study of the musical parameters of the above listed variants (except the first two since they depict only the text) reveals firm resemblance of the melodic contour and harmonic vertical. The variants of R. Chark'viani and I. Pilpani are almost identical. A couple of significant details concerning the ways of coupling the words and tune, as well as the vocable segments of the song also have been observed.
Most of the poetic lines without vocables are constructed by the versification model of "maghali shairi"(4+4) (sometimes "dabali (5+3) shairi") (Fig. 3). The peripheral vocabulary of the sung text appears to differ, however. For example, the variant recorded in 1936 starts with the word *arida* (Shanidze et al., 1939:200), whereas the one dated by 1950 (from Ushguli) says *zhare da* (Akhobadze, 1950). In the manuscript from 1959 we read *zhaleoda* (Sarishvili,1959), whereas R. Charkviani (from Ushguli by origin) also calls the song *zhareda* as he starts the song with it (https://lazardb.gbv.de/detail/8079).

Apart from this, as mentioned earlier, the first line of the song (which often represents the title of the song without any changes) is read/sung differently:

Arida… ushgul lasma khilghwazhaled (Shanidze et al., 1939, 200)
Zhare da … Ushgulas makhe da ghwazhareda (Akhobadze 1950)
Zhaleoda… Ushguliwo lasmasai khelghwaio zhale (Sarishvili, 1959)
Zharedai… Ushguliwo lasmawo khelghwaia zhare da (R. Charkviani (https://lazardb.gbv.de/detail/8079)

Fig. 4 demonstrates the process which the given line undergoes while musicalized. Exclamation marks label the musical accents which split the words and make linguistically illogical stress on syllables and a listener perceives a nonsense verbal unit as a word (Fig. 4 a, b). They mark the strong beat of the musical metre. The syllables given in black are non-grammatical segments, "fillers" (Fig.: 4a, 4b). The words split due to musical demands are highlighted respectively in green and blue (Fig. 4 a-c).
As a result, if we read the words as they are adapted with the musical phrases, we can discern that the accents occur on linguistically irrelevant syllables and therefore, the sentence makes no sense (Fig. 4c).
Although it is easier to correctly read the ending part of the phrase, transcription of the middle part is complicated because it becomes semantically very obscure due to the singling out of the segment ***lasma*** as a separate word. In fact, it is a product of the artificial union of the suffix of the previous word: **- *las*** and prefix of the word after: ***ma* - (las+ma)** (Fig. 4c).
Due to the dominance of the music, and the influence of musical metric accentuation, the spoken "guts" of the word dissolves into chaos, altering our perception of the literal meaning of the word. So, informants automatically name the song not by a separate title, but instead by the first words of the opening phrase of the song and thus, the song becomes "Ushgul lasma" – a title devoid of any real meaning.

[1] http://titus.uni-frankfurt.de/texte/etca/cauc/svan/svapo/svapolex.ht

[2] http://titus.uni-frankfurt.de/texte/etca/cauc/svan/svapo/svapolex.htm

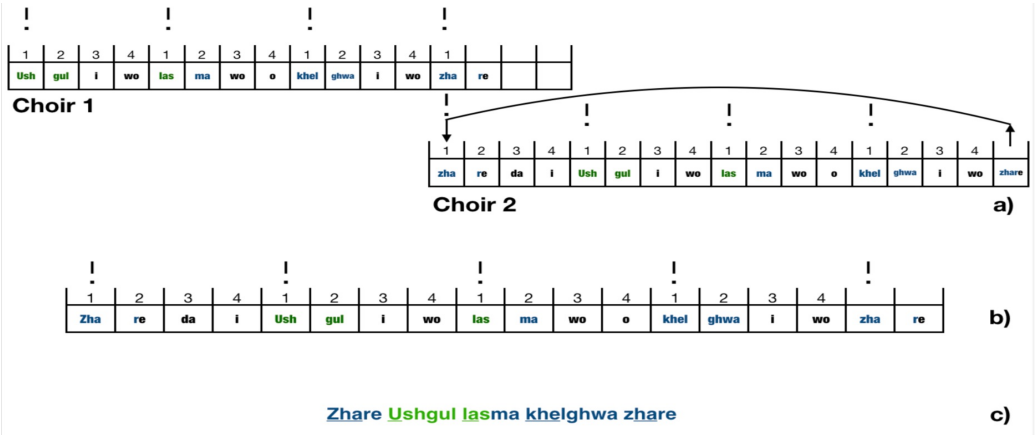


Fig. 4 Modification of text patterns in the lyrics of the song *Ushgulas makheghwazhare* during the musicalization.

Often, after a word gets fragmented due to musical demands, this fragment obtains totally musical function as it can repeatedly appear in the song. An example of this can be the word *zhareda*, which is typically sung at the end of a musical phrase. This has a strong definitive beat, with the accent on ***zha***. However, as mentioned above, ***zhar(e)*** is in fact the second segment of the actual word *makheghwa-**zhar(e)***. ***zhar(e)*** is therefore used in this context to distribute the text over the musical phrase. The given part of the word then becomes the refrain which is heard repetitively by both choirs and which although semantically degraded, gains its own independence as a euphonic tool (Fig. 4).

3. DISCUSSION AND CONCLUSIONS

The joint analysis of a wider singing repertoire of Svans made it possible to describe types of interrelation between words and music. In addition, songs featuring a similar type of language-music relationship, are also closely related musically and share the ethnological context. For example: in some sacred ritual hymns with extended recitative parts, words and music interrelate differently than those in dance songs with a clear rhythm and meter (Mzhavanadze, 2018).

The results of the present analysis challenge two scholarly hypotheses: a) that archaic features of the Svan language have survived in texts of Svan songs; b) that Svan poetry is of syncretic origin and represent song-poems.

Instead, we observe that musical imperative repeatedly relegates linguistic expression to a subordinating role. We illustrated how the status of a word can change by the addition of extra vowels (for example in "Q'ansaw Q'ipyane"). In this context, the analysis of the verbal text without considering its relation with the musical structure can lead to wrong conclusions. In this way too, textual patterns can mistakenly be assessed as a linguistic phenomenon (cf. words such as *p'ilare* or *topare* etc. which are wrongly seen simply as older forms of given words). In addition, there are often occasions where the music takes precedence to the text in even more dominant ways, using segments/fragments of words for euphonic, music structure forming purposes. In this case these fragments are ripped off their semantical meaning, with the resultant loss of any concrete message. Such is the inherent power of this process that even more extraordinarily, these fragmented segments, can turn into actual refrains. In the most extreme cases, they can re-appear as the very title of a song, such as e.g. "Ushgul lasma". Most of the Svan poetic texts are not of ontologically synthetic origin because on the one hand their poetic forms as well as the verbal units get distorted as they are sung, and, on the other hand, because without music they have their own versification mode.

On more general level, we have noted that the Svan intonation inventory is rather limited and its rhythmic-melodic constructs are multiply used. As a consequence, verbal texts are being modified and even distorted due to the shift of their rhythmic accents. This sometimes leads to morphological changes of words and to semantic ambiguity. In order to fill the gaps in the existing musical rhythmic-melodic constructions/phrases, multiple phonetic units (vowels, syllables) are inserted in the process of musicalization. This may have the effect that the resulting **musical constructs are more stable and longer lasting than the original textual ones.** Such a stability and "dominance" of musical features could be explained by the hypothesis that the musical frame would be of older origin than the current texts.

Apart from this, we feel that a comprehensive multidisciplinary study of Svan repertoire could further reveal peculiarities of the local musical language and help identify the rhythmic-melodic constructs which may have been imported from neighboring regions/dialects of Georgian traditional music. This, however, is the subject of future research.

4. ACKNOWLEDGEMENTS

Nana Mzhavanadze thanks Frank Scherbaum for the stimulating discussions and his help with generating the figures. She also gratefully acknowledges her funding by the German Research Foundation (DFG) through the project „Computational Analysis of Traditional Georgian Vocal

Music" [SCHE 280/20-1] (2018 – 2021). She also expresses her gratitude to FaRiG fund for its financial contribution to the field work in Svan eco-villages in 2014.

5. REFERENCES

Akhobadze, V. (1950). *Archive Materials (notated transcriptions) of The Folklore State Center of Georgia* (No. #1824). Tbilisi.

Akhobadze, V. (1957). *Kartuli khalkhuri simgherebis k'rebuli (Collection of Georgian folk songs)*. Tbilisi: T'eknik'a da shroma.

Barbakadze, T. (2011). *Kartuli leksmtsodneobis ist'oria (History of theory of Georgian poetry)*. Tbilisi: Lit'erat'uris inst'it'ut'is gamomtsemloba (Publishing house of the Institute of Georgian Literature).

Bardavelidze, J. (1960). *Kartuli khalkhuri lekstts'q'obis sak'itkhebi (Versification issues in Georgian folk poetry)*. Tbilisi: Georgian SSR Academy of Sciences.

Beradze, P. (1948). Hegzamet'ris sak'itkhisatvis (On the issue of hexameter). *Mnatobi*, (7), 140–148.

Bolle Zemp, S. (2001). Khmovnebi da ak'ordebi. Simghera zemo Svanetshi (Vowels and chords. Singing in Upper Svaneti). In R. Tsurtsumia (Ed.), *Sasuliero Da Saero Musik'is Mravalkhmianobis P'roblemebi (Problems of Polyphony in Sacred and Secular Music)* (pp. 292–303). Tbilisi: Tbilisi State Conservatoire (in Georgian with English summary).

Bright, W. (1963). Language and Music: Areas for Cooperation. *Ethnomusicology*, *7*(1), 26–32.

Gippert, J. (2002). TITUS; Svan Poetic Texts. Retrieved March 19, 2019, from http://titus.uni-frankfurt.de/texte/etca/cauc/svan/svapo/svapo.htm

Gippert, J. (2002). TITUS; Svan Prose Texts. Retrieved March 19, 2019, from http://titus.fkidg1.uni-frankfurt.de/texte/etca/cauc/svan/spto1/spto1.htm

Imedashvili, G. (1959). Kartuli k'lasik'uri sagaloblis p'oet'ik'is sak'itkhebi (Issues of poetic speech of a Georgian classical chant). In *Literaturuli Dziebani (literary investigations)* (pp. 177–193). Tbilisi.

Ingoroq'va, P. (1954). *Giorgi Merchule. Kartveli mts'erali meate sauk'unisa (Giorgi Merchule. A Georgian writer of the 10th century)*. Tbilisi: Sabch'ota Mts'erali.

Kalandadze-Makharadze, N. (1992). *Kartuli khalkhuri simgheris p'oet'uri da musik'aluri t'ekst'is shinaarsobrivi urtiertmimartebis p'roblema: glosolaliis sak'itkhisatvis (Problem of interrelation between of poetic and musical texts of Georgian folk songs: an issue of glossolalia)* (No. 5910). Tbilisi.

Kalandadze-Makharadze, N. (2003). Mravalkhmiani kartuli khalkhuri sasimghero met'q'velebis erti art'ik'ulatsiuri taviseburebis shesakheb (On one peculiarity of articulation in Georgian polyphonic singing). In J. Tsurtsumia, Rusudan and Jordania (Ed.), *The First International Symposium on Traditional Polyphony*. (pp. 340–349). Tbilisi: International Research Center for Traditional Polyphony of Tbilisi State Conservatoire.

Kalandadze-Makharadze, N. (1993). *P'oeturi da musik'aluri t'ekst'is semant'ik'isatvis kartul folk'lorshi (On semantics of poetic and musical texts in Georgian folklore)* (No. 5910). Tbilisi.

Kurdiani, M. (1998). *Saerto-Kartveluri versipik'atsiuli sist'ema da lekst'ts'q'obis zogadlingvist'uri teoria. (Common-Kartvelian versification system and general linguistic theory of versification)*. Academy of Sciences of Georgia. Arn. Chikobava Institute of Linguistics.

List, G. (1963). The Boundaries of Speech and Song. *Ethnomusicology*, *7*(1), 1–16. https://doi.org/10.2307/924141

Mzhavanadze, N. (2018). *Svanuri sak'ult'o rit'ualis musik'ologiur-antrop'ologiuri asp'ekt'ebi (Musicological and anthropological aspects of Svan aacred aitual)*. Ilia State University. Retrieved from https://drive.google.com/file/d/1-Q1-1a7SWLHKJrW2_XaAWVC4dvl6zPLH/view

Mzhavanadze, N., & Chamgeliani, M. (2015). On the Problem of Asemantic Texts of Svan Songs. *Kadmos*, *7*(8), 49–109.

Nakudashvili, N. (1996). *Himnograpiuli leksis st'rukt'ura (Structure of a hymnographic text*. Tbilisi: Metsniereba.

Palmer, C., & Kelly, M. H. (1992). Linguistic prosody and musical meter in song. *Journal of Memory and Language*, *31*(4), 525–542. https://doi.org/10.1016/0749-596X(92)90027-U

Patel, D. A. (2008). *Music, Language, and the Brain*. Oxford: Oxford University Press.

Powers, H. S. (1980). Language Models and Musical Analysis. *Ethnomusicology*, *24*(1), 1–60.

Sarishvili, E. (1959). *Archive Materials (notated transcriptions) of the Folklore State Center of Georgia* (No. # 4227). Tbilisi.

Scherbaum, F., Mzhavanadze, N., & Dadunashvili, E. (2018). A web-based, long-term archive of audio, video, and larynx-microphone field recordings of traditional Georgian singing, praying and lamenting with special emphasis on Svaneti. In *9. International Symposium on Traditional Polyphony,Oct 30 - Nov 3.* (p. Submitted).

Scherbaum, F., Mzhavanadze, N., Rosenzweig, S., & Mueller, M. (2019). Multi-media recordings of traditional Georgian vocal music for computational analysis. In *9th International Workshop on Folk Music Analysis, FMA 2019, Birmingham, July 2-4.* (p. Submitted).

Shanidze, Akaki. Topuria, Varlam. Gujejiani, M. (1939). *Svanuri p'oezia (Svan poetry)*. (M. Kaldani, Ed.) (2014th ed.).

Tserediani, D. (1968). *Svanuri khalkhuri leksebi. Irinola (Svan folk poems. Irinola)*. Tbilisi: Sabch'ota Sakartvelo.

Zhghenti, S. (1963). *Kartuli enis rit'mik'ul-melodik'uri st'rukt'ura (Rhythmic-melodic structure of Georgian language)*. Tbilisi: Tsodna.

Асафьев, Б. (1965). *Речевая интонация*. Ленинград: Гос. муз. издательство (Музгиз).

TOLEDO, ROME AND THE ORIGINS OF GREGORIAN CHANT - AN ALTERNATIVE HYPOTHESIS

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

1. INTRODUCTION

It has long been believed that Gregory the Great (540-604) created Gregorian chant. Since the restoration of this chant in the late nineteenth century, however, the Carolingian propaganda that created this myth has been unmasked. In the 1950's, the scholarly debate began to focus on the second half of the eighth century as the era of the origin of Gregorian chant. In that period, Roman chant was introduced in Francia, underwent some changes, and was exported throughout Europe as "Gregorian chant", still preserved in dozens of manuscripts with music notation since ca 900. Simultaneously the chant in Rome itself also changed, and was finally written down in manuscripts since the late eleventh century: this is referred to as "Old Roman chant" (Hiley, 1993). While most scholars agree on this general picture, this paper offers a new hypothesis based on computational evidence. In this hypothesis, the Carolingians deliberately created a new repertory out of Roman texts by setting them to Iberian melodies, thus replacing their own "Gallican" ones, that is, the local melodies of Francia.

2. CAROLINGIAN PROPAGANDA AND REALITY

Although the legend concerning Pope Gregory may have been unmasked, it seems possible that much of the Carolingian propaganda still lingers on. Since nothing about the Roman melodies before the year 800 is known with certainty, it is possible that these melodies no longer existed at all, and were basically reduced to only the chant texts, to be recited or sung to simple formulas, much as acclamations in modern times. Maybe only some ordinary chants survived. After all, the only evidence for Roman "music" is given by text sources such as the *ordines romani,* describing the mass, and the Roman (as opposed to the Gallican) psalter that formed the source for the Gregorian mass proper texts preserved in the ninth century sources of the *Sextuplex* (Hesbert, 1935). The loss of these melodies seems particularly plausible when there would have been a decline in the Roman liturgical tradition in the seventh or eighth centuries. Even when such a decline cannot be shown with certainty, the contrast between the well-documented rise of Toledo and the poorly documented history of Rome gives pause for thought.

It seems plausible that the Carolingian Renaissance was an effort to revitalize ancient Rome, including Gregory and the Apostles Peter and Paul -- for "thou art Peter, and upon this rock I will build my church" (Mt. 16:18). The Carolingian propaganda may have concealed an agreement with the Papacy to put Rome on the map again: protection of the Papacy in exchange for the Papacy helping the Carolingians to unify their realm through the liturgy. As is well known, Pippin the Short founded the Papal State in 754 and Charlemagne was crowned "Emperor of the Romans" in St. Peter's basilica by Pope Leo III on Christmas Day of the year 800. This coronation was probably one of the most significant events in Western history.

A mass antiphoner with music notation would have been an important vehicle for such an agreement. Although disputed, Kenneth Levy has convincingly argued the existence of a lost late eighth century Carolingian archetype of the Gregorian gradual with neumatic notation (Levy, 1998). Traces of early editing in graduals copied all over Europe presuppose a manuscript with music notation preceding the earliest preserved sources. A typical example is the difference between the nearly identical verses of the graduals *Excita Domine* and *Hodie scietis.* While the earliest sources give the complete verse of *Excita* with notation, the verse of the next gradual, *Hodie*, has only notation on the words *coram Ephraim*, marking the slightly different melos of *Hodie*'s verse (Maessen, 2008). Such details show that there must have been an authoritative source preceding the earliest surviving witnesses. Apart from Gregory as the author, the Carolingian propaganda may therefore have included the music as well. If so, where would this music have come from? In the hypothesis of this paper, Rome did not have music of its own, and the Gallican music should be replaced for the sake of Rome. The best place to look for this music seems the Iberian Peninsula, separated from the rest of Europe by the Muslim conquest.

One of Charlemagne's important advisors, Theodulf of Orleans (755-821), author of our Palm Sunday hymn, *Gloria laus et honor,* was a Visigoth, probably from Zaragoza, who admired Rome. Unlike Rome, however, Toledo, the centre of the Visigothic church, had been growing in importance since the time of Gregory's friend Leander of Seville (534-600). Leander and Gregory had met in Constantinople between 579 and 582. Leander then converted Ibe-

ria to Catholicism in 589. The Byzantine centre of Cartagena moved to Toledo in 610. Leander's brother Isidore (565-636) provided a detailed description of the Visigothic rite in the early seventh century and presided over the fourth council of Toledo (633), where he decreed a single order of praying and chanting for Iberia and Gaul. The seventh century saw increasing liturgical and musical activity in Toledo (even with different composers), continuing after the Muslim conquest of 711, at least until the end of the eighth century. It was Isidore who lamented the fact that the sound of the melodies would vanish, since there was no way to write it down: "If the sounds are not learnt by heart, they will perish, since they cannot be written" (Levy, 1998). Yet there are strong arguments that most of the lost melodies of the Visigothic/Mozarabic rite, as preserved in pitch-unreadable notation of tenth century manuscripts, already existed before the Muslim invasion (Randel, 1969). Studying the early tenth-century León antiphoner (E-L 8) we can easily see that these melodies must have been quite sophisticated (Maloy, 2014). In addition, computational analysis (based on n-gram language models of numbers of notes on syllables) shows that a significant part of these melodies is much closer to the Gregorian melos than to other preserved medieval chant traditions, including the Old Roman (Maessen & Van Kranenburg, 2018), suggesting that they may have been at the base of it.

Although the above hypothesis may seem provocative, it is less pervious to counter-arguments than one might surmise. Pfisterer's argument, for example (Pfisterer, 2002), based on text sources, that the earliest chants were created in Rome for the major feasts from the fifth century onwards, can easily be refuted as inconclusive. We need strong arguments against a Roman decline, or references to musical details in Roman sources from the seventh or eighth centuries. Such references barely exist. The best there is are general references to the liturgy and its chant, or references to specific chants in a general way. An example of the latter is the arguably only real "Gregorian" chant, *Deprecamur te Domine*, that in 597 was "sweetly sung" near Canterburry (Levy, 1998). However, there is no conclusive argument for a specific melody of this chant, simple formulas could also explain the story.

Another objection to the hypothesis may be found in the fact that the unification of monastic observance, replacing the *regula mixta* observances with the *Rule of Benedict*, was only realized after Charlemagne's death (814), under his son Louis the Pious. Since it seems easier to unify monastic observance than chant practice, a previous change in chant practice would seem unlikely. And yet, for the unification of his realm, Charlemagne may well have aimed at chant from the beginning, since chant touched everybody, not just the monks.

In the absence of more specific references, there is a distinct possibility that Pippin the Short and Charlemagne created a new repertory out of Roman texts set to Iberian melodies, replacing the Gallican ones. To accomplish this they may have preferred Iberian melodies set to the typical Iberian textual collages (Levy's "libretti"; Levy, 1998), because newly created chants based on these melodies, contrary to those with literal biblical citations, would less likely be perceived as Iberian chants. In the margin of their newly created "Gregorian" mass propers, some chants may have escaped the control of the Carolingian propaganda. Examples of this can be found in offertories like *Erit vobis* and *Oravi Deum,* that Baroffio and Levy argued to be of possible Gallican heritage (Levy, 1998). Significantly, these offertories are also found to be outliers in computational analysis (Maessen & Van Kranenbrug, 2018).

3. CONCLUSION

The hypothesis of this paper interprets so-called Old Roman chant as a local development of Gregorian chant. It argues that Leander may have contributed more to the Gregorian melos than Gregory. The complete repression in the eleventh century of the Beneventan and Mozarabic rites (Hiley, 1993) is seen as the ultimate result of a Carolingian agreement with the Papacy.

The hypothesis is based on computational evidence and is strengthened by the contrast between the well-documented rise of Toledo and the poorly documented history of Rome. What is at issue is the question what can be said at all about music in a period for which we have no musical witnesses and much of the circumstantial evidence is lacking. This paper shows that computational analysis of the available data can help answering this question.

4. REFERENCES

E-L 8: The León antiphoner; León, Cathedral Archive, MS 8; https://bit.ly/2KAGGrV

Hesbert, R. (1935). Antiphonale Missarum Sextuplex. Rome: Herder.

Hiley, D. (1993). Western Plainchant. A Handbook. Oxford: Clarendon Press.

Levy, K. (1998). Gregorian Chant and the Carolingians. Princeton: Princeton University Press.

Maessen, G. (2008). De tweede fase in de reconstructie van het gregoriaans. Amsterdam: Gregoriana Amsterdam.

Maessen, G. & Van Kranenburg, P. (2018). A Non-Melodic Characteristic to Compare the Music of Medieval Chant Traditions. In *Proceedings of the 8th International Workshop on Folk Music Analysis, 26-29 June 2018, Thessaloniki, Greece,* (pp. 78-79).

Maloy, R. (2014). Old Hispanic Chant and the Early History of Plainsong. In *Journal of the American Musicological Society 67-1* (pp. 1-76).

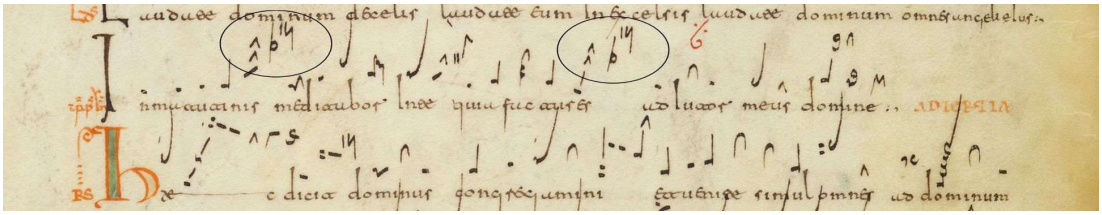
Pfisterer, A. (2002). Cantilena Romana. Untersuchungen zur Überlieferung des gregorianischen Chorals. Paderborn: Ferdinand Schöningh.

Randel, D. (1969). The Responsorial Psalmtones of the Mozarabic Office. Princeton: Princeton University Press.

ASPECTS OF MELODY GENERATION FOR THE LOST CHANT OF THE MOZARABIC RITE

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

ol-ol-ohh-oehl ol-ol-ohh-oehl



oh-ohhh-oell-ol-ol-ohl---o--oel-oehl---oh--ol--o---oh--o--ohh--ol--ohhll-oh-ol

Figure 1. Two lines from the early tenth-century León antiphoner (E-L 8, 111v6-7). At the bottom: the opening of the responsory *Haec dicit Dominus congregamini.* Below the manuscript image a representation of the neumatic notation on these first four words in contour letters. In the top line two occurrences of an *intra-opus* pattern with representation.

1. INTRODUCTION

In medieval Europe several textually and musically related monophonic liturgical chant traditions existed. Most famous is the Franco-Roman chant of the Roman rite, better known as Gregorian chant. Most other rites and traditions were abolished at some point in favor of the Roman rite and its chant (Hiley, 1993).

The Mozarabic rite existed from the end of the sixth till the end of the eleventh centuries on the Iberian peninsula. Its music (over 5000 chants) is preserved in pitch-unreadable neumatic notation. Figure 1 gives an example. The tradition was abolished in the time when pitch readable notation came in use. Therefore the intervals of most melodies are unknown. Only a handful of chants was ever found in pitch readable notation (Randel, 2001).

We have presented two computational methods of melody generation for the lost chant of the Mozarabic rite (Maessen & Conklin, 2018). To improve this generation we examined melodic aspects to be included, experimentally and in the literature (Gregoriana Amsterdam; Hiley, 1993; Troelsgård, 2014). Some aspects appear hard to quantify, for example, the meaning of the chant texts in relation to the liturgical calendar in which all chants have their specific places. Also problematic is the recent articulation of *musemes* that underline specific text passages (Lousberg, 2018). We found ten quantifiable aspects of the lost melodies that can (and should) be implemented in the generation. More aspects may emerge by using a third method, based on deep learning and neural networks.

2. QUANTIFIABLE ASPECTS

1. All generated melodies should agree with the *neumatic notation* in which the chants are preserved. The meaning of this notation (Rojo & Prado, 1929) should be represented in machine readable form. A basic way to do so is with contour letters. Each neume can be represented as a sequence of letters from the set L ={`o, h, l, e`}, the `o` representing the first note of a neume, the `h` a note higher than the previous one, the `l` a note lower, and the `e` a note of equal pitch. Figure 1 gives an illustration including Volpiano conventions to separate notes for different neumes with a dash `-`, for neumes on different syllables with two dashes `--`, and neumes on different words with three dashes `---` (Swanson, Bain, Helsen et al. 2016). However, since most neumes have several variants in their graphical forms, the representation in four letters needs improvement. Figure 1 shows e.g. three variants of the *pes* (`oh`) and three for the *clivis* (`ol`). Also, neumatic positions above the chant text may represent indications for melodic motion. The second neume (`ohhh`) in Figure 1 e.g. starts lower than the first (`oh`) and the third (`oell`) starts higher than the second. Including variants and positions in the representation could improve melody generation.

2. Some recurring patterns within single chants seem to represent the same melodic content, for example, the encircled neumes in Figure 1. For these *intra-opus patterns* melody generation should result in patterns with equal sequences of pitches (Conklin, 2010). Recurring patterns of five notes, such as `oh--ol--o`, should not always generate the same sequence of pitches. On the other hand there is a wide consensus that recurring patterns of twenty or more notes do represent the same sequence of pitches (Maloy & Hornby, 2012). Of importance also, is the precision of the representation. Representations using all pattern information should result more easily in equal pitch sequences than representations using only the letters of L.

3. Melodies generated for single chants should be constrained to a specific *range* or *ambitus.* Random melody

generation could not only exceed the limits of the human voice, but also the expected range of certain chant genres. Simple antiphons should be limited in their range, while more complex chants could have a wider range. In Gregorian chant the range of a chant depends at least on chant genre, mode, and the parts within genres. Ranges could be set manually, or trained on related traditions.

4. In some cases it will be desirable to define *specific pitches* of the generated melody beforehand, like the first and the last pitch. For some chants specific pitches may be known, as is the case with some responsory verses.

5. Melody generation should respect *cadences.* The melodic pacing of Mozarabic chant is determined by the grammatical phrases of the text and specific recurring patterns in the neumatic notation (Maloy & Hornby, 2012). These cadences should be included in the representation. Volpiano conventions make the numerals `6`, `3`, `4` and `5` respectively represent the end of a phrase, of a sentence, of a major part, and of the end of the piece. Longer melismas, also, can include cadences.

6. Several chants seem to have patterns in common. For these *inter-opus patterns* melody generation should result in equal sequences of pitches, or even better, intervals. We have similar problems here as with *intra-opus* patterns. *Inter-opus* patterns of 30 or more notes definitely should result in equal pitch sequences, but some specific patterns of only five notes should also do so. The 30-note pattern on the three opening words of the responsory in Figure 1 exists in six different chants (four responsories, a sacrificium and an alleluiaticus; E-L 8: 66r12, 94v03, 98r15, 111v07, 240r16 & 266r09) and therefore is an *inter-opus* pattern. In 458 responsories, the five-note pattern on *Dominus,* however, exists 39 times with the same neume variants on the same word, *Domin(us),* and only 14 times on other words. Unspecified, `oh--ol--o` exists 122 times. Therefore the specific pattern is a serious candidate for equal pitch sequences. *Inter-opus* patterns require the generation of a set of related melodies. Most chant traditions consist of such related melodies.

7. Several lost chants are related to chants on the same text in other traditions (Levy, 1998). For each chant, melody generation should be based on the *most related tradition* and within that tradition on the most related genre. We developed a method to find the most related tradition (Maessen & Van Kranenburg, 2018).

8. Most chants of the Franco-Roman tradition are associated with one of the eight church *modes.* Don Randel suggested the improbability of a well-defined concept of mode (or tonality) for the responsories of the Mozarabic Office (Randel, 2001). We are looking for ways to define melodic characteristics of subsets of large sets of chants, specific for those only preserved in neumatic notation.

9. Melody generation should handle *rhythm.* Equal pitch sequences with different rhythm appear to have different occurrence rates in chant. Also, there is a distinct mensuralistic interpretation for the tenth-century notation of Gregorian chant (Van Biezen, 2013). A similar interpretation can (partially) be given for Mozarabic notations.

10. Finally we consider *word accents* a quantitative aspect that should be implemented in melody generation. Word accents of medieval Latin are known and determine the melodic motion of chant (Randel, 2001).

3. CONCLUSION AND FUTURE WORK

In order to improve any method of melody generation for the lost chant of the Mozarabic rite there are at least ten quantifiable aspects, features or constraints, that should be implemented. Until now, the first five aspects have been partially implemented in our methods. We are working on the full implementation of these and the other aspects. Aspect 7 is already the subject of a publication. Currently we are focusing on aspect 8. Since September 2018 we are also experimenting with aspect 9 in performances.

4. REFERENCES

E-L 8: The León antiphoner; León, Cathedral Archive, MS 8; https://bit.ly/2KAGGrV

Conklin, D. (2010). Discovery of distinctive patterns in music. Intelligent Data Analysis, 14(5):547-554.

Gregoriana Amsterdam; www.gregoriana.nl

Hiley, D. (1993). Western Plainchant, A Handbook. Oxford: Clarendon Press.

Levy. K. (1998). Gregorian Chant and the Carolingians. Princeton: Princeton University Press.

Lousberg, L. (2018). Microtones According to Augustine - Neumes, Semiotics and Rhetoric in Romano-Frankish Liturgical Chant. PhD Thesis: Utrecht University.

Maessen, G. & Conklin, D. (2018). Two Methods to Compute Melodies for the Lost Chant of the Mozarabic Rite. In *Proceedings of the 8th International Workshop on Folk Music Analysis, Thessaloniki, Greece,* (pp. 31-34).

Maessen, G. & Van Kranenburg, P. (2018). A Non-Melodic Characteristic to Compare the Music of Medieval Chant Traditions. In *Proceedings of the 8th International Workshop on Folk Music Analysis, Thessaloniki, Greece,* (pp. 78-79).

Maloy, R. & Hornby, E. (2012). Toward a Methodology for Analyzing the Old Hispanic Responsories. In *International Musicological Society Study Group Cantus Planus, Papers read at the 16th meeting, Vienna, Austria, 2011* (pp. 242-249).

Randel, D. (2001). Mozarabic Chant, *The New Grove Dictionary of Music and Musicians (second edition).* http://www.oxfordmusiconline.com

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

Swanson, B., Bain, J., Helsen, K. Koláček, J., Lacoste, D. & Ignesti, A. (2016). Volpiano Protocols with Self-Test. http://cantus.uwaterloo.ca

Troelsgård, C. (2014). Byzantine Chant Notation - Written Documents in an Aural Tradition. *Lecture at Stanford University (California), 24 Feb 2014* (pp. 1-24).

Van Biezen, J. (2013). Rhythm, Metre and Tempo in Early Music. Diemen: AMB Publishers.

CONTENT-BASED MUSIC RETRIEVAL OF IRISH TRADITIONAL MUSIC VIA A VIRTUAL TIN WHISTLE

Pierre Beauguitte, Hung-Chuan Huang
School of Computer Science
Technological University Dublin
{pierre.beauguitte,hungchuan.huang}@mydit.ie

1. INTRODUCTION

We present a mobile phone application associating a virtual musical instrument (emulating a tin whistle) to a content-based music retrieval system for Irish Traditional Music (ITM). It performs tune recognition, following the architecture of the existing query-by-playing software Tunepal (Duggan & O'Shea, 2011). After explaining the motivation for this project in Section 2 and presenting some related work in Section 3, we describe our proposed application in Section 4. Section 5 discusses current shortcomings of our project and potential future directions.

2. MOTIVATION

Tunepal [1] (Duggan & O'Shea, 2011) has become a popular tool among practitioners of ITM, with more than 20 thousand monthly active users. It allows searching for a tune by playing a short excerpt on an instrument. Tunepal transcribes the audio into a sequence of notes, and attempts to identify the tune by finding the most similar sequence in a database of existing tune notations.

Obtaining a good transcription of the recorded audio is challenging, especially as Tunepal is often used in rather noisy environments (typically a pub where an Irish session is taking place). By offering a similar tune recognition system using a virtual instrument instead of a real one, thus requiring no audio input, we believe that our app can be a useful alternative or complement to Tunepal.

Several reasons guided our choice of the tin whistle. First, its dimensions and the simplicity of its fingering, consisting of combinations of six tone holes, make it a good fit for the limited input capabilities of a smartphone. Second, the tin whistle is among the most popular instruments in ITM (Vallely, 2011), and most practitioners have at least some rudiments of this instrument and will thus be able to use the app.

3. RELATED WORK

A number of projects have investigated the use of mobile devices as musical instruments. Essl & Lee (2017) offer a recent survey of existing projects. An early example of mobile musical instrument is the "Ocarina", released in 2008 (Wang, 2014). In addition to on-screen buttons simulating the finger holes, it uses the microphone as input so that breath controls articulation, and accelerometers so that motion controls vibrato. Our app does not aim at allowing expressivity, but merely at recognizing tunes. Hence, only the multi-touch screen is of use to us.

The websites Folk Tune Finder [2] and Musipedia [3] offer the possibility of query-by-playing using a virtual in-browser piano keyboard. Our app allows for a more portable solution, and more importantly emulates an instrument more familiar to ITM practitioners.

The SoundTracer app (Wallace & Jensenius, 2018) allows query-by-gesture. Accelerometers are used to record vertical motion of the device, by which a user imitates a pitch contour, that is then searched for within a database of automatically transcribed recordings of Norwegian folk music. The main difference with our approach is that the type of interaction chosen is not modelled on the playing of an existing instrument, and that the search space consists of automatic transcriptions of audio recordings.

4. APP STRUCTURE

This section describes the architecture of the app, dubbed "Virtual Flute", as illustrated in Figure 1.

4.1 Virtual instrument

The main interface displays 6 buttons disposed in the fashion of the tone holes of a tin whistle, as can be seen on Figure 2. A pitch-mapping following the standard fingerings is defined, and used both to record the played sequence and to play the corresponding pitch as feedback to the user. Non standard fingerings are ignored.

The form factor of modern smartphones allow for the buttons to be placed in a realistic layout. The distance between the centres of the top and bottom holes on a tin whistle was measured to be about 11 cm. A screen with ratio 16:9 and diagonal 5.5 in, found in a number of smartphone models, has a height of about 12.2 cm.

4.2 Tune recognition

A query is obtained from the touch screen interaction described above, in the form of a pitch contour. Then, the rest of the app functions in a similar manner to Tunepal

[1] https://tunepal.org

[2] https://www.folktunefinder.com

[3] https://www.musipedia.org/

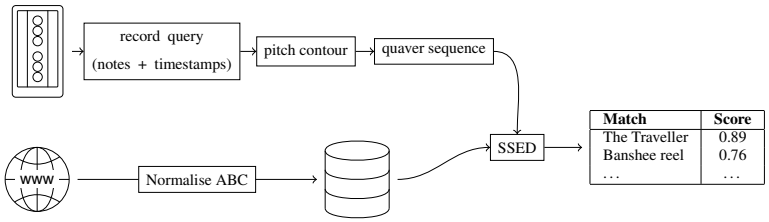


Figure 1: Architecture of the "Virtual Flute" app

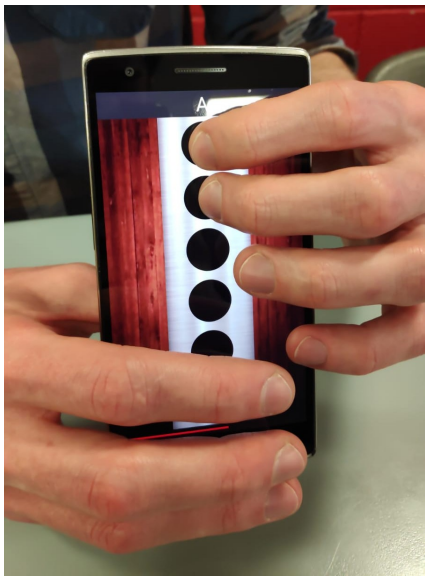


Figure 2: Photograph of a user holding the phone as a tin whistle to use the app. The text bar at the top indicates the note being played. The red bar at the bottom shows the recording progress.

(Duggan & O'Shea, 2011), itself based on the MATT2 algorithm described in (Duggan et al., 2009) which is now briefly described. The first step is to find the quaver duration q by finding the most common note duration in the recorded sequence. The pitch contour is then quantized into a quaver sequence: a note of duration d and pitch p is transformed into $round(d/q)$ quavers of note p.

A database of tunes is built from the collection from The Session.[4] The ABC notation used in this collection is normalised to sequences of quavers. The recorded query is compared, using substring edit distance (SSED), to all the tunes in the database. Results are sorted in order of ascending SSED, and the 15 closest ones are returned to the user. In its current state, the database is embedded in the app, allowing offline search. Future iterations could communicate with the servers of Tunepal or Musipedia.

[4] https://thesession.org

5. CONCLUSION AND FUTURE WORK

Although we do have a working prototype of the app, we have, as of yet, not asked for feedback from potential users. An important next step would be to follow approaches from User Centered Design (Tanaka et al., 2012).

Duggan et al. (2009) reports a retrieval accuracy of over 90% for MATT2, which is at the core of the app. Other melodic similarity measure than the SSED could be used (Janssen et al., 2017), and investigating the impact of this choice on the performance of the app would be worthwhile.

Our code is open source, and available along with a compiled APK file.[5] We intend to make it available from an app store when it reaches a further state of development.

6. REFERENCES

Duggan, B. & O'Shea, B. (2011). Tunepal: Searching a digital library of traditional music scores. *OCLC Systems & Services: International digital library perspectives*, *27*(4), 284–297.

Duggan, B., O'Shea, B., Gainza, M., & Cunningham, P. (2009). Compensating for expressiveness in queries to a content based music information retrieval system. In *International Computer Music Conference*, Montreal, Quebec, Canada.

Essl, G. & Lee, S. W. (2017). Mobile devices as musical instruments-state of the art and future prospects. In *International Symposium on Computer Music Multidisciplinary Research*, (pp. 525–539)., Porto & Matosinhos, Portugal.

Janssen, B., van Kranenburg, P., & Volk, A. (2017). Finding Occurrences of Melodic Segments in Folk Songs Employing Symbolic Similarity Measures. *Journal of New Music Research*, *46*(2), 118–134.

Tanaka, A., Parkinson, A., Settel, Z., & Tahiroglu, K. (2012). A Survey and Thematic Analysis Approach as Input to the Design of Mobile Music GUIs. In *New Interfaces for Musical Expression*, Ann Arbor, Michigan, USA.

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

Wallace, B. & Jensenius, A. R. (2018). SoundTracer: A brief project summary. Technical report, University of Oslo.

Wang, G. (2014). Ocarina: Designing the iPhone's magic flute. *Computer Music Journal*, *38*(2), 8–21.

[5] https://github.com/pierrebeauguitte/VirtualFlute

MODELLING OF LOCAL TEMPO CHANGE WITH APPLICATIONS TO LITHUANIAN TRADITIONAL SINGING

Rytis Ambrazevičius
Department of Audiovisual Arts, Kaunas University of Technology, Lithuania
Department of Ethnomusicology, Lithuanian Academy of Music and Theatre, Lithuania
rytis.ambrazevicius@ktu.lt

ABSTRACT

The present study aims to develop techniques of measurement, mathematical modelling, and evaluation of temporal irregularities (first of all, local tempo changes) in vocal performance, and to test the techniques on examples of Lithuanian traditional singing.

Methods of measurements of note durations (IOIs) in vocal performance are reviewed, their problems including the identification of perceptual attack time and adequate precision based on duration JND are discussed. Three folk song recordings are chosen for modelling of temporal irregularities. The performances are more or less *tempo giusto* so rhythm values are easily identified. Tempo curves of the chosen folk song performances are composed and analyzed: microtiming, in terms of LS/SL divisions of rhythm values, and local tempo changes in longer time spans are evaluated. Three measures of temporal unevenness are introduced; 1) the general unevenness, 2) the note-to-note unevenness, and 3) the unevenness of smoothed local tempo.

The designed model is applied to a set of 40 song recordings (10 songs from each of the 4 Lithuanian main ethnographic regions). The vocal dialects corresponding to the ethnographic regions differ noticeably in timing expressed in terms of microtiming and the three indices. Thus different combinations of the indices are characteristic of different musical dialects. This allows us to conclude that the different parameters of rhythm interpretation in vocal style can serve as more or less reliable markers of a musical dialect.

1. INTRODUCTION

There is a large number of studies on (micro)timing and research techniques, with some ethnomusicological applications (cf. Bengtsson, Gabrielsson, & Thorsén, 1969; Clarke & Cook, 2004; Danielsen, 2010; Ellis, 1991a; Friberg, Bresin, & Sundberg, 2006; Friberg, & Sundström, 2002; Gabrielsson, 1999; Ledang, 1967). Also, there is a considerable number of studies on perception of tempo changes and extraction of tempo changes from (mostly MIDI based) recordings (cf. Ellis, 1991b; Sheldon & Gregory, 1997; Dahl & Granqvist, 2003; Thomas, 2007; Müller et al., 2009; Dannenberg & Mohan, 2011; Yanagida & Yamamoto, 2017). nPVI (normalized Pairwise Variability Index) is an additional technique to eliminate the factor of local tempo changes and to study the small-scale timing phenomena (Grabe & Low, 2002). Yet, to our knowledge, there is a lack of studies on gradual tempo changes in vocal performance and their mathematical modelling.

Concerning the ethnomusicological studies on timing, they usually stop short of discussing patterns of rhythm categories (cf. Čiurlionytė, 1969; for Lithuanian sources). The microtime deviations are usually only presented as lengthening or shortening (microfermatas) of individual notes (cf. Bengtsson, 1974, p. 30; Sevåg & Sæta, 1992, p. 49; Četkauskaitė, 2007; see the examples of microfermatas in Figures 1 and 3) or short note groups (cf. Bengtsson, 1980, p. 303; Czekanowska, 1961; Ledang, 1967; Bartkowski, 1987, p. 69) in transcriptions.

2. METHODS AND PRECISION OF MEASUREMENTS

The analysis of microtiming bases on tempo curves composed from measurements of note durations (Inter-Onset-Intervals; IOIs). IOI measurements are simple and can even be automatized for keyboard performances (and, in general, in performances with short sound attacks), while vocal performances, frequently containing long and smooth attacks, pose some problems. It seems that perceptual attack time (PAT) comes somewhat earlier than the attack peak and there is no clear dependence of the lag on acoustical parameters (cf. Vos & Rasch, 1981; Gordon, 1987).

Therefore the measurements of IOIs in vocal performances can hardly be automatized or objectivized in any way. One should rely on his/her ability to detect PATs while listening to recordings and grasping the PATs from positions of the moving cursor. Fortunately, this technique is, nevertheless, precise enough compared to precision of time perception: in the cases of the best listening conditions, duration JND (just noticeable difference) can even amount 10% (Woodrow, 1951; Michon, 1964; Povel, 1981) while the readings of individual listeners usually differ not more than in 15 ms (Ambrazevičius, 2009). For crotchets and quavers performed in moderate tempo (MM=100 bpm), this correspondingly results in 2.5 and 5%. Thus precision of the IOI measurements can be considered adequate.

3. EXAMPLES OF TEMPO CURVES. LS AND SL CASES

Composition of tempo curves is well described elsewhere (cf. Clarke, 2004). Basically, all IOIs are normalized, i.e., one rhythm value (category) is chosen as a duration unit and other IOIs are recalculated. For instance, if a quaver is the unit, the duration of an individual crotchet in actual performance is divided by two; this gives the corresponding (effective) duration of the quaver. The sequences of the effective durations presented graphically constitute the tempo curves. Therefore, only (at least approximate) *tempo giusto* performances (i.e. which are characterized by identified rhythm values) can be employed for the composition of tempo curves. If listening to a performance results in ambiguous or alternative, different interpretations of rhythm and meter, the composition of tempo curves seems useless. Nevertheless, collation of the alternative tempo curves may facilitate recognition of more adequate interpretations.

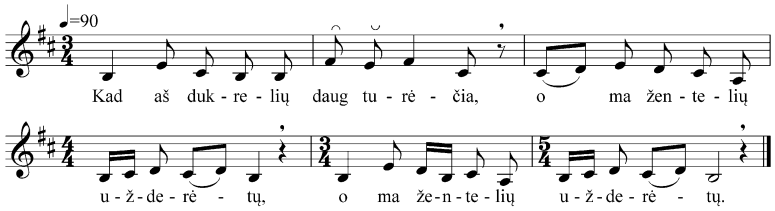


Figure 1. Transcription of the first verse of the song *Kad aš dukrelių daug turėčia* (Adelė Kazlauskienė; Gustaičiai, Prienai Dst. Recording: Četkauskaitė, 2002, N76).

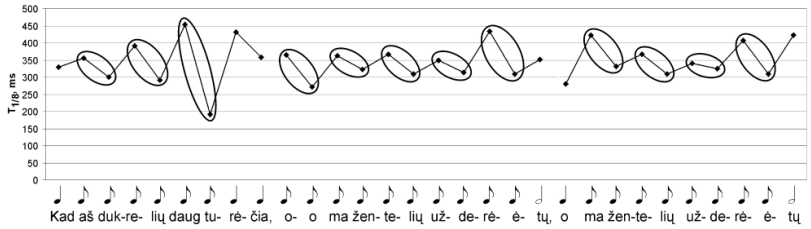


Figure 2. Example of a tempo curve (see the melody in Figure 1). Vertical axis: (effective) durations of eighth notes; LS tendency. Pairs of quavers are circled.

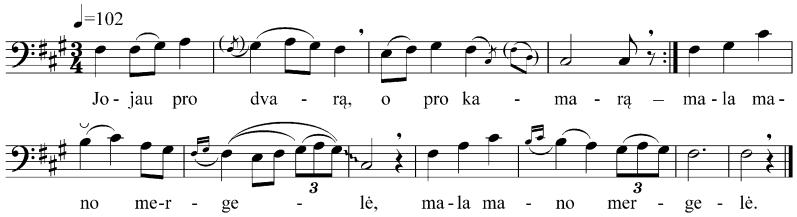


Figure 3. Transcription of the first verse of the song *Jojau pro dvarą* (Vincas Jurčikonis; Babrai, Lazdijai Dst. Recording: Četkauskaitė, 1995, N18).

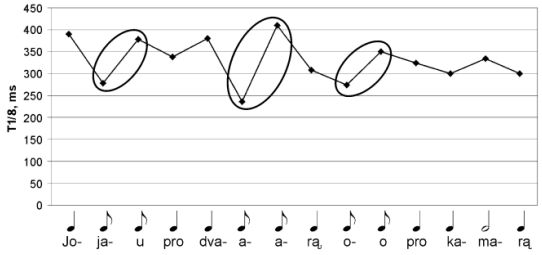


Figure 4. Example of a tempo curve (first four measures of the melody in Figure 3). Vertical axis: (effective) durations of eighth notes; SL tendency. Pairs of quavers are circled.

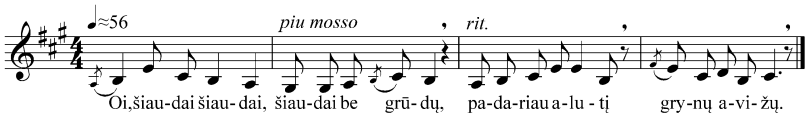


Figure 5. Transcription of the first verse of the song *Oi, šiaudai šiaudai* (Barbora Buivydaitė; Rūdaičiai, Kretinga Dst. Recording: Baranauskienė *etal*, 2006, N6).

Figures 2 and 4 show typical tempo curves composed from the IOI measurements of two Lithuanian traditional vocal *tempo giusto* performances (Figures 1 and 3). The structural notes were considered; durations of embellishments (appoggiaturas, etc.) were incorporated into the corresponding structural notes. Only three notes in the transcriptions are supplemented with microfermata marks (see the syllables *daug* and *tu-* in Figure 1, and the first *-no* in Figure 3). Consequently, only for three notes is the prolonging or shortening of the rhythm values clearly perceived. Yet a significant fluctuation of the durations is observed in the tempo curves; the performances present two opposite cases of *inégales*. Figure 2 shows a clear LS tendency (long-short division of crotchet into two quavers) whereas Figure 4 shows a reverse SL tendency; the performance is characterized by a somewhat "limping" rhythm. The median of *T*1/*T*2 ratios (ratios of quaver durations forming one crotchet) for the song *Kad aš dukrelių daug turėčia* equals 1.23 and the interquartile encompass the range 1.17-1.34. For the song *Jojau pro dvarą*, the median is 0.72 and the interquartile is 0.62-0.78.

4. MODELLING OF GRADUAL TEMPO CHANGE

The third song (Figure 5) is chosen for modelling of characteristics of temporal unevenness in longer time spans of a performance. Three measures of temporal unevenness are introduced; 1) the general unevenness ("general rubato index", R_{AAD}), 2) the note-to-note unevenness (the "nPVI rubato index" provided by nPVI technique, R_{nPVI}), and 3) the unevenness of smoothed local tempo ("tempo change index", TV). For the first measure, average absolute deviation of duration is used (Figure 6), instead of previously used standard deviation (Ambrazevičius, 2009; 2018); this facilitates better compatibility with nPVI measures as they also apply absolute deviation. R_{AAD} is evaluated as *AAD* normalized to the mean duration:

$$R_{AAD} = \frac{1}{n\bar{T}}\sum_{k=1}^{n}|T_k - \bar{T}|\,; \qquad (1)$$

where T_k is the duration of *k*th note, $\bar{T}$ is the mean duration, and n is the number of structural notes in the melody contour. For the examined performance, R_{AAD} = 0.122, i.e. the actual 1/8-durations deviate in 12.2% from the mean, on the average.

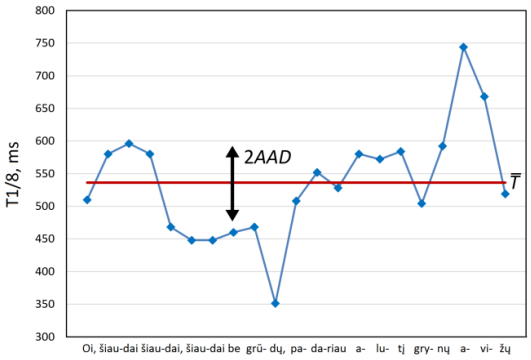


Figure 6. Example of tempo curve (see the melody in Figure 5). Vertical axis: (effective) durations of eighth notes. The horizontal line depicts average duration of eighth note.

However, if one needs to eliminate relatively slow gradual change of tempo and to evaluate note-to-note unevenness (the "jaggedness" of tempo curve), R_{nPVI} is applied instead. This index reflects average deviation from the changing local average of duration (average of two adjacent note durations; Figure 7). Grabe & Low (2002) introduced the "normalized Pairwise Variability Index" (*nPVI*):

$$nPVI = 100 \times \left[\sum_{k=1}^{n-1}\left|\frac{T_k - T_{k+1}}{(T_k + T_{k+1})/2}\right| / (n-1)\right]. \qquad (2)$$

Then the expression for R_{nPVI} follows:

$$R_{nPVI} = \frac{1}{n-1}\sum_{k=1}^{n-1}\frac{|T_k - T_{k+1}|}{T_k + T_{k+1}}. \qquad (3)$$

For the examined performance, R_{nPVI} = 0.057, i.e. the actual 1/8-durations deviate in 5.7% from the changing local mean, on the average. Naturally, $R_{nPVI} < R_{AAD}$ (in general; not only for the analyzed particular piece).

First approximation of tempo change in longer time spans can be made by substitution of duration in R_{nPVI} with moving duration average (period = 2; Figure 7):

$$TV_{(d)} = \frac{1}{n-2}\sum_{k=1}^{n-2}\frac{|T_{k+2} - T_k|}{\frac{T_k}{2} + T_{k+1} + \frac{T_{k+2}}{2}}. \qquad (4)$$

However, from the viewpoint of perception, interpretation of continuous tempo change (Figure 8) seems to be more adequate than interpretation of discrete

tempo change, in smaller or larger steps. (Of course, this statement is not valid for cases of sudden tempo changes, e.g. when tempo is clearly different in two structural parts of melody.) Following this approach, the "tempo change index" is designed again based on the nPVI logic; only summation of discrete IOI values in "nPVI rubato index" is substituted with integration of IOI equivalents in the continuous smoothed curve of local tempo change:

$$TV_{(c)} = \frac{1}{2(n-1)} \int_{t_i}^{t_f} \frac{|dT(t)|}{T(t)}. \quad (5)$$

After some mathematical procedures we get

$$TV_{(c)} = \frac{1}{n-1} \left[ln \frac{\prod T_{max}}{\prod T_{min}} \pm \frac{lnT_i}{2} \pm \frac{lnT_f}{2} \right]; \quad (6)$$

here T_i and T_f stand for the initial and final values ($T(t_i)$and $T(t_f)$) of function $T(t)$, and T_{max} and T_{min} are its (local) maximum and minimum values. Plus and minus signs in the expression are applied if T_i, T_f present, correspondingly, local maxima or minima. For instance, for the examined performance, minus signs are applied for both $lnT_i/2$ and $lnT_f/2$. For this performance, $TV_{(c)} = 0.032$, i.e. the local tempo changes in 3.2% from note to note, on the average. The performance example chosen here for modelling is characteristic of large local tempo change. Later we will see that usually tempo changes to a lesser extent (see the values for *TV* in Figure 10).

Certainly, the chosen technique of smoothing affects the values in $TV_{(c)}$ expression. Nevertheless, the discrepancies are not significant. For instance, if even an alternative technique of smoothing gives a descending line from the first T_{max} to the second T_{min} in Figure 8 (no first T_{min} and second T_{max}), $TV_{(c)} = 0.0305$ instead of 0.032. Thus the technique is still adequate for generalized evaluations.

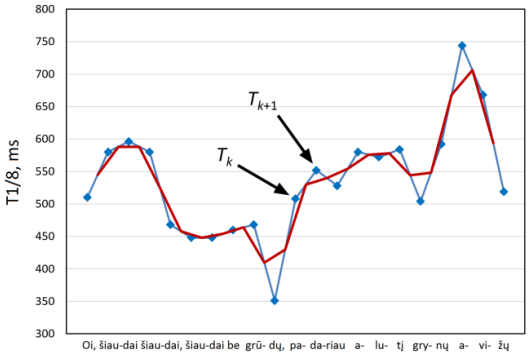


Figure 7. Example of tempo curve, as in Figure 6. The red line depicts changing average duration of an eighth note (average of two adjacent note durations).

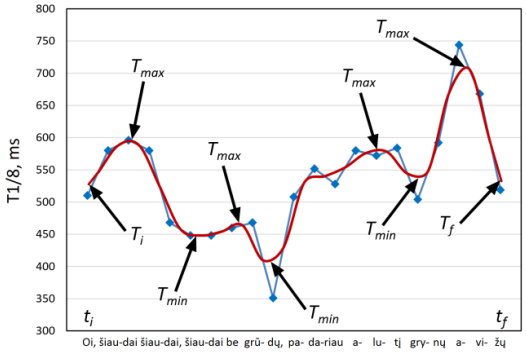


Figure 8. Example of tempo curve as in Figure 6. The smoothed line depicts local change of duration of an eighth note. Characteristic durations for calculation of $TV_{(c)}$ ("tempo change index") are marked.

5. TECHNIQUE TESTING ON SETS OF EXAMPLES OF LITHUANIAN TRADITIONAL SINGING

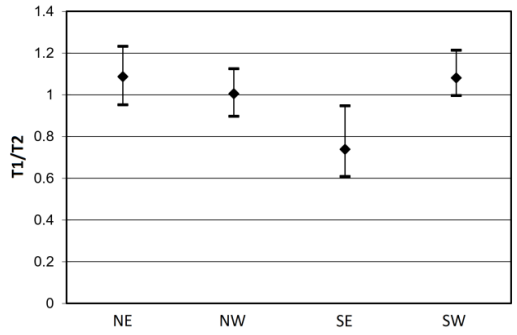


Figure 9. Generalized *T*1/*T*2 values for the samples representing four Lithuanian musical dialects. Diamonds and vertical lines mark the medians and interquartiles.

The designed model is applied to the set of 40 song recordings (10 songs from each of 4 Lithuanian main ethnographic regions; Aukštaitija, Dzūkija, Suvalkija, and Žemaitija; NE, SE, SW, and NW, correspondingly) used in a previous study (Ambrazevičius, 2018). The vocal dialects corresponding to the ethnographic regions differ noticeably in timing expressed in terms of microtiming and the three indices (Figures 9 and 10). The mean *T*1/*T*2 ratios range from 1.09 (NE) to .77 (SE) ($p_{SE\text{-}NE}$, $p_{SE\text{-}NW}$, $p_{SE\text{-}SW}$ < .001, $p_{NE\text{-}NW}$ = .035, $p_{SW\text{-}NW}$ = .026), mean R_{AAD} values range from .14 (SE) to .09 (SW), R_{nPVI} values range from .08 (SW) to .12 (SE), and *TV* values range from .005 (SW) to .013 (NW).

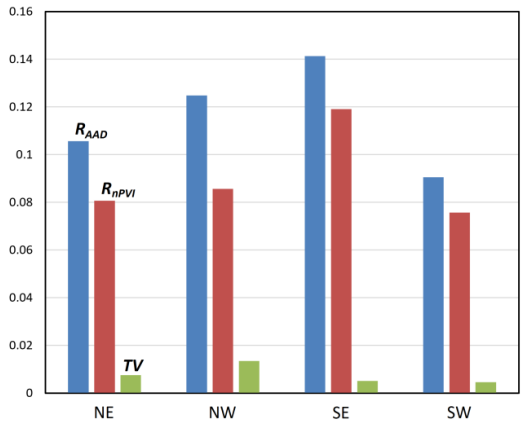


Figure 9. Generalized R_{AAD}, R_{nPVI}, and *TV* values (averages) for the samples representing four Lithuanian musical dialects.

Thus different combinations of the indices are characteristic of different musical dialects. For instance, large note-to-note rhythm unevenness combines with quite negligible changes in overall tempo in SE, whereas noticeably less note-to-note rhythm unevenness combines with large changes in overall tempo in NW ($p_{SE\text{-}NW} < .001$). This allows us to conclude that the different parameters of rhythm interpretation in vocal style can serve as more or less reliable markers of a musical dialect. The positive correlation between R_{nPVI} values in spoken and vocal dialects found in Ambrazevičius, 2018, shows the impact of linguistic rhythm onto musical rhythm.

6. DISCUSSION

One of the most important results of the present study is the proposed and derived index for evaluation of gradual tempo changeability (*TV*). Together with the other two indices (R_{AAD} and R_{nPVI}) and *T*1/*T*2, it constitutes a system of quantitative parameters for description of general temporal characteristics of a musical performance. It is demonstrated how the proposed system can be applied for revelation of stylistic traits of a performance and, in turn, how different styles can be compared, in terms of temporal performances.

Further the model could be developed in several directions. First, different techniques of tempo curve smoothening could be examined, the most adequate ones could be identified, and techniques for automated extraction of *TV* could be derived. Second, the model could be extended for the analysis of the phenomena characteristic of intermediate time spans (basic rhythm values, measures, etc.). For instance, in certain cases, it would be interesting to know whether the temporal movement is based on the basic rhythm values or rather on their subdivisions (e.g. whether the movement in crotchets is more stable than the movement in quavers). Third, the developed model could be applied for the evaluation of stability of performance repetitions (e.g. similarity of melostrophes). Fourth, the model could be modified for the study of temporally more complicated (non 'tempo giusto') performances.

Finally, the model could be applied not only for the study of traditional vocal performance, but also for other vocal styles and instrumental music.

7. REFERENCES

Ambrazevičius, R. (2009). Rhythm interpretation in vocal performance: Lithuanian examples. In *XXV ESEM. 'Performance'. Abstracts* (p. 10). Milton Keynes, UK: The Open University.

Ambrazevičius, R. (2018). Aspects of timing in Lithuanian traditional singing. In R. Parncutt & S. Sattmann (Eds.), *ICMPC15/ESCOM10: Abstract book (electronic)* (p. 34). Graz, Austria: Centre for Systematic Musicology, University of Graz.

Baranauskienė, V., Zakarienė, V., Juciutė, L., & Ivanauskaitė, V., eds. (2006). *Dainų karaliai ir karalienės. Barbora Buivydaitė*. Vilnius: Lietuvos muzikos ir teatro akademija.

Bartkowski, B. (1987). *Polskie śpiewy religijne w żywej tradicji. Style i formy*. Kraków: Polskie Wydawnictwo Muzyczne.

Bengtsson, I. (1974). On notation of time, signature and rhythm in Swedish polskas. In G. Hilleström (Ed.), *Studium instrumentorum musicae popularis. B. III*. (pp. 22-31). Stockholm: Nordiska Musikförlaget.

Bengtsson, I. (1980). *Folkmusikboken*. Stockholm: Bokförlaget Prisma.

Bengtsson, I., Gabrielsson, A., & Thorsén, S. M. (1969). Empirisk rytmforskning. *Swedish Journal of Musicology, 51*, 49-118.

Četkauskaitė, G., ed. (1995). *Lietuvių liaudies muzika*. Vilnius: 33 Records.

Četkauskaitė, G., ed. (2002). *Lietuvių liaudies muzika. V. 3. Suvalkiečių dainos. Pietvakarių Lietuva*. Vilnius: Lietuvos muzikos akademija.

Četkauskaitė, G., ed. (2007). *Lietuvių liaudies dainų antologija*. Vilnius: Lietuvos muzikos ir teatro akademija.

Čiurlionytė, J. (1969). *Lietuvių liaudies dainų melodikos bruožai*. Vilnius: Vaga.

Clarke, E. F. (2004). Empirical methods in the study of performance. In E. Clarke & N. Cook (Eds.), *Empirical musicology. Aims, methods, prospects* (pp. 77–102). New York: Oxford University Press.

Czekanowska, A. (1961). *Pieśni biłgorajskie*. Wrocław. Polskie Towarzystwo Ludoznawcze.

Dahl, S., & Granqvist, S. (2003). Looking at perception of continuous tempo drift - a new method for estimating internal drift and just noticeable difference. In R. Bresin (Ed.), *Proceedings of the Stockholm Music Acoustics Conference* (pp. 595-598). Stockholm: KTH Speech, Music and Hearing.

Danielsen, A., ed. (2010). *Musical rhythm in the age of digital reproduction*. Surrey, Burlington: Ashgate.

Dannenberg, R. B., & Mohan, S. (2011). Characterizing tempo change in musical performances. In *Proceedings of the International Computer Music Conference* (pp. 650-656). Huddersfield: University of Huddersfield.

Ellis, M. C. (1991). An analysis of "swing" subdivision and asynchronization in three jazz saxophonists. *Perceptual and Motor Skills*, 73, 707-713.

Ellis, M. C. (1991). Research note. Thresholds for detecting tempo change. *Psychology of Music, 19*, 164-169.

Friberg, A., Bresin, R., & Sundberg, J. (2006). Overview of the KTH rule system for musical performance. *Advances in Cognitive Psychology, 2*, 145-161.

Friberg, A., & Sundström, A. (2002). Swing ratios and ensemble timing in jazz performance: Evidence for a common rhythmic pattern. *Music Perception*, 19, 333–349.

Gabrielsson, A. (1999). The performance of music. In D. Deutsch (Ed.). *Psychology of music* (2nd ed.; pp. 501-602). San Diego, London: Academic Press.

Gordon, J. W. (1987). The perceptual attack time of musical tones. *Journal of the Acoustical Society of America, 82*, 88-105.

Grabe, E., & Low, E. (2002). Durational variability in speech and the rhythm class hypothesis. In C. Gussenhoven & N. Warner (Eds.), *Papers in Laboratory Phonology*, 7. Cambridge: Cambridge University Press.

Ledang, O. K. (1967). *Song syngemåte og stemmekarakter*. Oslo: Universitetsforlaget.

Michon, J. A. (1964). Studies on subjective duration: I. Differential sensitivity in the perception of repeated temporal intervals. *Acta Psychologica, 22*, 441-450.

Müller, M., Konz, V., Scharfstein, A., Ewert, S., & Clausen, M. (2009). Towards automated extraction of tempo parameters from expressive music recordings. Retrieved from: https://www.researchgate.net/publication/47863781_Towards_Automated_Extraction_of_Tempo_Parameters_from_Expressive_Music_Recordings

Povel, D.-J. (1981). Internal representation of simple temporal patterns. *Journal of Experimental Psychology: Human Perception and Performance, 7*, 3-18.

Sevåg, R., & Sæta, O. (1992). *Norsk Folkemusikk. Serie II. Slåtter for vanlig fele. B. I. Oppland*. Oslo: Universitetsforlaget.

Sheldon, D. A., & Gregory, D. (1997). Perception of tempo modulation by listeners of different levels of educational experience. *Journal of Research in Music Education, 3*, 367-379.

Thomas, K. (2007). Just noticeable difference and tempo change. *Journal of Scientific Psychology, 2*, 14-20.

Vos, J., & Rasch, R. (1981). The perceptual onset of musical tones. *Perception & Psychophysics, 29*, 323-335.

Woodrow, H. (1951). Time perception. In S. S. Stevens (Ed.). *Handbook of experimental psychology* (pp. 1224-1236). New York: Wiley.

Yanagida, M., & Yamamoto, S. (2017). Effects of the mode of tempo change on perception of tempo change. *Proceedings of Meetings on Acoustics, 28*, (1-10).

Towards Singing Perception Universals

Polina Proutskova
Queen Mary, University of London
proutskova@googlemail.com

1. INTRODUCTION

How do we talk about singing? When describing a difference between two performers singing the same song, we may mention the mood they create, their skill, list musical choices they make (tempo, dynamics, etc.). Yet when we try to characterise what they sound like, we are left to invoking metaphors and comparisons to other areas of life. It seems that there is no widely understood vocabulary about vocal production in singing, not even within an established and theorised tradition such as Western classical music (Proutskova, 2018; Garnier et al., 2007; McGlashan, 2013). Or is there? If we learn that someone's singing is somewhat husky or really twangy, wouldn't we understand similarly? Are there any terms we would interpret in the same way?

2. DEFINITION AND IMPACT

I would like to introduce a notion of a *singing perception universal* – a descriptor of vocal production that would be understood similarly by a large group of people. In contrast to musical universals (Brown and Jordania, 2013; Harwood, 1976), *singing perception universals* are not about being present or absent in a large number of cultures; rather they are about being recognised and interpreted similarly e.g. by people coming from different cultures. The advantage of this approach is that it does not rely on a "correct" description of a vocalisation; it merely measures the agreement between listeners' judgements.

Singing is the most widespread artistic activity (e.g. 37 million choral singers in Europe, [Bartel and Cooper, 2015]) yet it is one that is less well understood. The lack of vocabulary impedes further research. In MIR, research on singing recordings lags far behind other types of data, due to lack of annotations. But which characteristics should be annotated in absence of commonly understood vocabulary? With the advance of deep learning, there is a possibility that automatic classification and computer generation of vocals would soon emerge from the leading digital music content holders. Yet their models will be biased in the same way as their data: if they possess pop music alone, only pop music will be generated. In order to represent the whole variety of human vocalisation in AI models, we should be looking for ways to capture that variety. Understanding how we understand singing would help us to grasp how a consensus is formed about matters which are considered difficult to verbalise, having a "mystery" about them. Also, in previous studies vocal production has been linked to societal traits such as status of women in a society. Such hypotheses cannot be independently validated without a more rigorous approach to semantic formalisation.

3. PREVIOUS WORK, PLAUSIBILITY

The motivation to this strand of investigation comes from my PhD in which I intended to formalise the language about singing on the basis of objective physiological traits (Proutskova, 2018). In my investigative study 13 experts annotated singing samples from a wide range of cultures using my ontology of vocal production: there was a tendency to agreement for only 2 out of 19 dimensions they were rating. This result indicates that experts cannot generally hear physiology underlying vocal sounds. Yet there were two descriptors about which they agreed – *larynx height* and *narrowness of the vocal tract* (often referred to by its main vocal function – the *twang*). These two descriptors could possibly be easier to agree about for vocal production experts and therefore constitute *singing perception universals* for this group, though further studies with more participants would have to confirm that.

While there is little research on cross-cultural vocal production, one experiment stands out – the Cantometrics project. Alan Lomax and Victor Grauer designed a parametrisation system for cross-cultural singing performance practice, consisting of 36 parameters, some of which reflect vocal production (Lomax, 1977). They chose characteristics for a rough subjective description of a singing sample, which were self-explanatory or easy to teach to non-experts: volume, rasp, nasality, accent, enunciation, vocal blend, glottal shake, etc. In their experiment they had 5000 singing recordings rated by three raters each, and the agreement between raters was good.

Why was the Cantometrics approach more successful in finding consensus between human listeners than my study? Cantometrics used subjective, perceptual descriptors rather than physiological ones. That allowed non-specialists, not familiar with vocal physiology, to become raters. In fact research in voice pathology confirms that experts tend to disagree more than amateurs when analysing voices (Kreiman et al., 1993).

For example, in Cantometrics, *vocal width* is the dimension stretching between a wide, open, relaxed and reso-

nant sound on the one hand, and on the other hand a narrow, tense squeezed singing. I demonstrated that this definition from the physiological point of view was flawed, lumping together three distinct dimensions of the vocal sound which are ambiguous and not well correlated: narrow/wide, constricted/relaxed and more resonant/less resonant (Proutskova, 2018). Then how was it possible that Lomax found his raters agreeing about *vocal width* while I observed no agreement on its dimensions in my study? Was it due to the fact that a general look at a wider phenomenon, without going into much detail, allows for a better consensus? Or did it reflect the extensive training procedure that all Cantometrics raters were subject to?

Related to Cantometrics *vocal width* is the notion of *open throat*, which is widely referred to in Western countries, associating open throat vocalisation with a big, relaxed and aesthetically pleasing sound, whereas the absence of open throat is often considered as incorrect singing. Mitchell et al. (2003) showed that singing teachers project different meanings on the term. Yet the term persists – could that be an indication of a general consensus about a more rough definition of an *open throat*?

4. METHOD

I suggest to approach the choice of possible candidates for *singing perception universals* from two different angles. Firstly, the terms which in previous studies were shown to lead to agreement among raters should be independently investigated: the Cantometrics descriptors (Lomax, 1977), *larynx height* and *twang* from my study (Proutskova, 2018), widely used terms from established vocal schools (see e.g. Granier, 2007, for an analysis of Western classical vocabulary). For these terms an experiment in the form of an online game should be developed, collecting participants' ratings alongside their cultural background and singing proficiency. The agreement can then be examined for a variety of demographic groups.

Secondly, a similar online game approach can be used for an agnostic study, in which semantic dimensions are not pre-defined. In contrast to absolute ratings on a scale collected in the first case, here participants will be asked to provide relative similarity estimations of singing samples through odd-one-out choice (triadic comparisons, see Farrugia et al., 2016). Multidimensional scaling (Weller and Romney, 1988) can then be used to extract the most salient factors accounting for dissimilarity of singing samples. These factors would be the perfect candidates for *singing perception universals*. The advantage of this approach is that no prior knowledge of semantic categories is assumed and no training of participants is required. The disadvantage lies in the fact that many more ratings will be required.

Some study design questions will have to be examined beforehand to achieve transparency and replicability: e.g. whether raters should be pre-trained for the semantic categories they are supposed to rate. The training can anchor them to the same scale (zero and extremes) and thus boost agreement. Yet over-training can lead to "overfitting".

5. CONCLUSION

Should *singing perception universals* be found, they might prove to be the key to analysing vocal style and its change in time and space, the geographic spread may provide insights into human migration and cultural evolution (Grauer 2006); they will also inform our singing perception models. If, on the other hand, no such characteristics emerge, vocal production would make a perfect test case for automated methods to surpass human classification ability; and the singing voice would retain the aura of mystery continuing to enthrall us with its versatility and expressivity, finding its way directly to our hearts.

6. REFERENCES

Bartel, R. and Cooper, C. F. (2015). Singing Europe. Technical report, European Choral Association.

Brown, S. and Jordania, J. (2013). Universals in the world's musics. *Psychology of Music*, 41(2):229–248.

Farrugia, N., Allan, H., Müllensiefen, D., and Avron, A. (2016). Does it sound like progressive rock? a perceptual approach to a complex genre. In Gonin, P., editor, *Prog Rock in Europe: Overview of a persistent musical style*, pages 197–212. Editions Universitaire, Dijon.

Garnier, M., Henrich, N., Castellengo, M., Sotiropoulos, D., and Dubois, D. (2007). Characterisation of voice quality in western lyrical singing: from teachers' judgements to acoustic descrip- tions. *Journal of interdisciplinary music studies*, 1(2):62–91.

Grauer, V. A. (2006). Echoes of our forgotten ancestors. The World Of Music, 48(2).

Harwood, D. L. (1976). Universals in music: A perspective from cognitive psychology. *Ethnomusicology*, 20(3):521–533.

Kreiman, J., Gerratt, B. R., Kempster, G. B., Erman, A., and Berke, G. S. (1993). Perceptual evaluation of voice quality: review, tutorial, and a framework for future research. *Journal of Speech, Language, and Hearing Research*, 36(1):21–40.

Lomax, A. (1977). Cantometrics: A Method of Musical Anthropology (audio-cassettes and handbook). Berkeley: University of California Media Extension Center.

McGlashan, J. (2013). What descriptors do singing teachers use to describe sound ex- amples? *Presented at PEVOC 10 (Pan-European Voice Conference)* Prague, Czeck Republic.

Mitchell, H. F., T. Kenny, D., Ryan, M., and Davis, P. J. (2003). Defining 'open throat' through content analysis of experts' pedagogical practices. *Logopedics Phoniatrics Vocology*, 28(4):167–180.

Proutskova, P. (2018). Investigating the Singing Voice: Quantitative and Qualitative Approaches to Studying Cross-Cultural Vocal Production. PhD thesis, Goldsmiths University of London.

Weller, S. C. and Romney, A. K. (1988). Systematic data collection, volume 10. Sage publications.

AUTOMATIC COMPARISON OF HUMAN MUSIC, SPEECH, AND BIRD SONG SUGGESTS UNIQUENESS OF HUMAN SCALES

Jiei Kuroyanagi*[1], Shoichiro Sato*[1], Meng-Jou Ho[1], Gakuto Chiba[1], Joren Six[2], Peter Pfordresher[3], Adam Tierney[4], Shinya Fujii[1], Patrick E. Savage[1]**

[1]Keio University, Japan, [2]Ghent University, Belgium, [3]University at Buffalo, NY, USA, [4]Birbeck, University of London, UK

*Equal contribution, **Correspondence to: psavage@sfc.keio.ac.jp

ABSTRACT

The uniqueness of human music relative to speech and animal song has been extensively debated, but rarely directly measured. We applied an automated scale analysis algorithm to a sample of 86 recordings of human music, human speech, and bird songs from around the world. We found that human music throughout the world uniquely emphasized scales with small-integer frequency ratios, particularly a perfect 5th (3:2 ratio), while human speech and bird song showed no clear evidence of consistent scale-like tunings. We speculate that the uniquely human tendency toward scales with small-integer ratios may relate to the evolution of synchronized group performance among humans.

1.BACKGROUND

The origins of music and language have been debated for centuries. Both music and language are human universals found in all known societies, but language appears to be unique to humans while music has many parallels in non-human species such as songbirds and whales [6]. Why might this be, and what - if anything - is unique about human music?

Comparative analyses of music have produced conflicting results. Some studies argue that certain aspects of music like simple scales and rhythms are unique to human music and may have evolved to bond people together [19, 20]. Others argue against such ideas on the grounds that such features are not cross-culturally universal [5, 15] or that they are not specific to human music, but instead a byproduct of more general constraints on acoustic perception and production that are shared with speech and/or animal song [23].

The degree to which music and language share similar features has seen vigorous debate in the recent literature [9, 17]. An improved understanding of such boundaries requires empirical research to better understand their similarities and differences. However, while many have compared speech vs. song, human song vs. bird song, and human language vs. bird song [9, 10, 17, 21, 23], we know of only one previous empirical study that has simultaneously compared musical aspects of human music, human speech, and non-human vocalizations [24]. We decided to use automated scale analysis software to analyse and compared global samples of human music, human speech, and bird song. Because definitions of music, language, and animal song are controversial, we did not define and collect samples ourselves but used pre-existing databases (see Methods for details).

Previously, we used automated scale analysis software to demonstrate a strong cross-cultural tendency for human music to use scales containing pitches separated by intervals that approximate simple integers, particularly a perfect 5th (e.g., 700 cents, ~ 3:2 frequency ratio) [8]. If perfect fifths also predominate in bird song or human speech as well as human music, then they are likely a consequence of perceptual/motor constraints, whereas if they are specific to human music, this suggests that they could be an adaptation specifically for human music [23].

2.METHOD

2.1. Audio Samples

For this preliminary analysis, we aimed to assemble globally distributed samples of approximately 30 recordings each of 1) human music, 2) human speech, and 3) bird song. We were only able to identify 26 recordings of human speech, giving a total sample of 86 recordings (Fig. 1, Table 1). For all samples, only monophonic recordings were used to enable accurate automatic transcription.

(1) 30 music recordings from nine regions were obtained from the *Garland Encyclopedia of World Music* [16]. The audio files were recorded in diverse regions, covering a diverse mix of traditional genres (e.g., healing, love, religious). From the 124 monophonic Garland recordings assessed as usable, we randomly selected 3-4 recordings for this preliminary analysis from each of the following 9 regions designated by *Garland's* editors: Africa, South America, North America, Southeast Asia, South Asia, Middle East, East Asia, Europe, and Oceania. Our sample included both instrumental and vocal music.

(2) We obtained 26 recordings of human speech from the Linguistic Data Consortium [12] spoken language sampler. Because we were unable to locate 30 or more samples, we used all available samples without sampling equally by region as we did for the human music samples. The samples mainly consist of recorded telephone conversations. Each sample was edited to

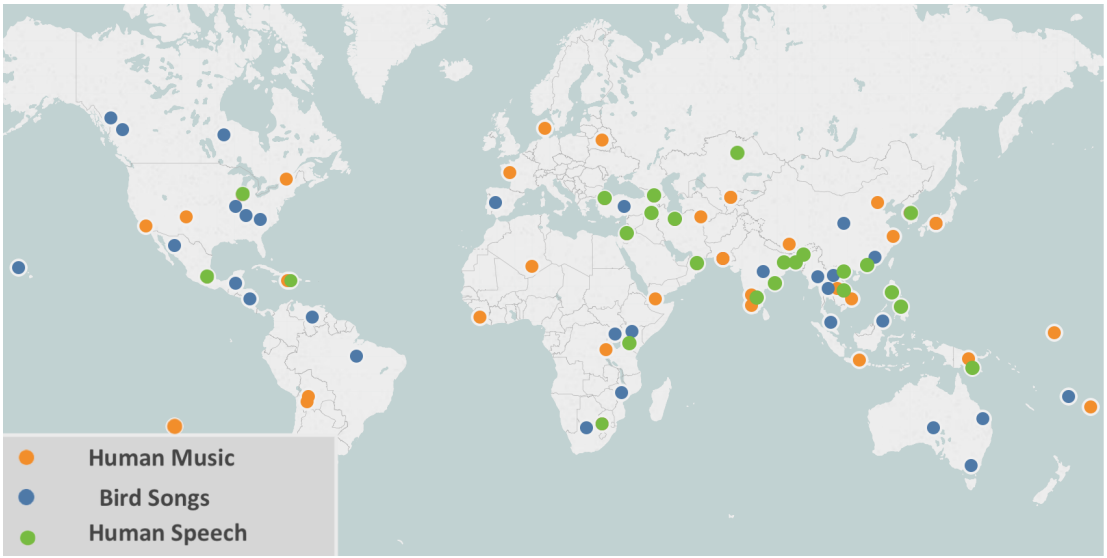


Figure 1. A map of the 86 recordings analysed. Human music (n=30), bird songs (n=30) and human speech (n=26) are respectively coloured orange, blue and green.

Table 1. Metadata about musical genre, bird species, or language name for the 86 recordings analysed. See original publications for additional metadata [8, 11, 18].

Human Music	Bird Species	Human Speech
Inanga Chuchotee (Whispered Inanga)	*Cinclus mexicanus*	American English
Tuareg Tihadanaren	*Icterus galbula*	Georgian
Somali caayar "dhaanto," excerpt 1	*Poecile atricapillus*	Haitian
"Kulwa" (Kupla, cobla)	*Monarcha melanopsis*	Cebuano
Lichiwayu notch flutes	*Tchagra senegala*	Kazakh
Rara instrumental music	*Nectarinia kilimensis*	Levantine Arabic
"La Finada Pablita"	*Prunella fulvescens*	Malto
"Sabá Medley"	*Nilaus afer*	Mexican Spanish
"El Pájaro Verde" ("The Green Bird")	*Catherpes mexicanus*	Pashto
Chinese-Thai sizhu ensemble piece "Chung we meng" ("Moon Shining Brightly in the Spring")	*Bombycilla cedrorum*	Persian
Jarai gong ensemble with song "Yong Thoach" ("Brother Thoach, Please Come Back")	*Dendroica pensylvanica*	Tagalog
East Javanese "Srempeg, pelog patet wolu"	*Psophodes occidentalis*	Gulf Arabic
Rgvedic recitation by Nambudiri Brahmins	*Acridotheres tristis*	South Korean
South Indian devotional kriti, "Girirājasuta" 'Son of the mountain king's daughter, In rāga bangāla, desdi tāla	*Aegithina tiphia*	Japanese
Benjo 'keyed zither' perfoemance in Balochistan	*Galerida cristata*	Swahili
Persian narrative song, Sayyed Mohammad Khan	*Aethopyga siparaja*	Lao
Uzbek classical song, Sarabaxi Ōröm-I Jōn 'Peace of the Soul' in maqām dugāh (or dugōh), saraxbōr (4/4) rythm	*Sturnella magna*	Bengali
Uzbek classical instrumental dance piece, Oynasin Dugah	*Spizella pusilla*	Assamese
Jewish-Yemenite liturgical、The song of the Sea	*Dicaeum ignipectus*	Cantonese
"Moonlight On The Ching Yang (Xunyang) River"	*Gymnomyza viridis*	Tamil
"The Revenge" (Lyrical area from Qi yaun bao)	*Regulus satrapa*	Telgu
"Song Foe Reparing Water Channels in Barley Fields"	*Pachycephala pectoralis*	Tok Pisin
Nozakimura	*Eminia lepida*	Turkish
Denmark: sailor's lovesong "Sode pige, du er sa laught fra mig" (Dear girl, I am so far away')	*Vireolanius pulchellus*	Vietnamese
France: male solo song, "Mon pére a fait faire un étang"	*Garrulax canorus*	Zulu
Belarus: polyphonic harvest song	*Oriolus kundoo*	Kurmanji Kurdish
Rapa Nui string-figure song (pāta' uta' u)	*Moho braccatus Mimus*	
Tongan formal dance-song (lakalaka)	*Chloropsis cyanopogon*	
"Anawa anawa" Kiribati women's hip-shaking dance (kabuti)	*icrurus remifer*	
Usarufa blood-song(naa-imá)	*Ramphocaenus melanurus*	

approximately 5 to 10 seconds in order to capture only a single speaker, and noises that could affect the result of the analysis were removed (using Logic's Noise Gate function). The edited recordings were then slowed down by a factor of five to avoid under-sampling the rapidly changing pitch (in the future we aim to modify our software to allow for increasing the sampling rate to resolve this issue without having to manually slow down recordings).

(3) We obtained 30 bird song recordings by selecting a subset of recordings without considerable noise from 80 previously analysed recordings of taxonomically diverse songbirds [23 One of the limitations of our

automatic analysis software is that polyphonic melodies or audio with considerable noise cannot be properly extracted. Moreover, exceedingly high pitches or short sounds are unanalysable. Thus, human speech and bird songs had to be slowed down (by 5x) and the bird songs had to be transposed two octaves lower and have noise removed. In the future we aim to modify our software to allow for increasing the sampling rate, noise filtering, and transposing to resolve this issue without having to manually edit recordings.

2.2. Pitch Class Histogram

We used Tarsos [22] to extract and compare musical scales, since it was designed for automatic quantitative analysis of any music from around the world. Thus it allows us to analyse audio data in a way that allows comparison of frequency ratio relationships across cultures and even species. Most of the analysis is executed using pitch histograms and octave reduced pitch class histograms. Tarsos first extracts the pitch histogram, then combines this pitch histogram across octaves to a pitch class histogram which is expressed in cents [5] ranging from 0-1200. We used Tarsos's default YIN pitch estimation algorithm [3]. In the future, we plan to explore the effects of using pitch histograms without assuming octave equivalence, and of using newer algorithms such as pYIN [14] and CREPE [11]. However, we note that such algorithms contain additional assumptions that may not be appropriate for cross-cultural/cross-species analyses.

2.3. Normalisation and comparison of averaged pitch class histograms

Normalisation is required to compare scales between different songs that have different keys or tuning systems. If we were confident that we could identify a tonal center for different recordings of human music, human speech, and bird song by, for example, selecting the final pitch of a recording, this might be a useful method of normalizing. However, because the idea of final notes as tonal centre is not necessarily applicable across cultures or species, we chose to normalise all recordings by setting the most frequent pitch class to 0. In a separate study, we have validated this approach by directly comparing results of normalizing human music using the final pitch vs. the most frequent pitch, finding that there is almost no difference between the two methods, leading us to use most frequent as it can be calculated more objectively (e.g., in cases where music fades out before the end) [4]. Figure 2 shows example analyses.

In addition, the raw count of pitch annotations is converted to a percentage so that longer recordings will not be weighted more than shorter ones. After normalizing, pitch class histograms were averaged across recordings separately for human music, human speech, and bird songs to determine whether there were any tuning intervals that were consistent across each sample.

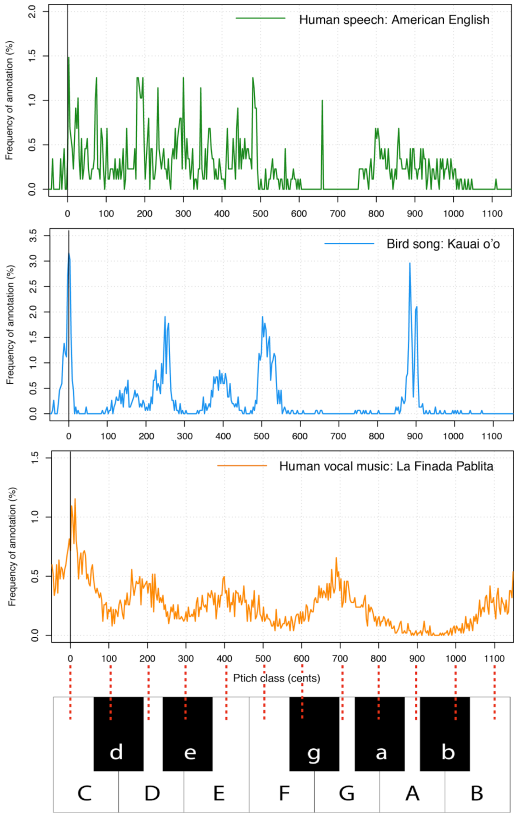


Figure 2. Bottom: A pitch class histogram of "La finada Pablita", a Mexican-American narrative song, aligned against a keyboard octave for comparison. The vertical axis represents how often a given pitch class occurs in the audio. The horizontal axis is plotted in cents over an octave range, from 0-1200, where the most common pitch class (~1.2% of annotations) is set to 0. In this figure the second most frequent scale degree appears a perfect fifth (~700 cents, equivalent to G in key of C) above the most frequent note. **Middle:** Pitch class histogram analysis of a bird song (Extinct bird in Kauai called "Kauaʻi ʻōʻō). **Top:** Pitch class histogram analysis of human speech (North American English).

3. RESULTS

Figure 3 shows the average pitch class histograms for human music, human speech, and bird song, plotted on the same axis for comparison. By definition, all samples show a peak at 0 cents, because 0 was defined as the most frequent note for each recording. Human speech and bird songs show no other clear peaks. Bird song does shows a stronger peak at the most frequent pitch

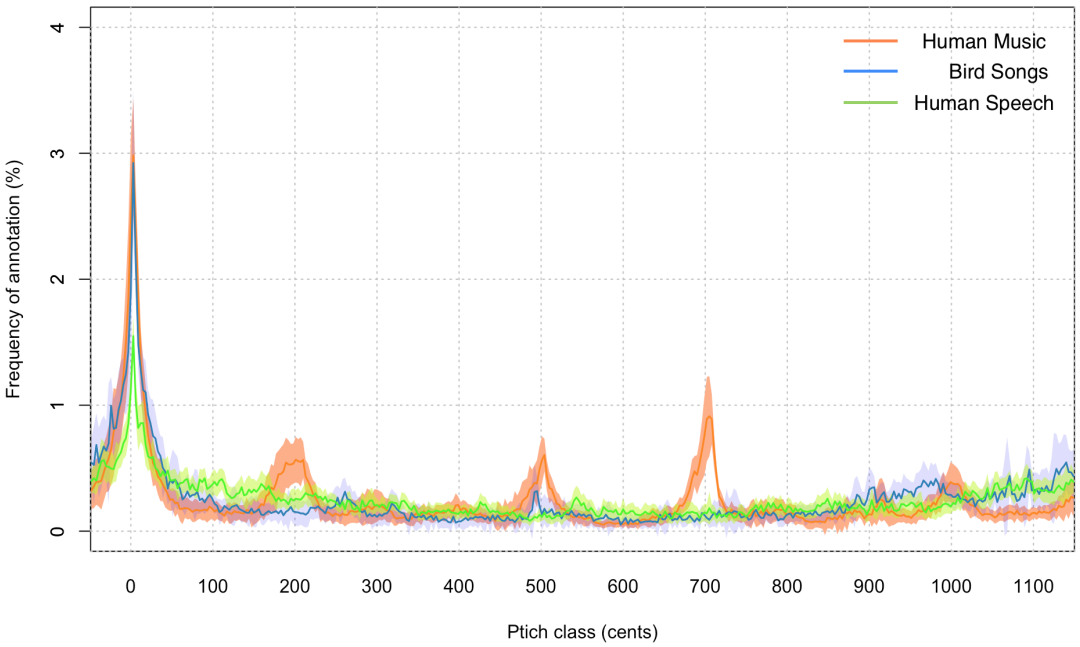


Figure 3. Averaged pitch class histograms for human music (n=30), bird song (n=30), and human speech (n=26). Shading indicates 95% confidence intervals.

(0 cents) and possible peaks at approximately 250, 500 and 1,000 cents, while human speech shows a possible peak at approximately 550 cents. These are small enough that it is not clear if they are true peaks or artefacts of the small sample size. However, human music shows much stronger peaks at intervals of approximately 700, 500, and 200 cents. These correspond approximately to small-integer ratios of 3:2 (perfect 5th), 4:3 (perfect 4th), and major 2nd (9:8), respectively.

4. DISCUSSION

Our results show that scale tunings in human music uniquely tend to emphasize intervals with small-integer ratios - particularly the perfect 5th (~ a 3:2 ratio) - while no consistent intervals emerge when the same analyses are applied to human speech or bird song (cf. Fig. 3).

Our previous analyses breaking up human music into sub-samples based on region and instrumentation [4, 7] showed that of these ratios, only the perfect 5th - the smallest possible integer-ratio within the octave - consistently predominated across all sub-samples. Taken together, these studies suggest that the perfect 5th uniquely predominates throughout the world's music but not in speech or bird song.

Many scholars have proposed that there is something special about small-integer ratios in human music, with most explanations centering around the psychoacoustics of harmonic overtone structure. Whenever an object resonates to produce a fundamental pitch, it also can produce a series of "overtones" which appear at integer ratios above the fundamental pitch due to the physics of how objects vibrate. While birds and many other animals tend to produce pure tone vocalizations without complex harmonic structure, human vocalizations tend to have a rich harmonic structure emphasizing many overtones, and this has been proposed to explain preferences for small-integer ratios in music via statistical learning through exposure to the speech of other humans that contains such harmonic structure [1, 7]. However, this "vocal similarity" hypothesis does not explain why we find small-integer ratios in human music but not human speech.

We speculate that human music may be unique because it evolved to be performed in synchronized groups, possibly to bond group members [19, 20]. This may have selected for integer-ratio frequencies because the resulting harmonies are more likely to perceptually fuse and sound like one large auditory event [2, 20]. Neither speech nor bird song are regularly performed in synchronized groups, and thus harmonicity does not result in any perceptual fusion.

While our preliminary data suggest that human scales are unique to music, there are also intriguing suggestions of similarities between bird song and human music as distinct from human speech. In particular, both human music and bird song show a stronger tendency to prefer a single most frequent note that remains relatively stable

throughout a performance, and bird songs may suggest a possible peak at a perfect 4th (4:3 ratio, approximately 500 cents). We hope to investigate this further with larger samples and through perceptual experiments in humans and birds.

5. FUTURE WORK

The main challenge for future work is to expand the sample to include several hundred recordings and perform statistical testing to formally evaluate the degree to which the pitch class histograms depart from distributions that would be expected by chance. We also intend to perform sub-sample analyses to explore the degree to which any average trends are consistent across geographic regions and instrumentation (particularly vocal vs. instrumental music). Expanding the sample of human music to the full set of 124 monophonic recordings has confirmed that intervals of approximately 200, 500, and 700 cents tend to predominate throughout the world, and that 700 cents is particularly common across almost all world regions [4]. We aim to similarly expand our samples of human speech and animal song, as the current results must be considered preliminary due to their small sample sizes and somewhat uneven geographic distribution.

Our analysis methods are well-suited for analyzing scales with a stable tonal center throughout a recording, but in the future we would like to explore ways of analyzing pitch relationships at interval-by-interval and phrase-by-phrase levels, which will particularly help in cases such as unaccompanied singing in which tonal centers can drift over time [18]. We also plan to implement additional features in Tarsos to allow us to automate the analysis to make it applicable to larger samples and to make the pre-analysis process more objective.

While the universality or cultural specificity of scale intonation has been debated for centuries, there remains little cross-species and cross-domain data to identify musical features that may represent unique adaptations for human music. Our analysis, while limited in scale, provides suggestive evidence that scales with small-integer ratios may represent a candidate for such an adaptation, and allow us to speculate why they may have evolved. Future studies with larger samples should help to clarify the robustness of our present findings.

6. AUTHOR CONTRIBUTIONS

P.E.S., S.F., A.T., J.S., and P.Q.P. designed the study; J.K., S.S., M.-J.H., and G.C. analysed the data, supervised by P.E.S.; J.K. and P.E.S. drafted the manuscript.

7. ACKNOWLEDGMENTS

This work was supported by a Grant-in-Aid for Young Scientists from the Japan Society for the Promotion of Science, Keio Research Institute at SFC Startup Grant, and a Keio Gijuku Academic Development Fund grant to P.E.S.

8. REFERENCES

[1]. D. L. Bowling and D. Purves, “A biological rationale for musical consonance,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 112, no. 36, pp. 11155–11160, 2015.

[2]. S. Brown, “Contagious heterophony: A new theory about the origins of music,” *Music. Sci.*, vol. 11, no. 1, pp. 3–26, 2007.

[3]. A. de Cheveigne and H. Kawahara, “YIN, a fundamental frequency estimator for speech and music,” *Journal of the Acoustical Society of America*, vol. 111, no. 4, pp. 1917– 1930, 2002.

[4]. G. Chiba, M.-J. Ho, S. Sato, J. Kuroyanagi, J. Six, P. Pfordresher, A. Tierney, S. Fujii, and P. E. Savage, “Small-integer ratio scales predominate throughout the world’s music.” 2019. *PsyArXiv.* Preprint doi: 10.31234/osf.io/5bghm

[5]. A. J. Ellis, “On the Musical Scales of Various Nations,” *Journal of the Society of Arts,* vol.23, no.1688, pp. 435-527, 1885.

[6.] W. T. Fitch, “The biology and evolution of music: A comparative perspective,” *Cognition*, vol. 100, no. 1, pp. 173–215, 2006.

[7]. K. Z. Gill and D. Purves, “A biological rationale for musical scales,” *PLOS ONE*, vol. 4, no. 12, p. e8144, 2009.

[8]. M. Ho, S. Sato, J. Kuroyanagi, J. Six, S. Brown. S Fujii, P. E. Savage. “Automatic analysis of global music recordings suggests scale tuning universals,” in *Extended abstracts for the Late-Breaking Demo Session of the 19th International Society for Music Information Retrieval Conference (ISMIR 2018)*, 2018.

[9]. H. Honing (Ed.). *The origins of musicality*. Cambridge, MA: MIT Press, 2018.

[10]. E. D. Jarvis, “Learned birdsong and the neurobiology of human language,” *Annals of the New York Academy of Science*, vol. 1016, no. 1, pp. 749-777, 2004.

[11]. Kim, J. W., Salamon, J., Li, P., & Bello, J. P. “CREPE: A convolutional representation for pitch estimation”. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 161-165), 2018.

[12]. M. Lieberman, Ed., “*Linguistic Data Consortium*,” 2019. [Online]. Available: https://www.ldc.upenn.edu. [Accessed: 20-Mar-2019].

[13]. A. Lomax (Ed.). *Folk song style and culture.* Washington, DC: American Association for the Advancement of Science, 1968.

[14]. M. Mauch and S. Dixon, “pYIN: A fundamental frequency estimator using probabilistic threshold distributions,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 659–663, 2014.

[15]. J. H. McDermott, A. F. Schultz, E. A. Undurraga, and R. A. Godoy, “Indifference to Dissonance in Native Amazonians Reveals Cultural Variation in Music Perception,” *Nature*, vol. 535, pp. 547–550, 2016.

[16]. B. Nettl, R. Stone, J. Porter, and T. Rice, Eds., *The Garland encyclopedia of world music* [10 volumes; 9 CDs]. New York: Garland Pub., 1998-2002.

[17]. A. D. Patel, *Music, language and the brain*. Oxford: Oxford University Press, 2008.

[18]. P. Q. Pfordresher and S. Brown, “Vocal mistuning reveals the origin of musical scales,” *J. Cogn. Psychol.*, vol. 29, no. 1, pp. 35–52, 2017.

[19]. P. E. Savage, S. Brown, E. Sakai, & T. E. Currie, ”Statistical universals reveal the structures and functions of human music,” *Proceedings of the National Academy of Sciences of the United States of America*, vol.112, no.29, pp. 8987–8992, 2015.

[20]. P. E. Savage, P. Loui, B. Tarr, A. Schachner, L. Glowacki, S. J. Mithen, and W. T. Fitch, “Music as a coevolved system for social bonding.” In preparation.

[21]. P. E. Savage, A. T. Tierney, A. D. Patel. “Global music recordings support the motor constraint hypothesis for human and avian song contour,” *Music Perception*. 34. 327-334, 2017.

[22]. J. Six, O. Cornelis, and M. Leman. “Tarsos, a Modular Platform for Precise Pitch Analysis of Western and Non-Western Music,” *Journal of New Music Research*, vol.42, no.2, pp. 113-129, 2013.

[23]. A. T. Tierney, F. A. Russo, and A. D. Patel, “The motor origins of human and avian song structure,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 108, no. 37, pp. 15510–15515, 2011.

[24]. A. T. Tierney, F. A. Russo, and A. D. Patel, “Empirical comparisons of pitch patterns in music, speech, and birdsong,” in *Proceedings of the Acoustics’08 Paris conference*, 2008.

AUTOMATIC COMPARISON OF GLOBAL CHILDREN'S AND ADULT SONGS SUPPORTS A SENSORIMOTOR HYPOTHESIS FOR THE ORIGIN OF MUSICAL SCALES

Shoichiro Sato[1], Joren Six[2], Peter Pfordresher[3], Shinya Fujii[1], Patrick E. Savage*[1]
[1]Keio University, Japan, [2]Ghent University, Belgium, [3]University at Buffalo, NY, USA
*Correspondence to: psavage@sfc.keio.ac.jp

ABSTRACT

Music throughout the world varies greatly, yet some musical features like scale structure display striking cross-cultural similarities. Are there musical laws or biological constraints that underlie this diversity? The "vocal mistuning" hypothesis proposes that cross-cultural regularities in musical scales arise from imprecision in vocal tuning, while the integer-ratio hypothesis proposes that they arise from perceptual principles based on psychoacoustic consonance. In order to test these hypotheses, we conducted automatic comparative analysis of 100 children's and adult songs from throughout the world. We found that children's songs tend to have narrower melodic range, fewer scale degrees, and less precise intonation than adult songs, consistent with motor limitations due to their earlier developmental stage. On the other hand, adult and children's songs share some common tuning intervals at small-integer ratios, particularly the perfect 5th (~3:2 ratio). These results suggest that some widespread aspects of musical scales may be caused by motor constraints, but also suggest that perceptual preferences for simple integer ratios might contribute to cross-cultural regularities in scale structure. We propose a "sensorimotor hypothesis" to unify these competing theories.

1. INTRODUCTION

Music exists in many different forms among almost all human societies, yet some common musical features are shared throughout the world. Although humans can discriminate more than 240 pitches within one octave, most tonal systems around the world incorporate scales with only seven or fewer pitches [1, 6, 12], but the precise tunings and combinations of these pitches used vary greatly throughout the world [5]. Theories of the origin of scales have been based on perceptual theories involving mathematical ratios since the time of Pythagoras [2]. Pitched notes on harmonic instruments produce multiple frequencies called harmonics or overtones, which resonate at integer multiples of the fundamental frequency (e.g., a note with a fundamental frequency of 100 Hz produces overtones at 200 Hz, 300 Hz, etc.). The intervals that have traditionally been considered "consonant" in Western music theory are also those with the simplest integer-ratios (e.g., 2:1 = octave, 3:2 = Perfect 5th, 4:3 = Perfect 4th; cf. Fig. 1).

Interval Name	Abbr.	Cents	Approx. frequency ratio	Example on keyboard
Perfect unison	**P1**	**0**	**1 : 1**	P1
Minor second	m2	100	16 : 15	m2
Major second	M2	200	9 : 8	M2
Minor third	m3	300	6 : 5	m3
Major third	M3	400	5 : 4	M3
Perfect fourth	**P4**	**500**	**4 : 3**	P4
Tritone	tt	600	7 : 5	tt
Perfect fifth	**P5**	**700**	**3 : 2**	P5
Minor sixth	m6	800	8 : 5	m6
Major sixth	M6	900	5 : 3	M6
Minor seventh	m7	1000	9 : 5	m7
Major seventh	M7	1100	15 : 8	M7
Perfect octave	**P8**	**1200**	**2 : 1**	P8

Figure 1. The equal tempered chromatic scale system. Boldface shows the diatonic ratios which have the simplest integer ratios.

However, these theories are generally based on tunable instruments that often use equal-tempered intervals that only approximate pure ratios (e.g., a pure perfect 5th is 702 cents, but a perfect 5th on a piano is produced at 700 cents), and some are skeptical as to whether this theory applies to vocal song, recognized as the most ancient and universal instrument of human music [11]. One alternative theory is that scales do not arise from perceptual constraints regarding integer ratios but instead due to production constraints on how precisely the voice can generate pitches [11]. This can be seen as a special case of the "motor constraint hypothesis", which proposes that many musical universals are not evolutionary adaptations for human music but simply byproducts of constraints on the way music is produced [13, 18]. The vocal mistuning theory argues that the universal tendency to use sparse scales with 7 or fewer scale degrees is because singing is too imprecise to allow

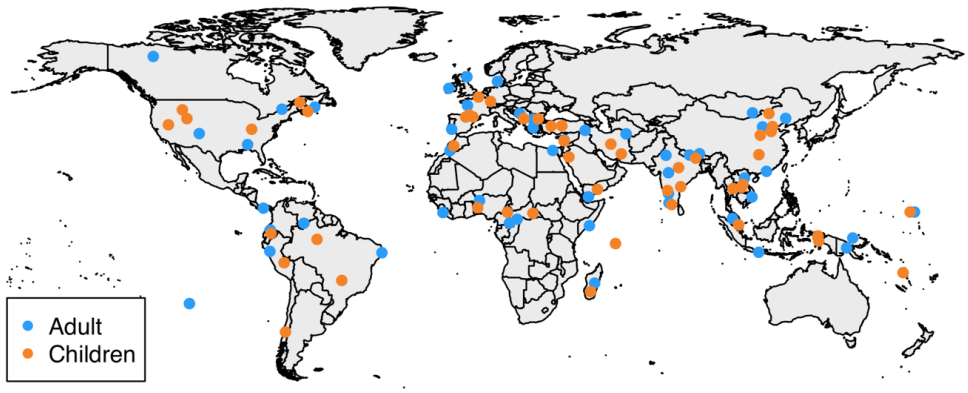


Figure 2. Map showing the approximate geographical distribution of 50 adult songs and 50 children songs used in this study

accurate production of scales using more than 7 scale degrees [11]. It predicts a negative correlation between tuning precision and number of scale degrees across musical genres. For example, children's songs should use sparser scales than adult songs because children are less able to precisely tune fine details of scales.

Although many studies of children's musical perception have been performed, there are only a handful of quantitative musicological studies of children's songs. In previous research, scales, melodic ranges, and other aspects of 100 children's songs from around the world were analyzed using the Cantometrics classification scheme [10, 15]. However, this study was limited to manual analysis and children's songs were not directly compared with adult songs.

In recent years, researchers have increasingly used large cross-cultural music corpora to address the relationship between human capacities and musical systems, as opposed to analyses that focus on specific musical cultures [14]. Thus, in the present study, we conduct automatic analyses to directly compare children's and adult's songs objectively using a matched global sample to examine the relationship between human vocal mistuning and musical scale structure.

In order to test the "vocal mistuning" hypothesis of scale origins, here we use an automatic method to compare children's songs and adult songs from around the world. Since children's vocal-motor control system are still developing, children should have smaller melodic ranges, less precise pitches, and sparser scales (i.e., fewer and more widely spaced scale degrees) than adult songs, even if both types of songs draw from the same overall tuning system (e.g., a given children's song might only utilize a sparser 4- or 5-note subset from within a broader 7-note tuning system used by adults). Meanwhile, in order to test the integer ratio hypothesis, we compared average scale degree tunings for children's and adult songs. If perceptual preferences for integer ratios contribute to scale structure, we predict common tunings on average across all songs, whereas if only motor constraints contribute to scale structure, we expect that no common trend will be found in this comparison.

2. METHODS

2.1 The automatic pitch extraction tool - Tarsos

Conventional musical scale analysis has been dominated by manual transcription methods. Since different cultures have different tonal systems, it has generally been considered difficult to compare cross-cultural scale structures. Especially when transcribing non-Western folk song with Western annotation, it often happens that each pitch is fitted to a specific tuning system, such as equal-temperament, based on the perception the transcriber, which may not accurately reflect the original music. Therefore, in this study, we used the automatic pitch extraction tool Tarsos [16], because it was designed explicitly for automatic analysis of any scales from around the world without imposing any culture-specific theories. Tarsos has several pitch extraction modules such as pitch distribution filters, audio feedback tools, and scripting tools for batch processing of large databases of musical recordings for analyzing the pitch distribution.

2.2 Song Samples

In order to explore general tendencies of human music, we constructed a globally balanced sample of audio recordings (Fig. 2). First we chose 50 children's songs from *Le chant des enfants du monde* [4], the "Lullabies and Children's Songs" CD from the *UNESCO Collection* [17], and *Mama Lisa's World International Music & Culture* [19], selecting 5-6 songs each from the nine regions designated by the *Garland encyclopedia of world music* [9] (Africa, South America, North America, Southeast Asia, South Asia, East Asia, Middle East, Europe, and Oceania). After selecting children's songs, we used the same regional sampling process to select 50 traditional adult songs from the *Garland Encyclopedia of World Music* [9], *UNESCO Collection of Traditional Music* [17] and *Folkways*

Records that were as geographically matched as possible with the children's songs.

Most samples from these CD sets were recordings by ethnomusicologists of folk songs transmitted by oral tradition in relatively small societies that were minimally influenced by Western music. "Adult songs", as defined here, are songs sung by adults, while "children's songs" are songs sung by children (generally 8~16 years old), including nursery rhythms and gaming songs, but not lullabies or other genres sung by adults to children. Gender differences, musical functions, performance context, etc. were not considered for this study.

Before choosing samples from the CD sets, we manually checked Tarsos's automatic pitch extraction by ear in order to determine whether the songs work sufficiently with pitch detection; usable songs were added to the corpus and songs with instrumental accompaniment, polyphonic singing, or heavy background noise were excluded due to the inaccuracy of the pitch extraction algorithm on polyphonic recordings.

2.3 Comparison of melodic range measurements

First, we measured the melodic ranges to check how much vocal development affects music expression. Melodic range was calculated based on Tarsos's pitch annotation output function by subtracting the lowest pitch from the highest pitch to appear in the song, manually excluding transcription errors due to noise or overtones from the calculation (Fig. 3).

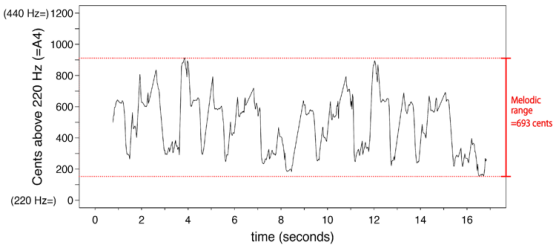


Figure 3. An example of melodic range analysis for an excerpt from "Plou i fa so" from Catalogne, Spain.

2.4 Comparison of number of scale degrees and vocal imprecision

2.4.1 *Number of Scale Degrees*

We examined the number of scale degrees to investigate how pitch precision influences scale structure. Tarsos first extracts the pitch histogram by detecting each pitch and number of annotations from the recording in real time. Next, another pitch distribution filter combines this pitch histogram across octaves to create a pitch class histogram, visualizing how pitch classes are distributed within one octave. This is expressed in cents [5] ranging from 0 to 1200 (see Fig. 1). Peak picking (see Fig. 4) is performed almost automatically, yet in order to enable extraction with the highest accuracy for all songs, we manually adjusted the parameter values for window (minimum distance between two peaks) and threshold (minimum occurrence necessary to be considered a "peak") depending on the songs. In order to automate these processes, we plan to evaluate optimal window and threshold values to allow objective peak picking without the need for manual adjustment in the future.

We used Tarsos's default YIN pitch estimation algorithm in this analysis [16]. Figure 4 shows examples of a children's song that has a relatively imprecise distribution, and an adult song that has more precise peaks.

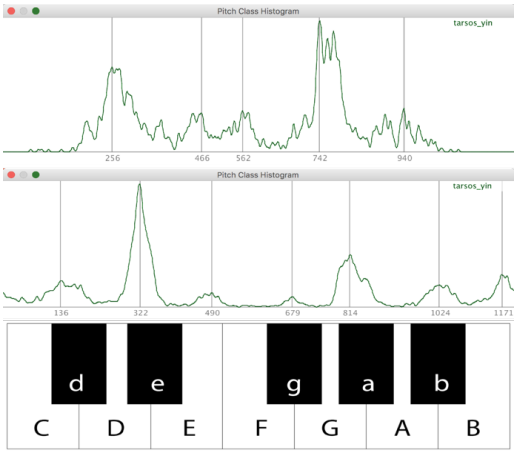


Figure 4. Automatic analysis of traditional West African scales: "Sigereti Fe Bara" (top, children's song, Pentatonic) and "Vai Call to Prayer" (bottom, adult song, Heptatonic). Vertical lines represent automatically detected scale degrees. The horizontal axis of the pitch class histogram shows pitch class across one octave (0 to 1200 cents) and the vertical axis shows the relative frequency of annotations.

2.4.2 *Calculating vocal imprecision*

We developed a novel formula (1) for quantifying the vocal imprecision (*I*). After obtaining pitch class histogram data, we took a width of ± 50 cents from each peak (*P*) and got the intersecting points ($qi_{a,}$ q_{ib} represent the frequency of occurrence for pitch classes that span a quarter note (50 cents) above and below each peak (P_i), respectively. First, we calculated the average imprecision of each peak by dividing the frequency a quarter tone away from the peak by the frequency of occurrence for the peak pitch class; $((qi_a + q_{ib}) /2/ P)$ and then averaged this imprecision across all pitch classes, such that 1 represents maximum imprecision (essentially no peaks), while 0 represents maximum precision (where scale degrees never fall more than one quarter tone from the peak; see Fig. 5 for details). We performed this calculation for all 100 songs and evaluated the correlation between tuning imprecision and number of scale degrees (Fig. 8C).

$$I = \frac{1}{N}\sum_{i=1}^{N} \left(\frac{q_{ia}+q_{ib}}{2}\right) \Big/ P_i \qquad (1)$$

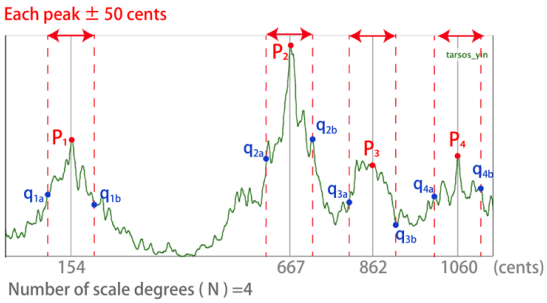


Figure 5. Visualization of formula for calculating imprecision, using an example children's song (see text for details).

2.5 Comparison of normalized scale analysis

Finally, the scale normalization to the tonal center (= tuning into same key) was performed to verify whether there is a tendency toward common intervals relative to the tonal centre of the scales. In order to compare scales across cultures, we attempted to normalize each scale by tuning the final pitch class to be 0 cents, following the method attempted in our previous study [7]. When the ending note is not detectable because of fade out or excerpts in a song, we instead normalized to the most frequent note as a tonal center (recent analyses [3] suggest that there is little difference between these two methods, leading us to propose consistently normalizing to the most frequent note in the future). Figure 6 shows an example of two normalized pitch class histograms. As shown in the figure, the tonal center was normalized to 0/1200 cents. All counts of pitch annotations are calculated as a percentage of pitch frequency, so as to be able to compare across recordings regardless of recording length.

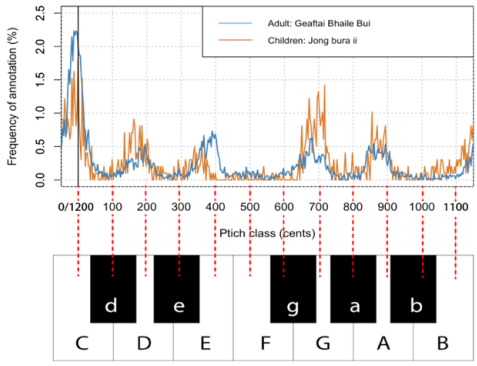


Figure 6. Examples of pitch class histograms for a children's song (Indonesian folk song "*Jong bura ii*", orange) and an adult song (Irish folk song "*Geaftai Bhaile Bui*", blue). Pitch class histograms are normalized so that the final/most frequent note is set to 0 cents. Both normalized pitch class histogram demonstrates similar scale structure:

3. RESULTS

3.1 Comparison of melodic range measurement

Figure 7 shows the distribution of melodic range measurements for 100 children's and adult songs. The mean absolute melodic range of the 50 children's songs was 941 cents (i.e., more than a minor 6th), while that of the 50 adult's songs was 1362 cents (i.e., more than a minor 9th): almost 50 % greater than children's songs (t(98) = 5.8, p =1.2x10^{-7}). This result suggests that developmental constraints on vocal production influence musical expression. While children's songs tended to fall within one octave (~ 1200 cents), some adult songs showed a range of up to two octaves (~ 2400 cents). The average of a minor 9th range for adult songs is surprising given that previous manual research using a similar sample found that adult music throughout the world consistently tended to have a range of less than one octave [12]. This may be due to our automated analysis method overestimating melodic range due to being overly sensitive to outliers or to octave errors in transcription. In the future we intend to explore other algorithms besides the YIN algorithm and other measures that are less sensitive to outliers, such as interquartile range.

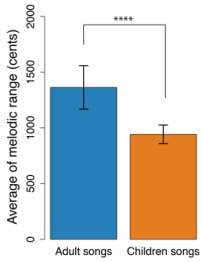
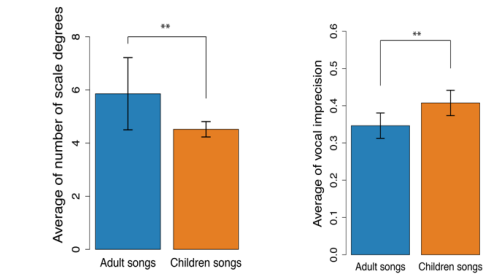


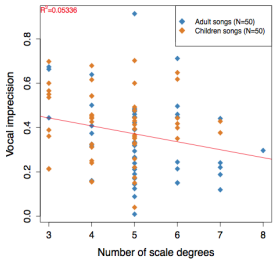
Figure 7. Comparison of average melodic range for children's and adult songs. Error bars = 95% confidence intervals.

3.2 Number of scale degrees and vocal imprecision

Figure 8A shows the average of the number of scale degrees for children's and adult songs. The scales of all songs were composed of seven or fewer scale degrees, except for one adult song with an 8-note scale. The use of pentatonic scales was the most dominant, accounting for approximately 1/3 of all songs (36% of adult corpus, 30% of children corpus). The average number of scale degrees in adult sample of 5.8 was significantly higher than that of the children's sample of 4.5 (t(98) = 1.98, p = .007; Fig. 8A). The mean vocal imprecision for children's songs of 0.41 was significantly higher than the average imprecision for adult songs of 0.34 (t(98)=1.96, p=.009; Fig.8B). There was a negative correlation between the number of scale degrees and vocal imprecision (r=-0.23, p=.001; Fig. 8C).



A) B)



C)

Figure 8. (A) The comparison of average number of scale degrees between children and adult songs. (B) The comparison of average vocal imprecision between children and adult songs. (C) Scatter plot and regression line showing the relationship between the number of scale degrees and vocal imprecision between children songs and adult songs. Error bars = 95% confidence intervals.

3.3 Normalized scale analysis

Figure 9 shows average scale tunings across children's and adult songs. Both children's and adult songs displayed peaks at the same four approximate intervals: perfect 5th (3:2 ratio, ~700 cents), perfect 4th (4:3 ratio, ~500 cents), major 3rd (5:4 ratio, ~400 cents), and major 2nd (9:8 ratio, ~200 cents), although the peaks were more precise for the adult songs. This result suggests that some aspects governing scale structure are consistent regardless of developmental stage. Note that tonality was not used as a criterion in sampling songs, so the predominance of major thirds over minor thirds reflects that major tonalities are more common cross-culturally.

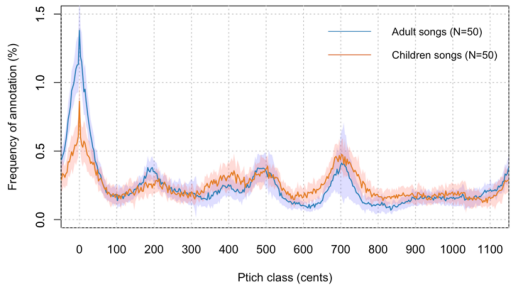


Figure 9. Average scale tunings across children's and adult songs. The x-axis begins and ends at 1150 cents in order to show the distribution around the tonal center (set to 0 cents). The transparent portions along the line of the pitch class histogram represent the 95% confidence intervals.

4. DISCUSSION AND FUTURE WORK

4.1 Discussion

These comparative analyses measuring melodic range, number of scale degrees, imprecision, and scale tuning, were conducted to test the hypothesis that cross-cultural regularities in pitch structure are determined by developmental constraints, particularly regarding vocal mistuning. We found that adult songs throughout the world consistently employ wider melodic ranges and denser scales with more precise tuning than children's songs, as predicted by the vocal mistuning hypothesis [11]. However, we also found similar small-integer ratio intervals appear consistently in both children's and adult songs, which is predicted by perceptual hypotheses based on small-integer ratio consonance [2, 6] but not by motor constraint hypotheses such as the vocal mistuning hypothesis.

Based on these results, we believe neither the perceptually-based integer ratio hypothesis nor the production-based vocal mistuning hypothesis alone fully explains cross-cultural regularities in scale structure [11]. We instead propose a more nuanced "sensorimotor hypothesis" for the origin of musical scales that combines these previous two hypotheses. This sensorimotor hypothesis argues that scale structure is determined by a balance between optimizing interval size for accurate production and optimizing ratios among scale degrees for maximal consonance in group music-making. Among other things, this hypothesis predicts that motor constraints will play a stronger role in governing solo and monophonic music, while perceptual constraints will play a larger role in polyphonic and group music.

4.2 Future work

In future studies, we plan to expand the scope of our methods to better test our sensorimotor hypothesis and other hypotheses for the origins of musical structure. This includes expanding the sample of human songs as well as including instrumental music, speech, and animal song for comparison [8]. To do so, we need to refine and further automate our process (e.g., peak-picking, noise removal for melodic range analysis, quantification of imprecision, improved automatic pitch extraction to accommodate polyphonic music) to be able to analyze larger samples while making fewer manual judgments. It will also be necessary to test our predictions through cross-cultural and cross-species behavioral experiments at different developmental stages in addition to corpus studies such as this.

Comprehensively addressing such improvements will require substantial investment to go beyond the current state-of-the-art in musical information retrieval, music cognition, and ethnomusicology, but hold the promise for

understanding how and why music has evolved to hold such universal power, and how we might harness that power for a more harmonious future.

5. AUTHOR CONTRIBUTIONS

S.S., P.E.S., S.F., J.S., and P.Q.P. designed the study; S.S. analyzed the data, supervised by P.E.S.; S.S. and P.E.S. drafted the manuscript.

6. ACKNOWLEDGMENTS

This work was supported by a Grant-in-Aid for Young Scientists from the Japan Society for the Promotion of Science, Keio Research Institute at SFC Startup Grant, and a Keio Gijuku Academic Development Fund grant to P.E.S.

7. REFERENCES

[1] Brown, A., Jordania, J. (2013). Universals in the world's musics, *Psychology of Music*, *41*(2), 229–248.

[2] Bowling, L. D., Purves, D., & Gill, Z. K. (2018). Vocal similarity predicts the relative attraction of musical chords. *Proceedings of the National Academy of Sciences of the U. S. A.,115*(1), 216-221.

[3] Chiba, G., Ho, M.-J., Sato, S., Kuroyanagi, J., Six, J., Pfordresher, P. Q., Tierney, A. T., Fujii, S., & Savage, P. E. (2019). Small-integer ratio scales predominate throughout the world's music. *PsyArXiv* preprint. https://doi.org/10.31234/osf.io/5bghm

[4] Corpataux, F. (1993-2018). *Le Chant des enfants du monde* [57 CDs]. ARION.

[5] Ellis, J. A. (1885). On the musical scales of various nations. *Journal of the Society of Arts, 23*(1688), 435-527.

[6] Gill, Z. K., & Purves, D. (2009). A biological rationale for musical scales. *PLOS ONE*, 4(12), e8144.

[7] Ho, M.-J., Sato, S., Kuroyanagi, J., Six, J., Brown, S., Fujii, S., & Savage, P. E. (2018). Automatic analysis of global music recordings suggests scale tuning universals. *Extended Abstracts for the Late-Breaking Demo Session of the 19th International Society for Music Information Retrieval Conference*.

[8] Kuroyanagi, J., Sato, S., Ho, M.-J., Chiba, G., Six, J., Pfordresher, P. Q., Tierney, A. T., Fujii, S., & Savage, P. E. (2019). Automatic comparison of human music, speech, and bird song suggests uniqueness of human scales. *Proceedings of the 9th International Workshop on Folk Music Analysis*. Preprint: https://doi.org/10.31234/osf.io/zpv5w

[9] Nettl, B., Stone, R., Porter, J., & Rice, T., eds. (1998–2002). *The Garland encyclopedia of world music* (Garland, New York).

[10] Pai, J. S. (2009). *Discovering musical characteristics of children's songs from various parts of the world.* MA thesis. University of British Columbia.

[11] Pfordresher, Q. P., & Brown, S. (2017). Vocal mistuning reveals the origin of musical scales. *Journal of Cognitive Psychology*, *29*(1), 35–52.

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

[13] Savage, P. E., Tierney, T. A., & A. D. Patel, (2017). Global music recordings support the motor constraint hypothesis for human and avian song contour. *Music Perception*, *34*(3), 327–334.

[14] Savage, P. E., & Brown, S., (2013). Toward a new comparative musicology. *Analytical Approaches to World Music*, *2*(2), 148–197.

[15] Savage, P. E. (2018). Alan Lomax's Cantometrics Project: A comprehensive review," *Music & Science.*, 1, 1–19, https://doi.org/10.1177/2059204318786084.

[16] Six, J., Cornelis, O., & Leman, M. (2013). Tarsos, a modular platform for precise pitch analysis of Western and non-Western music. *Journal of New Music Research*, *42*(2), 113-129.

[17] Smithsonian Institution Archives Record Unit (1961-2006), "*UNESCO Collection of traditional music*" [123 CD set].

[18] Tierney, T. A., Russo, A. F., & Patel, D. A. (2011). The motor origins of human and avian song structure. *Proceedings of the National Academy of Sciences of the U. S. A.*, *108*(37), 15510–15515.

[19] Yannucci, L. (2019). "*Mama Lisa's World International Music & Culture*". https://www.mamalisa.com.

Country Classification with feature selection and network construction for folk tunes

Cornelia Metzig[1], Roshani Abbey[2], Mark Sandler[1], Caroline Colijn[3]

[1]Centre for digital music, Queen Mary University of London, UK
[2]Royal Academy of Music, London, UK
[3]Simon Fraser University, Burnaby, Canada

June 28, 2019

Abstract

We explore two approaches to quantify folk song similarity. In the first part of this paper, we investigate to what extent the a folk tunes differ by country of origin. If it is possible for a human with a limited set of tunes in mind to guess the country of origin correctly, then there must be a signal that can be detected with automatic classification. This question has been addressed in the literature, where classifiers were trained both on global and local features [6, 5, 3].

In this paper, we aim at predicting country of origin based on extracted features We use a large number local features (n-grams of note successions and rhythm successions), for a MIDI dataset of songs from England, Scotland, Ireland, Germany, USA, Spiritual songs and African songs. Each group contains 80 songs, although larger groups have been used for comparison. These extracted features contain indirectly information of global features such as mode and time signature. Since we use a high number of initially extracted features ($\approx$ 7000 that are not 0) we reduce this number with statistical filtering methods, before performing random forest classification on them. In addition we use feature selection, based on the importances of the random forest classifier, which strongly increases accuracy. Our method allows us to predict country of origin with up to 91% accuracy, depending also on the degree of similarity of the two countries that were used for training. We use 10-fold cross validation to reduce overfitting, and compare results to a classifier trained on the same data with randomly assigned country labels. The importances of the features are interesting since they reveal typical patterns of tunes for each country. Interestingly, rhythm grams and melodic grams of different lengths are all relevant. The results depend strongly on the country pair. They confirm established results from musicology but also have the potential to complement them, due to the systematic way they are found.

However, country of origin is not the only relevant aspect when studying song similarity. We construct similarity networks of songs, based on these extracted features. This approach can reveal song families that follow other patterns than countries, and may contain interesting information regarding the history of songs, for example many links between Scottish and American songs, or between African songs and Spirituals. To construct a network, we first calculated Hamming and Euclidean distances between the song vectors of extracted features. Whenever the distance between two songs falls below a certain threshold, a link between those songs is created. This simple method bears the problem of so-called "hubness" [1, 2], which is an aggregation artefact due to the fact that the distances are calculated in very high-dimensional spaces. Songs without apparent similarity will appear close (and therefore connected in our network), and highly connected hub songs will emerge. We explore two methods correcting for this: (i) to use fewer dimensions, which are selected based on random forest importances in the classification. This reduces hubness and creates plausible networks, however some hghly connected songs remain, unless the number of dimensions is reduced to a number where the interesting similarities disappear as well. We compare our feature selection approach with a method of reducing hubness in audio similarity introduced by [4], the so-called mutual proximity (MP). Based on the idea of using only the closest songs and constructing a symmetric distance from it, MP reduces hubness, however loses some of the interesting information as well. We propose a combination of MP and selecting important dimensions. Although it is ultimately a matter of choice which songs to consider to be similar, the proposed method yields the most plausible results. We find strong clusters that belong within one country, and also that songs from certain countries

of origin (e.g. Germany) are much more distant than others. Also heterogeneity within one country group differs strongly. Genre seems to be much less informative than country, e.g. drinking songs and Christmas carols of one country are often connected.

The classification method has the potential to be applied to smaller geographic regions than countries, as a tool to help location, or to melodies from different composers, to identify authorship.

References

[1] Jean-Julien Aucouturier and Francois Pachet. A scale-free distribution of false positives for a large class of audio similarity measures. *Pattern recognition*, 41(1):272–284, 2008.

[2] Arthur Flexer, Dominik Schnitzer, and Jan Schlüter. A mirex meta-analysis of hubness in audio music similarity. In *ISMIR*, pages 175–180, 2012.

[3] Ruben Hillewaere, Bernard Manderick, and Darrell Conklin. Global feature versus event models for folk song classification. In *ISMIR*, volume 2009, page 10th. Citeseer, 2009.

[4] Dominik Schnitzer, Arthur Flexer, Markus Schedl, and Gerhard Widmer. Using mutual proximity to improve content-based audio similarity. In *ISMIR*, volume 11, pages 79–84, 2011.

[5] Peter van Kranenburg, Anja Volk, and Frans Wiering. A comparison between global and local features for computational classification of folk song melodies. *Journal of New Music Research*, 42(1):1–18, 2013.

[6] Anja Volk and Peter Van Kranenburg. Melodic similarity among folk songs: An annotation study on similarity-based categorization in music. *Musicae Scientiae*, 16(3):317–339, 2012.

ON THE SINGER'S FORMANT IN LITHUANIAN TRADITIONAL SINGING

Robertas Budrys
Department of Audiovisual Arts, Kaunas University of Technology, Lithuania
robudrys@gmail.com

Rytis Ambrazevičius
Department of Audiovisual Arts, Kaunas University of Technology, Lithuania
Department of Ethnomusicology, Lithuanian Academy of Music and Theatre, Lithuania
rytis.ambrazevicius@ktu.lt

ABSTRACT

While the singer's formant is in fact obligatory for the unamplified male operatic voice – to be heard in the context of the whole orchestra, the question of this voice quality in other vocal techniques remains open. The present pilot study aims to analyze possible manifestations of the singer's formant in Lithuanian traditional singing and to review techniques of evaluation of the singer's formant.

Two examples of traditional singing (recordings of vocal performances) are chosen for the analysis; a male hay making song and a female rye harvesting song. Both examples represent so-called 'field' genres performed outdoors and characterized by of esonant and loud voices; thus appearance of the singer's formant is more likely compared to 'indoor' genres. A set of parameters proposed in previous studies and indicating presence of the singer's formant is applied; singing power ratio (SPR), energy ratio (ER), L3–L1, and level of the singer's formant (L_{SF}). Although reliability of the parameters is to some extent disputable, the intense spectral bands characteristic for the singer's formant are detected. The singer's formant in both examined performances (especially in the female example) is less prominent compared to the case of the male operatic voice.

1. INTRODUCTION

The singer's formant, a specific formant occurring at approximately 2 kHz – 4 kHz frequencies, is widely discussed, starting from the seminal studies published several decades ago (cf. Sundberg, 1970; 1972; 1973; 1974; Dmitriev & Kiselev 1979; Shutte & Miller, 1985) and continuing up to the present (cf. Sundberg, 1995; Millhouse, Clermont & Davis, 2002; Ferguson et al., 2006; Reid et al., 2007). Presence of the singer's formant is considered in fact an obligatory requirement for operatic male voices. In addition, more recent studies found possible manifestations of the singer's formant in operatic female or castrato voices as well, though to a lesser extent (cf. Sundberg, 2007; Lee et al., 2008). Several techniques were proposed for evaluating intensity of the singer's formant and for differentiation of the true formant from other phenomena responsible for some intensification of spectra in the 2 kHz – 4 kHz range (Frøkjaer-Jensen & Prytz, 1976; Omori et al., 1996; Sundberg, 2001; Kenny & Mitchell, 2006; Ternström, Bohman, & Sodersten, 2006; Watts et al., 2006).

This peculiar voice quality is important, first of all, for large acoustical spaces without artificial sound reinforcement. Most probably, it is not that urgent for smaller chambers and for contemporary environments with sound reinforcement. Also it is not clear whether the singer's formant is relevant for other than operatic (or academic) vocal styles. For instance, traditional and generally non-Western singing is barely studied in terms of the singer's formant (cf. Sengupta, 1990; Delviniotis, 1998; Kovačić, Boersma, & Domitrović, 2003; Joshi & Raju, 2016). One could expect that, at least for certain acoustical conditions and certain styles of traditional singing, the technique that is normally applied to operatic singing could be applied as well.

2. TWO EXAMPLES OF VOCAL PERFORMANCES

Various cases of vocal performances with or without the singer's formant may be found in Lithuanian vocal tradition. It seems that the singer's formant could be investigated, first of all, in songs performed outdoors, the 'field song genres' characterized by resonant voices required for communication between (usually) quite distant groups of singers. Among these genres, rye harvesting and hay making songs are the most popular and the largest in number. For instance, according to the catalogue of Lithuanian folk songs (Misevičienė, 1972), rye harvesting and hay making songs contain, correspondingly, 348 and 152 variants of lyrics whereas the numbers for oat harvesting, buckwheat pulling, and flax and cannabis harvesting are, correspondingly, 25, 6, and 93.

Two typical traditional song performances belonging to the discussed quality of performance are chosen for the investigation. The first one, a hay making song (see the transcription in Figure 1), comes from Aukštaitija (Northeastern Lithuania). Hay making songs of this particular type (*valiavimai*) were usually performed by male singers-haymakers. The examined example is a recorded performance of 4 male singers; one of them presents the leading part and the others add the lower part (actually some women's voices are also heard in the background, yet they are very faint and their impact on the results of acoustical measurements is expected to be negligible). Thus the analysis would lead to certain

averaging of effects of the singer's formant, possibly somewhat different for individual singers. The second song, a rye harvesting song (see the transcription in Figure 2) comes from Dzūkija (Southeastern Lithuania). Rye harvesting was women's work and, naturally, songs of this type were performed by female singers. The chosen recording contains antiphonal performance of three women, i.e. they sing in succession, one after another, verse after verse. Thus they imitate the situation of communication of singers standing at a certain distance one from another. The performance of one singer, Ona Jauneikienė (2nd, 5th,.. verses), is chosen for the analysis as her voice seems to be the most resonant.

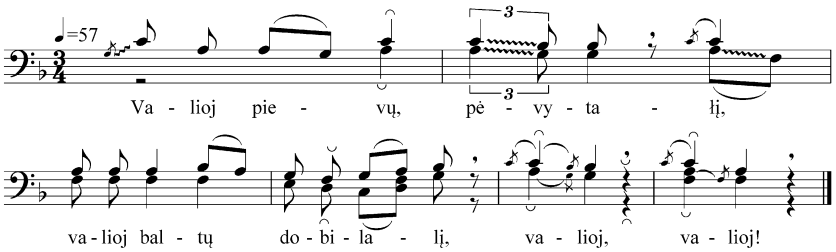


Figure 1. Transcription of the first verse of the song *Valioj pievų, pėvytalį* (Puponiai singers, Kupiškis Dst. Recording: Četkauskaitė, 1998, N18).

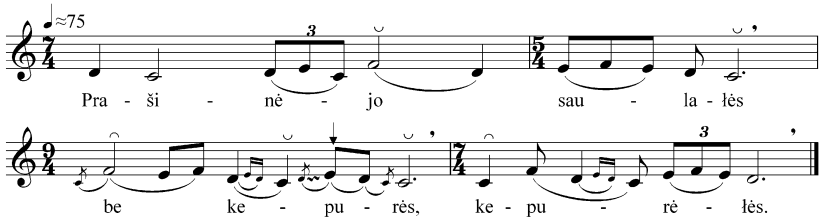


Figure 2. Transcription of the second verse of the song *Vaikštinėjo tėvulis* (Ona Jauneikienė; Masališkės, Varėna Dst. Recording: Četkauskaitė, 1995, N14).

3. METHODS

Praat software is applied for the acoustical measurements, composing of smoothed spectra, and calculations of different parameters. Five techniques of evaluation of the singer's formant are employed:

- **Initial evaluation**. The peak corresponding to frequency of the singer's formant (F_S) in the region around 3 kHz is visually evaluated and the frequency band around this peak (from the preceding dip to the corresponding, roughly symmetrical frequency) is manually selected. The ratio between the energy of the band and the total energy of the spectrum is calculated.
- The **singing power ratio (SPR)** "is calculated by measuring the ratio of the peak intensities between the 2–4-kHz and 0–2-kHz frequency bands in the context of sustained vowels or vocalic segments in sung/spoken samples" (Watts et al., 2006).
- **Energy Ratio (ER)** is a ratio between the total energy in the 0–2 kHz and 2–4 kHz bands (Kenny & Mitchell, 2006).
- **L3–L1**. It is the difference between the levels of the third and the first formant (applied by Sundberg, 2001).
- **Level of the singer's formant (L_{SF})**. It is the difference between observed and expected L3–L1 values. The measure of L_{SF} takes into account the effect of different vowels, i.e. the influence of formant frequencies on formants levels (Sundberg, 2001).

For the visual evaluation of a singer's formant, LTAS spectra are also composed, in the manner used in various studies noted above. Time spans of 20–30 seconds are considered, as in Sundberg, 2001.

4. RESULTS AND DISCUSSION

Melostrophes 1–4 (33 occurrences) of the male song *Valioj pievų, pėvytalį* and melostrophes 2, 5, 8, and 11 (30 occurrences) of the female song *Vaikštinėjo tėvulis* were considered. The results are presented in Figures 3–8 and in Tables 1–2.

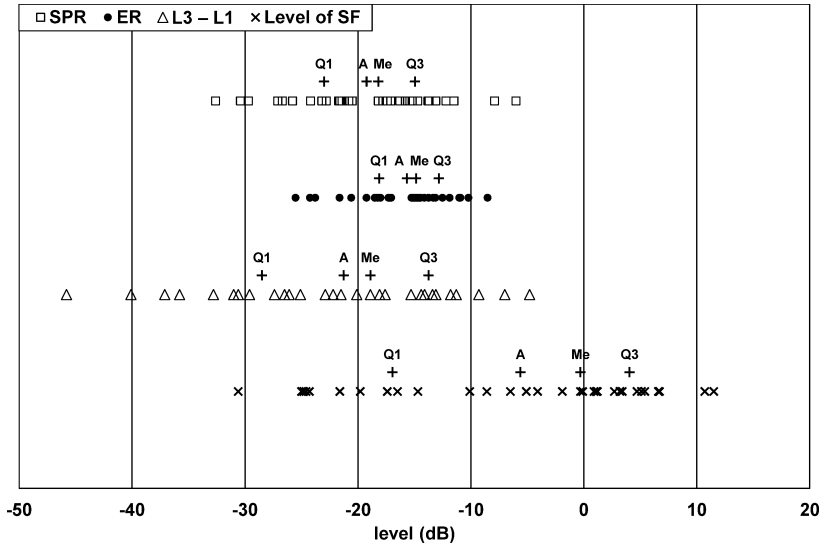


Figure 3. Values of four parameters for evaluation of a possible singer's formant; the male song *Valioj pievų, pėvytalį*. Averages, medians, and interquartiles (A, Me, Q1–Q3) are marked.

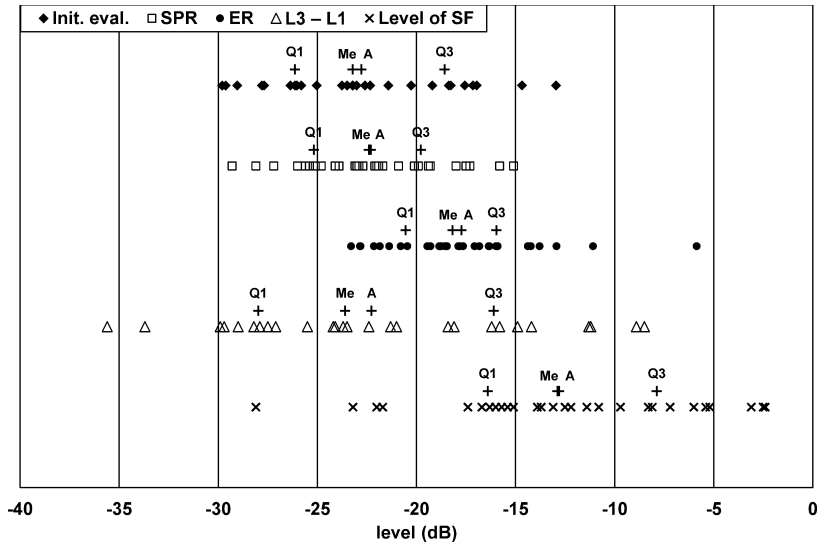


Figure 4. Values of five parameters for evaluation of a possible singer's formant; the female song *Vaikštinėjo tėvulis*. Averages, medians, and interquartiles (A, Me, Q1–Q3) are marked.

	SPR	**ER**	**L3–L1**
L_{SF}	.663	.719	.909
L3–L1	.821	.883	
ER	.935		

Table 1. Correlation between the parameters; male song *Valioj pievų, pėvytalį*.

	Init. eval.	**SPR**	**ER**	**L3–L1**
L_{SF}	.368	.288	.280	.815
L3–L1	.619	.484	.617	
ER	.856	.783		
SPR	.699			

Table 2. Correlation between the parameters; female song *Vaikštinėjo tėvulis*.

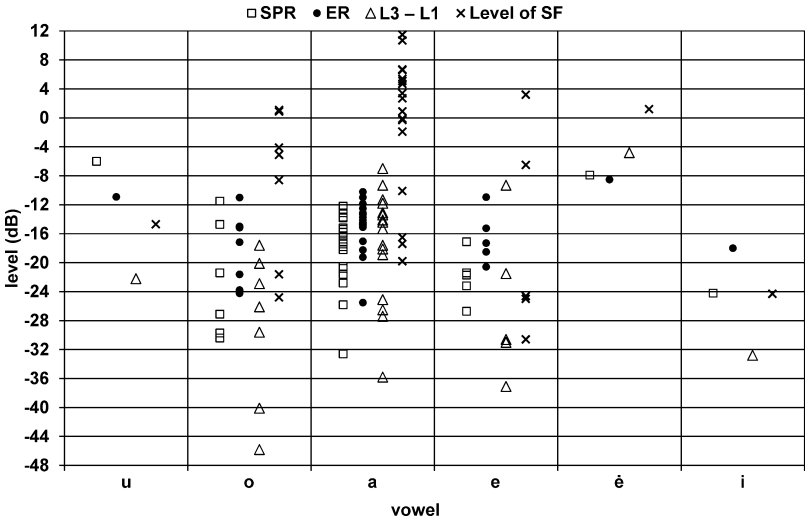


Figure 5. The same as in Figure 3. Occurrences of different vowels are grouped.

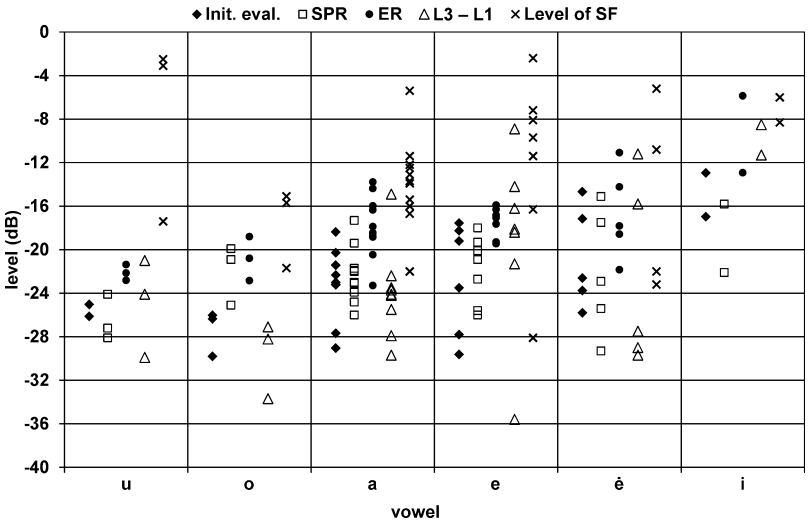


Figure 6. The same as in Figure 4. Occurrences of different vowels are grouped.

The obtained results show more or less prominent intensification of the spectra in the 2–4 kHz range. As expected, this quality is more distinct in the case of the male voice. Incidentally, the female example is characterized by two spectral peaks (Figure 8). The same tendency was observed by Seidner et al. (1985); they found that for female singers, two peaks can occur at frequency ranges of 2.5–3 kHz and 3–4 kHz.

Aural impression suggests that both examples are characteristic of resonant voices and possibly F_S is responsible for this. The problem is that there are no clear and undisputed referential values showing presence or absence of the singer's formant; the techniques for evaluation of the singer's formant proposed in the discussed previous studies and their results are quite ambiguous or even contradictory. For instance, Omori et al., 1996, presented the following values of SPR: roughly –14 dB for professional vocalists with S_F and roughly –23 dB for non-vocalists without S_F (the numbers were somewhat larger for males and smaller for females), while the evaluations by Watts et al., 2006, were, respectively, –23 dB and –31 dB. Kenny & Mitchell (2006) examined singing of advanced vocal students and found that their SPR ranged from –33 to –11 dB (–19,7 dB, on the average)

and ER ranged from –29 to –9 dB (–16,8 dB, on the average). However, their subjects were sopranos and mezzo-sopranos, thus naturally the values tend to be lower. Sundberg (2001) found L3–L1 ranging from –19 to –8, for basses and baritones, and proposed $L_{SF} > 0$ condition for quantification of presence of the singer's formant. Yet he did not recommend the use of this parameter for sopranos. L_{SF} technique proposed by Sundberg is not sufficiently reliable since it is also dependent on other factors, such as vocal volume (Sundberg, 2001) and chosen theoretical bandwidths of formants (Millhouse et al., 2002). This may be the reason this technique was avoided in later studies by other authors. Obviously, differences of F1 and F2 for various vowels can influence the results as well.

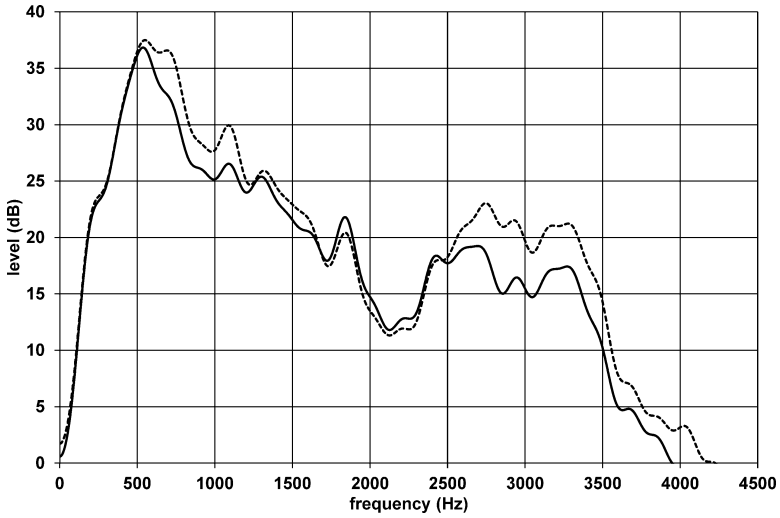


Figure 7. Smoothed LTAS spectrum; the male song *Valioj pievų, pėvytalį,* melostrophes 2 and 4.

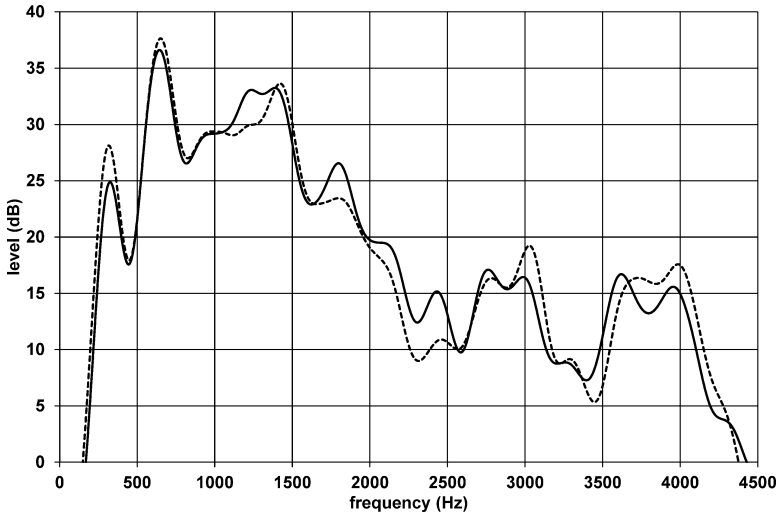


Figure 8. Smoothed LTAS spectrum; the female song *Vaikštinėjo tėvulis*, melostrophes 2 and 5.

Female singing (Figure 6) shows quite prominent differences of the parameter values for different vowels whereas the differences are not that clear for male singing (Figure 5). This might be attributed to the differences of male and female voices but, most probably, the differences of vocal dialects are at work: Aukštaičiai singing is characteristic of stronger voice 'covering', i.e., the vowels

tend to be acoustically more similar (Ambrazevičius, 2001).

The present study is, in part, oriented towards evaluation and comparison of several techniques for detection of the singer's formant. Correlations of SPR, ER ir L3–L1 (Tables 1 and 2) show that these parameters estimate the examined occurrences quite similarly; they also correspond to the initial evaluations fairly well.

At any rate, intensification of the spectra in the discussed frequency range found for the examined vocal performances suggests that discussed vocal quality is somewhat specific and partly marks a certain quality of the resonant voice.

5. REFERENCES

Ambrazevičius, R. (2001). Vocal technique in aukštaičiai and dzūkai male solo singing. *Lietuvos muzikologija, 2*, 169-179.

Četkauskaitė, G., ed. (1995). *Lietuvių liaudies muzika*. Vilnius: 33 Records.

Četkauskaitė, G., ed. (1998). *Lietuvių liaudies muzika. V. 2. Aukštaičių dainos. Šiaurės rytų Lietuva*. Vilnius: Lietuvos muzikos akademijos muzikologijos instituto etnomuzikologijos skyrius.

Delviniotis, D. S. (1998). A classification of Byzantine singing voices based on singer's formant. Retrieved from: https://ieeexplore.ieee.org/document/7089751

Dmitriev, L., & Kiselev, A. (1979). Relationship between the formant structure of different types of singing voice and the dimension of supra-glotal cavities. *Folia Phoniatrica, 31*, 231-241.

Ferguson, S., Kenny, D. T., & Cabrera, D. (2006). Effects of training on time-varying spectral energy and sound pressure level in nine male classical singers. *Journal of Voice, 24*, 39-46.

Frøkjaer-Jensen, B., & Prytz, S. (1976). Registration of voice quality. *Bruel & Kjaer, Technical Revue, 3*, 3-17.

Joshi, N., & Raju, M. A. (2016). Singer's formant in Hindustani classical singers. *Journal of Laryngology & Voice, 6*, 7-13.

Kenny, D. J., & Mitchell, H. F. (2006). Acoustic and perceptual appraisal of vocal gestures in the female classical voice. *Jounal of Voice, 22*, 55-70.

Kovačić, G., Boersma, P., & Domitrović, H. (2003). Long-term average spectra in professional folk singing voices: a comparison of the *klapa* and *dozivački* styles. *Institute of Phonetic Sciences, University of Amsterdam, Proceedings, 25*, 53-64.

Lee, S.-H., Kwon, H.-J., Choi, H.-J., Lee, N.-H., Lee, S.-J., & Jin, S.-M. (2008). The singer's formant and speaker's ring resonance: a long-term average spectrum analysis. *Clinical and Experimental Otorhinolaryngology, 1*, 92-96.

Millhouse, T., Clermont, F., & Davis, P. (2002). Exploring the importance of formant bandwidths in the production of the singer's formant. In C. Bow (Ed.), *Proceedings of the 9th Australian International Conference on Speech Science & Technology, Melbourne, December 2 to 5, 2002* (pp. 373-378). Melbourne: Australian Speech Science and Technology Association.

Misevičienė, V. (1972). *Lietuvių liaudies dainų katalogas. Darbo dainos. Kalendorinių apeigų dainos*. Vilnius: Vaga.

Omori, M., Kacker, A., Carroll, L., Riley, W., & Blaugrund, S. (1996). Singing power ratio: quantitative evaluation of singing voice quality. *Journal of Voice, 10*, 228-235.

Reid, K. L. P., Davis, P., Oates, J., Cabrera, D., Ternström, S., Black, M., & Chapman, J. (2007). The acoustic characteristics of professional opera singers performing in chorus versus solo mode. *Journal of Voice, 21*, 35-45.

Seidner, W., Schutte, H. K., Wendler, J., & Rauhut, A. (1985). Dependence of the high singing formant on pitch and vowel in different voice types. In A. Askenfelt et al. (Eds.), *Proceedings of the Stockholm Music Acoustics Conference 1983* (pp. 261–268).

Sengupta, R. (1990). Study of some aspects of the `singer's formant' in North Indian classical singing. *Journal of Voice, 4*, 29-34.

Shutte, H., & Miller, R. (1985). Individual parameters of the singer's formant. *Folia Phoniatrica, 37*, 31-35.

Sundberg, J. (1970). The level of the 'singing formant' and the source spectra of professional bass singers. *STL-QPSR, 4*, 21-39.

Sundberg, J. (1972). An articulatory interpretation of the 'singing formant'. *STL-QPSR, 13*, 45-53.

Sundberg, J. (1973). The sourcce spectrum in professional singing. *Folia Phoniatrica, 25*, 71-90.

Sundberg, J. (1974). Articulatory interpretation of the "singing formant". *Journal of the Acoustical Society of America, 55*, 838-844.

Sundberg, J. (1995). The singer's formant revisited. *STL-QPSR, 36*, 83-96.

Sundberg, J. (2001). Level and center frequency of the singer's formant. *Journal of Voice, 15*, 176-186.

Sundberg, J. (2007). Sopranos with a singer's formant? Historical, physiological, and acoustical aspects of castrato singing. *TMH-QPSR, KTH, 49*, 1-6.

Ternström, S., Bohman, M., & Sodersten, M. (2006). Loud speech over noise: some spectral attributes, with genre differences. *Journal of the Acoustical Society of America, 119*, 1648-1665.

Watts, Ch., Barnes-Burroughs, K., Estis, J., & Blanton, D. (2006). The singing power ratio as an objective measure of singing voice quality in untrained talented and nontalented singers. *Journal of Voice, 20*, 82-88.

PhonoViz: Chroma Feature Visualization for Hindustani Classical Music

Sauhaarda Chowdhuri
Resoniq Research
sauhaarda@resoniq.com

Abstract

Hindustani Classical music is an improvisational form of music based on melodic frameworks called ragas which are passed down from teacher to student in a fading oral tradition. PhonoViz aims to provide live visualization and feedback for a singer's treatment of a raga in addition to digital documentation of Hindustani classical music. To accomplish this, a deep convolutional network is trained to predict the raga from approximately two-minute chunks of Hindustani classical vocal music. The PhonoViz algorithm provides a method for visualizing the saliency of various melodic phrases as they relate to the network's prediction. This visualization method when evaluated on a convolutional network for raga prediction demonstrates a 72.8% signal to noise isolation performance according to a new proposed metric and clearly identifies characteristic melodic phrases in validation audio input.

1 Introduction

Hindustani classical music is a fading art form which focuses predominantly on improvisations on specific melodic modes, or ragas. There are hundreds of Indian ragas of different forms, but because the music is passed down through an ancient oral tradition, much of the rich knowledge of Indian ragas is dying as the guru-shishya oral tradition fades [14].

Although each raga is confined to a specific scale in which only certain notes can be used. Multiple ragas may overlap and have the same scale. Ragas are truly defined by a specific set of improvisational patterns, referred to as the *chalan* of a raga. The chalan includes specific transitions and ornamentations given to notes, known as *gamakas*. In the same raga, the chalan may call for a smooth *meend*, or glide, between a pair of notes, while using a *khatka*, or broken, ornamentation between another set of notes. These subtle differences in the treatment of notes are critical when distinguishing ragas.

The infinitesimal variations of these ornamentations and the improvisational nature of the Indian ragas make ragas very difficult to consistently record in a distributable and universal written format. Therefore, preserving information regarding the nature of each raga is a difficult task.

To solve this problem, we propose PhonoViz, a deep learning based system for the visualization of key chromagram based features in audio data, which allows for the preservation and documentation of complex improvisational music sources digitally. The system consists of a deep convolutional network for raga prediction paired with a saliency visualization algorithm which allows for extraction and visualization of salient audio features from an audio input.

2 Background

Existing research has been done on the prediction of raga from audio data using machine learning methods, like SVM, clustering, and deep neural networks [2–4, 7–13, 15]. Although these works are effective at the prediction of raga information from audio data, the raga prediction itself is not a useful tool for the preservation or documentation of the Indian music.

Significant research has been done into visualizing the saliency of convolutional neural networks for image processing tasks, like autonomous driving and image classification [1, 16, 18]. One common approach is the visualization of activations or first-layer weights [6]. This technique yields very sparse activation maps which can be difficult to interpret. Another strategy is to visualize the weights or filters of a convolutional neural network. The weights become less interpretable deeper into a convolutional neural network, and the first layer weights tend to visualize into simple patterns like edge detection filters common between multiple tasks. A more effective approach is a technique in which inputs are organized by how they activate a particular ReLU within the model [5]. This approach is limited by the fact that ReLU neurons do not have inherent semantic meaning in themselves, reducing the specific impact and interpretation of the visualizations. A final approach is an occlusion method in which a patch of zeros is moved across the input image, occluding parts of the input data [18]. If the resultant network prediction is mapped across various occlusions, a heat map can be generated to isolate the most important audio segments. Even this occlusion method has limitations when applied to the scope of audio inputs, as an audio occlusion visualization can easily identify important segments of a longer piece but cannot identify specific movements, transitions, and notes within a small interval which contribute to the network's prediction.

A related work, Bojarski et al. [1], describes a visualization method for increasing confidence in autonomous driving convolutional deep neural networks. This algorithm uses scaled copies of receptive fields of each of the convolutional layers of the network to compute a saliency "overlay" on top of the input image, to ensure the deep network is identify key objects. Although a similar visualization technique is used, the PhonoViz method differs from this work. The first obvious difference is that PhonoViz applies to audio chromagram data which differs significantly from the multi-channel RGB input to the NVIDIA driving network. Next, PhonoViz attempts the problem of feature extraction, rather than computation of a saliency overlay. Due to these different applications, the PhonoViz algorithm includes a final Hadamard product computation between the original audio input and computed saliency activation, which is not described in Bojarski et al. [1], allowing for the isolation of only features which are originally present in the original input. Finally, our work provides a mathematical definition for the visualization method, explaining beyond simple intuition the function of the visualization algorithm.

3 Methodology

3.1 Chromagram Input

To provide saliency information in a human-readable format, the audio input to the convolutional neural network is provided as an input chromagram. Chromagrams divide an audio sample into twelve frequency bins, one for each note in the equal-tempered scale. The values in the chromagram represent the relative intensity of one of the twelve pitch classes at a given interval of time. Chromagrams are computed using a Short Term Fourier Transform (STFT).

One limitation of the chromagram is its lack of embedded octave information. While octave, or *saptak*, information is important in Indian classical music, it is secondary to pitch transitions and frequencies within a single octave. Additionally, due to the octave equivalency effect, notes separated by octaves are perceived as roughly equivalent so the loss of octave information does not significantly hinder the understanding of the raga.

The full audio chromagrams from the Hindustani Music Raga Recognition Dataset were generated and broken into 150 second chunks for processing by our convolutional neural network. This was done to ensure a consistent input chromagram size to the networks that is limited in it's size in the temporal domain.

3.2 Network Architecture and Training

The PhonoViz prediction network used has four convolutional blocks and one fully connected layer as shown in Figure 1. A visualization of the saliency visualization algorithm described in Section 3.4 is also shown in this figure.

This network was trained for 800 epochs on the chunked chromagram data. This network achieved an accuracy of 78.9 % on a reserved 10% validation set from the Hindustani Music Raga Recognition Dataset when tasked with predicting one of 30 ragas from a 150 second chunk audio input. This trained network is the basis for all of the visualizations discussed in this work.

3.3 Network Definition

First we define a mathematical representation of the deep convolutional neural network. Let x be a chromagram which represents the input to the network. The prediction network may be represented as an approximation function:

$$F(x) = (FC_n \circ FC_{n-1} \circ \dots FC_1) \circ (C_m \circ C_{m-1} \circ \dots C_1)(x)$$

where C_i represents the i^{th} convolutional block with a block including Convolution, Batch Normalization, Pooling, and a Rectified Linear Unit (ReLU). FC_i represents the i^{th} fully connected layer. In the PhonoViz prediction network, there are $n = 1$ fully connected layer and $m = 4$ convolutional blocks.

3.4 Saliency Visualization Algorithm

Our interest in this work is the computation of the saliency, or the relative influence each part of the input x has on the resulting output of the approximator $F(x)$. In this way, we can identify a set of salient values $x_{sal} \in x$. Our work aims to visualize the output of the chromagram using a technique which examines the relative impact each individual pixel has on an output prediction at a given layer.

Instead of directly calculating the saliency of the inputs to $F(x)$, computing $S(F, x)$, this work instead attempts approximate this value by mapping the saliency of the input to the convolutional blocks of the network, $C(x)$. $S(F, x) \approx S(C, x)$ is a reasonable assumption as the convolutional section of a deep convolutional network is designed primarily for feature extraction with filters, thus the input saliency into this section of the network will give a good insight into the overall saliency of the deep network.

To compute $S(C, x)$, we employ a recursive sum, scale, and multiply operation. This may be represented as follows:

$$S(C_{1-m}, x) = S(C_{1-(m-1)}, x) \odot U(\overline{C_{1-m}(x)})$$

where $U(x)$ is the nearest-neighbor upsampling and $\overline{C_{1-m}(x)}$ represents the activation (mean over channel axis) for the m^{th} convolutional layer. Upsampling is done to ensure the two activation maps inputted to the Hadamard product are of the same size. In this way, when the full recursive equation is evaluated, the final output will have the same size as the original input x.

If $S(C, x)$ is recursively evaluated in this way, an image will be obtained giving the relative importance of each pixel to the output of the convolutional block $C(x)$. This image can then be multiplied with the original input chromagram via a hadamard product to remove saliency from pixels not activated in the input chromagram.

3.5 Key Phrase Identification

Even though the input to our networks are chunked into 150 second segments, it is still difficult to visualize and read chromagrams greater than 10 seconds in length. Thus, before visualizing the extracted features from a chunk of audio data using the saliency visualization algorithm, the PhonoViz system

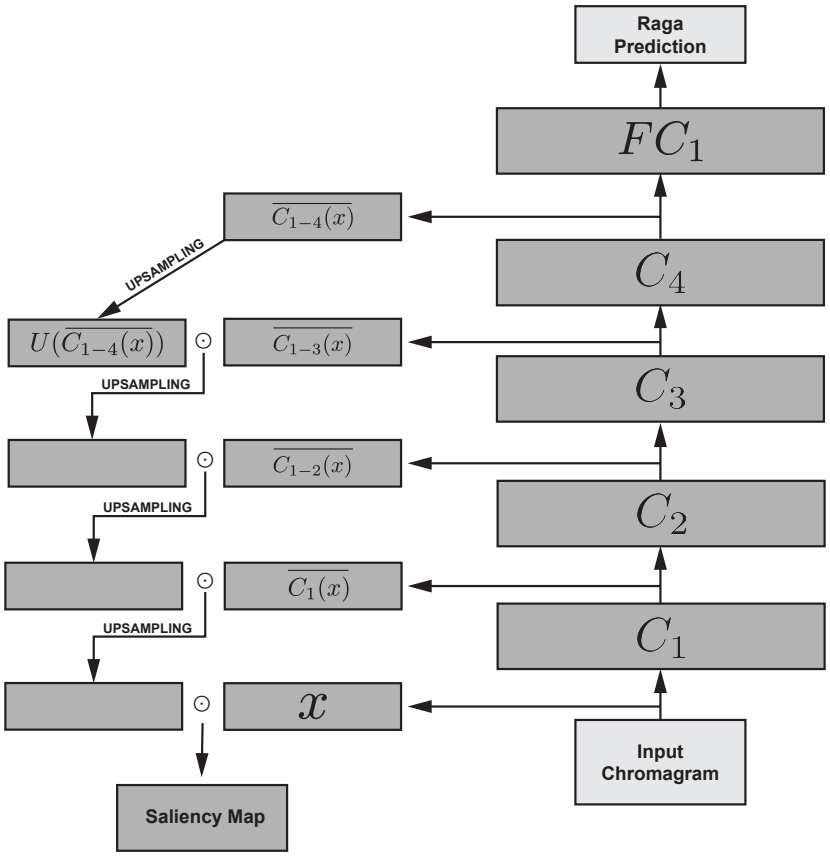


Figure 1: PhonoViz Network and Algorithm Diagram

includes a key phrase identification algorithm to extract key segments of audio data from the input chromagram chunk.

Key phrases are detected using an occlusion method, similar to that described in Zeiler and Fergus [18]. Instead of a small occlusion filter that is slid both vertically and horizontally across the image, the PhonoViz occlusion method instead zeroes ten second phrases from the input chromagram and evaluates the percentage loss in prediction accuracy when a segment is occluded. In this way, the segment whose removal results in the greatest loss of accuracy is the most key phrase necessary for prediction of the raga. The visualizations shown in Section 4 are shown for sections of audio that the key phrase identification occlusion method has determined to have an accuracy loss of at least 50%.

4 Results

4.1 Qualitative Visualization Analysis

The PhonoViz system was evaluated on several sets of validation data from the Hindustani Music Raga Recognition Dataset as well as from unseen samples of Indian classical performances from the Dunya Compmusic Hindustani Corpus [17]. This section will analyze a few of the visualizations produced by the PhonoViz system to better understand the algorithm's capabilities.

Figure 2 depicts the key phrase with the highest occlusion accuracy drop from a song in Raga Bhupali by Pt. KG Ghinde. This specific phrase is notable as it includes an interesting feature of many Indian classical performances: audience participation and applause during the performance itself. In the 10 second audio clip, the last three seconds includes this applause which makes the performer inaudible. The PhonoViz visualization system is able to recognize this noise from the data, and exhibits negligible activation during this section. The visualization algorithm is also able to isolate several key transitions in the Raga Bhupali from a rapid melodic movement, or *taan*, earlier within the phrase. The movement from the tonic note (Sa) to the next note (Re) moves slightly up smoothly touching the third note (Ga) before settling on Re. This transition can be seen in the visualization by a slight activation in Ga between Sa and Re at approximately 6 seconds.

Figure 3 depicts another key phrase from a song in Raga Yaman Kalyan by Pt. Bhimsen Joshi. This visualization may

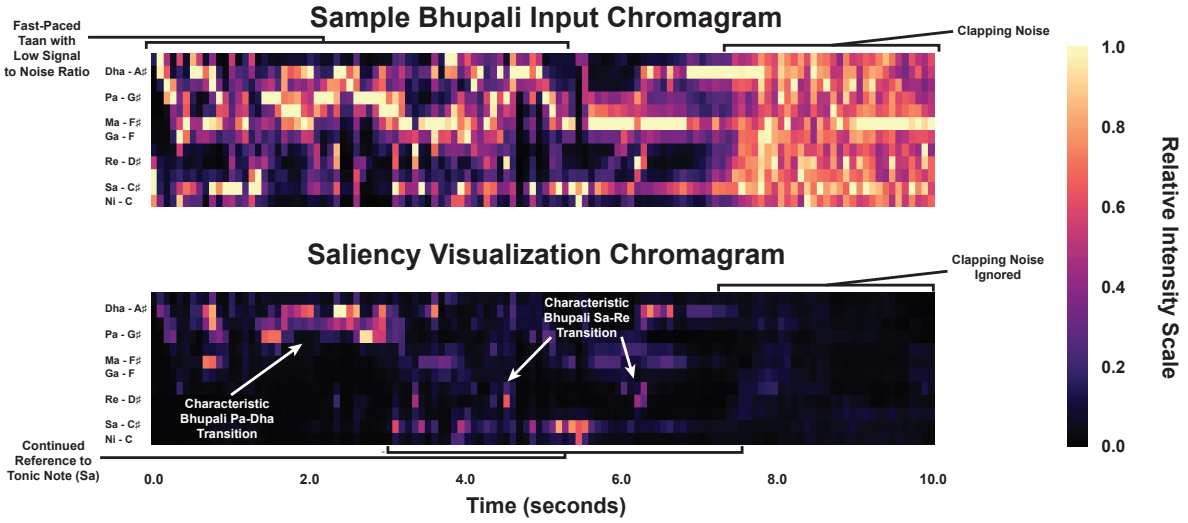


Figure 2: PhonoViz Bhupali Visualization

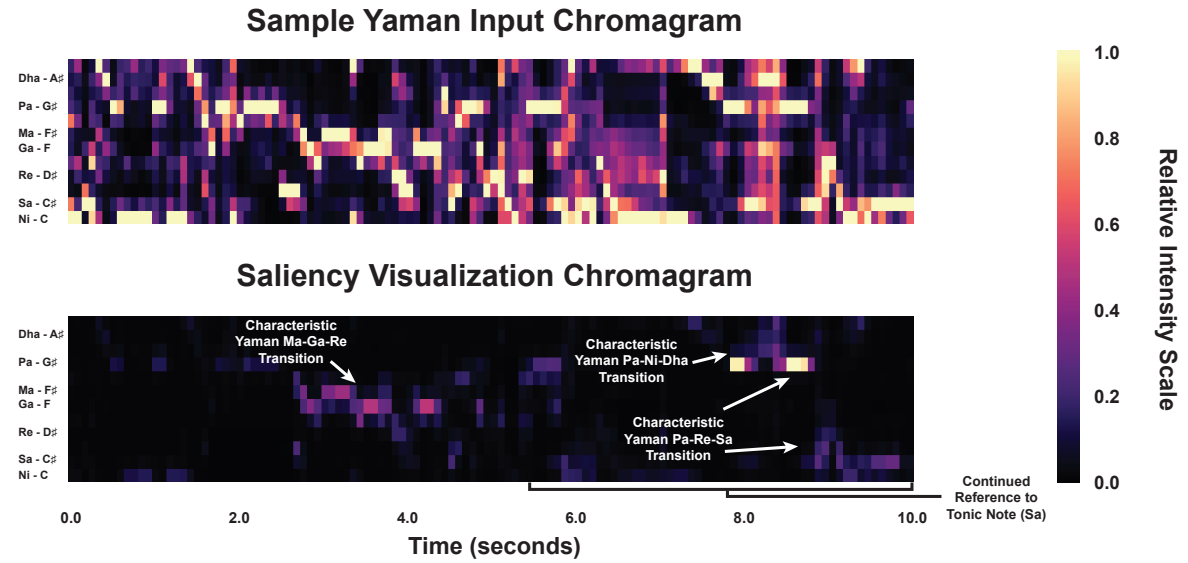


Figure 3: PhonoViz Yaman Visualization

be noted for it's isolation of characteristic phrases of the audio sequence. The known chalan of this popular raga includes a phrase highlighted by our visualization algorithm: Pa Ni Dha Pa Re Sa. In addition to highlighting this characteristic series of notes, the algorithm is also able to visualize the subtle transitions between these pitches, something difficult to accomplish in a written format.

4.2 Noise Reduction

In our qualitative analysis, it was clear that the PhonoViz algorithm can make a distinction between noisy and salient data. To quantify this ability, a test was performed in which half of the chromagram input to the network was filled with random Gaussian noise. The entire validation set was then run through the PhonoViz algorithm and the percent activation of the noise section of chromagram input and the audio input section of the chromagram was compared. Our results

clearly indicate the PhonoViz system's noise reduction capability, as it is able to attune to real audio 72.8% more than the generated noise. Thus, this system is quite promising for isolation of relevant audio features.

5 Conclusion

The PhonoViz visualization technique proposed in this work can effectively highlight important transitions characteristic to specific ragas of Indian classical music and demonstrates a 72.8% difference in it's activation for real audio clips in comparison to generated noise, making the system robust in its isolation of important audio features. The proposed system is easily adaptable and can be extended to other audio processing tasks involving the chromagram representation. This work focused only on Hindustani Classical vocal music. Future work can extend the scope of the PhonoViz algorithm to the South Indian Carnatic Classical Music and additional world music genres.

References

[1] M. Bojarski, P. Yeres, A. Choromanska, K. Choromanski, B. Firner, L. D. Jackel, and U. Muller. Explaining how a deep neural network trained with end-to-end learning steers a car. *CoRR*, abs/1704.07911, 2017. URL http://arxiv.org/abs/1704.07911.

[2] S. Chakraborty, G. Mazzola, S. Tewari, and M. Patra. *Computational musicology in Hindustani music*. Springer, 2014.

[3] V. N. Degaonkar and A. V. Kulkarni. Automatic raga identification in indian classical music using the convolutional neural network. *Journal of Engineering Technology*, 6(2):564–576, 2018.

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

[5] R. B. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. *CoRR*, abs/1311.2524, 2013. URL http://arxiv.org/abs/1311.2524.

[6] G. Hinton, S. Osindero, M. Welling, and Y.-W. Teh. Unsupervised discovery of nonlinear structure using contrastive backpropagation. *Cognitive science*, 30(4):725–731, 2006.

[7] R. Joseph and S. Vinod. Carnatic raga recognition. *Indian Journal of Science and Technology*, 10(13), 2017.

[8] P. Kirthika and R. Chattamvelli. A review of raga based music classification and music information retrieval (mir). In *Engineering Education: Innovative Practices and Future Trends (AICERA), 2012 IEEE International Conference on*, pages 1–5. IEEE, 2012.

[9] V. Kumar, H. Pandya, and C. Jawahar. Identifying ragas in indian music. In *Pattern Recognition (ICPR), 2014 22nd International Conference on*, pages 767–772. IEEE, 2014.

[10] S. S. Manjabhat, S. G. Koolagudi, K. S. Rao, and P. B. Ramteke. Raga and tonic identification in carnatic music. *Journal of New Music Research*, 46(3):229–245, 2017. doi: 10.1080/09298215.2017.1330351. URL https://doi.org/10.1080/09298215.2017.1330351.

[11] G. Pandey, C. Mishra, and P. Ipe. Tansen: A system for automatic raga identification. In *IICAI*, pages 1350–1363, 2003.

[12] J. C. Ross, A. Mishra, K. K. Ganguli, P. Bhattacharyya, and P. Rao. Identifying raga similarity through embeddings learned from compositions'notation. In *Proceedings of the 18th International Society for Music Information Retrieval Conference, ISMIR 2017, Suzhou, China*, pages 515–522, 2017.

[13] J. Salamon, S. Gulati, and X. Serra. A multipitch approach to tonic identification in indian classical music. In *Gouyon F, Herrera P, Martins LG, Müller M. ISMIR 2012: Proceedings of the 13th International Society for Music Information Retrieval Conference; 2012 Oct 8-12; Porto, Portugal. Porto: FEUP Ediçoes; 2012.* International Society for Music Information Retrieval (ISMIR), 2012.

[14] H. Schippers. The guru recontextualized? perspectives on learning north indian classical music in shifting environments for professional training. *Asian music*, pages 123–138, 2007.

[15] S. Shetty and K. Achary. Raga mining of indian music by extracting arohana-avarohana pattern. *International Journal of Recent Trends in Engineering*, 1(1): 362, 2009.

[16] K. Simonyan, A. Vedaldi, and A. Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

[17] A. Srinivasamurthy, G. K. Koduri, S. Gulati, V. Ishwar, and X. Serra. Corpora for music information research in indian art music. In *International Computer Music Conference/Sound and Music Computing Conference*, pages 1029–1036, Athens, Greece, 14/09/2014 2014. URL http://hdl.handle.net/10230/35356.

[18] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. *CoRR*, abs/1311.2901, 2013. URL http://arxiv.org/abs/1311.2901.