Report from Dagstuhl Seminar 19052

Computational Methods for Melody and Voice Processing in Music Recordings

Edited by

Meinard Müller[1], Emilia Gómez[2], and Yi-Hsuan Yang[3]

1 **Universität Erlangen-Nürnberg, DE,** meinard.mueller@audiolabs-erlangen.de
2 **UPF – Barcelona, ES,** emilia.gomez@upf.edu
3 **Academia Sinica – Taipei, TW,** yang@citi.sinica.edu.tw

Abstract

In our daily lives, we are constantly surrounded by music, and we are deeply influenced by music. Making music together can create strong ties between people, while fostering communication and creativity. This is demonstrated, for example, by the large community of singers active in choirs or by the fact that music constitutes an important part of our cultural heritage. The availability of music in digital formats and its distribution over the world wide web has changed the way we consume, create, enjoy, explore, and interact with music. To cope with the increasing amount of digital music, one requires computational methods and tools that allow users to find, organize, analyze, and interact with music–topics that are central to the research field known as *Music Information Retrieval* (MIR). The Dagstuhl Seminar 19052 was devoted to a branch of MIR that is of particular importance: processing melodic voices (with a focus on singing voices) using computational methods. It is often the melody, a specific succession of musical tones, which constitutes the leading element in a piece of music. In the seminar we discussed how to detect, extract, and analyze melodic voices as they occur in recorded performances of a piece of music. Gathering researchers from different fields, we critically reviewed the state of the art of computational approaches to various MIR tasks related to melody processing including pitch estimation, source separation, singing voice analysis and synthesis, and performance analysis (timbre, intonation, expression). This triggered interdisciplinary discussions that leveraged insights from fields as disparate as audio processing, machine learning, music perception, music theory, and information retrieval. In particular, we discussed current challenges in academic and industrial research in view of the recent advances in deep learning and data-driven models. Furthermore, we explored novel applications of these technologies in music and multimedia retrieval, content creation, musicology, education, and human-computer interaction. In this report, we give an overview of the various contributions and results of the seminar. We start with an executive summary, which describes the main topics, goals, and group activities. Then, we present a more detailed overview of the participants' contributions (listed alphabetically by their last names) as well as of the ideas, results, and activities of the group meetings, the demo, and the music sessions.

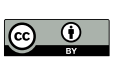
Seminar January 27–February 1, 2019 – http://www.dagstuhl.de/19052

2012 ACM Subject Classification Information systems → Music retrieval, Applied computing → Sound and music computing

Keywords and phrases Acoustics of singing, audio signal processing, machine learning, music composition and performance, music information retrieval, music perception and cognition, music processing, singing voice processing, sound source separation, user interaction and interfaces

Digital Object Identifier 10.4230/DagRep.9.1.125

Edited in cooperation with Frank Zalkow



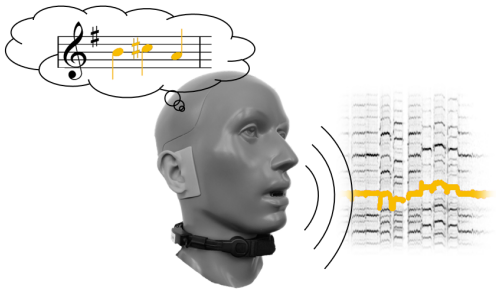
Computational Methods for Melody and Voice Processing in Music Recordings, *Dagstuhl Reports*, Vol. 9, Issue 1, pp. 125–177
Editors: Meinard Müller, Emilia Gómez, and Yi-Hsuan Yang

Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

1 Executive Summary

Meinard Müller (Universität Erlangen-Nürnberg, DE)
Emilia Gómez (UPF – Barcelona, ES)
Yi-Hsuan Yang (Academia Sinica – Taipei, TW)

In this executive summary, we give an overview of computational melody and voice processing and summarize the main topics covered in this seminar. We then describe the background of the seminar's participants, the various activities, and the overall organization. Finally, we reflect on the most important aspects of this seminar and conclude with future implications and acknowledgments.



Overview

When asked to describe a specific piece of music, we are often able to sing or hum the main melody. In general terms, a *melody* may be defined as a linear succession of musical tones expressing a particular musical idea. Because of the special arrangement of tones, a melody is perceived as a coherent entity, which gets stuck in a listener's head as the most memorable element of a song. As the original Greek term *melōidía* (meaning "singing" or "chanting") implies, a melody is often performed by a human voice. Of course, a melody may also be played by other instruments such as a violin in a concerto or a saxophone in a jazz piece. Often, the melody constitutes the leading element in a composition, appearing in the foreground, while the accompaniment is in the background. Sometimes melody and accompaniment may even be played on a single instrument such as a guitar or a piano. Depending on the context and research discipline (e. g., music theory, cognition or engineering), one can find different descriptions of what may be meant by a melody. Most people would agree that the melody typically stands out in one way or another. For example, the melody often comprises the higher notes in a musical composition, while the accompaniment consists of the lower notes. Or the melody is played by some instrument with a characteristic timbre. In some performances, the notes of a melody may feature easily discernible time–frequency patterns such as vibrato, tremolo, or glissando. In particular, when considering performed music given in the form of audio signals, the detection, extraction, separation, and analysis of melodic voices becomes a challenging research area with many yet unsolved problems. In the following, we discuss some MIR tasks related to melody processing, indicating their relevance for fundamental research, commercial applications, and society.

The problem of detecting and separating melodic voices in music recordings is closely related to a research area commonly referred to as *source separation*. In general, audio signals are complex mixtures of different sound sources. The sound sources can be several people talking simultaneously in a room, different instruments playing together, or a speaker talking in the foreground with music being played in the background. The general goal of source separation is to decompose a complex sound mixture into its constituent components. Source separation methods often rely on specific assumptions such as the availability of multiple channels, where several microphones have been used to record the acoustic scene from different directions. Furthermore, the source signals to be identified are assumed to be independent in a statistical sense. In music, however, such assumptions are not applicable in many cases. For example, musical sound sources may outnumber the available information channels, such as a string quartet recorded in two-channel stereo. Also, sound sources in music are typically highly correlated in time and frequency. Instruments follow the same rhythmic patterns and play notes which are harmonically related. This makes the separation of musical voices from a polyphonic sound mixture an extremely difficult and generally intractable problem.

When decomposing a music signal, one strategy is to exploit music-specific properties and additional musical knowledge. In music, a source might correspond to a melody, a bass line, a drum track, or a general instrumental voice. The separation of the melodic voice, for example, may be simplified by exploiting the fact that the melody is often the leading voice, characterized by its dominant dynamics and by its temporal continuity. The track of a bass guitar may be extracted by explicitly looking at the lower part of the frequency spectrum. A human singing voice can often be distinguished from other musical sources due to characteristic time–frequency patterns such as vibrato. Besides such acoustic cues, score-informed source separation strategies make use of the availability of score representations to support the separation process. The score provides valuable information in two respects. On the one hand, pitch and timing of note events provide rough guidance within the separation process. On the other hand, the score offers a natural way to specify the target sources to be separated.

In this seminar, we discussed source separation techniques that are particularly suited for melodic voices. To get a better understanding of the problem, we approached source separation from different directions including model-based approaches that explicitly exploit acoustic and musical assumptions as well as data-driven machine learning approaches.

Given a music recording, melody extraction is often understood in the MIR field as the task of extracting a trajectory of frequency values that correspond to the pitch sequence of the dominant melodic voice. As said before, melody extraction and source separation are highly related: while melody extraction is much easier if the melodic source can be isolated first, the source separation process can be guided if the melodic pitch sequence is given a priori. However, both tasks have different goals and involve different challenges. The desired output of melody extraction is a trajectory of frequency values, which is often sufficient information for retrieval applications (e. g., query-by-humming or the search of a musical theme) and performance analysis. In contrast, for music editing and audio enhancement applications, source separation techniques are usually needed.

In the seminar, we addressed different problems that are related to melody extraction. For example, the melody is often performed by a solo instrument, which leads to a problem also known as *solo–accompaniment separation*. The estimation of the fundamental frequency of a quasi-periodic signal, termed *mono-pitch estimation*, is a long-studied problem with applications in speech processing. While mono-pitch estimation is now achievable with

reasonably high accuracy, the problem of *multi-pitch estimation* with the objective of estimating the fundamental frequencies of concurrent periodic sounds remains very challenging. This particularly holds for music signals, where concurrent notes stand in close harmonic relation. For extreme cases such as complex orchestral music where one has a high level of polyphony, multi-pitch estimation becomes intractable with today's methods.

Melodic voices are often performed by singers, and the singing voice is of particular importance in music. Humans use singing to create an identity, express their emotions, tell stories, exercise creativity, and connect while singing together. Because of its social, cultural, and educational impact, singing plays a central role in many parts of our lives, it has a positive effect on our health, and it creates a link between people, disciplines, and domains (e. g., music and language). Many people are active in choirs, and vocal music makes up an important part of our cultural heritage. In particular in Asian countries, karaoke has become a major cultural force performed by people of all age groups. Singing robots, vocaloids, or synthesizers such as Hatsune Miku[1] have made their way into the mass market in Japan. Thanks to digitization and technologies, the world wide web has become an important tool for amateur and professional singers to discover and study music, share their performances, get feedback, and engage with their audiences. An ever-increasing amount of music-related information is available to singers and singing enthusiasts, such as music scores[2] as well as audio and video recordings.[3] Finally, music archives contain an increasing number of digitized audio collections of historic value from all around the world such as Flamenco music, Indian art music, Georgian vocal music, or Bejing Opera performances.

Due to its importance, we placed in our seminar a special emphasis on music technologies related to singing. This involves different research areas including singing analysis, description, and modeling (timbre, intonation, expression), singing voice synthesis and transformation, voice isolation/separation, and singing performance rating. Such research areas require a deep understanding of the way people produce and perceive vocal sounds. In our seminar, we discussed such issues with researchers having a background in singing acoustics and music performance.

Over the last years, as is also the case for other multimedia domains, many advances in music and audio processing have benefited from new developments in machine learning.

In particular, deep neural networks (DNNs) have found their way into MIR and are applied with increasing success to various MIR tasks including pitch estimation, melody extraction, sound source separation, and singing voice synthesis. The complex spectro-temporal patterns and relations found in music signals make this domain a challenging testbed for such new machine learning techniques. Music is different from many other types of multimedia. In a static image, for example, objects may occlude one another with the result that only certain parts are visible. In music, however, concurrent musical events may superimpose or blend each other in a more complicated way. Furthermore, as opposed to static images, music depends on time. Music is organized in a hierarchical way ranging from notes, bars, and motifs, to entire sections. As a result, one requires models that capture both short-term and long-term dependencies in music.

---

[1] https://en.wikipedia.org/wiki/Hatsune_Miku

[2] For example, the Choral Public Domain Library currently hosts free scores of at least 24963 choral and vocal works by at least 2820 composers, see http://www.cpdl.org/

[3] See, for example, the material hosted at platforms such as YouTube or SoundCloud

In the seminar, we looked at the new research challenges that arise when designing music-oriented DNN architectures. Furthermore, considering the time-consuming and labor-intensive process of collecting human annotations of musical events and attributes (e. g., timbre, intonation, expression) in audio recordings, we addressed the issue of gathering large-scale annotated datasets that are needed for DNN-based approaches.

Participants and Group Composition

In our seminar, we had 32 participants, who came from various locations around the world including North America (4 participants from the U.S.), Asia (4 participants from Japan, 2 from Taiwan, 2 from Singapore, 1 from Korea, 1 from India), and Europe (18 participants from France, Germany, Netherlands, Spain, United Kingdom). More than half of the participants came to Dagstuhl for the first time and expressed enthusiasm about the open and retreat-like atmosphere. Besides its international character, the seminar was also highly interdisciplinary. While most of the participating researchers are working in the field of music information retrieval, we also had participants with a background in musicology, acoustics, machine learning, signal processing, and other fields. By having experts working in technical as well as in non-technical disciplines, our seminar stimulated cross-disciplinary discussions, while highlighting opportunities for new collaborations among our attendees. Most of the participants had a strong musical background, some of them even having a dual career in an engineering discipline and music. This led to numerous social activities including singing and playing music together. In addition to geographical locations and research disciplines, we tried to foster variety in terms of seniority levels and presence of female researchers. In our seminar, 10 of the 32 participants were female, including three key researchers (Anja Volk, Emilia Gómez, and Johanna Devaney) from the "Women in Music Information Retrieval" (WiMIR)[4] initiative.

In conclusion, by gathering internationally renowned scientists as well as younger promising researchers from different research areas, our seminar allowed us to gain a better understanding of the problems that arise when dealing with a highly interdisciplinary topic such as melody and voice processing–problems that cannot be addressed by simply using established research in signal processing or machine learning.

Overall Organization and Schedule

Dagstuhl seminars are known for having a high degree of flexibility and interactivity, which allows participants to discuss ideas and to raise questions rather than to present research results. Following this tradition, we fixed the schedule during the seminar asking for spontaneous contributions with future-oriented content, thus avoiding a conference-like atmosphere, where the focus tends to be on past research achievements. After the organizers gave an overview of the Dagstuhl concept and the seminar's overall topic, we started the first day with self-introductions, where all participants introduced themselves and expressed their expectations and wishes for the seminar. We then continued with a small number of short (15 to 20 minutes) stimulus talks, where specific participants were asked to address some critical questions on melody and voice processing in a nontechnical fashion. Each of

[4] https://wimir.wordpress.com/

these talks seamlessly moved towards an open discussion among all participants, where the respective presenters took over the role of a moderator. These discussions were well received and often lasted for more than half an hour. The first day closed with a brainstorming session on central topics covering the participants' interests while shaping the overall schedule and format for the next day. On the subsequent days, we continued having stimulus tasks interleaved with extensive discussions. Furthermore, we split into smaller groups, each group discussing a more specific topic in greater depth. The results and conclusions of these parallel group sessions, which lasted between 60 to 90 minutes, were then presented and discussed with the plenum. This mixture of presentation elements gave all participants the opportunity for presenting their ideas while avoiding a monotonous conference-like presentation format. On the last day, the seminar concluded with a session we called "self-outroductions" where each participant presented his or her personal view on the seminar's results.

Additionally to the regular scientific program, we had several additional activities. First, we had a demo session on Thursday evening, where participants presented user interfaces, available datasets, and audio examples of synthesized singing voices. One particular highlight was the incorporation of singing practice in the seminar. In particular, we carried out a recording session on Wednesday afternoon, where we recorded solo and polyphonic singing performed by Dagstuhl participants. The goal of this recording session was to contribute to existing open datasets in the area of music processing. The singers were recorded with different microphone types such as throat and headset microphones to obtain clean recordings of the individual voices. All participants agreed that the recorded dataset should be made publicly available for research purposes. As preparation for these recordings, we assembled a choir consisting of ten to twelve amateur singers (all Dagstuhl participants) covering different voice sections (soprano, alto, tenor, bass). In the lunch breaks and the evening hours, the group met for regular rehearsals to practice different four-part choral pieces. These musical activities throughout the entire week not only supported the theoretical aspects of the seminar but also had a very positive influence on the group dynamics. Besides the recordings, we also had a concert on Thursday evening, where various participant-based ensembles performed a variety of music including classical music and folk songs.

Conclusions and Acknowledgment

Having a Dagstuhl seminar, we gathered researchers from different fields including information retrieval, signal processing, musicology, and acoustics. This allowed us to approach the problem of melody and voice processing by looking at a broad spectrum of data analysis techniques (including signal processing, machine learning, probabilistic models, user studies), by considering different domains (including text, symbolic, image, audio representations), and by drawing inspiration from the creative perspectives of the agents (composer, performer, listener) involved. As a key result of this seminar, we achieved some substantial progress towards understanding, modeling, representing, and extracting melody- and voice-related information using computational means.

The Dagstuhl seminar gave us the opportunity for having interdisciplinary discussions in an inspiring and retreat-like atmosphere. The generation of novel, technically oriented scientific contributions was not the main focus of the seminar. Naturally, many of the contributions and discussions were on a conceptual level, laying the foundations for future projects and collaborations. Thus, the main impact of the seminar is likely to take place in the medium and long term. Some more immediate results, such as plans to share research

data and software, also arose from the discussions. In particular, we plan to make the dataset recorded during the Dagstuhl seminar available to the research community. As further measurable outputs from the seminar, we expect to see several joint papers and applications for funding.

Beside the scientific aspect, the social aspect of our seminar was just as important. We had an interdisciplinary, international, and very interactive group of researchers, consisting of leaders and future leaders in our field. Many of our participants were visiting Dagstuhl for the first time and enthusiastically praised the open and inspiring setting. The group dynamics were excellent with many personal exchanges and common activities. Some scientists expressed their appreciation for having the opportunity for prolonged discussions with researchers from neighboring research fields–something that is often impossible during conference-like events.

In conclusion, our expectations for the seminar were not only met but exceeded, in particular concerning networking and community building. We want to express our gratitude to the Dagstuhl board for giving us the opportunity to organize this seminar, the Dagstuhl office for their exceptional support in the organization process, and the entire Dagstuhl staff for their excellent service during the seminar. In particular, we want to thank Susanne Bach-Bernhard, Annette Beyer, Michael Gerke, and Michael Wagner for their assistance during the preparation and organization of the seminar.

2 Table of Contents

Working Groups

3 Stimulus Talks and Further Topics

3.1 Challenges in Melodic Similarity

Rachel Bittner (Spotify – New York, US)

Many methods exist that generate reasonably accurate melody estimates from polyphonic music, particularly for vocal melodies. Looking beyond the problem of estimation itself, how do we measure the similarity between two estimated melodic sequences? In MIR we have most often chosen to estimate melody as a sequence of f0 curves. However, the vast majority of work on melodic similarity has been done in the symbolic domain and focused on perceptual similarity. Work on query-by-humming in some sense measures melodic similarity, but focuses on shorter excerpts in a retrieval setting. This raises a number of questions:

- Should melodic similarity be computed in the symbolic domain? If so, what is the state of f0-to-note-conversion algorithms?
- Can we compute melodic similarity directly in the f0 domain? How do we enforce robustness to differences in pitch curve characteristics and estimation errors?
- How do we build upon and incorporate the previous work on perceptual melodic similarity?

3.2 Singing Voice, Speech, or Something in Between

*Estefanía Cano Cerón (Fraunhofer IDMT, DE & A*STAR – Singapore, SG)*

Joint work of Estefanía Cano Cerón, Antonio Escamilla, Gustavo López Gil, Fernando Mora Ángel, José Ricardo Zapata

In the context of the ACMus research project, we are investigating automatic techniques for annotation and segmentation of digital archives of non-western music. In particular, we are focusing on a collection of traditional Colombian music compiled in the Músicas Regionales archive at the Universidad de Antioquia in Medellín, Colombia. Of particular interest are a series of recordings of vocal expressions of indigenous cultures native to different regions of the country. These vocal expressions can either be very close to speech (almost as reciting something), can include some melodic elements, or can be closer to the concept of singing voice from a Western music perspective (exhibiting a defined melodic line). From an automatic classification point of view, the traditional binary discrimination between speech and singing voice falls short, calling for a more general characterization of vocal expressions. For example, one may use a range $[0, 1]$ with 0 being pure speech, 1 being singing voice (from a Western music perspective), and allowing everything that falls in between these bounds. In this context, interesting research questions arise: Is the degree of melodic elements in these vocal expressions informative for the region where a recording was made? Can we conclude the functional aspects of these recordings (ritual, prayer, social, playful, healing) based on this characterization? Could these categories be shared between different cultures?

3.3 Frequency Measurements and Perceived Pitch in Vocal Productions

Michèle Castellengo (Sorbonne University – Paris, FR)

The research developed in the MIR community aims to automatically extract the fundamental frequency of sounds and match it with a notated musical score. In general, it is assumed that the perceived pitch matches the measured frequency. In the case of singing, this postulate raises several questions related to the acoustic characteristics of the vocal source (such as unstable emission; rich and complex harmonic), the cognitive criteria of the "human voice" category (male or female, vowel listening), as well as the mental scale of reference of the listener (in this case the equal temperament). In my presentation, I focused on three examples that illustrate the importance of interactions between acoustic, psychophysical, and cognitive aspects involved in the interpretation of results.

- Perceived pitch accuracy of bel canto vocal techniques (vibrato versus trill, pitch of very short vibrated notes), see [1].
- Competition between the perception of pitch due to the periodicity of the signal and the spectral pitches coming from the vowel formants (octavian singing, diphonic singing), see [2].
- Perceptual emergence of a voice without a real existence (the Sardinian quintina), see [3].

Furthermore, I discussed issues related to traditional music and interval measurement. The confrontation with traditional polyphonies questions the unconscious mental references that condition our musical listening. In particular measuring intervals in cents, if perfectly adapted to piano music, can mislead the researcher when it comes to natural intervals. As an example, one can find in [4, pp. 403–404] the interval of a natural minor third (315.61 cents) in a song from Cameroon. Could we imagine a spectral analysis method based on the search for the harmonics common to both sounds of an interval [4, pp. 55]? See also [4, pp. 418–420] for analysis of stable multiphonic sounds.

References

1 Christophe d'Alessandro and Michèle Castellengo. *The Pitch of Short-Duration Vibrato Tones*. Journal of the Acoustical Society of America, 95 (3), 1994, pp. 1617–1630.
2 Michèle Castellengo and Nathalie Henrich Bernardoni. *Interplay Between Harmonics and Formants in Singing: When Vowels Become Music*. Proceedings of the International Symposium on Musical Acoustics (ISMA), Le Mans, France, 2014, pp. 433–438.
3 Michèle Castellengo, Bernard Lortat-Jacob, and Gilles Léothaud. *Pitch Perception: Five Voices with Four Sardinian Singers*. Proceedings of the International Symposium on Musical Acoustics (ISMA), Perugia, Italy, 2001, pp. 351–354.
4 Michèle Castellengo. *Écoute Musicale et Acoustique*. Eyrolles, Paris, France, 2015.

3.4 Generative Models for Singing Voice

Pritish Chandna (UPF – Barcelona, ES)

In the demo session of the Dagstuhl seminar, I presented a novel methodology for the synthesis of expressive singing voices using the NUS-48E corpus [1]. The approach is based on a generative adversarial network (GAN) for synthesizing the singing voice [2]. Furthermore, I discussed strategies for evaluating the quality of the synthesized results, see also the discussion of the working group on "Subjective Evaluation and Objectives Metrics for Singing Voice Generation and Separation" (Section 4.5).

References

1 Zhiyan Duan, Haotian Fang, Bo Li, Khe Chai Sim, and Ye Wang. *The NUS Sung and Spoken Lyrics Corpus: A Quantitative Comparison of Singing and Speech*. Proceedings of the Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, Kaohsiung, Taiwan, 2013, pp. 1–9.
2 Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. *Generative Adversarial Networks*. Proceedings of the International Conference on Neural Information Processing Systems, Montreal, Canada, 2014, pp. 2672–2680.

3.5 Measuring Interdependence in Unison Choral Singing

Helena Cuesta (UPF – Barcelona, ES)

Choral singing is probably the most widespread type of singing [1], but there is still little research on the topic, especially from a computational perspective. In a choir, singers are influenced by other singers' performances in terms of pitch and timing. In [3], we investigate the synchronization between singers by analyzing fundamental frequency (f0) envelopes using two different features: the derivative of the f0 curves [2] and the deviation from the target pitch specified by the score. In the first case, synchronization is assumed to be linear and measured using the Pearson correlation coefficient; in the second case, we consider the mutual information measure. Results suggest that the mutual information is a better metric for this task because of its non-linear nature; however, the question of which is the most suitable metric to measure this aspect of ensemble singing remains open for discussion.

References

1 Johan Sundberg. *The Science of the Singing Voice*. Northern Illinois University Press, 1987.
2 Helena Cuesta, Emilia Gómez, Agustín Martorell, and Felipe Loáiciga. *Analysis of Intonation in Unison Choir Singing*. Proceedings of the International Conference on Music Perception and Cognition, 2018. http://doi.org/10.5281/zenodo.1319597.
3 Helena Cuesta and Emilia Gómez. *Measuring Interdependence in Unison Choir Singing*. Late Breaking Demo of the International Conference on Music Information Retrieval (ISMIR), Paris, France, 2018.

3.6 Can Listening Tests Help Understanding and Improving Data Models for Vocal Performance?

Johanna Devaney (Brooklyn College, US)

The flexibility of the singing voice affords its great expressivity, but this flexibility also makes it challenging to model it computationally. In my presentation, I described an experiment on singer identification based on note-level descriptors (related to pitch, timing, loudness, and timbre) that compared the results of an SVM-based computational model and a listening experiment [1]. In this experiment, I considered not only the results but also the general implications for singing research and music information retrieval research in general.

References

1 Johanna Devaney. *Inter- Versus Intra-Singer Similarity and Variation in Vocal Performance.* Journal of New Music Research. 45(3), 2016, pp. 252–264.

3.7 Measuring and Modelling Intonation and Temperaments

Simon Dixon (Queen Mary University of London, GB)

Joint work of Simon Dixon, Matthias Mauch, Dan Tidhar, Klaus Frieler, Emmanouil Benetos, Tillman Weyde, Jiajie Dai

Melody and voice are themes running through several current research projects. In the context of the Trans-Atlantic Platform Digging into Data Challenge project "Dig that Lick: Analysing Large-scale Data for Melodic Patterns in Jazz Performances," we are investigating the use of melodic patterns in jazz improvisation in a large corpus spanning a significant proportion of the recorded history of jazz. The research involves automatic extraction of the main melody voice [1], the recognition of musical structures, and their linkage to historical and social metadata [2]. This project informs the study of the transmission of musical ideas–in this instance "licks" and patterns, but more generally musical style [3]–across time and location.

This leads to another strand of research, on the singing voice, which began with a focus on intonation and the effects of musical context on singers' ability to sing in tune [4, 5, 6, 7]. Wider questions are now being addressed to aid the study of performance and style, including the development of suitable features and representations for singing which capture and allow the comparison of continuous pitch trajectories, their segmentation, the use of dynamics, articulation, timbre (including vowel sounds and phonation modes [8]). Supported by the EU H2020 project "New Frontiers in Music Information Processing," we are collaborating with DoReMIR Music Research on singing transcription for their ScoreCloud transcription service[5]. We are also developing source separation methods for voice [9], which will enable the analysis of accompanied singing.

Based on my previous work, I presented several studies on pitch, intonation, and singer interaction. Pitch is perhaps the most essential characteristic of musical sounds, being

[5] www.scorecloud.com

the basis of both melody and harmony. Western music theory provides a framework for understanding pitch relationships in terms of intervals, scales and chords, expressed as a first approximation in indivisible units of semitones. Common music notation reflects this world-view. At the same time, it has been recognized since the time of Pythagoras that it is not possible for all theoretically consonant intervals to be perfectly "in tune", and this has led to many theoretical and practical approaches to intonation, the realization of pitch, in music performance.

I presented investigations of intonation at two extremes of musical practice: a fixed-pitch instrument, the harpsichord, where the tuner determines the intonation of each pitch before the performance, and a variable-pitch instrument, the human voice, which can adjust the pitch of each note to the musical context and vary the pitch over the note's duration. In each case, we have developed software tools for (semi-)automatic analysis of the pitch content from audio recordings.

We analyzed a collection of solo harpsichord CDs, estimated the inharmonicity and temperament of the harpsichord for each movement, and compared the measured temperaments with those given in the CD sleeve notes. The observed differences illustrate the tension between temperament as a theoretical construct and as a practical issue for professional performers and tuners. We conclude that "ground truth" is not always scientific truth and that content-based analysis has an essential role in the study of historical performance practice.

The second study investigates intonation and intonation drift in unaccompanied solo singing and proposes a simple intonation memory model that accounts for many of the effects observed. Singing experiments were conducted with 24 singers of varying ability. Over the duration of the recordings, approximately 50 seconds, a median absolute intonation drift of 11 cents was observed, which was smaller than the median note error (19 cents) but was significant in 22 percent of the recordings. Drift magnitude did not correlate with other measures of singing accuracy or singing experience. Neither a static intonation memory model nor a memoryless interval-based intonation model can account for the accuracy and drift behavior observed. The proposed causal model provides a better fit.

The third study looked at how pitch is negotiated by imperfect unaccompanied singers when they sing in pairs and the factors that influence pitch accuracy. Two singing conditions (unison versus 2-part harmony) were compared, along with an experimental condition varied which singers could hear their partners, measured in terms of pitch and interval errors. We found the following:

- Unison singing is more accurate than singing harmony.
- Singing solo is more accurate than singing with a partner.
- Singers adjust pitch to mitigate their partner's error and preserve harmonic intervals at the expense of melodic intervals and absolute pitch.
- Other factors influencing pitch accuracy include score pitch, score harmonic interval, score melodic interval, musical background, vocal part, and individual differences.

References

1 Dogac Basaran, Slim Essid and Geoffroy Peeters. *Main Melody Extraction with Source-filter NMF and C-RNN*. International Society for Music Information Retrieval Conference, Paris, France, 2018, pp. 82–89.

2 Klaus Frieler, Frank Hoeger, Martin Pfleiderer and Simon Dixon. *Two Web Applications for Exploring Melodic Patterns in Jazz Solos*. International Society for Music Information Retrieval Conference, Paris, France, 2018, pp. 777–783.

3 Maria Panteli, Rachel Bittner, Juan Pablo Bello and Simon Dixon. *Towards the Characterization of Singing Styles in World Music*. IEEE International Conference on Acoustics, Speech and Signal Processing, New Orleans, USA, 2017, pp. 636–640.
4 Matthias Mauch, Klaus Frieler and Simon Dixon. *Intonation in Unaccompanied Singing: Accuracy, Drift and a Model of Reference Pitch Memory*. Journal of the Acoustical Society of America, 136 (1), 2014, pp. 401–411.
5 Jiajie Dai and Simon Dixon. *Analysis of Vocal Imitations of Pitch Trajectories*. International Society for Music Information Retrieval Conference, New York, USA, 2016, pp. 87–93.
6 Jiajie Dai and Simon Dixon. *Analysis of Interactive Intonation in Unaccompanied SATB Ensembles*. International Society for Music Information Retrieval Conference, Suzhou, China, 2017, pp. 599–605.
7 Jiajie Dai and Simon Dixon. *Pitch Accuracy and Interaction in Unaccompanied Duet Singing*. Journal of the Acoustical Society of America, 145 (2), 2019, pp. 663–675.
8 Daniel Stoller and Simon Dixon. *Analysis and Classification of Phonation Modes in Singing*. International Society for Music Information Retrieval Conference, New York, USA, 2016, pp. 80–86.
9 Daniel Stoller, Sebastian Ewert and Simon Dixon. *Wave-U-Net: A Multi-Scale Neural Network for End-to-End Audio Source Separation*. International Society for Music Information Retrieval Conference, Paris, France, 2018, pp. 334–340.
10 Simon Dixon, Matthias Mauch and Dan Tidhar. *Estimation of Harpsichord Inharmonicity and Temperament from Musical Recordings*. Journal of the Acoustical Society of America, 131 (1), 2012, pp. 878–887.
11 Dan Tidhar, Simon Dixon, Emmanouil Benetos and Tillman Weyde. *The Temperament Police*. Early Music, 42 (4), 2014, pp. 579–590.

3.8 What Makes Singing Unique?

Zhiyao Duan (University of Rochester, US)

Singing is arguably the most popular kind of music throughout human history. People love singing or listening to singing on various occasions, and numerous styles of singing music have been composed, performed, recorded, and enjoyed. The reasons for this popularity, in my opinion, are manifold. First, singing is a musical behavior that is easier to learn and to perform compared to instrumental music. Second, singing often presents lyrics and is an enriched form of storytelling, which is at the core of human civilization. Third, singing voices generally show more flexibilities in pitch, dynamics, and timbre than musical instruments, which greatly enhance their expressiveness. Fourth, compared to musicians playing an instrument, singers have more freedom in expressing their emotions through facial expressions and body gestures, which significantly helps engage audiences.

Research on computational models for singing voices, in my opinion, needs to consider these unique properties of singing voices. For example, melody generation for singing voices may need to consider matching certain rhythms and tones of the natural speech of the underlying lyrics. This match is quite common in opera and folk songs of tonal languages (e. g., Chinese). Also, melody transcription may need to go beyond the standard piano-roll notation for instrumental music due to the extensive use of pitch glides and microtonality.

Raw pitch contours, however, seem to not bear enough abstraction for high-level processing. Some symbolic representations in between might be ideal. Furthermore, the visual aspects of singing, which have been neglected in the MIR community for a long time, may be crucial for singing analysis and synthesis. Based on these thoughts, I propose the following two novel research directions.

- **Lyrics-Informed Melody Generation for Chinese Opera.** Given lyrics (e. g., a poem or short story) and its speech intonation (e. g., Mandarin, Cantonese, other dialects), I propose to generate a melody for the lyrics automatically. The pitch contour of the melody should match with certain aspects of that of the speech intonation. It is an interesting question to investigate through a large corpora study which aspects should be considered for the match and which aspects allow for more flexibility.
- **Audio–Visual Analysis of Singing.** I propose to analyze the emotion and expressiveness of singing from the audio and visual modalities jointly. In particular, facial expressions and body movements should be analyzed and correlated with the audio signals. A related problem is audio–visual singing voice separation. Visual signals, especially the lip movements, provide cues about the singing activity and content, although some cues can be ambiguous or misleading depending on the singing style.

3.9 From Science to Engineering to Deep Learning in Singing Processing

Emilia Gómez (UPF – Barcelona, ES)

Joint work of Emilia Gómez, Merlijn Blaauw, Jordi Bonada, Pritisy Chadna and Helena Cuesta
Main reference Emilia Gómez, Merlijn Blaauw, Jordi Bonada, Pritisy Chadna, Helena Cuesta: "Deep Learning for Singing Processing: Achievements, Challenges and Impact on Singers and Listeners".Keynote speech, 2018 Joint Workshop on Machine Learning for Music. The Federated Artificial Intelligence Meeting (FAIM), a joint workshop program of ICML, IJCAI/ECAI, and AAMAS, 2018.
URL https://arxiv.org/abs/1807.03046

The singing voice is probably the most complex "instrument" to analyze and synthesize, and a vast amount of research literature has contributed to understanding the mechanisms behind singing acoustics and perception. In addition to its complexity, the human voice is also the most popular musical instrument, as everybody sings and listens to vocal music. This has led to many practical engineering applications based on computational tasks such as singing assessment (linked to pitch, timbre and timing description), voice separation, singing synthesis, singer identification, singing style characterization, and query by humming/singing. Recent advancements in machine learning, in particular deep learning (DL), have provided a boost in performance in the mentioned computational tasks. However, these methods are still limited in terms of pitch resolution, separation quality, and naturalness of synthesized singing. In fact, the success of DL methods depends on three main factors: data, computing resources, and selected architectures.

- In terms of data, the performance of DL is directly connected to the amount and quality of available annotations. While data is more widely available in industrial settings, open datasets are still scarce due to privacy issues and industrial interest, the research community lacks of an agreed methodology for data gathering and annotation, and data needs are highly dependent on the computational task.

- Concerning computing, GPU availability is crucial for training complex architectures with high amounts of data. Computing resources are not always mentioned in scientific publications, and they provide some competitive advantages to researchers in large labs.
- Finally, in terms of DL architectures, there are still few studies on singing-specific algorithms that take advantage of knowledge on singing acoustics and perception.

Given the current research landscape, we need to address two crucial issues for our future research. First, we need to assess if the traditional link between singing processing and singing acoustics/perception research is lost because of the recent focus on deep learning. We believe that the efforts into explainable and optimized DL models might bring back this link in the future. Second, we need to consider if this boost in performance will have a different impact on listeners and singers, e. g. if singing synthesis becomes indistinguishable from human singing or if good quality singer impersonation is possible from polyphonic and noisy recordings. In this respect, our community has to be more aware of the social impact that our technologies might bring in the future.

3.10 Singing Information Processing

Masataka Goto (AIST – Tsukuba, JP)

As music information research has continued to develop, research activities related to singing have become more vigorous. Such activities are attracting attention not only from a scientific point of view but also from the standpoint of commercial applications. Singing-related research is highly diverse, ranging from basic research on the features unique to singing to applied research such as that on singing synthesis, lyrics recognition, lyrics alignment, singer identification, retrieval of singing voices, singing skill evaluation, singing training, singing voice conversion, singing impression estimation, and the development of singer robots. I have named this broad range of singing-related studies "singing information processing" in 2008 (see [1] and [2]).

Since singing is one of the most important elements of music, singing information processing has a significant impact on society from the viewpoints of industry and culture. In fact, automatic pitch-correction technology for vocals is already used on a routine basis in the production of commercial music (popular music, in particular). It has become essential for correcting pitch at points in a song where the singer is less than skillful and for achieving a desired artificial effect. A function for evaluating (scoring) a person's singing in the karaoke industry is also popular. More recently, singing-synthesis systems have become widely used, and people actively enjoy songs with synthesized singing voices as the main vocals.

During the Dagstuhl seminar, I gave an overview of this attractive research field. In particular, I covered the following examples for singing-related research:

- Singing synthesis including text-to-singing synthesis (VOCALOID), speech-to-singing synthesis (SingBySpeaking) and singing-to-singing synthesis (VocaListener).
- Robot singer (VocaListener + VocaWatcher).
- Singer identification.
- Retrieval of singing voices (VocalFinder).
- Creating hyperlinks between phrases in lyrics (Hyperlinking Lyrics).
- Lyrics alignment (LyricSynchronizer and Lyric Speaker).

I then discussed grand challenges such as ultimate singing analysis, superhuman singing synthesis, and perfect singing voice conversion. I finally addressed open questions to inspire an open discussion with participants from the viewpoint of technical issues and social impacts that singing technologies could give in the future.

References

1 Masataka Goto, Takeshi Saitou, Tomoyasu Nakano, and Hiromasa Fujihara. *Singing Information Processing Based on Singing Voice Modeling*. Proceedings of the 2010 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Dallas, USA, 2010, pp. 5506–5509.
2 Masataka Goto. *Singing Information Processing*. Proceedings of the 12th IEEE International Conference on Signal Processing (ICSP), Hangzhou, China, 2014, pp. 2431–2438.

3.11 Melody and Voice Processing – Some Thoughts on the Seminar

Frank Kurth (Fraunhofer FKIE – Wachtberg, DE)

There is a multitude of real-life audio signals containing components with melody- or voice-like characteristics. With a research background from music audio, bioacoustics and speech processing, my interest has always been to explore synergies and cross-domain relationships in signal representation and analysis. Particularly for signal analysis, my interest in the seminar was to discuss how strong the impact of deep learning affects a research field like melody and voice processing, where generations of researchers have grown up with handcrafted and interpretable features (or, more generally, interpretable signal models). Relevant (and maybe provoking) questions were: Are interpretable signal features any longer necessary? How do learning-based approaches change the way we analyze signals and extract patterns? Are there subtasks in this field of MIR where machine learning already or in the future will outperform humans?

3.12 Try, Try, Try Again: Rehearsals as a Data Source

Cynthia Liem (TU Delft, NL)

(Comparative) analysis of musical performance and musical expression has, on the one hand, focused on commercially available recordings, and on the other hand on carefully conditioned experimental situations. In the first case, we usually only have an audio signal at our disposal which reflects the ultimate artistic intent of a performer and producer, although this intent is not always articulated explicitly. In the latter case, richer data is collected, but experimental conditions may not necessarily have created a 'naturalistic' environment. Whenever a musician is in progress of mastering a piece, multiple rehearsals are needed before the mastering is achieved. Within these rehearsals, multiple realizations of musical intent will be present; over time, they should ideally converge. As the musician progresses through the rehearsal, she also will likely go through various physical and mental states

of well-being, which may be evidenced in data that can be acquired during rehearsals. In summary, rehearsals may be a very interesting and data-rich source of information on musical expression, musician well-being, and developmental progress. At the Dagstuhl seminar, I discussed current data-acquisition and research activities at my lab that focus on this, both touching upon comparative performance analysis and physiological monitoring.

3.13 Singing Voice Separation: Recent Breakthroughs with Data-Driven Methods

Antoine Liutkus (Inria, University of Montpellier, FR)

In my presentation, I gave an overview of different approaches that have been undertaken for the separation of vocals from music recordings in the past 40 years. First, I recalled that until recently, research on this topic largely focused on singing voice models based on physiological, acoustic, or musical aspects. In particular, researchers considered interpretable approaches that could be understood in terms of features of the singing voice and the accompaniment. Such model-based approaches, however, have never really met performance. Of course, it is always possible to find examples for which some given model is appropriate, but experience shows that any model proves inappropriate for the overwhelming majority of the other recordings.

Second, I explained that in this context, a new data-driven trend of research arose in the past five years in conjunction with deep learning. The singing voice is not described any more by some human-understandable model, but rather only through examples. In this setting, source separation systems are seen merely as mappings between mixtures and vocals. Training such systems has been made possible by the recent availability of dedicated datasets, where both isolated vocals and accompaniment music are available.

In the third part of my presentation, I demonstrated how the current state of the art impressively outperforms model-based approaches, based on the latest results from the international signal separation evaluation campaign (SiSEC). From these results followed a discussion about the current challenges on this topic, as well as how we should evaluate and compare contributions. Topics for discussion included the relative importance of models and their interpretability over trainable systems that loose interpretability but are much more effective at solving particular problems.

References

1 Estefania Cano, Derry Fitzgerald, Antoine Liutkus, Mark Plumbley, and Fabian-Robert Stöter. *Musical Source Separation: An Introduction*. IEEE Signal Processing Magazine, 36(1), 2019, pp. 31–40.
2 Zafar Rafii, Antoine Liutkus, Fabian-Robert Stöter, Stylianos Ioannis Mimilakis, Derry FitzGerald, and Bryan Pardo. *An Overview of Lead and Accompaniment Separation in Music*. IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP), 26(8), 2018, pp. 1307–1335.
3 Fabian-Robert Stöter, Antoine Liutkus, and Nobutaka Ito. *The 2018 Signal Separation Evaluation Campaign*. International Conference on Latent Variable Analysis and Signal Separation, Springer, 2018, pp. 293–305.

3.14 Interactive Interfaces for Choir Rehearsal Scenarios

Meinard Müller (Universität Erlangen-Nürnberg, DE), Sebastian Rosenzweig (Universität Erlangen-Nürnberg, DE), and Frank Zalkow (Universität Erlangen-Nürnberg, DE)

Joint work of Meinard Müller, Sebastian Rosenzweig, Frank Zalkow, Johannes Graulich (Carus-Verlag, DE)
Main reference Frank Zalkow, Sebastian Rosenzweig, Johannes Graulich, Lukas Dietz, El Mehdi Lemnaouar, Meinard Müller: "A Web-Based Interface for Score Following and Track Switching in Choral Music". Demos and Late Breaking News of the International Conference on Music Information Retrieval (ISMIR), 2018
URL https://www.audiolabs-erlangen.de/resources/MIR/2018-ISMIR-LBD-Carus

Choral music is an essential part of our musical culture. Most choral singers practice their parts with traditional material, such as printed sheet music and CD recordings. Given recent advances in music information retrieval (MIR), important research questions are how and in which way new interfaces may enhance the rehearsal experience in particular for amateur choral singers. For example, score-following technology makes it possible to present audio and sheet music modalities synchronously. Furthermore, audio decomposition techniques may be useful to separate or enhance a specific voice (e. g., corresponding to the soprano part) that is relevant for a singer. Then, a choral singer's rehearsal experience may be enhanced by switching between audio tracks of a multitrack recording. In collaboration with the Carus publishing house, a leading music publisher for religious and secular choral music with headquarters in Stuttgart, we explore the potential of such MIR technologies by building web-based prototypes for the interactive navigation and access of choral music recordings [4]. In particular, such interfaces should include personalization strategies that allow users to structure and analyze music recordings according to their specific needs, expectations, and requirements. Additionally, the integration of real-time feedback mechanisms concerning, e. g., rhythm, interaction, or intonation of the singers' voices, may be useful components to make choir rehearsal preparations more effective [1].

In this context, a fundamental research topic is to investigate how and to what extent (partially) automated procedures can help to simplify the process of linking, decomposing, analyzing multimedia content. Even though there has been significant progress in MIR [2], the results of automatic alignment, voice separation, or music analysis procedures are still far from being perfect. As for commercial music applications, users are very critical regarding the quality of the presented music content. As for sheet music, for example, users often expect a visually appealing layout, where even small inaccuracies or distortions in the appearance of musical symbols may not be tolerable. Therefore, using software for automatically rendering sheet music such as Verovio [3] is often problematic when layout issues are of high importance. Similarly, when playing back a music recording along with showing a synchronized musical score, already small temporal asynchronies between the audio position and corresponding sheet music symbols may confuse the listener. Furthermore, feedback mechanisms on performance and intonation should work with high accuracy to satisfy a user's expectation. Through close collaboration with the Carus publishing house, we explore the benefits and limitations of current MIR technologies in practical applications. This collaboration also offers numerous cross-connections to fields such as music education and musicology, which stimulates further interdisciplinary cooperations.

References

1 Helena Cuesta, Emilia Gómez, Agusín Martorell, and Felipe Loáiciga. *Analysis of Intonation in Unison Choir Singing*. Proceedings of the International Conference on Music Perception and Cognition (ICMPC), Graz, Austria, 2018.

2 Meinard Müller. *Fundamentals of Music Processing*. Springer Verlag, 2015.
3 Laurent Pugin, Rodolfo Zitellini, and Perry Roland. *Verovio: A library for Engraving MEI Music Notation into SVG*. Proceedings of the International Society for Music Information Retrieval Conference (ISMIR), Taipei, Taiwan, 2014, pp. 107–112.
4 Frank Zalkow, Sebastian Rosenzweig, Johannes Graulich, Lukas Dietz, El Mehdi Lemnaouar, and Meinard Müller. *A Web-Based Interface for Score Following and Track Switching in Choral Music*. Demos and Late Breaking News of the International Conference on Music Information Retrieval (ISMIR), 2018.

3.15 Singing Interfaces and Visualizations

Tomoyasu Nakano (AIST – Tsukuba, JP)

Technologies to automatically understand music and singing voices make it possible to develop systems enriching activities based on music. In such systems, interaction and visualization techniques play important roles. In this seminar, I introduced studies on interface development and information visualization based on signal processing and machine learning technologies. First, I discussed strategies to overcome errors introduced by automatic recognition approaches. One such strategy is based on human–computer interaction. For example, VocaListener, which imitates human singing expressions such as the fundamental frequency [1], integrates an interaction mechanism that lets a user easily correct lyrics-to-singing synchronization errors just by pointing them out. Furthermore, I discussed approaches for the effective visualization of music and singing. As an example of such research, I introduced TextTimeline which simultaneously visualizes words (lyrics) and acoustic features (intensity) [2]. By visualizing the time axis of the sound in a direction orthogonal to characters (vertical direction for horizontal text), TextTimeline can visualize them without changing the characters' display position and without stretching the time axis of the sound.

References

1 Tomoyasu Nakano and Masataka Goto. *VocaListener: A Singing-to-Singing Synthesis System Based on Iterative Parameter Estimation*. Proceedings of the 6th Sound and Music Computing Conference (SMC), Porto, Portugal, 2009, pp. 343–348.
2 Tomoyasu Nakano, Jun Kato, and Masataka Goto. *TextTimeline: Visualizing Acoustic Features and Vocalized Timing along Display Text*. Proceedings of the 11th IEEE Pacific Visualization Symposium (IEEE PacificVis), Kobe, Japan, 2018.

3.16 Social Voice Processing

Juhan Nam (KAIST – Daejeon, KR)

Singing is a healthy activity that promotes not only physical conditions through exercise but also social or mental status by connecting people as an auditory medium. However, unless people are confident of their singing skills, they are reluctant to sing. Also, even

if the singing skill is good, recorded voices are often not satisfactory, compared against professionally processed voices. How can we encourage people to sing with more confidence, fun, and satisfaction?

We propose "social voice processing" as a concept of voice signal processing to transform input voice utilizing different renditions of voices for a given song in terms of singing skills and recording quality. Everyone has a different voice and a different level of singing skills. Therefore, every rendition of singing for a given piece of music is unique in terms of timbre, tempo, pitch, and dynamics. Even for the same singer, repeated performances have small differences. By leveraging karaoke apps or vast amount of singing voice recordings on online music content platforms, we can exploit such different singing performances for a given song to transform one's voice as a digital audio effect.

One direction is mixing the multiple renditions. A traditional example is voice doubling as a recording technique. Mixing different voices can emulate a choir effect [1]. This can not only enrich the timbre but also suppresses singers' concern about accurate pitch by diluting a prominent voice. Another direction is modifying the voice directly. A recent attempt is singing expression transfer that allows for exchanging musical expressions such as timing, pitch contours or energy between two voices [2]. This can improve one's voice using skilled singers' voices or make it worse for fun using intentionally poor voices. The two voices can be paired between multiple renditions from oneself, between friends, or between fans and professional singers [3]. There will be more possibilities by combining multiple singing voices and transferring different combinations of expressions among them. Furthermore, singing voice mixed with background music (e. g. commercial pop music) can be used along with melody extraction or source separation algorithms.

References

1 Mark Barry Dolson. *A Tracking Phase Vocoder and its Use in the Analysis of Ensemble Sounds*. Ph.D. thesis, California Institute of Technology, 1983.

2 Sangeon Yong and Juhan Nam. *Singing Expression Transfer from One Voice to Another for a Given Song*. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, Canada, 2018, pp. 151–155.

3 Masahiro Hamasaki, Masataka Goto, and Tomoyasu Nakano. *Songrium: A Music Browsing Assistance Service with Interactive Visualization and Exploration of A Web of Music*. Proceedings of the International Conference on World Wide Web (WWW), Seoul, Korea, 2014, pp. 523–528.

3.17 Automatic Singing Transcription for Music Audio Signals

Ryo Nishikimi (Kyoto University, JP)

Automatic singing transcription (AST) refers to estimating musical notes of a sung melody. Since the melody is the most salient part of music, the transcribed notes are useful for many MIR tasks such as singing voice generation, score-informed source separation, query-by-humming, and musical grammar analysis.

Since musical notes are represented as semitone-level pitches, onset score times, and note values, a singing voice should be quantized in the frequency and temporal directions. Most conventional methods aim to estimate a piano-roll representation by quantizing only pitches,

and note-value quantization (a.k.a rhythm transcription) has been studied independently. To integrate these tasks, it is necessary to associate frame-level spectrograms with tatum-level note values. To solve the AST problem, it is also necessary to accurately separate singing voice or directly estimate musical notes from music audio signals.

In this seminar, I introduced a statistical method of AST using hidden Markov models and demonstrated the transcriptions of recordings taken during the seminar. I plan to extend the model by integrating pitch and note-value quantizations and singing voice separation.

3.18 Dominant Melody Estimation and Singing Voice Separation

Geoffroy Peeters (Telecom ParisTech, FR)

In my presentation, I discussed recent research on dominant melody estimation and singing voice separation. Both topics have in common the use of Convolutional Neural Networks and skip-connections (U-Net). Furthemore, I reviewed possible input representations such as waveform, STFT, and HCQT representations [2, 3] as well as source/filter models [1]. Furthermore, I discussed how to automatically create large datasets annotated with pitch and lyrics using a student-teacher paradigm [4].

References

1 Dogac Basaran, Slim Essid, and Geoffroy Peeters. *Main Melody Extraction with Source-Filter NMF and C-RNN.* Proceedings of the International Society for Music Information Retrieval, Paris, France, 2018, pp. 82–89.
2 Alice Cohen-Hadria, Axel Roebel, and Geoffroy Peeters. *Improving Singing Voice Separation Using Deep U-Net and Wave-U-Net with Data Augmentation.* submitted to the IEEE International Conference on Acoustics, Speech, and Signal Processing, Brighton, UK, 2019.
3 Guillaume Doras, Philippe Esling, and Geoffroy Peeters. *On the Use of U-Net for Dominant Melody Estimation in Polyphonic Music.* Proceedings of the International Workshop on Multilayer Music Representation and Processing, Milan, Italy, 2019.
4 Gabriel Meseguer Brocal, Alice Cohen-Hadria, and Geoffroy Peeters. *Dali: A Large Dataset of Synchronized Audio, Lyrics and Pitch, Automatically Created Using Teacher-Student.* Proceedings of the International Society for Music Information Retrieval, Paris, France, 2018, pp. 431–437.

3.19 Russian/Ukrainian Traditional Polyphony: Musicological Research Questions and Challenges for MIR

Polina Proutskova (Queen Mary University of London, GB)

As a performer and a collector of Russian/Ukrainian traditional polyphonic music, I have a practical application in mind which I would love to have and for which I would like to draw on the collective mind of the honored Dagstuhl participants.

These traditions are generally polyphonic but the relationship between the parts, particularly in older genres, is not harmonic but heterophonic: while in European traditions

the parts build consonant chords (are coordinated vertically), in Russian heterophony the voices are mainly constrained by the mode and are improvised in this mode, they are not coordinated vertically but conceived as "melodies" in their own right.[6]

Unfortunately, in a recording of more than three singers, it is difficult to make out all the parts by ear. Ethnomusicologists have used multi-channel recordings (where one voice is dominant) to capture the parts. See the website of the Polyphony Project[7] for an example (scroll down for the multi-channel player). Yet such multi-channel recordings have been produced for only a small subset of the repertoire.

The inventiveness of folk singers is often breathtaking, and different ensembles have their unique styles. Can we model an ensemble's style–the melodic lines created by each singer–based on multi-channel recordings of that ensemble? The goal would be to re-construct multi-channel recordings based on a recording of the whole ensemble. It would help me and my ensemble to learn songs based on archival recordings, for which no multi-channel takes were made.

In order to achieve that, our models will have to automatically transcribe the dominant voice in a messy choral recording, to learn the improvisational habits of a given singer, and to fit possible melodic lines for each of the singers so that together they build a construct closest to the recording of the whole ensemble. Also, our models would be able to produce a new voicing each time, true to the improvisational nature of the tradition.

If such a tool could be built, my ensembles IZBA (London) and Polynushka (Berlin) will be its grateful users and will help turn its output into real Russian folk songs. Conversely, the tool would help reconstruct a tradition nearly extinct, and bring archival recordings to life. And it might become a tool for more than reconstructing an old tradition, but for establishing a new one based on the principles of the old.

3.20 Aspects of Melodic Similarity

Preeti Rao (Indian Institute of Technology Bombay, IN)

Joint work of Preeti Rao, Kaustuv Kanti Ganguli

The computation of "melodic similarity" remains central to MIR applications. We consider audio-based similarity involving the typical pipeline of deriving a suitable melodic representation from an audio "query" signal which is then compared with the reference representation to obtain a computed similarity concerning the reference. While acknowledging the importance of somehow modeling human similarity judgments, computational methods typically involve one of exact match, approximate match (i. e., exact matching applied to reduced forms) or categorical matching [1, 2]. Challenges also arise in the evaluation of melodic similarity measures. A compelling case is that of music that is naturally categorized based on a geographical basis such as folk tunes, or on a musicological basis such as the music of the raga and makam traditions. The similarity of pieces drawn from these genres depends on the local match of melodic motifs where the frequency of appearance of recognized motifs influences global similarity ratings [3]. Thus modeling of categorical matching at the time

[6] Example: https://www.youtube.com/watch?v=pP9p1p63BGE
[7] https://www.polyphonyproject.com/uk/song/BMI_UK18060360

scale of phrases or short melodic motifs appears relevant. In our work on raga music, we consider melodic motifs represented by continuous pitch curves. While trained musicians easily label the motifs in terms of solfege notation, it is all too common to find that the symbolic sequence representations are inadequate in discriminating ragas where identical sequences occur. With the raga considered to be a class of melodies described by tonal hierarchy and characteristic phrases, we find that the melodic motifs exhibit both intra-class and inter-class variations. A clustering study on a dataset of vocal concert recordings across two ragas with identical motifs in terms of solfege sequences reveals specific discriminating acoustic cues between the motifs. The cues derived from the dataset of pre-segmented phrases are consistent with musicological knowledge about the raga difference and, more importantly, serve to quantify the difference and make it exploitable for MIR. Perception experiments with trained musicians confirm the perceptual importance of the discovered cues [4, 5, 6]. The presented work points to fundamental questions around the segmentation of motifs, similarity measures between the melodic shapes (should this be some form of pitch distance or some higher-level cue-based comparison?), the process of learning cues from data, and finally applications to MIR and music education.

References

1 Emilios Cambouropoulos, Tim Crawford, and Costas S. Iliopoulos. *Pattern Processing in Melodic Sequences: Challenges, Caveats and Prospects*. Computers and the Humanities 35, 2001, pp. 9-21.
2 Alan Marsden. *Interrogating Melodic Similarity: A Definitive Phenomenon or the Product of Interpretation?* Journal of New Music Research, 41(4), 2012, pp. 323–335.
3 Anja Volk and Peter Van Kranenburg. *Melodic Similarity Among Folk Songs: An Annotation Study on Similarity-Based Categorization in Music*. Musicae Scientiae, 16(3), 2012, pp. 317–339.
4 Kaustuv Kanti Ganguli and Preeti Rao. *Towards Computational Modeling of the Ungrammatical in a Raga Performance*. Proceedings of the 18th International Society for Music Information Retrieval (ISMIR), 2017, pp. 39–45.
5 Kaustuv Kanti Ganguli and Preeti Rao. *Exploring Melodic Similarity in Hindustani Classical Music Through the Synthetic Manipulation of Raga Phrases*. Workshop on Cognitive Music Information Retrieval (CogMIR), New York, 2016.
6 Kaustuv Kanti Ganguli and Preeti Rao. *Imitate or Recall: How Do Musicians Perform Raga Phrases?* Proceedings of Frontiers of Research on Speech and Music (FRSM), Rourkela, India, 2017.

3.21 Extraction Techniques for Harmonic and Melodic Interval Analysis of Georgian Vocal Music

Sebastian Rosenzweig (Universität Erlangen-Nürnberg, DE), Meinard Müller (Universität Erlangen-Nürnberg, DE), and Frank Scherbaum (Universität Potsdam – Golm, DE)

Polyphonic singing plays a vital role in many musical cultures. One of the oldest forms of polyphonic singing can be found in Georgia, a country located in the Caucasus region of Eurasia. The traditional three-voice chants are acknowledged as Intangible Cultural Heritage by the UNESCO. Being an orally transmitted culture, most of the sources are available as field recordings, which often are of rather poor audio quality. Musicologists typically research

on Georgian vocal music on the basis of manually created transcriptions of the recorded material. Such approaches are problematic since important tonal cues, as well as performance aspects, are likely to get lost in the transcription process. Within an interdisciplinary project, we apply and develop computational methods to support musicological research on traditional Georgian vocal music.

In this context, the non-tempered nature of the Georgian tuning is of particular interest and subject to many controversial discussions among musicologists [1, 2]. We try to contribute to this discussion by measuring melodic (horizontal) and harmonic (vertical) intervals of field recordings. From these two types of intervals, one then may obtain cues on the tonal organization of Georgian vocal music, as indicated in [3]. In our approach, we compute these intervals in three steps. First, we estimate the fundamental frequency (F0) trajectories for all three voices from the audio recordings. Second, we clean the extracted F0 trajectories using filtering techniques. Third, we determine harmonic and melodic intervals from suitable parts of the trajectories. This overall approach faces several challenges due to the polyphonic nature of the audio material. We report on several experiments to illustrate, handle, and circumvent some of these challenges. First, we show how one can improve the F0 estimation step by using informed and semi-automated (with user feedback) approaches [4]. Furthermore, we show how the F0 estimation problem can be significantly simplified when using headset and throat microphones additionally to conventional microphone types. We also discuss various filtering approach to remove unstable and unreliable parts of the extracted F0 trajectories, which often correspond to pitch slides at the beginning and end of sung notes. Finally, we discuss the effect of the various F0 extraction and processing strategies on the derived harmonic and melodic intervals. In particular, we look at various interval distributions, which may serve as a basis for further musicological studies.

Besides contributing to the "Georgian scale controversy" our goal is to gain a deeper understanding of the interdependence of different sensor types, F0 extraction methods, and filtering techniques and their influence on the computed interval statistics.

References

1 Malkhaz Erkvanidze. *The Georgian Musical System*. Proceedings of the 6th International Workshop on Folk Music Analysis, Dublin, Ireland, 2016, pp. 74–79.

2 Zaal Tsereteli and Levan Veshapidze. *On the Georgian Traditional Scale*. Proceedings of the first International Symposium Traditional Polyphony, Tbilisi, Georgia, 2014, pp. 288–295.

3 Frank Scherbaum. *On the Benefit of Larynx-Microphone Field Recordings for the Documentation and Analysis of Polyphonic Vocal Music*. Proceedings of the 6th International Workshop on Folk Music Analysis, Dublin, Ireland, 2016, pp. 80–87.

4 Meinard Müller, Sebastian Rosenzweig, Jonathan Driedger, and Frank Scherbaum. *Interactive Fundamental Frequency Estimation with Applications to Ethnomusicological Research*. Proceedings of the AES International Conference on Semantic Audio, Erlangen, Germany, 2017.

3.22 Some Uncomfortable Statements about Melody Extraction

Justin Salamon (Adobe Research, US)

In my Dagstuhl presentation, I raised two fundamental questions through the lens of melody extraction, even though these issues can be considered in the broader context of MIR research. Some of the statements in my talk might have annoyed you ... some might be wrong ... some might be true! What do *you* think?

- **Question: Is melody extraction "done right"?**
  Melody extraction has been an active topic of research in MIR for decades now, and yet there is still no consensus as to what a melody is. To illustrate this, I reviewed various definitions of melody found in musicology and music history, as well as definitions proposed by members of the MIR community, based on [1]. This review led to my first uncomfortable statement: *In MIR, a melody is whatever the annotations contain in the dataset I am using for my research.* That is, as a community, we often resort to "definition-by-annotation" to circumvent the challenge of estimating a musical concept that is inherently ambiguous. Should we change the way we think of, and evaluate melody extraction?
  Next, I gave an overview of the existing datasets and evaluation metrics for melody extraction, their limitations, and the various efforts that have been made over the past few years to address these limitations. Importantly, I argued that the community has by-and-large ignored these efforts, and continues to use outdated datasets and evaluation metrics for melody extraction research. This led to my second uncomfortable statement: *Existing datasets for melody extraction are still (mostly) too small/artificial/homogenous, and (most) metrics in use have severe limitations ..., but we use them anyway!*
- **Question: What's the point of melody extraction anyway?**
  In the second part of the talk, I presented some of the concepts and methodologies coming from the "lean startup" movement and contrasted them with equivalent processes in MIR research. In particular, I highlighted how the lean-startup methodology begins with a thorough customer discovery stage. This stage aims at identifying real problems, shared by a significant number of people, that require solutions. MIR research, in contrast, is sometimes driven by problems that may be interesting and challenging and may have "potential" applications, but that in practice have seen little applications outside of research. This led to my third uncomfortable statement: *There is a disconnect between MIR research (on melody extraction) and potential users of MIR technologies.*

References

1 Justin Salamon. *Melody Extraction from Polyphonic Music Signals*. Ph.D. Thesis, Universitat Pompeu Fabra, Barcelona, Spain, 2013.

3.23 Computational Analysis of Traditional Georgian Vocal Music

Frank Scherbaum (Universität Potsdam – Golm, DE), Nana Mzhavanadze (Universität Potsdam – Golm, DE), Sebastian Rosenzweig (Universität Erlangen-Nürnberg, DE), Daniel Vollmer (Universität Potsdam – Golm, DE), Vlora Arifi-Müller (Universität Erlangen-Nürnberg, DE), and Meinard Müller (Universität Erlangen-Nürnberg, DE)

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. In our contribution, we consider a computational approach, where we focus on the use of new and partially unconventional recording and analysis techniques to document and analyze this type of music. To circumvent the source separation problem for multiple singing voices in a natural singing environment, we explored the potential of recordings of skin vibrations close to the larynx (using larynx microphones) and the mastoid (using NAM microphones). In combination with conventional audio recordings (e. g., using high-quality headset microphones), these vibrational signals turned out very useful for subsequent computational analysis regarding a multitude of aspects including pitch tracking, tuning analysis, analysis of voice interaction, as well as for documentation and archiving purposes [1, 2, 3, 4, 5].

In rare cases, such as the Tbilisi State Conservatory Recordings of master chanter Artem Erkomaishvili in 1966 (which were recorded in an overdubbing mode), historical recordings can also be used to separate the individual voices using signal processing techniques and to investigate questions of historical performance practice, scales, and tuning, using computational techniques from the field of music information retrieval [6, 7]. Due to the existence of conventional transcriptions into Western staff notation by Shugliashvili [8], it became possible to compare different approaches to digitally represent this unique set of recordings statically (as images) as well as dynamically (as movies). One of the most exciting findings in the context of this work was the detection of numerous instances of melodic tuning adjustments (intonation changes to achieve particular harmonic intervals), in particular for harmonic fifths, ninths, octaves, and unison, the systematic investigation of which is now a topic of ongoing work.

References

1 Frank Scherbaum, Wolfgang Loos, Frank Kane, and Daniel Vollmer. *Body Vibrations as Source of Information for the Analysis of Polyphonic Vocal Music*. Proceedings of the International Workshop on Folk Music Analysis (FMA), Paris, France, 2015, pp. 89–93.

2 Scherbaum Frank. *On the Benefit of Larynx-Microphone Field Recordings for the Documentation and Analysis of Polyphonic Vocal Music*. Proceedings of the International Workshop on Folk Music Analysis (FMA), Dublin, Ireland, 2018, pp. 80–87.

3 Frank Scherbaum and Nana Mzhavanadze. *A New Archive of Multichannel-Multimedia Field Recordings of Traditional Georgian Singing, Praying, and Lamenting with Special Emphasis on Svaneti*. LaZAR-Database. https://lazardb.gbv.de

4 Frank Scherbaum, Nana Mzhavanadze, and Elguja Dadunashvili. *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*. Proceedings of the International Symposium on Traditional Polyphony, 2018.

5 Frank Scherbaum, Sebastian Rosenzweig, Meinard Müller, Daniel Vollmer, and Nana Mzhavanadze. *Throat Microphones for Vocal Music Analysis*. Demos and Late Breaking News of the International Society for Music Information Retrieval Conference (ISMIR), Paris, France, 2018.
6 Meinard Müller, Sebastian Rosenzweig, Jonathan Driedger, and Frank Scherbaum. *Interactive Fundamental Frequency Estimation with Applications to Ethnomusicological Research*. Proceedings of the AES Conference on Semantic Audio, Erlangen, Germany, 2017. https://www.audiolabs-erlangen.de/resources/MIR/2017-GeorgianMusic-Erkomaishvili
7 Frank Scherbaum, Meinard Müller, and Sebastian Rosenzweig. *Analysis of the Tbilisi State Conservatory Recordings of Artem Erkomaishvili in 1966*. Proceedings of the International Workshop on Folk Music Analysis (FMA), Málaga, Spain, 2017, pp. 29–36.
8 David Shugliashvili. *Georgian Church Hymns, Shemokmedi School*. Georgian Chanting Foundation & Tbilisi State Conservatory, 2014.

3.24 Measuring Intonation via Dissonance in Polyphonic Choral Singing

Sebastian J. Schlecht (Universität Erlangen-Nürnberg, DE), Christof Weiß (Universität Erlangen-Nürnberg, DE), Sebastian Rosenzweig (Universität Erlangen-Nürnberg, DE), and Meinard Müller (Universität Erlangen-Nürnberg, DE)

One of the central challenges in the performance of polyphonic choral singing is to sing in tune. Excellent intonation (i. e., a singer's proper realization of pitch accuracy) requires many years to master. Especially in the challenging context of polyphonic singing, we see potential in supporting this learning task with computational techniques. Intonation may be defined either relative to an absolute pitch scale (such as equal temperament) or relative to other active harmonic sounds. Our current working hypothesis is that good intonation is achieved when the overall perceived tonal dissonance is minimized. We tested a dissonance measure developed by Sethares [1] based on the psychometric curves by Plomp and Levelt [2]. Sethares' dissonance measure is computed from the relative frequency distance of all active partials and can be directly retrieved from audio recordings with partial tracking techniques [3]. Furthermore, we experimented with a dissonance- or tuning-measure based on the 12-tone equal-tempered scale. Inspired by [4], we accumulate the overall deviation of partial frequencies from an idealized 12-tone grid followed by a suitable normalization. As a test scenario, we compiled a small but diverse dataset of Anton Bruckner's Gradual "Locus iste" in different performances, see [5] and the contribution described in Section 3.30.

At the Dagstuhl seminar, we raised questions about the potential of this approach in comparison to alternative methods. Further open questions included the separation of the composer's intended dissonance from unintended dissonance introduced by the performance, the preferred quality of intonation in high dissonance harmonic contexts, and the role of dissonance in ensemble performances.

References

1 William A. Sethares. *Local Consonance and the Relationship Between Timbre and Scale*. Journal of the Acoustical Society of America, 94(3), 1993, pp. 1218–1228.
2 Rosina Plomp and Willem J. M. Levelt. *Tonal Consonance and Critical Bandwidth*. Journal of the Acoustical Society of America, 38(4), 1965, pp. 548–560.

3 Julian Neri and Philippe Depalle. *Fast Partial Tracking of Audio with Real-Time Capability Through Linear Programming*. Proceedings of the International Conference on Digital Audio Effects (DAFx), Aveiro, Portugal, 2018, pp. 326–333.
4 Tomoyasu Nakano, Masataka Goto, and Yuzuru Hiraga. *An Automatic Singing Skill Evaluation Method for Unknown Melodies Using Pitch Interval Accuracy and Vibrato Features*. Proceedings of the International Conference on Spoken Language Processing (Interspeech), 2006, pp. 1706–1709.
5 Helena Cuesta, Emilia Gómez, Agusín Martorell, and Felipe Loáiciga. *Analysis of Intonation in Unison Choir Singing*. Proceedings of the International Conference on Music Perception and Cognition (ICMPC), 2018.

3.25 Signal Processing for Multiple Fundamental Frequency Estimation

Li Su (Academia Sinica – Taipei, TW)

There are various competing methods for multiple fundamental frequency estimation (MF0E) of polyphonic recordings. Saying they are 'competing' does not imply that some of them will survive while others become useless in the future. Actually, what the different MF0E approaches compete for is not merely the accuracy achieved, but also their generalization capability and their applicability in real-world scenarios. By considering all these aspects, one also gets a better understanding of the benefits and limitations of recent machine learning methods (in particular, deep learning) and more traditional hand-crafted approaches in solving the MF0E problem.

In brief, machine learning and data-driven approaches for MF0E are driven by large-scale and well-annotated datasets. Hand-crafted or rule-based approaches, on the other hand, are driven by music theories and domain knowledge. Even though hand-crafted approaches might not be as competitive as the ones based on machine learning when evaluated against standard MIR datasets, traditional approaches may be still more useful in the wild, especially for case-dependent usage such as transcribing a class of non-Western music where training data is rare or even unavailable.

Both data-driven and rule-based approaches require techniques from signal processing. While rule-based MF0E usually employ signal-processing techniques to simplify the problem, data-driven methods need signal processing for generating suitable input (feature) representations as well for post-processing. For example, with a pure machine-learning framework for MF0E (framed as a classification problem), it may be hard to obtain F0 estimates at a high frequency resolution. Knowing the role of signal processing in MF0E, I discussed in my presentation the following two challenges.

- **Challenge: Instantaneous frequency (IF) estimation in general polyphonic music signals.** The definition of instantaneous frequency (IF) is itself an oxymoron. The Fourier paradigm assumes the stationarity of the analyzed signals, while the term instantaneous implies the signals are never stationary. In fact, in the sinusoidal model, the amplitude and frequency are not uniquely defined, if the stationary condition is not imposed. Analyzing multi-component signals is even more tricky because of the interference between overlapped components, which is unavoidable according to the Gabor-Heisenberg uncertainty principle. Voice and singing signals usually challenge the

basic assumptions of Fourier-based signal processing. For example, in choir singing, every note is actually an ensemble of sounds with nearby frequencies; it is neither a single tone nor several components with separable frequencies. How to break the quasi-stationary paradigm is a key to describe what the IF is in musical signal, and this topic is highly related to fundamental signal processing theories.

Another noteworthy advantage of signal processing is that smart use of signal processing can greatly reduce the data or computation resources in model training. A classic example is the pitch detection method combining frequency and periodicity representations, which was proposed by Peeters in [1]. It uses the temporal features to help suppress the unwanted harmonics of a component and enhances true F0 peaks. Such an approach is found not only useful for music signals but also for the MF0E tasks in biomedical signals [2, 3]. An extension of this approach is the recently proposed multi-layer cepstrum (MLC), which performs the Fourier transform, nonlinear scaling and a high-pass filter recursively to achieve iterative purification of F0 information [4]. Preliminary studies also show its potential in analyzing choir singing.

- **Challenge: Use of signal processing in transcribing non-Western music, taking Taiwanese aboriginal music 'Pasibutbut' as an example.** There are more than ten aboriginal tribes living in Taiwan. Studies on Taiwanese aboriginal music started in the mid-20th century. According to the studies, most of the tribes' choir music is heterophonic, and some tribes such as Bunun and Amis even have the tradition of polyphonic choir singing. Pasibutbut (meaning "Praying for a Rich Harvest") is a classic example of polyphonic singing of the Bunun tribe. The leading voice (named 'Mahusngas') gradually increases the pitch while the other three voices follow the leading voice singing lower than the leading voice by a minor third, perfect fourth, and perfect fifth, respectively. In the experiments, we found that the MLC method outputs the pitch contours of each part in a resolution better than humans can do. Processing the contours by an unsupervised clustering method such as DBSCAN results in note transcriptions without any labeled training data. In conclusion, a smart signal processing strategy can reduce efforts in labeling large amounts of data while giving satisfactory results for this type of non-Western music.

References

1 Geoffroy Peeters. *Music Pitch Representation by Periodicity Measures Based on Combined Temporal and Spectral Representations*. IEEE Proceedings of the International Conference on Acoustics Speech and Signal Processing Proceedings (ICASSP), Toulouse, France, 2006, pp. 53–56.

2 Li Su and Yi-Hsuan Yang. *Combining Spectral and Temporal Representations for Multipitch Estimation of Polyphonic Music*. IEEE/ACM Transactions Audio, Signal Language Processing (TASLP), 23(10), 2015, pp. 1600–1612.

3 Chen-Yun Lin, Li Su, and Hau-tieng Wu. *Wave-Shape Function Analysis – When Cepstrum Meets Time-Frequency Analysis*. Journal of Fourier Analysis and Applications, 24(2), 2018, pp. 451–505.

4 Chin-Yun Yu and Li Su. *Multi-layered Cepstrum for Instantaneous Frequency Estimation*. Proceedings of the IEEE Global Conference on Signal and Information Processing (GlobalSIP), Anaheim, USA, 2018.

3.26 Augmented Vocal Production towards New Singing Style Development

Tomoki Toda (Nagoya University, JP)

Singers can produce attractive singing voices by expressively controlling pitch, dynamics, rhythm, and voice timbre. However, individual singers have their own limitations to control these components widely owing to physical constraints in speech production. For instance, it is physically difficult to change their own voice timbre into that of another specific singer. Moreover, if they suffered from vocal disorder, they would be unable to produce singing voices. If singers could freely produce singing voices as they want beyond their physical constraints, it would open up entirely new ways to express a greater variety of expression.

Towards the development of techniques to augment our speech production mechanism, I have studied a real-time statistical voice conversion technique for more than 15 years. Statistical voice conversion is a technique based on machine learning to modify a speech waveform for converting non- or para-linguistic information while keeping linguistic information unchanged. Its real-time implementation has been successfully achieved by incorporating real-time signal processing. This technique makes it possible for us to produce speech and singing voices beyond our physical constraints, and therefore, it has great potential to develop various applications to break down the existing barriers in our speech production. I have named this technique "augmented speech production" [1], and have developed some applications for augmenting vocal production, such as vocal effector [2] and singing-aid for laryngectomees [3].

In the Dagstuhl seminar, I gave an overview of augmented speech production techniques, showed some applications in singing voice production including a demo system of vocal effector, and addressed open questions to inspire an open discussion with participants.

Acknowledgements: This work was partly supported by JST, PRESTO Grant Number JPMJPR1657.

References

1 Tomoki Toda. *Augmented Speech Production Based on Real-Time Statistical Voice Conversion*. Proceedings of the IEEE Global Conference on Signal and Information Processing (GlobalSIP), Atlanta, USA, 2014, pp. 755–759.

2 Kazuhiro Kobayashi, Tomoki Toda, and Satoshi Nakamura. *Intra-Gender Statistical Singing Voice Conversion with Direct Waveform Modification Using Log-Spectral Differential*. Speech Communication, Vol. 99, 2018, pp. 211–220.

3 Kazuho Morikawa and Tomoki Toda. *Electrolaryngeal Speech Modification Towards Singing Aid System for Laryngectomees*. Proceedings of the Asia-Pacific Signal and Information Processing Association Annual Summit and Conferenc (APSIPA ASC), Kuala Lumpur, Malaysia, 2017, pp. 610–613.

3.27 Useless Evaluation versus User-less Evaluation

Julián Urbano (TU Delft, NL)

Performance metrics are one of the keystones in our everyday research. They (should) serve us as abstractions of how our algorithms are used in real use cases, and the numbers we compute with them are the evidence we present to judge the merit of our research and decide what works and what does not. However, do they tell us anything useful? Most of our metrics are algorithm-oriented, and that makes sense. Many of our algorithms are just smaller parts in bigger systems and so it makes sense to treat them in a purely system-oriented fashion. However, very often we have real people at the other end, and how they perceive the output of our algorithms, or how useful it is for them, cannot be measured in a system-oriented fashion. For instance, probably not all kinds of mistakes are perceived the same, and two outputs with the same accuracy may be perceived entirely differently depending on how errors are arranged throughout the piece. How do real users perceive the output from our algorithms? Should we start devising new user-oriented measures? Can we do it without constant human intervention?

3.28 Computational Modeling of Melody

Anja Volk (Utrecht University, NL)

Joint work of Anja Volk, Marcelo Rodriguez-Lopez, Peter van Kranenburg, Iris Yuping Ren

In contrast to harmony and rhythm, melody is not considered a "basic musical structure" [11] in music theory. Accordingly, while there exist many theories on harmony and rhythm, theories on melodies are sparse. The computational modeling of melody in MIR and computational musicology has focused on topics such as modeling melodic similarity [11], melodic segmentation [10], the stability of melodic features over the course of oral transmission in folk music [3], modeling the role of melodies for listeners' expectations [8], and discovering prototypical contours and patterns [2, 8].

Discovering repeated patterns which are musically meaningful is a specifically challenging task, as usually algorithms discover much more patterns than humans [5]. However, repeated patterns have been shown to be crucial for important aspects of melody such as similarity [13, 1] and segmentation [10], and can be used for discovering repeated sung phrases in audio corpora [4]. I consider it an interesting direction of research on how the computational modeling of melody, specifically the aspect of melodic patterns, in symbolically annotated music and in recordings can cross-pollinate each other in order to improve challenging tasks such as melodic contour extraction. Integrating audio and symbolic approaches to modelling melodic aspects would not only contribute to solving specific MIR tasks related to melody, but might contribute to a broader theoretization of the phenomenon of melody also in the context of musicology and cognition.

References

1 Peter Boot, Anja Volk, and W. Bas de Haas. *Evaluating the Role of Repeated Patterns in Folk Song Classification and Compression*. Journal of New Music Research, 45(3), 2016, pp. 223–238.

2 Kaustuv Ganguli and Preeti Rao *Discrimination of Melodic Patterns in Indian Classical Music*. National Conference on Communications, IIT Bombay, India 21 (1), 2015, pp. 1-6.

3 Berit Janssen, Peter van Kranenburg, and Anja Volk. *Finding Occurrences of Melodic Segments in Folk Songs Employing Symbolic Similarity Measures*. Journal of New Music Research, 46(2), 2017, pp. 118–134.

4 Nadine Kroher, Aggelos Pikrakis, and Jose-Miguel Díaz-Báñez, *Discovery of Repeated Melodic Phrases in Folk Singing Recordings*. IEEE Transactions on Multimedia 20(6), 2018, pp. 1291–1304.

5 Iris Yuping Ren, Hendrik Vincent Koops, Anja Volk, and Wuoter Swierstra. *In Search of the Consensus Among Musical Pattern Discovery Algorithms*. Proceedings of the 18th International Society for Music Information Retrieval Conference, Suzhou, China, 2017, pp. 671–678.

6 Iris Yuping Ren, Anja Volk, Wouter Swierstra, and Remco C. Veltkamp. *Analysis by Classification: A Comparative Study of Annotated and Algorithmically Extracted Patterns in Symbolic Music Data*. Proceedings of the 19th International Society for Music Information Retrieval Conference, Paris, France, 2018, pp. 539–546.

7 Iris Yuping Ren, Hendrik Vincent Koops, Dimitrios Bountouridis, Anja Volk, Wouter Swierstra, and Remco C. Veltkamp. *Feature Analysis of Repeated Patterns in Dutch Folk Songs using Principal Component Analysis*. Proceedings of the 8th International Workshop on Folk Music Analysis, Thessaloniki, Greece, 2018, pp. 86–87.

8 Marcelo E. Rodríguez-López and Anja Volk. *Symbolic Segmentation: A Corpus-Based Analysis of Melodic Phrases*. Proceedings of the 10th International Symposium on Computer Music Modeling and Retrieval (CMMR), Laboratoire de Mecanique et d'Acoustique, Marseille, France, 2013, pp. 548–557.

9 Marcelo E. Rodríguez-López and Anja Volk: *Melodic Segmentation Using the Jensen-Shannon Divergence*. Proceedings of the 11th International Conference on Machine Learning and Applications, Boca Raton, USA, 2012, pp. 351–356.

10 Marcelo E. Rodríguez López and Anja Volk. *Location Constraints for Repetition-Based Segmentation of Melodies*. Proceedings of the 5th International Conference in Mathematics and Computation in Music (MCM), Springer, 2015, pp. 73–84.

11 David Temperley, *The Cognition of Basic Musical Structures*. MIT Press, 2001.

12 Peter van Kranenburg, Anja Volk, Frans Wiering, and Remco C. Veltkamp. *Musical Models for Folk-Song Melody Alignment*. Proceedings of the International Society on Music Information Retrieval Conference, Kobe, Japan, pp. 507–512.

13 Anja Volk and Peter van Kranenburg. *Melodic Similarity Among Folk Songs: An Annotation Study on Similarity-Based Categorization in Music*. Musicae Scientiae, 16(3), 2012, pp. 317–339.

3.29 Singing Voice Modelling for Language Learning

Ye Wang (National University of Singapore, SG)

Singing is a popular form of entertainment, as evidenced by the millions of active users of Karaoke apps like Smule's Sing! and Tencent's Quanmin K Ge. Singing is presumed to be the oldest form of music making and can be found in human cultures around the world. However, singing can be more than just a source of entertainment: parents sing nursery rhymes to their young children to help them learn their first language, music therapists use singing to help aphasia patients speak again, and medical studies have revealed that singing, in general, has many health benefits. Consequently, computational methods for singing analysis have emerged as an active research topic in the music information retrieval community.

Pedagogical research has shown that actively singing in a foreign language helps with pronunciation, vocabulary acquisition, retention, fluency, and cultural appreciation [1]. Inspired by this scientific discovery, we have developed a novel multi-language karaoke application called SLIONS (Singing and Listening to Improve Our Natural Speaking), designed to foster engaging and joyful language learning [2]. We followed a user-centered design process that was informed by conducting interviews with domain experts and by conducting usability tests among students. The key feature of SLIONS is an automatic speech recognition (ASR) tool used for objective assessment of sung lyrics, which provides students with personalized, granular feedback based on their singing pronunciation.

During its proof of concept phase, SLIONS employed Google's ASR technology to evaluate sung lyrics. However, this solution lacks technical depth and has several critical limitations. First, Google ASR is proprietary technology and is effectively a black box. As a result, it is impossible for us to understand precisely why it succeeds in evaluating certain sung lyrics but fails in others. This not only prevents us from gaining insights into the underlying models but also affects SLIONS' value in real-world applications. It is also impossible for us to modify Google's ASR technology even though we wish to use SLIONS for widely varying applications that range from language learning to melodic intonation therapy. Google ASR technology is designed for speech recognition and is suboptimal for analyzing singing voice, as the characteristics of sung utterances differ from those of spoken utterances. Therefore it is desirable to investigate better and more versatile computational methods for objective assessment of sung lyrics. Furthermore, it is also useful to address some important human–computer interaction (HCI) questions. For example, while previous studies have shown that singing can help with learning pronunciation, the critical question of which factors are essential for not only improving pronunciation but also maintaining engagement during singing exercises remains. It is vital to design interface/interaction features that support both learning and engagement aspects.

Although speech and singing share a common voice production organ and mechanism, singing differs from speech in terms of pitch variations, possible extended vowels, vibrato, and more. It is interesting to exploit the similarities between speech and singing in order to employ existing methods/tools and datasets in the relatively mature ASR field while also developing new methods to address the differences. To this end, we have created and published the NUS Sung and Spoken Lyrics Corpus, a small phonetically annotated dataset of voice utterances [3]. Furthermore, we have also attempted to address the problem of lyrics and singing alignment [4, 5, 6, 7], evaluation of sung lyrics [8, 9], and intelligibility of sung lyrics [10]. While many challenges remain as to adequately modeling and analyzing singing

voice for real-world applications such as language learning, our efforts are already pointing the way towards a robust, versatile model that can enable the automatic evaluation of sung utterance pronunciation.

References

1 Arla J. Good, Frank A. Russo, and Jennifer Sullivan. *The Efficacy of Singing in Foreign-Language Learning*. Psychology of Music, 43(5), 2015, pp. 627–640.

2 Dania Murad, Riwu Wang, Douglas Turnbull, and Ye Wang. *SLIONS: A Karaoke Application to Enhance Foreign Language Learning*. Proceedings of the ACM International Conference on Multimedia, Seoul, Korea, 2018, pp. 1679–1687.

3 Zhiyan Duan, Haotian Fang, Bo Li, Khe Chai Sim, and Ye Wang. *The NUS Sung and Spoken Lyrics Corpus: A Quantitative Comparison of Singing and Speech*. Proceedings of the IEEE Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA), Kaohsiung, Taiwan, 2013, pp. 1–9.

4 Chitralekha Gupta, Rong Tong, Haizhou Li, and Ye Wang. *Semi-Supervised Lyrics and Solo-Singing Alignment*. Proceedings of International Society for Music Information Retrieval Conference (ISMIR), Paris, France, 2018, pp. 600–607.

5 Min-Yen Kan, Ye Wang, Denny Iskandar, Tin Lay Nwe, and Arun Shenoy. *LyricAlly: Automatic Synchronization of Textual Lyrics to Acoustic Music Signals*. IEEE Transactions on Audio, Speech, and Language Processing, 16(2), 2008, pp. 338–349.

6 Denny Iskandar, Ye Wang, Min-Yen Kan, and Haizhou Li. *Syllabic Level Automatic Synchronization of Music Signals and Text Lyrics*. Proceedings of the ACM International Conference on Multimedia, Santa Barbara, USA, 2006, pp. 659–662.

7 Ye Wang, Min-Yen Kan, Tin Lay Nwe, Arun Shenoy, and Jun Yin. *LyricAlly: Automatic Synchronization of Acoustic Musical Signals and Textual Lyrics*. Proceedings of the ACM International Conference on Multimedia, New York, USA, 2004, pp. 212–219.

8 Chitralekha Gupta, Haizhou Li, and Ye Wang. *Automatic Pronunciation Evaluation of Singing*. Proceedings of the International Conference on Spoken Language Processing (Interspeech), Hyderabad, India, 2018, pp. 1507–1511.

9 Chitralekha Gupta, David Grunberg, Preeti Rao, and Ye Wang. *Towards Automatic Mispronunciation Detection in Singing*. Proceedings of the International Society for Music Information Retrieval Conference (ISMIR), Suzhou, China, 2017, 390–396.

10 Karim M. Ibrahim, David Grunberg, Kat Agres, Chitralekha Gupta, and Ye Wang. *Intelligibility of Sung Lyrics: A Pilot Study*. Proceedings of the International Society for Music Information Retrieval Conference (ISMIR), Suzhou, China, 2017, pp. 686–693.

3.30 Analyzing and Visualizing Intonation in Polyphonic Choral Singing

Christof Weiß (Universität Erlangen-Nürnberg, DE), Sebastian J. Schlecht (Universität Erlangen-Nürnberg, DE), Sebastian Rosenzweig (Universität Erlangen-Nürnberg, DE), and Meinard Müller (Universität Erlangen-Nürnberg, DE)

Unaccompanied vocal music constitutes the nucleus of Western art music and the starting point of polyphony's evolution. Despite an increasing number of exciting studies [1, 2, 3], many facets of polyphonic a cappella singing are yet to be explored and understood. In some preliminary experiments, we made first attempts to investigate and visualize intonation aspects in choral singing. At this moment, many questions arise that are highly interrelated:

Which effects occur at different temporal levels (global reference tuning, pitch drifts, local interval errors)? What are the influences of singing conditions (solo singers, multiple singers per part, feedback and interaction), level of training (professional singers vs. amateurs), and acoustic conditions (room size, reverb)? What is the role of the musical composition and how does it affect tuning adjustment (harmony, complexity of chords, interactions of parts, and the influence of overtones)? Finally, how are these effects perceptually relevant to listeners?

In order to systematically study such questions, we compiled a small but diverse dataset of Anton Bruckner's Gradual "Locus iste" (WAB 23) in different performances. Our examples comprise a 16-singer multi-track recording from the Choral Singing Dataset [3] as well as several commercial and non-commercial performances. Furthermore, we generated sine-tone renditions of the piece with different pitch accuracy using random pitch deviations. We experimented with several types of visualization for investigating pitch trajectories in relation to the score in an intuitive way. Furthermore, we tested more general measures of consonance and tuning quality, which do not require score information (see also Section 3.24). Such visualizations may have high potential as supporting tools for choir training and rehearsals.

References

1 Jiajie Dai and Simon Dixon. *Analysis of Interactive Intonation in Unaccompanied SATB Ensembles*. Proceedings of the International Society for Music Information Retrieval Conference (ISMIR), Suzhou, China, 2017, pp. 599–605.
2 Matthias Mauch, Klaus Frieler, and Simon Dixon. *Intonation in Unaccompanied Singing: Accuracy, Drift and a Model of Reference Pitch Memory*. Journal of the Acoustical Society of America, 136(1), 2014, pp. 401–411.
3 Helena Cuesta, Emilia Gómez, Agusín Martorell, and Felipe Loáiciga. *Analysis of Intonation in Unison Choir Singing*. Proceedings of the International Conference on Music Perception and Cognition (ICMPC), 2018.

3.31 Melody Extraction for Melody Generation

Yi-Hsuan Yang (Academia Sinica – Taipei, TW)

Recent years have witnessed a growing interest in applying machine learning algorithms to automatically compose or generate music in the symbolic domain [1, 4, 5, 6]. As melody is a core part of a music piece, many prior approaches have focused on generating a melody, considered as a sequence of notes (specified at a semitone resolution) [2]. Recently, this is usually done by collecting a large number of melodies in a symbolic format such as XML, finding a way to represent the melodies computationally, and then training a neural network to learn to generate melodies, either with conditions (e. g., chords [4]) or no conditions.

However, we note that there are two significant limitations of the approach above. First, not all the existing melodies in the world have been or can be digitally stored in a symbolic format, and are available to us. Relying on melodies in the symbolic format restricts the quantity as well as the diversity of the training dataset. Second, melodies in semitones lack performance-level attributes such as variations in pitch, dynamics, and note duration. Accordingly, to render realistic audio, one needs to either invite a human musician to play the melody, or to build another machine learning model to synthesize the audio from the given melody.

The PerformanceRNN [3] proposed by Google Magenta, and some follow-up research, attempts to address the second issue mentioned above by generating music with performance-level attributes in a single pass. However, this has been limited to piano music only, and not the melody for general music.

We propose here the idea of using the result of audio-domain melody extraction (e. g., [7]) to learn to compose/generate melodies. Such audio-domain melody extraction algorithms can be applied to any music piece as long as there are audio recordings, thereby bypassing the difficulty of obtaining symbolic data. Moreover, the target output of such algorithms is pitch specified in Hertz (rather than in semitones), comprising performance-level attributes in pitch and note duration. This makes it possible to learn to generate melodies of diverse styles and to learn to generate melodies with expressive qualities.

One can further apply note tracking algorithms (e. g., [8]) to convert the melody contour in Hertz to a melody note sequence in semitones, and use both the Hertz and semitone versions to train melody composition and generation models.

The PerformanceNet [9] is a convolutional neural network model we recently proposed to convert a piano-roll-like symbolic-domain musical score to the magnitude spectrogram, which can then be rendered to audio. Following similar ideas, one can build a model that generates the melody in semitones first, and then, based on that, to generate the melody in Hertz.

The idea can be more broadly considered as a "transcription first, and then generation" approach. It is a promising use case of audio-domain melody extraction algorithms, and it might lead to new progress in music generation research.

References

1. Bob L. Sturm, João Felipe Santos, Oded Ben-Tal, and Iryna Korshunova. *Music Transcription Modelling and Composition Using Deep Learning*. Proceedings of the Conference on Computer Simulation of Musical Creativity, Huddersfield, UK, 2016.
2. Elliot Waite. *Project Magenta: Generating Long-Term Structure in Songs and Stories*. 2016. https://magenta.tensorflow.org/2016/07/15/lookback-rnn-attention-rnn
3. Ian Simon and Sageev Oore. *Performance RNN: Generating Music with Expressive Timing and Dynamics*. 2017. https://magenta.tensorflow.org/performance-rnn/
4. Li-Chia Yang, Szu-Yu Chou, and Yi-Hsuan Yang. *MidiNet: A Convolutional Generative Adversarial Network for Symbolic-domain Music Generation*. Proceedings of the International Society for Music Information Retrieval Conference, Suzhou, China, 2017, pp. 324–331.
5. Hao-Wen Dong, Wen-Yi Hsiao, Li-Chia Yang, and Yi-Hsuan Yang. *MuseGan: Multi-Track Sequential Generative Adversarial Networks for Symbolic Music Generation and Accompaniment*. Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, USA, 2018, pp. 34–41.
6. Jean-Pierre Briot, Gaëtan Hadjeres, and François Pachet. *Deep Learning Techniques for Music Generation: A Survey*. arXiv preprint, 2017. http://arxiv.org/abs/1709.01620
7. Justin Salamon, Emilia Gómez, Daniel P. W. Ellis, Gaël Richard:. *Melody Extraction from Polyphonic Music Signals: Approaches, Applications and Challenges*. IEEE Signal Processing Magazine, 31(2), 2014, pp. 118–134.
8. Yuan-Ping Chen, Ting-Wei Su, Li Su, and Yi-Hsuan Yang. *TENT: Technique-Embedded Note Tracking for Real-World Guitar Solo Recordings*. Transactions of the International Society for Music Information Retrieval, 2019.
9. Bryan Wang and Yi-Hsuan Yang. *PerformanceNet: Score-To-Audio Music Generation with Multi-Band Convolutional Residual Network*. Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, USA, 2019.

3.32 Finding Musical Themes in Western Classical Music Recordings

Frank Zalkow (Universität Erlangen-Nürnberg, DE), Stefan Balke, and Meinard Müller (Universität Erlangen-Nürnberg, DE)

Many pieces from Western classical music contain short melodies or musical gestures that are especially prominent and memorable, so-called musical themes. Finding such themes in audio recordings is a challenging cross-modal retrieval scenario. In such a setting, the query is a symbolic encoding of the theme's melody, and the database is a collection of classical music recordings. In this scenario, we face several problems: the difference in modality, differences in tuning, transposition, and tempo, as well as the difference in polyphony between a query and database document [1].

Usually, one employs common mid-level representations for performing retrieval tasks with different modalities. In particular, one may use chroma features, which measure the energy in the twelve chromatic pitch class bands. The difficulty due to the difference in polyphony could be bypassed with accurate melody extraction methods [4]. Those methods often work on so-called salience representations, which are time–frequency representations with enhanced tonal frequency components [2, 3]. However, for polyphonic classical music, melody extraction often fails. Therefore, we propose not to extract the melodies, but to directly map the salience representations to chroma features before performing the retrieval [6]. As an alternative, we suggest learning common mid-level representations for both query and database with a data-driven approach, e. g., using deep learning with the triplet loss [5].

References

1 Stefan Balke, Vlora Arifi-Müller, Lukas Lamprecht, and Meinard Müller. *Retrieving Audio Recordings Using Musical Themes*. Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Shanghai, China, 2016, pp. 281–285.

2 Meinard Müller. *Fundamentals of Music Processing*. Springer Verlag, 2015.

3 Justin Salamon and Emilia Gómez. *Melody Extraction from Polyphonic Music Signals Using Pitch Contour Characteristics*. IEEE Transactions on Audio, Speech, and Language Processing, 20(6), pp. 1759–1770, 2012.

4 Justin Salamon, Emilia Gómez, Daniel P. W. Ellis, and Gaël Richard. *Melody extraction from polyphonic music signals: Approaches, applications, and challenges*. IEEE Signal Processing Magazine, 31(2), pp. 118–134, 2014.

5 Florian Schroff, Dmitry Kalenichenko, and James Philbin. *FaceNet: A Unified Embedding for Face Recognition and Clustering*. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 2015, pp. 815–823.

6 Frank Zalkow, Stefan Balke, and Meinard Müller. *Evaluating Salience Representations for Cross-Modal Retrieval of Western Classical Music Recordings*. Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Brighton, United Kingdom, 2019.

4 Working Groups

4.1 Data Sets: What is Missing and How do We Get It?

Participants of Dagstuhl Seminar 19052

This working group considered the data needs of singing research, specifically the lack of isolated vocal data for singing voice synthesis and other singing voice research. We proposed that the first step in addressing the issue of data paucity is to establish what data we already have and where exactly the holes are. To this end, we discussed developing a shared document (perhaps on Google) that lists existing data. We envision this as a bootstrapped (and free) version of the Linguistic Data Consortium[8]. We also considered the various ways in which data could be collected (including recordings by researchers, collaborating with karaoke services and other relevant companies, and using source-separation on existing recordings), how we may deal with related copyright issues, and how much data we need for different tasks. This led us to consider how diverse our data needs to be in terms of pitch range, vocabulary, singer type, and genre as well as how it should be annotated with respect to musical and lyrical content. We concluded that, in addition to creating a resource for listing available data sets, we should also work on creating community-sourced guidelines on best practices on curating and annotating singing data sets.

4.2 Deep Learning Versus Acoustic Models

Participants of Dagstuhl Seminar 19052

In this working group, we discussed how different types of systems can represent the acoustical and physical properties of singing. The main focus was on the relationship between implicit knowledge, as encoded in a Deep Neural Network (DNN) trained on annotated data, and explicit knowledge, as built into hand-crafted signal processing approaches.

As the first central question, we discussed to which extent current Deep Learning (DL) approaches exploit acoustic models of singing production in the design of their model architecture. We noticed that there is only a small number of studies that explicitly exploit such knowledge. For example, in the work by Basaran et al. [1], a source–filter model is used as an input to a CNN. Another example is the WaveNet vocoder [2]. For the majority of works, knowledge is only given implicitly to the system via training on suitable data. In this context, data augmentation strategies play a crucial role to enforce that a system possesses certain invariance properties.

As the second major question, we wondered what kind of knowledge about singing production could potentially be derived from trained systems. For example, can we learn an acoustic or physical model of singing production from a DL system? We agreed that deriving any explicit knowledge from a trained model is generally hard. The architecture itself

[8] https://catalog.ldc.upenn.edu

provides little insights; indeed many architectures are used successfully for many different problems within MIR. Further points that arose in our discussion touched the challenge of data collection. As one strategy, semi-supervised training procedures were proposed, for example, by exploiting information from synchronized videos or by making use of suitable embeddings.

References

1 Dogac Basaran, Slim Essid, and Geoffroy Peeters. *Main Melody Estimation with Source-Filter NMF and CRNN*. Proceedings of the International Society for Music Information Retrieval Conference, Paris, France, 2018, pp. 82–89.
2 Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. *Wavenet: A generative model for raw audio*. Electronic preprint (arXiv:1609.03499), 2016.

4.3 Insights

Participants of Dagstuhl Seminar 19052

In this working session, we addressed the question of what it means to provide "insights" into tasks, techniques, data, and any aspect related to our research area. We started by discussing how one may improve and evaluate the insights provided by ISMIR papers. We agreed on the usefulness of community guidelines that may help authors, reviewers, and meta-reviewers to improve and evaluate the usefulness of scientific contributions. The following list contains some of our suggestions:

- A paper should go beyond stating something like "we got x% accuracy on this dataset."
- The aspect of "insights" could be added in author guidelines. Furthermore, it could be added in the ISMIR paper template.
- Good/bad examples of result and conclusion sections should be provided. Such examples could be collected via "crowdsourcing."
- Guidelines could be provided on the ISMIR website as early as possible so that potential authors are ready and prepared.
- As for reviewing, the review form could be expanded by a specific question about the insights of a submission. Reviewers should explicitly comment on technical, conceptual and other insights provided by a submission.
- Meta-reviewers should evaluate a paper according to its insights. Such insights may refer to different aspects, e. g. machine learning insights or general MIR insights.
- The discussion could be continued within the broader ISMIR community, e. g., by means of a shared Google doc or an email thread.

In the second half of our group discussion, we had a controversial debate about the importance of sharing research code as part of publications. In particular, we discussed whether open source code is useful or necessary for improving the insight provided by MIR papers. A variety of opinions and issues were mentioned, which can be roughly summarized as follows:

- Code must be shared. Otherwise, the research is not reproducible, and thus provides little insight. Without sharing the code, how can the community be confident the insight provided by a paper is correct?

- Code sharing should be encouraged but not mandatory. Sharing the code is its own reward, as it increases impact. Code sharing increases reproducibility, which in turn increases impact.
- Code sharing is nice, but far less important than providing a good description of the method and analysis of the results. In other words, sharing code in lieu of a good description of the method is not acceptable. Papers should not be rejected for not sharing code.
- Code sharing is irrelevant; the explanations given in a paper should be sufficient. Open-source should not play any role in the review process.
- Will research that does not come bundled with open-source code be relevant in 10 years?
- Should MIR students learn open-source coding skills to increase their job prospects?

We agreed that source code has a different degree of importance in different subareas of MIR research. In some areas of MIR research (e. g., deep learning) the line between research and software engineering is blurring. As a community, we should maintain a broad view of these issues accepting people on either side of the debate. A further debate may be informed by a recent publication on the topic of open-source software for MIR research, see [1].

References

1 Brian McFee, Jong Wook Kim, Mark Cartwright, Justin Salamon, Rachel M. Bittner, Juan Pablo Bello. *Open-Source Practices for Music Signal Processing Research: Recommendations for Transparent, Sustainable, and Reproducible Audio Research*. IEEE Signal Processing Magazine, 36(1), 2019, pp. 128–137.

4.4 Singing Assessment and Performance Evaluation

Participants of Dagstuhl Seminar 19052

The role of music technology in music education is visible, but not as much as one may guess or hope for. The reason for this gap may be attributed to the fact that MIR and music education are two different communities with their distinct outlooks and goals. For example, while singing teachers are likely to believe that abstract aesthetic judgments play the major role in their assessment of learners, MIR researchers often assume that the rapid improvements in automatic extraction of pitch, intonation, rhythm, and dynamics from audio signals herald the coming of powerful teaching tools [1]. In this working group, we reflected on such issues and discussed the potential and limitations of specific technologies for music education. The following list gives an impression on the topics covered:

- Slowing down of recorded musical pieces for superior listening and understanding has been a prevalent but technologically simple tool.
- Separation of sources in a mixed recording in order to enable easy listening to a chosen part, or to eliminate a part in order to use the accompaniment to practice along with, has also been sought after. While source separation technology is not robust enough for use with arbitrarily generated mixes, special recordings have been used to facilitate this.
- Automatic feedback may be a boon to self-learning and practice where learners do not have to rely on their perception to guide them. Computer feedback can also be preferred at times as being less intrusive. The type of feedback that can be reliably provided is

about accuracy in pitch and rhythm concerning a reference (i. e., learning by imitation scenario). Some karaoke apps do precisely this based on a very basic transcription of the reference song. Given the difficulty in characterizing the pitch of a note given the continuously varying pitch in the singing of lyrics and the presence of consonants, such feedback is based on some gross measures only. Even feedback on onsets is not completely reliable given the variety of sung phonemes and singing styles.

- Instead of explicit feedback, it may be preferred to provide a visual representation of the detected continuous pitch superposed on the expected melody including events such as vibrato as in the MiruSinger singing interface [2]. Even scoring an imitation is not easy but such visual comparisons can help a large category of learners who aim to imitate chosen reference songs as closely as possible.
- To make learning by imitation easier, it may help further to create audio tracks that render the reference song in the learner's pitch range and with similar voice quality.
- At the opposite extreme, it would be interesting to predict the actual ratings of human experts via regression methods. This might be expected to involve a more holistic examination of the audio recording with possibly multiple musical attributes taken into account simultaneously. Considering that there are probably several distinct ways to sound right, it may be better to focus on identifying no-gos and designing algorithms to trap such forbidden events in the singing.
- Apart from pitch and timing, vocal dynamics and voice texture or timbre are of great interest, especially to the more advanced singers and performers. Expressiveness in performance, for example, owes itself to intonation, dynamics and voice quality. Again, visual representations of these parameters that facilitate comparisons with a reference may be the best option.
- Recently, feedback in the context of choir singing has been of much interest. Choir directors can benefit from automatic detection of dissonance and of tuning drift with time. Singers can also benefit from source separation tools and the ability to get automatic accompaniment of the other parts in order to practice choir singing away from the choir.
- Non-Western music presents distinct challenges based on the specific genre. In classical Indian music, the rendering of transitions between notes and of ornaments is as important as note accuracy. Salient features are extracted from complex pitch movements to create a perceptually meaningful space for comparisons [3]. More research is needed to understand what constitutes out-of-tune singing in this context and how other attributes such as loudness/dynamics and vocal timbre might play a role.

References

1 Christian Dittmar, Estefanía Cano, Jakob Abeßer, and Sascha Grollmisch. *Music information retrieval meets music education*. In Dagstuhl Follow-Ups (Vol. 3). Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2012, pp. 95–120.

2 Tomoyasu Nakano, Masataka Goto, and Yuzuru Hiraga. *MiruSinger: A Singing Skill Visualization Interface Using Real-Time Feedback and Music CD Recordings as Referential Data*. Proceedings of the IEEE International Symposium on Multimedia Workshops (ISMW), Beijing, 2007, pp. 75–76.

3 Chitralekha Gupta and Preeti Rao. *Objective Assessment of Ornamentation in Indian Classical Singing*. In Speech, Sound and Music Processing: Embracing Research in India. Springer, Berlin & Heidelberg, Germany, 2011, pp. 1–25.

4.5 Subjective Evaluation and Objective Metrics for Singing Voice Generation and Separation

Participants of Dagstuhl Seminar 19052

In light of recent deep learning approaches based on generative models in the fields of source separation and voice synthesis, it is important to discuss an evaluation strategy that is standardized and well accepted. In recent years, there has been some discussion [3] on the effectiveness of objective and perceptually inspired metrics such as SDR, SAR and SIR [1, 2]. For voice synthesis, it has been shown that objective measures like the Mel-Cepstral Distortion do not correlate well with subjective evaluation via listening tests [4]. In this working group, we addressed these issues and proposed the idea of formulating guidelines for publishing evaluation results. In particular, we agreed on the following suggestions:

- Existing objective metrics should always be provided.
- Additionally, subjective evaluations should be conducted.
- In case of a discrepancy between objective metrics and subjective evaluations, a preliminary explanation should be provided, along with supportive and exceptional examples, possibly via a supplementary website.
- GitHub and/or Zenodo should be used for sharing sound examples of the results. Furthermore, if possible, source code or detailed explanations on how the examples were generated should be provided.

With consideration to these guidelines for authors, we also considered the possibility of finding effective objective metrics which can bridge the gap between subjective and objective evaluation. While this seems a daunting task, the possibility of using discriminators based on generative adversarial networks (GANs) was discussed.

References

1 Emmanuel Vincent, Rémi Gribonva, and Cédric Févotte. *Performance Measurement in Blind Audio Source Separation*. IEEE Transactions on Audio, Speech and Language Processing, 14(4), pp. 1462–1469.

2 Emmanuel Vincent. *Improved Perceptual Metrics for the Evaluation of Audio Source Separation*. International Conference on Latent Variable Analysis and Signal Separation. Springer, Berlin & Heidelberg, Gemany, 2012, pp. 430–437.

3 Dominic Ward, Hagen Wierstorf, Russell D. Mason, Emad M. Grais, and Mark D. Plumbley. *BSS Eval or PEASS? Predicting the Perception of Singing-Voice Separatio*. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, Canada, 2018, pp. 596–600.

4 Merlijn Blaauw and Jordi Bonada. *A Neural Parametric Singing Synthesizer*. Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech), Stockholm, Sweden, 2017, pp. 4001–4005.

5 Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. *Generative Adversarial Nets*. Advances in neural information processing systems, Montreal, Canada, 2014, pp. 2672–2680.

4.6 Symbolic Domain Melody Estimation

Participants of Dagstuhl Seminar 19052

In this working group, we considered the task of identifying the melody notes from single-track (e. g., piano) or multi-track symbolic music representation. For melody estimation from multi-track music, this also involves identifying the musical track that contains the melody for each time instance. We call this task "symbolic-domain melody estimation." To our knowledge, this is a task that has been only sporadically addressed in the literature, and the focus of recent work is mostly on voice separation (i. e., the separation of symbolic music into perceptually independent streams of notes such as voices or melodic lines) [1, 2, 4, 5, 6].

There are a few possible reasons why the task has not received full attention in the community. First, over the past two decades, most research activities in the MIR community are concerned with the analysis, processing, and retrieval of audio-based rather than symbolic representations [3]. Second, for people working with symbolic domain data, melody estimation is typically assumed as a pre-processing step rather than the primary research focus; people tend to assume that the melody is given and work on something else that interests them more. There is no strong demand, or use case, for symbolic-domain melody estimation.

We found a new use case for symbolic-domain melody estimation: the support AI-based music generation. Currently, most research on music generation is on melody/chord generation. Here, symbolic data formats such as the ABC format [7] or MusicXML of lead sheets [8], where the melody track is specified, are used. A natural extension of such research is to generate the accompaniment, which may be composed of multiple musical tracks other than the melody and chords. To learn to generate accompaniment tracks, one can use MIDI files as the training data, as done by Dong et al. [9]. However, as people typically do not specify which track in a MIDI file is the melody track, the MuseGAN model presented by Dong et al. [9] can generate only the accompaniment, not the combination of melody and accompaniment. Liu et al [10] attempted to address this issue by using chord-related representations to connect lead sheets and MIDI files (since we can compute chord-related features from both data). The model can generate not only the melody but also the arrangement for the music of 4-bar long. However, it falls short when it comes to generating more extended music, as that requires paired data of lead sheets and MIDI files that are both long enough.

We proposed that symbolic domain melody estimation, in particular identifying the melody track from a multi-track MIDI file, can be an essential building block toward generating multi-instrument music. With symbolic domain melody estimation, we can learn better the relationship between melody and accompaniment, without the need for paired data of lead sheets and MIDI files.

We also proposed that a possible approach to encourage research along this line is to build a MIDI data set with the melody tracks labeled (for example, from the Lakh MIDI data set [11]). This dataset can then be used to train a supervised model for symbolic domain melody estimation. With the labeled MIDI files, we can evaluate how well unsupervised, rule-based methods for symbolic domain melody estimation works, and whether there is gain by combining supervised and unsupervised methods.

Symbolic domain melody estimation is an interesting problem on its own, and it has strong applications in music generation. We hope this extended abstract can call for more attention toward this research problem.

References

1 David Rizo, Pedro J. Ponce de León, Antonio Pertusa, Carlos Pérez-Sancho, and José Manuel Iñesta Quereda. *Melody Track Identification in Music Symbolic Files*. Proceedings of the International Florida Artificial Intelligence Research Society Conference, Melbourne Beach, USA, 2006, pp. 254–259.

2 Dimitrios Rafailidis, Alexandros Nanopoulos, Yannis Manolopoulos, and Emilios Cambouropoulos. *Detection of Stream Segments in Symbolic Musical Data*. Proceedings of the International Conference on Music Information Retrieval (ISMIR), Philadelphia, USA, 2008, pp. 83–88.

3 Meinard Müller. *Fundamentals of Music Processing – Audio, Analysis, Algorithms, Applications*. Springer Verlag, 2015.

4 Patrick Gray, Razvan C. Bunescu. *A Neural Greedy Model for Voice Separation in Symbolic Music*. Proceedings of the International Conference on Music Information Retrieval (ISMIR), New York City, USA, 2016, pp. 782–788.

5 Andrew McLeod and Mark Steedman. *HMM-Based Voice Separation of MIDI Performance*. Journal of New Music Research, 2016, 45(1), pp. 17–26.

6 Wei-Tsung Lu and Li Su. *Deep Learning Models for Melody Perception: An Investigation on Symbolic Music Data*. Proceedings of the Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, Honolulu, USA, 2018, pp. 1620–1625.

7 Bob L. Sturm, João Felipe Santos, Oded Ben-Tal, and Iryna Korshunova. *Music Transcription Modelling and Composition Using Deep Learning*. Proceedings of the Conference on Computer Simulation of Musical Creativity, 2016.

8 Li-Chia Yang, Szu-Yu Chou, and Yi-Hsuan Yang. *MidiNet: A Convolutional Generative Adversarial Network for Symbolic-domain Music Generation*. Proceedings of the International Society for Music Information Retrieval Conference, Suzhou, China, 2017, pp. 324–331.

9 Hao-Wen Dong, Wen-Yi Hsiao, Li-Chia Yang, and Yi-Hsuan Yang. *MuseGAN: Multi-track sequential generative adversarial networks for symbolic music generation and accompaniment*. Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, USA, 2018, pp. 34–41.

10 Hao-Min Liu and Yi-Hsuan Yang. *Lead Sheet Generation and Arrangement by Conditional Generative Adversarial Network*. Proceedings of the International Conference on Machine Learning and Applications (ICMLA), Orlando, USA, 2018, pp. 722–727.

11 Colin Raffel. *Learning-Based Methods for Comparing Sequences, with Applications to Audio-to-MIDI Alignment and Matching*. Ph.D. Thesis, Columbia University, 2016.

4.7 Transcription beyond Western Music

Participants of Dagstuhl Seminar 19052

In this working group, we discussed aspects related to transcribing and representing music beyond standard Western music notation. Before defining how such a representation should look, it is essential to consider the intended use case.

One use case could be ethnomusicological research and archival purposes. For example, in the context of Georgian vocal music [1], we discussed a software tool for displaying a piano-roll representation, with further visual cues, like Gaussian approximations of sung pitches, annotations for sung lyrics, and so on. In this scenario, many important issues remain. First, it is unclear how to enrich such representations with timbre or tuning attributes. Annotations

for lyrics are very difficult to obtain for this kind of music. The automatic alignment of audio and lyrics would be of great help. An idea that emerged out of this working group was to approach this problem through clustering techniques of MFCC features [2].

Another important use case for music representations is performance practice. We agreed that representations for such a purpose have to be similar to Western score representations to a certain extent. The reason is that musicians are well-trained in reading this notation. Thus, all attempts of replacing such standardized representations did not succeed in the history of music. In general, performance practice representations need to be much more abstract than representations for archival purposes.

We also discussed several historical and modern representations of music and prosody, like neumes, a web representation for Indian classical music[9], rough categories of sung notes in Western music [3], Japanese music[10], ancient Chinese music[11], contemporary music notation [4], conventions for transcribing and annotating the prosody of speech[12], the Music Encoding Initiative [5], and the linking of performance data with score information [6]. A general treatment of music notation is given by Read [7].

Furthermore, we also discussed how to technically approach the problem of creating such representations. One of the most import questions is how reliable the extraction of fundamental frequency (F0) trajectories from audio recordings is. Also, the segmentation of F0 trajectories into notes is a difficult problem. A further aspect is the automatic detection of breathing sounds in singing voice music [8].

In summary, the participants agreed on a great interest in generalized music representations, which lie in between F0 trajectories and note representations.

References

1 Frank Scherbaum, Meinard Müller, and Sebastian Rosenzweig. *Rechnergestützte Musikethnologie am Beispiel historischer Aufnahmen mehrstimmiger georgischer Vokalmusik*. Proceedings of the GI Jahrestagung, 2017.

2 Ben Gold, Nelson Morgan, and Dan Ellis. *Speech and Audio Signal Processing*. John Wiley & Sons, Hoboken, USA, 2011.

3 Jiajie Dai. *Modelling Intonation and Interaction in Vocal Ensembles*. Ph.D. thesis, Queen Mary University of London, to appear.

4 Jonathan Feist. *Berklee Contemporary Music Notation*. Berklee Press, Boston, USA, 2017.

5 Andrew Hankinson, Perry Roland, and Ichiro Fujinaga. *The Music Encoding Initiative as a Document-Encoding Framework*. Proceedings of the International Society for Music Information Retrieval Conference, Miami, USA, 2011, pp. 293–298.

6 Johanna Devaney and Hubert Léveillé Gauvin. *Representing and Linking Music Performance Data with Score Information*. Proceedings of the International Workshop on Digital Libraries for Musicology, New York City, USA, 2016, pp. 1–8.

7 Gardner Read. *Music Notation*. Crescendo Publishing, Boston, USA, 1964.

8 Tomoyasu Nakano, Jun Ogata, Masataka Goto, and Yuzuru Hiraga. *Analysis and Automatic Detection of Breath Sounds in Unaccompanied Singing Voice*. Proceedings of the International Conference on Music Perception and Cognition, Sapporo, Japan, 2008, pp. 387–390.

[9] https://autrimncpa.wordpress.com

[10] In this tradition, extra symbols exist that indicate how to express lyrics, called “Shigin.”

[11] There exists a traditional musical notation method called “Gongche” notation.

[12] A set of conventions is called tones and break indices (ToBI).

4.8 Demo Session

Participants of Dagstuhl Seminar 19052

On Thursday evening, we had a demo and late-breaking news session. The contributions included scientific ideas, graphical user interfaces, data sets, and audio examples. The following list gives an overview of the contributions:

- Sebastian Schlecht, Sebastian Rosenzweig, Christof Weiß: Realtime Dissonance Detection
- Ryo Nishikimi: Automatic Singing Transcription
- Zhiyao Duan: AIR Lab Demo
- Yi-Hsuan Yang: Song Mixer
- Yi-Hsuan Yang: Latent Inspector
- Justin Salamon: Crepe
- Estefanía Cano Cerón: Columbian Music Archive
- Pritish Chandna: Generative Models for Singing Voice Synthesis
- Simon Dixon: Tony
- Juhan Nam: Piano Re-Performance
- Juhan Nam: VirtuosoNet: Expressive Piano Performance Rendering
- Juhan Nam: Singing Voice Synthesis Using Conditional GAN

5 Music and Recording Sessions

As one major overall topic, polyphonic singing was discussed over the course of the seminar. In particular, participants mentioned their interest in Western choral singing in several stimulus talks and submitted abstracts. In this context, pitch, intonation, and singer interaction constitute essential aspects. Detailed studies on such aspects typically require multitrack recordings comprising one or several tracks per singer as well as manual annotations, e. g., in terms of an aligned musical score and lyrics. However, recording multitrack audio in the choir context is challenging, since singers can hardly be recorded in separation and conventional microphones suffer from bleeding between different voices. The lack of suitable recordings and the joy of singing led to the idea of forming a small choir of Dagstuhl participants in order to record a multitrack dataset of choir singing using different types of microphones and sensors that capture both individual voices and the overall acoustic impression.

5.1 Choir Rehearsals

Christof Weiß (Universität Erlangen-Nürnberg, DE), Sebastian Rosenzweig (Universität Erlangen-Nürnberg, DE), Helena Cuesta (UPF – Barcelona, ES), Frank Scherbaum (Universität Potsdam – Golm, DE), Emilia Gómez (UPF – Barcelona, ES), Meinard Müller (Universität Erlangen-Nürnberg, DE)

In several rehearsals (see Figure 1) taking place during breaks and in the evenings, we tried a number of choir pieces collected in advance of the seminar. After several tries, we selected Anton Bruckner's Gradual "Locus iste" (WAB 23) in Latin (Figure 2). This 3-minute long choir piece is musically interesting, contains several melodic and harmonic challenges, and covers a large part of each voice's tessitura. Beyond that, this piece is part of the *Choral singing dataset* [1], thus allowing for interesting comparative studies. As further works, we considered the piece "Tebe poem" by the Bulgarian composer Dobri Hristov, the Catalan traditional song "El rossinyol," and "Otche nash" by the Russian composer Nikolai Kedrov, among others. All pieces were written for SATB choir in four parts, each of which we could perform with at least two singers (see Table 1 for details). The rehearsals usually started with a vocal warm-up led by Polina Proutskova, who works as a singing teacher. Composer Christof Weiß took the role of the choir director for practicing and performing the pieces. Over the first three days of the seminar, the choir steadily improved and reached a reasonable level of musical quality, which might be considered representative of a good amateur choir. Despite the varying level of singing training, the choir could produce a quite homogeneous sound.

The recording session mostly focused on "Locus iste" and included different variations of singer configuration and expressions. Beyond several runs performed by the full choir, we also recorded this piece in two different quartet versions with only one singer for each part (Table 1). For studying the effect of unison singing, we captured the beginning part with the basses only, thereby varying the number of singers (1–5 bass singers). Additionally, we recorded "Tebe poem" with the full choir and a small number of intonation exercises [2] with one of the quartets (*Quartet II*).

Figure 1 Choir rehearsal.

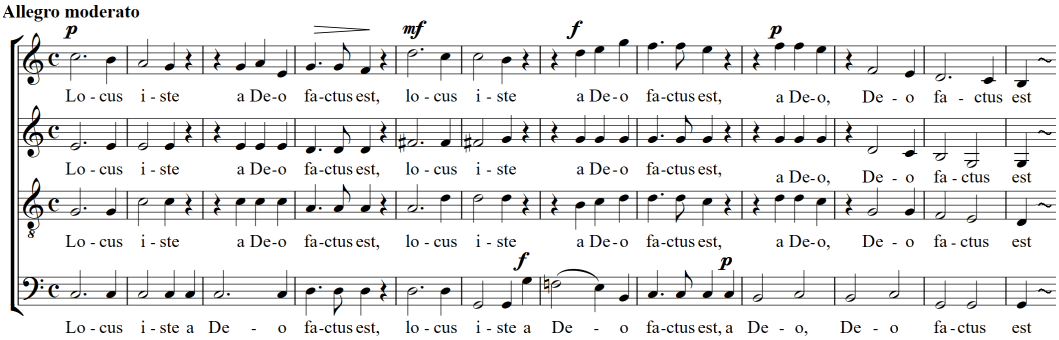


Figure 2 First measures of Anton Bruckner's "Locus iste" (WAB 23).

References

1 Helena Cuesta, Emilia Gómez, Agusín Martorell, and Felipe Loáiciga. *Analysis of Intonation in Unison Choir Singing*. Proceedings of the International Conference on Music Perception and Cognition (ICMPC), 2018.

2 Per-Gunnar Alldahl. *Choral Intonation*. Gehrmans, Stockholm, Sweden, 2008.

5.2 Recording Documentation

Sebastian Rosenzweig (Universität Erlangen-Nürnberg, DE), Helena Cuesta (UPF – Barcelona, ES), Frank Scherbaum (Universität Potsdam – Golm, DE), Christof Weiß (Universität Erlangen-Nürnberg, DE), Emilia Gómez (UPF – Barcelona, ES), Meinard Müller (Universität Erlangen-Nürnberg, DE)

The recording session took place on the third day of the seminar in room "Kaiserslautern." The room was equipped with recording devices, several microphones and many cables, organized and operated by Sebastian Rosenzweig, Helena Cuesta, and Frank Scherbaum. To precisely study the individual singers' performances and behaviors, we recorded selected singers with

Table 1 Singers participating in the choir recordings.

	Quartet I	*Quartet II*	*Full choir (additional singers)*
Soprano	Emilia Gómez	Polina Proutskova	
Alto	Rachel Bittner	Cynthia Liem	
Tenor	Justin Salamon	Simon Dixon	Zhiyao Duan, Tomoyasu Nakano
Bass	Meinard Müller	Frank Zalkow	Frank Kurth, Sebastian Schlecht, Li Su

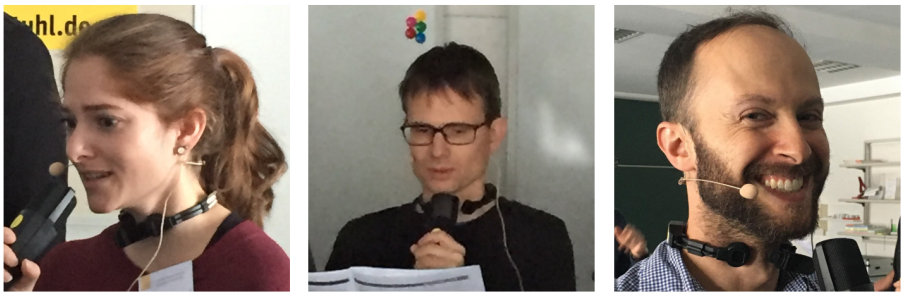


Figure 3 Microphone setup for three singers.

multiple close-up microphones such as throat, headset, and dynamic microphones. Throat microphones, which capture the vibrations of a singer's throat, have shown to be particularly useful for such studies thanks to their robustness to cross-talk from other singers [1]. The microphone setup for the singers is demonstrated in Figure 3. With this setup, we recorded two singers per voice section in the full choir and each singer in the quartets. Additionally, we recorded each setting with a stereo room microphone placed in a distance of about three meters from the singers to capture the overall impression.

In parallel to the multichannel audio recordings, we documented the session with videos and pictures. Furthermore, we equipped one singer with an ambulatory monitoring system to acquire behavioral and physiological data during the performances.[13] Another singer was equipped with binaural microphones to record the choir performances from a singer's perspective.

After the seminar, we plan to collect, synchronize, cut, and annotate the recorded material with the goal to create a publicly accessible dataset that may serve various research purposes for melody and voice processing. All singers already provided their consent to release the dataset along with all metadata and annotations for research.

References

1 Frank Scherbaum. *On the Benefit of Larynx-Microphone Field Recordings for the Documentation and Analysis of Polyphonic Vocal Music*. Proceedings of the International Workshop on Folk Music Analysis, Dublin, Ireland, 2016, pp. 80–87.

[13] http://www.vu-ams.nl/

Participants

- Rachel Bittner
Spotify – New York, US
- Estefanía Cano Cerón
Fraunhofer IDMT, DE &
A*STAR – Singapore, SG
- Michèle Castellengo
Sorbonne University – Paris, FR
- Pritish Chandna
UPF – Barcelona, ES
- Helena Cuesta
UPF – Barcelona, ES
- Johanna Devaney
Brooklyn College, US
- Simon Dixon
Queen Mary University of
London, GB
- Zhiyao Duan
University of Rochester, US
- Emilia Gómez
UPF – Barcelona, ES
- Masataka Goto
AIST – Tsukuba, JP
- Frank Kurth
Fraunhofer FKIE –
Wachtberg, DE
- Cynthia Liem
TU Delft, NL
- Antoine Liutkus
INRIA, University of
Montpellier, FR
- Meinard Müller
Universität Erlangen-
Nürnberg, DE
- Tomoyasu Nakano
AIST – Tsukuba, JP
- Juhan Nam
KAIST – Daejeon, KR
- Ryo Nishikimi
Kyoto University, JP
- Geoffroy Peeters
Telecom ParisTech, FR
- Polina Proutskova
Queen Mary University of
London, GB
- Preeti Rao
Indian Institute of Technology
Bombay, IN
- Sebastian Rosenzweig
Universität Erlangen-
Nürnberg, DE
- Justin Salamon
Adobe Research, US
- Frank Scherbaum
Universität Potsdam – Golm, DE
- Sebastian J. Schlecht
Universität Erlangen-
Nürnberg, DE
- Li Su
Academia Sinica – Taipei, TW
- Tomoki Toda
Nagoya University, JP
- Julián Urbano
TU Delft, NL
- Anja Volk
Utrecht University, NL
- Ye Wang
National University of Singapore,
SG
- Christof Weiß
Universität Erlangen-
Nürnberg, DE
- Yi-Hsuan Yang
Academia Sinica – Taipei, TW
- Frank Zalkow
Universität Erlangen-
Nürnberg, DE