VocalNotes: Investigating the Perception of Note Pitch and Boundaries through Varying Transcriptions of Vocal Performances from Five Musical Cultures

Polina Proutskova (BBC/Queen Mary University of Lonson, UK, proutskova@googlemail.com)
Gakuto Chiba (Keio University, Japan, gane1222@sfc.keio.ac.jp)
Miranda Crowdus (Concordia University, Canada, miranda.crowdus@concordia.ca)
Yulia Nikolaenko (independent researcher, UK, mitida173@gmail.com)
Yuto Ozaki (Keio University, Japan, yuto_ozaki@keio.jp)
Lawrence Shuster (Cornell University, USA, lbs239@cornell.edu)
Olga Velichkina (French Society for Ethnomusicology, France, olga.velichkina@gmail.com)
Yannick Wey (Lucerne University of Applied Sciences and Arts, Switzerland, yannick.wey@hslu.ch)
Zhaoxin Yu (Shandong College of Arts, China, 943469149@qq.com)
Wei Yue (Shandong College of Arts, China, yuewei1981zgyy@126.cm)
Gabriel Zuckerberg (Brown University, USA, gabriel_zuckerberg@brown.edu)
Yukun Li (Queen Mary University of London, UK, yukun.li@qmul.ac.uk)
Andrew Killick (University of Sheffield, UK, a.killick@sheffield.ac.uk)
John M. McBride (Center for Algorithmic and Robotized Synthesis, Institute for Basic Science, South Korea, jmmcbride@protonmail.com)
Elizabeth Phillips (McMaster University, Canada, phille10@mcmaster.ca)
Patrick E. Savage (Keio University, Japan & University of Auckland, New Zealand, psavage@sfc.keio.ac.jp)

Abstract

The VocalNotes project investigated how expert traditional music listeners conceive of notes in vocal performances by studying similarities and differences in their transcriptions. Teams of experts from five musical traditions (Japanese folk song, Chinese *bangzi* opera, Russian traditional village singing, Alpine yodelling, and Romaniote Jewish chanting) each transcribed ~10 minutes of vocal recordings from their culture, where manual transcription consisted of segmentation and note pitch correction, starting from an automatically extracted pitch curve. The experts then compared their independent transcriptions and looked for factors which could have led to disagreements.

Western staff notation is not suitable for investigating such variances, because it does not represent sufficiently fine gradations of pitch and timing. We therefore used tools that allowed more precise annotations, namely Tony for segmentation, and Sonic Visualiser for note pitch correction and transcription comparison.

We found that overall agreement was prevalent and the concept of note was generally applicable for analysis of vocal performances. Yet in some contexts disagreements were abundant, with the note concept reaching its limits. We identified four primary contexts which led to disagreements across several musical cultures: 1) differences in cultural knowledge between the transcribers, 2) differences in interpreting syllabic boundaries, 3) intra-syllabic pitch changes, and 4) “voice splash” - abrupt pitch changes caused by vocal techniques or used as an expressive device.

The VocalNotes dataset, containing the audio of the musical fragments, annotations, and song documentation, has been published for replicability and further research.

Motivation

Providing a sufficiently broad yet useful definition of “music” has challenged music researchers for decades (Nettl 2015: 19-30; Jacoby et al. 2020). Many would agree that, for the most part, music is an intentional temporal ordering of sounds following some cultural or stylistic form. A majority of musical analysis across fields has taken musical “sounds” to be “notes,” each with a defined onset, duration, and pitch, but for many musical cultures and styles this model can be hard to apply. Are the ornaments of Baroque music distinct notes, or components of the main note that they decorate? How does one define a note among the continuously gliding pitches and timbres that comprise much of contemporary electronic music? One could argue that these are special musical cases. Yet, the most ancient and universal form of human music is song (Brown and Jordania 2013; Patrick E Savage et al. 2015; Mehr et al. 2019), and conspecific vocalizations (in our case, the human voice) are among the most privileged audio signals in our cognition. Thus the issues that arise when analyzing and defining song are due to fundamental limitations on our ability to analyze and define music. Across musical cultures and styles, the analysis of song presents several difficulties.

The human vocal apparatus is very agile. It has a large number of movable parts of various forms, sizes, and functions, which are in constant movement and mutual adjustment during vocal production; the voice produces unlimited gradations of pitch and more extensive timbral variation than most instruments, as vocal formants change with every syllable (Sundberg 1987; Titze and Martin 1998; Steinhauer, McDonald, and Estill 2017). This flexibility causes the voice to produce very complex signals with continuously fluctuating fundamental frequency curves (Fig.1). This makes vocal analysis more difficult to formalise (Proutskova 2019) than the analysis of most instrumental music. Additionally, many singing techniques use continuous changes in pitch (e.g. vibrato, embellishments, glides) or harmonic irregularities (glottal sounds and rasp) which contribute to song’s analytical ambiguity. When presented with vocal music, both humans and machines have a hard time coming to consensus about its constituent parts.

Fig. 1 Pitch curve of a Russian traditional song “Vy kumushki kumitesia” from Poozerie region, played on a keyboard and sung by the traditional performer Olga Sergeeva (Razumovskaya 1997)

Difficulties can arise at multiple stages in the cognitive processes of transcription, including perception and representation. There is some evidence that the perception of singing is more permissive to small differences in pitch than the perception of instrumental (non-human) sounds (aka vocal generosity effect (Pfordresher et al. 2010; Hutchins, Roquet, and Peretz 2012; Pfordresher and Brown 2017; Ozaki et al. 2024)). On the other hand, humans (even infants) can be remarkably sensitive to minute changes in vocal pitch, timing, and timbre, which are important cues in emotional communication (Scherer 2003). This can lead even expert listeners to hear objectively different vocalizations as equivalent, and vice versa.

Ethnomusicologists have long recognised that two different but equally competent transcribers are liable to notate the same music differently (List 1963; Herzog 1964; England et al. 1964; List 1974; Alekseev 1990; Stanyek 2014): "Two different transcriptions of the same piece do not necessarily indicate varying competence; they may reflect differences in the purpose of the task at hand, in the conception of what constitutes a piece of music" (Nettl 2015: 76). While perceptual differences between transcribers play a role, this has been difficult to separate from differences of interpretation and representation, all the more so when a notation system originating in one musical culture is adapted to others.

Western staff notation is the most common system for music transcription, including in non-Western or cross-cultural contexts. However, Western staff notation is not optimal for representing the inherent complexity of singing that this project investigates. In particular, because it does not readily allow sufficiently fine gradations of pitch and timing, it is not able to capture minute differences between the perceptions of different transcribers. More nuanced, “close-to-data” tools are needed, which can document how transcribers construct “notes” while listening to the audio stream of a vocal performance (Ozaki et al. 2021).
Multiple fields stand to benefit from better understanding our perception and transcription of song. Human transcriptions are used for training machine learning models of automated music transcription (that is, note-based segmentation, including but not limited to staff notation). It has been shown that automated methods perform poorly when transcribing global songs (Ozaki et al. 2021). Our project demonstrates that there is inherent subjectivity

in the perception of singing, expressed in transcription differences, and therefore there is ambiguity in the ground truth used for training automated music analysis models. Improving these automatic models requires improved understanding and formalisation of human disagreements. Moreover, high-quality datasets of singing are a rarity, which is a barrier in computational music analysis and has held back research on singing. Even rarer are cross-cultural song datasets with high quality annotations of real-life performances.

Alongside providing a high-quality dataset, our exploratory project opens up avenues for more targeted experiments in singing perception. In musical analysis, it provides the basis for further studies of musical modes, rhythm and entrainment. We demonstrate the advantages and the limitations of the existing note segmentation and pitch correction tools for a variety of contexts and repertoires. Crucially, we provide the data, musical contexts and the practical experience to frame the advantages and limitations of a basic musical concept: the note.

Previous work

VocalNotes in the context of ethnomusicology and notation

The VocalNotes project sits at a particular juncture in the history of ethnomusicology and its methodologies of transcription and analysis. From the beginning, ethnomusicology and its parent discipline comparative musicology placed great emphasis on transcribing unwritten (or differently written) musics into a visual notation that would allow systematic analysis and comparison (Abraham and Hornbostel 1909; translated in 1994). The notation used was almost always based on Western staff notation, which necessarily involved breaking the sound down into discrete notes. But questions have long been raised about the validity of using a “prescriptive” notation from one tradition as a “descriptive” notation for others (Seeger 1958; Hood 1971: 62-92), and even about the assumption that music necessarily consists of “notes” at all. Veteran ethnomusicologist Bruno Nettl wrote: “The concept of the articulated note works well for certain musics, especially instrumental…. In other kinds of music, perhaps singing most of all, notes are useful prescriptive devices, but they are not particularly descriptive. Lines may be preferable, providing opportunities to show glides and other ornaments” (Nettl 2015: 82).

Dissatisfaction with transcription into staff notation led to a search for alternatives that would (supposedly) bypass both the limitations of a note-based system and the cultural bias of a human transcriber by using technology to generate graphs of sound automatically (Metfessel 1928). The “scientific” aspirations of these efforts are evident in Charles Seeger’s wording: “As a descriptive science, musicology is going to have to develop descriptive music-writing that can be written and read with maximum objectivity. I believe that the graphing devices and techniques… show the way towards such an end” (Seeger 1958: 194). By minimising the role of human interpretation in the transcription process, this “objective,” “scientific” musicology would be differentiated from the “interpretative” methods of existing historical musicology and music criticism. Something of the same desire for scientific replicability is implied in Wim van der Meer’s introduction to his manual on graphing melodies with the current software Praat: “When a musicologist makes an analysis – for instance a

transcription (yes, a transcription is an analysis), no one really knows how the transcription is made…. When it is done by a computer program we *can* at least know how it is done, especially since the source code is public" (Meer 2023).

Meanwhile, the drawbacks of automatic sound graphs, as of transcriptions in staff notation, have long been recognized: they can be difficult to read, they often fail to distinguish different "layers" in a musical texture, and even if they capture an "objective" record of the sound waves, this may not tell us much about the sound as experienced by human beings (Jairazbhoy 1977). Thus, recent extensions of the "melographic" idea, including Meer's own work, have retained an element of interpretation in the manual annotation of computer-generated graphs (Music in Motion 2023) or, in the case of Andrew Killick's Global Notation (Killick 2020), allowing for manual as well as automatic production of line-based notation.

The transcriptions produced for the VocalNotes project, in fact, work similarly to most applications of Global Notation in presenting human-determined note pitches, onsets, and continuations within a visual space that (unlike that of conventional staff notation) is pitch-proportional and time-proportional. Indeed, VocalNotes is just the kind of study that Killick anticipated when he suggested that Global notation could "represent differences in the way the same passage is perceived by different listeners," discovered through "ethnographic and/or experimental research" (2020: 249). However, as there is not yet a purpose-made software program for writing and playing back Global Notation, we have found it more efficient to adopt and adapt the existing tools Tony and Sonic Visualiser.

VocalNotes uses sound-graphing technology not to eliminate human interpretation, but to study it. In this case, the focus is on how different listeners, all of them familiar with the musical styles they are hearing, interpret a melody differently in terms of how it breaks down into notes. Like staff notation, this does assume that melodies consist of "notes"; but the project uses sound analysis software to allow a more precise description of the boundaries, pitch, and inflections of the perceived notes than staff notation can provide.

Thus, in the context of ethnomusicology, VocalNotes represents a new intersection between manual and automatic transcription practices. Whether listeners necessarily experience music as an arrangement of "notes" is perhaps more a question for music cognition research.

VocalNotes in the context of music perception and cognition

Any analysis – especially close analysis from a subjective experience like music listening – will inevitably reflect the cognitive state of the analyst (Zbikowski 2002). The intersection between ethnomusicology, theoretical analysis, and cognition has long been noted: in 1988, John Baily wrote "What is the cognitive role of music theory? [...] Such questions take us beyond anthropology and into the domain of psychology" (Baily 1988: 114). The VocalNotes project presents a rare opportunity to explore how individual differences in the processes underlying music cognition may affect musical transcription and analysis.

The most basic and essential process of music cognition is auditory perception. In the VocalNotes project, pitch discrimination and rhythmic segmentation are the most crucial perceptual skills; timbre perception, while necessary, is not as relevant for the current

investigation. Pitch discrimination abilities have been widely studied, ranging from people with amusia (or tone deafness) to those with perfect pitch (Reis et al. 2021). A common test is finding an individual's "just-noticeable difference" (JND) pitch threshold, which is the smallest difference in pitch (or smallest interval) that they can detect. JNDs depend on numerous factors, including the task demands themselves, but for melodic intervals starting on a pure tone of 528 Hz (roughly C5), musicians had an average JND of 3 Hz (0.59% or ~10 cents), whereas non-musicians had an average JND of 12 Hz (2.23% or ~38 cents) (Arndt, Schlemmer, and Van Der Meer 2020). In the timing domain, JNDs as low as 10 ms have been reported when participants indicate the absolute duration of a tone (Levitin, Grahn, and London 2018).

If we have such precise perception of pitch and timing, how can disagreements ever arise between musical transcriptions? Individual differences ranging from a participant's age to their primary instrument can affect pitch perception. Musical training – both generally and in terms of genre and instrument – has been shown to affect music perception in profound ways (Besson et al. 2007; Tervaniemi et al. 2009; Kühnis et al. 2013). More broadly, several studies have investigated the effect of enculturation, especially of language learning, on the perception of both pitch and timing (Grannan-Rubenstein, Grannan-Rubenstein, and Thibodeau 2014; Hannon and Trainor 2007).

Moreover, low-level perception is not the only task involved in musical transcription. VocalNotes was particularly interested in the process of segmentation, which is the process that translates a continuous audio stream into distinct units. This task relies on a higher-level process known as categorical perception, where the listener chunks the incoming sensory information into sensory "objects" – these can be discrete notes with a pitch, onset, and offset, but need not be – and then looks for patterns, especially patterns that are already familiar to them (Deutsch 2012; Dowling and Harwood 1987). Exposure to a series of sensory objects leads to the emergence of rhythmic and tonal structure models, which form the listener's expectations of how the music will continue (Lerdahl and Jackendoff 1983; Huron 2006; Margulis 2014). The emergence of pitch and rhythm models is also underpinned by a number of factors, including the listener's enculturation and musical expertise; experts are more efficient at segmentation and categorization, and our cultural exposure drives what musical patterns we are familiar with and thus what categories we are likely to form (Hannon 2009). More flexible cognitive states, such as the listener's attention and familiarity with the song, can further filter which aspects of the audio stream are focused on and which are ignored (Lerdahl and Jackendoff 1983; Huron 2006; Elhilali et al. 2009). So, even if two people could perceive the pitch and timing of a song the same way, they may not have the same cognitive interpretation, or internal experience, of the song.

Transcribers must then also face the task of representation, which externalises this internal experience. There are numerous cognitive tasks underpinning the process of visually representing an auditory melody in Tony and Sonic Visualiser (the digital tools used in the VocalNotes project). The field of embodied cognition has made it clear that we think through our tools, and *they* shape *our* cognition as well (Kirsh 2013). For example, although categorical perception occurs in both the auditory and the visual domain, the mechanisms differ (Goldstone and Hendrickson 2010). Wearing glasses inherently changes how one perceives visual stimulation; likewise, listening to audio while seeing its representation in

Tony and Sonic Visualiser, or directly manipulating that representation, inherently changes how it is perceived.

In order to draw meaningful conclusions, the VocalNotes methodology attempted to control for many of the above variables, such as enculturation, musical training and expertise, the effect of the tools on the transcription, and the differences between the audio and visual domains (see Methodology paper in this issue). Digital tools allowed not only for a fine-grained documentation of these outcomes, but also for generating data, making these outcomes available for further analysis with computational approaches.

The VocalNotes project is an exploratory study which investigated disagreements in transcriptions as reflections of differences in the cognition (or at least representation) of notes between transcribers. In the course of this exploration, the authors were simultaneously the subjects and the explorers, documenting the outcomes of their own cognitive processes through the medium of transcription, and analysing and comparing these transcriptions to reflect back on their cognitive processes.

VocalNotes in the context of computational music analysis and MIR

Computational music analysis (computational musicology) is an interdisciplinary research area at the intersection of musicology and computer science (Meredith 2015; Mor, Garhwal, and Kumar 2020). Music Information Retrieval or Music Information Research (MIR) investigates music as data with the goal of improving our understanding of music and providing automated tools for music-related tasks, including music transcription, musical instrument recognition and separation, and music classification (Downie 2003; Casey et al. 2008; Serra et al. 2013; Müller 2015). The VocalNotes project uses MIR tools Tony (Mauch et al. 2015) and Sonic Visualiser (Cannam, Landone, and Sandler 2010) for pitch extraction, note segmentation and note pitch correction. A new high-quality cross-cultural dataset was created as part of the VocalNotes project (Proutskova et al. 2023), contributing to music-as-data investigations, whereas the studies conducted by the teams fall into the area of computational music analysis.

Pitch extraction

Pitch is defined as a subjective quality of perceived sounds that closely corresponds to the fundamental frequency (f0) of a pure or complex tone (Hartmann 1997). There are deviations from the exact f0 and the pitch percept, but pitch and f0 are often used interchangeably outside psychoacoustical studies (Kim et al. 2018). Pitch extraction refers to the process of determining the fundamental frequency of a harmonic sound; in our case the sounds are a cappella singing sampled every 10 msec (Babacan et al. 2013). Pitch, or more precisely, the fundamental frequency, can be visually represented as a curve in the time/frequency space (Fig. 1), which we call a pitch curve or f0 curve. While there are numerous pitch extraction algorithms, the VocalNotes project is limited to using the pYIN algorithm as it is directly implemented in Tony and Sonic Visualiser. pYIN is a variant of YIN (De Cheveigné and Kawahara 2002), an autocorrelation-based f0 extraction method of the previous generation, with the addition of a Hidden Markov Model to decode the most probable sequence of pitch values. pYIN has been shown to have comparable performance

to other state-of-the-art algorithms on estimating f0 from unaccompanied monophonic singing (Devaney 2020; Rosenzweig, Scherbaum, and Muller 2021).

Onset detection and note segmentation

To identify a note, an algorithm needs to determine the beginning (onset) and end (offset) of the note as well as the note pitch. Automatic onset detection turns out to be a particularly difficult task for singing voice: according to the MIREX 2018 audio onset detection competition, the best F1-score (a statistical measure of accuracy) of singing onset detection is only 61.94%, which is at least 10% lower than the onset F1-scores of other musical instruments (X. Wang et al. 2022). Previous singing transcription systems were based on hidden Markov models, relying on musical features (such as pitch, amplitude or metre) within the voiced regions (Mauch et al. 2015; Ryynänen and Klapuri 2004; Viitaniemi, Klapuri, and Eronen 2003). Note pitch is usually determined automatically as the median of the pitch values within the segment. However, these methods performed poorly on notes with soft onsets and offsets, pitch oscillations within notes and glides between temporally adjacent pitches (Li et al. 2021; X. Wang et al. 2022). Despite recent improvements using deep learning approaches (Fu and Su 2019; X. Wang et al. 2022), algorithms still perform worse than human experts (Ozaki et al. 2021).

The main focus in the development of new transcription algorithms is comparing performance with previous models on a few baseline datasets (Benetos et al. 2019). These datasets are skewed towards Western music, involving fixed-pitch instruments, and are lacking culturally-diverse singing data. This might be the reason why such systems have not been adopted by musicologists. Holzapfel et al. (2022) asked 18 musicologists from 5 European universities to transcribe 8 excerpts of *sousta*, a traditional Greek instrumental dance genre, either from scratch or starting from an automatic transcription, finding no quantitative advantage of using automatic music transcription (Holzapfel et al. 2022). Although computer-assisted transcription studies exist (Gómez and Bonada 2013), recent reviews by musicologists argued that computational tools for musical analysis are useful for only low-level analysis and not widely adopted within mainstream musicology (Cottrell 2018; Tilley 2018). Similarly, the VocalNotes methodology, while relying on automatically extracted fundamental frequency curves as the first pass, does not make use of automatic note segmentation suggestions: segmentation is performed manually from scratch.

Ground truth and datasets

A major challenge in MIR is the lack of annotated data, especially for singing. Recently, large datasets based on crowd-sourced non-expert annotations, improved by deep neural networks, have been introduced (DALI (Meseguer-Brocal, Cohen-Hadria, and Peeters 2018), MIR-ST500 (J.-Y. Wang and Jang 2021)), butthe resulting quality of the annotations is low or difficult to assess. While these datasets comprise mainly Western popular music, new high-quality corpora of non-Western vocal traditions have emerged more recently, including for Georgian (Rosenzweig et al. 2020), Korean (Choi et al. 2020), and Chinese songs (Gong, Repetto, and Serra 2017). Cross-cultural datasets have also been published outside the MIR field, e.g. by the Many Voices project (Ozaki et al. 2024). We expand on this by

creating and publicly releasing a dataset of vocal performances from five different traditions with expert pitch and note annotations (Proutskova et al. 2023).

In MIR research, algorithms are typically developed with the aim of reproducing a correct, “ground truth” annotation. One of the pitfalls is the assumption that there is one “correct” way to transcribe music, which, as described above, has been refuted by ethnomusicologists; this assumption has also been questioned by MIR researchers (Sturm and Flexer 2023). Singing is particularly difficult to transcribe due to inherently unstable pitch curves (Fig. 1) and vocal drift (Mauch, Frieler, and Dixon 2014). Following Bittner et al. (2021) and Ozaki (2021), this project demonstrates that human annotations of singing show considerable disagreement, thus questioning the singularity of vocal annotations as ground truth. We provide a larger high-quality dataset with the advantage of multiple independent expert annotations for each excerpt, exploring the concept of a variable ground truth, which we hope will aid the development of more flexible automated transcription algorithms that can deal with ambiguities.

The VocalNotes project

The VocalNotes project investigated how expert traditional music listeners conceive of notes in vocal performances by studying similarities and differences in their transcriptions. Teams of experts from five musical traditions (Japanese folk song, Chinese Hebei *bangzi* opera, Russian traditional village singing, Alpine yodelling, and Romaniote Jewish chanting) each annotated approximately 10 minutes of ethnomusicological recordings in their culture using Tony (Mauch et al. 2015) and Sonic Visualiser (Cannam, Landone, and Sandler 2010) for segmentation and note pitch correction, starting from an automatically extracted fundamental frequency curve. The experts then compared their independent transcriptions and looked for factors which led to disagreement.

Fig 2 Map showing the VocalNotes teams and where the members were located.

A pilot project run by two co-authors, Polina Proutskova and Olga Velichkina, preceded the VocalNotes project. They attempted a corpus study on modes in Russian traditional singing. After standard MIR techniques for automatic mode extraction failed, they turned to manual transcription, using Tony and Sonic Visualiser to create ground truth for supervised machine learning models. They discovered numerous differences in their annotations, even in songs they knew well and sang together. It became obvious that these differences were not errors, but reflections of divergences in their perception of the song.

Was this difference of interpretation specific to Russian village singing? Would transcribers in other cultures experience similar disagreements? Are some vocal styles more difficult to transcribe (leading to more disagreements) than others? To answer these questions, they invited teams of ethnomusicologists to participate in a comparative study applying similar methodology to different traditions. The project started in December 2021 with eight teams, five teams reaching the finish (Fig. 2).

This project posed several organisational challenges: The participants are situated across 14 time zones, from Japan to North America. Some of our participants do not speak English. This project did not receive specific funding and participants contributed either in their free time or as part of their ongoing research. Teams were affected, and some had to drop out, due to the wars in Ukraine and in Israel/Gaza. Although these difficulties certainly delayed the overall completion, the success of the project is manifested in the publications of this special issue and the release of the VocalNotes dataset (Proutskova et al. 2023).

VocalNotes is an example of a project that differs from “helicopter research” where “researchers from high-income settings, or who are otherwise privileged, conduct studies in lower-income settings or with groups who are historically marginalised, with little or no involvement from those communities or local researchers in the conceptualization, design, conduct or publication of the research” (Nature Editors 2022; cf. Jacoby et al. 2020; Sauvé et al. 2023; P. E. Savage 2022). In VocalNotes, all teams were equal in their involvement in the methodology, the transcription process, the synthesis, the dataset and the publications (Polina Proutskova et al. 2024 (in preparation), see also the Author Contributions statement at the end of this paper). Each team’s work contributed to their own research questions of interest (see the Outcomes section of this paper for a summary), while the project provided technological and organisational support and a platform for the synthesis of the outcomes.

Methodology

Transcription was conducted in three phases: (i) Automated pitch curve extraction (Fig 3a) using the pYIN algorithm (Mauch and Dixon 2014) in Tony (Mauch et al. 2015) followed by manual correction (Fig 3b). (ii) Manual segmentation by annotating note onsets and offsets (Fig 3c). (iii) Note pitch correction of automated suggestions (Fig 3d). This was followed by a qualitative analysis where different transcriptions of the same performances were compared (Fig 4), uncovering the musical, linguistic and cultural contexts which led to disagreements. A more detailed description of our methodology and the lessons we learned through our

exploration are documented in the Methodology paper in this Special Issue (Polina Proutskova et al. 2024 (in preparation)).

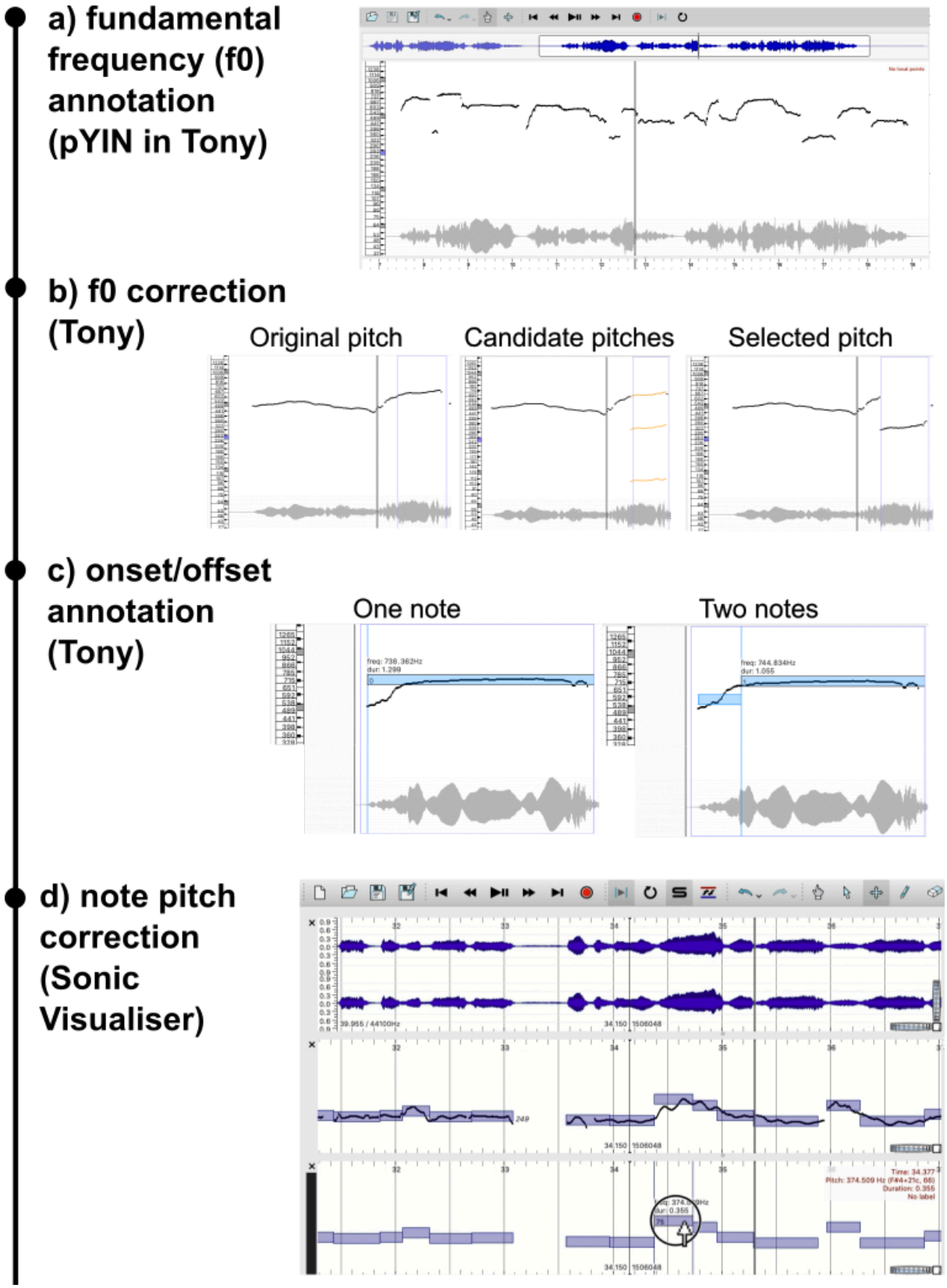

Fig. 3 **a)** Tony automatically extracts the fundamental frequency curve using the pYIN algorithm. pYIN works quite well for solo singing, though errors are still possible. **b)** Pitch

curve correction: the selected area of a yodel song was wrongly interpreted by pYIN to be in the same octave as preceding pitches. This octave error can be easily corrected in Tony by choosing one out of many candidate pitches suggested by pYIN. **c)** Segmentation: a “note” (a blue rectangle spans from the onset to the offset) can be easily added and further edited in Tony. Note pitch is automatically determined as the median of the pitch within the segment; there is no editing mechanism for note pitch in Tony. **d)** Note pitch correction: in Sonic Visualiser, note pitch can be edited by moving the blue note bar up or down (highlighted by the circle).

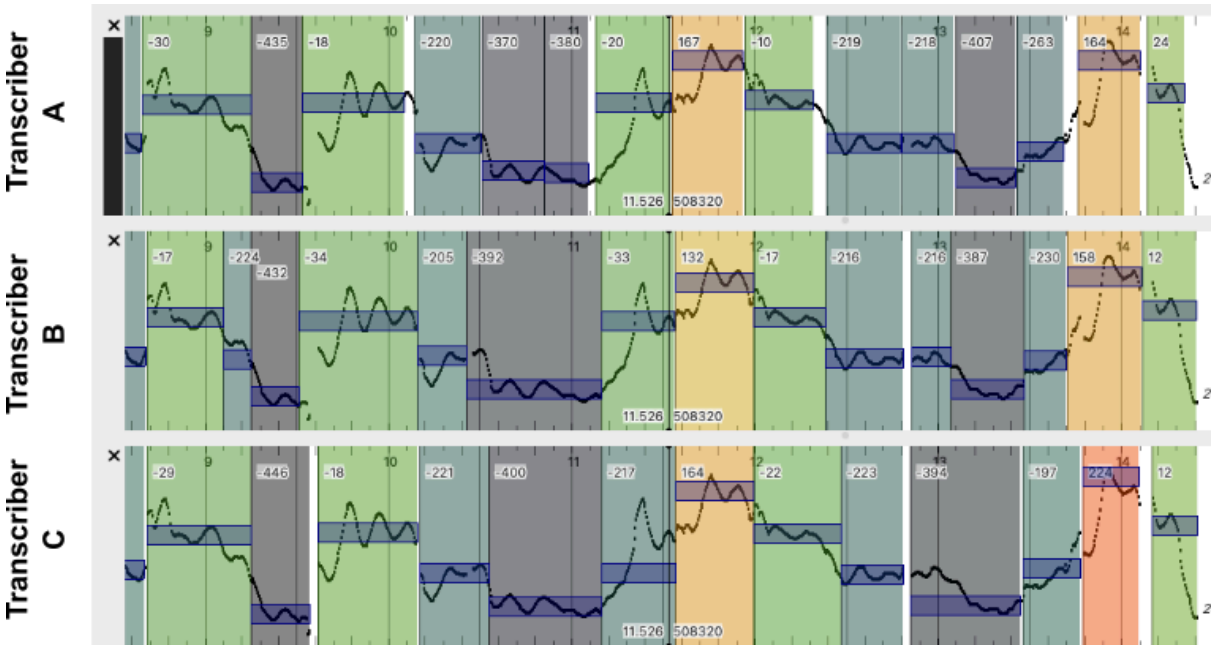

Fig. 4 Comparing transcriptions. The area above and below notes is coloured by note pitch for easy comparison; note pitch is also shown in cents (relative to an automatically extracted tonal centre). This representation allows us to easily estimate intervals between notes.

It must be noted that what is referred to here as transcription is not the same as musical transcription in traditional music analysis but is in fact a semi-automated annotation process which relies on automatically extracted pitch curve and automatically rendered note pitch suggestions. The annotators (who we also call transcribers for simplicity here) perform the segmentation and note pitch correction on the basis of these.

Outcomes

We find, perhaps unsurprisingly, that transcriptions by experts were mostly in agreement. Therefore, we can claim that in the majority of contexts the note concept was a reliable tool for analysing our repertoires. However disagreements were quite common. They arose in every transcription, and in some cases disagreements were as numerous as agreements (this of course depends on how strictly one evaluates “agreement”). To give one example, in Fig. 5a transcribers disagreed about the location of the boundary between the first two notes, and disagreed on whether the ensuing sound was one or two notes. Particularly where transcribers disagreed about the number of notes they heard, the usefulness of the note concept was at its limits.

We found that some repertoires and individual performances were “notier” than others: they were easier to transcribe and led to fewer disagreements between transcribers. Specifically, the notes in Alpine yodel (Fig. 5b) were much easier to agree about than in other analysed traditions (Fig 5c). Within the traditions, singers varied considerably in the “notiness” of their individual singing style; this could be clearly seen in Russian and Japanese collections which included different performances of the same songs (Fig 5d).

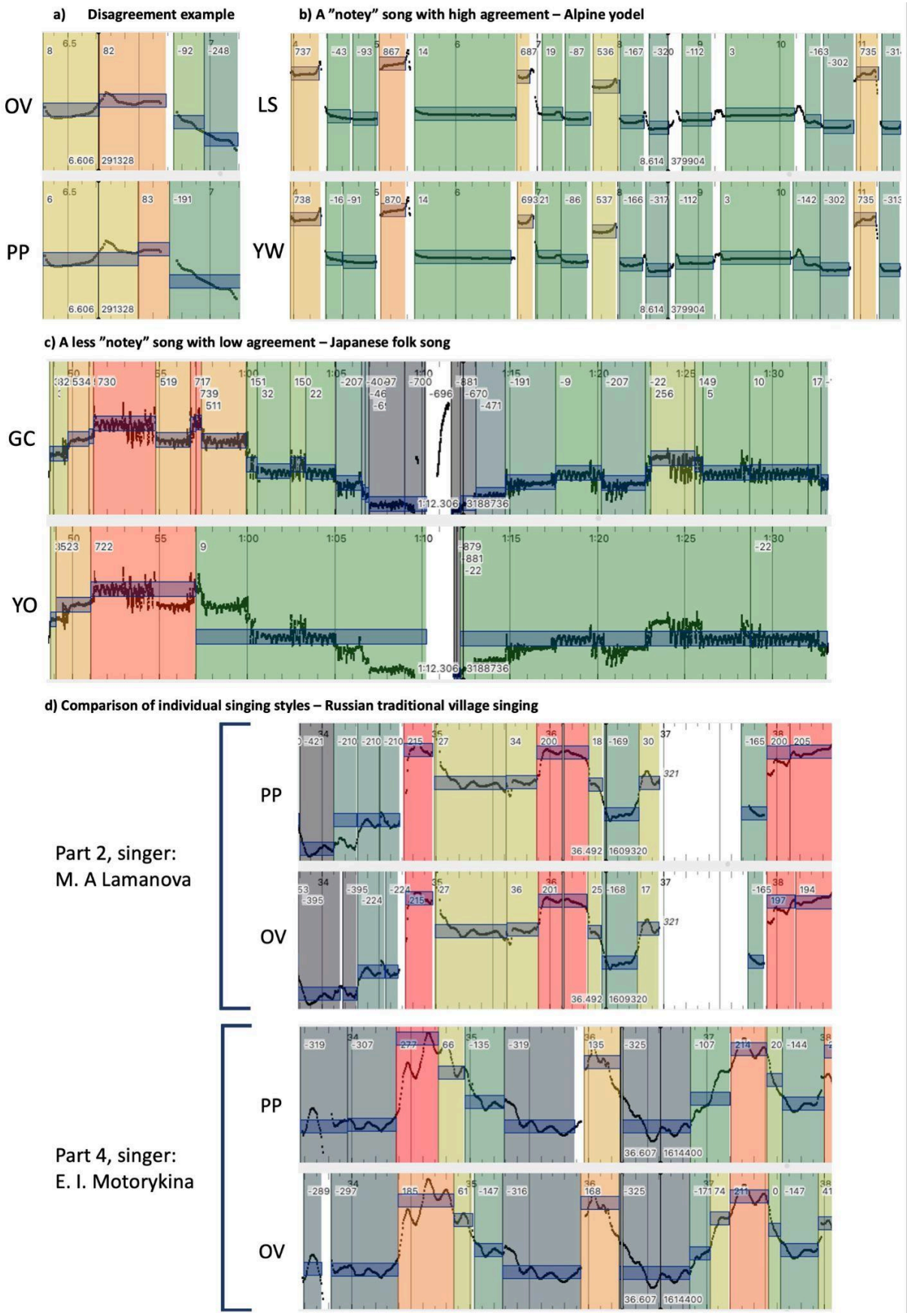

Fig. 5 Examples of agreement and disagreement in transcription. **a)** Transcribers disagree about the boundary separating the first two notes, and on whether the final descending pitch

is one or two notes. **b)** In this example of Alpine yodel, there is comparatively strong agreement on where notes begin and end. Overall, the Alpine repertoire is "notey", with a high level of agreement between transcribers, compared to other analysed repertoires. **c)** Japanese folk songs are less "notey" than Alpine yodel, displaying more disagreements between transcriptions. d) Within a tradition, individual singing styles can differ in how "notey" they are. Here, transcriptions of two Russian singers singing the song "Po zoriushke" are compared: Lamanova's singing style is "notier" than Motorykina's, the latter presenting more disagreements between transcribers than the former.

Summary of Findings from Individual Groups

Each team will present an individual paper describing the process and the outcomes of the project for their specific culture, in the context of the team's specific research question related to their tradition.

The Japanese team (Chiba, Ozaki, and Savage 2024) analysed 9 recordings (3 different singers x 3 different folk songs), concluding that the transcriber with extensive performance experience in the tradition transcribed much more detail than the transcriber with only listening experience.

The Russian team put together a corpus of ethnographic solo and group singing recordings (with one channel per singer) from a variety of local traditions and genres. Their findings demonstrate the subjectivity of mode perception in Russian traditional singing, challenging the established practice of how the mode of a song is determined.

The Jewish team transcribed 10 excerpts from the few existing recordings of Romaniote Torah cantillation. A strategic selection of the recitation of the same biblical texts chanted by different Romaniote practitioners (Genesis 1:1-11; Deuteronomy 6:5-9) allowed the team to analyse the transcription within a comparative framework; they identified similarities in cantorial practices in different diasporic locations and conditions while also identifying and analysing what they have understood as *idiolects*: each practitioner's personal aesthetic particularities within the flexible affordances of their traditional practice.

The Alpine team transcribed nine excerpts from the central European yodelling tradition. They analysed and interpreted the relationship between vocal expression, or "utterance" on the one hand, and its notated representation on the other. Drawing on Milton Babbitt's model of the threefold representation of sound (acoustic, auditory, and graphemic), the differences in two independent, semi-automatized transcriptions are discussed. The onset and end of notes lead to little disagreement and enable establishing a broader hypothesis of the use of consonants as start and transition points in singing without lexical words. The symbolic representation for music not built along hierarchical models of meter like yodel remains controversial and a semi-automatized visualization can offer a more balanced solution.

The Chinese research team transcribed 11 excerpts from Hebei *bangzi* (clapper opera), a genre of Chinese traditional opera, covering a range of tempos including slow, moderate, fast, and rubato patterns. Both transcribers agreed that their transcriptions shared notable similarities overall, though they found that a greater familiarity with the vocal style and a

higher level of reliance on auditory perception led one transcriber to create more detailed transcriptions, capturing subtle variations in pitch, timbre, and loudness.

Summary of Disagreements

Here we outline a few contexts where disagreements arose. Although this list is not completely exhaustive, we strive to present a cohesive overview of disagreements which were encountered by more than one team while analysing their repertoires. We group disagreements in the following section as due to “cultural knowledge,” “syllabic interpretation,” “intra-syllabic pitch change,” or a broad category of utterance that we term “voice splash.” Note that there can be overlap between groups, as some examples may fall into more than one category.

Cultural Knowledge

Disagreements due to cultural knowledge arise when the transcribers’ awareness of local dialects, musical practices, social context etc. differs, thus leading to differences in their interpretation of the singing. Such disagreements were relatively rare, but offer an important view into the reasons why transcriptions might diverge even when transcribers’ perception is in full agreement. We here present two cases (Fig. 6). In an example from Japan (Fig. 6a), the singer is given call cues by another person.The transcriber who can perform this song excluded the calls from the transcription, because he would not be making them as a singer, whereas the other transcriber with only listening experience included the calls. In a case from Jewish Romaniote chant (Fig. 6b), transcribers disagreed over an utterance because in the local dialect of the singer, of which one of the transcribers was aware, the number of syllables differs from the dialect spoken by the other transcriber.

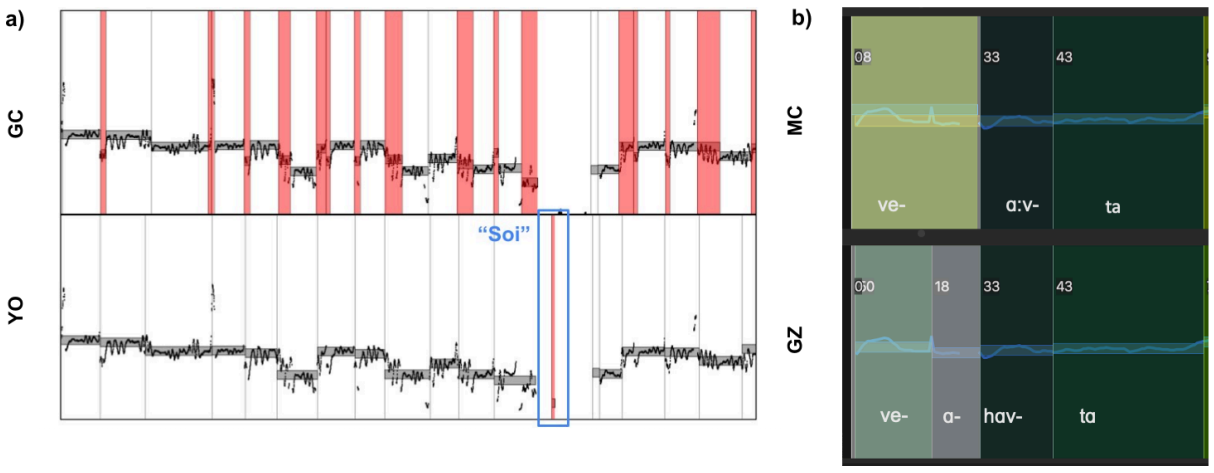

Fig. 6 Difference in cultural knowledge: **a)** Red is used to highlight areas where one transcriber has an extra note compared to the other. In this Japanese folk song, Esashi-Oiwake, GC annotated many more notes, except for one point. This call, “Soi”, is uttered by a different person than the singer. While both transcribers perceived the call, only YO transcribed it, not being aware of the original ensemble context; GC, as a singer, would

not make the call, so did not consider it to be a note (video). **b)** Romaniote Jewish dialect in Greece is different from that in NYC. In this example of cantor Levis, Romaniote Hebrew's [h] > Ø leads to vowel elision, such as in /vəɑhɑvtɑ/ pronounced as [vəɑ:vtɑ]. MC segments phonetically with elision as [vəɑ:vtɑ] and GZ segments phonemically (or without elision) as [vəɑ.ɑvtɑ]. (video)

Syllabic Interpretation

Overall we found that a syllabic change almost always triggers a note boundary (see Fig. 7c for a single exception). Yet in some contexts, what constitutes a syllabic change can be interpreted differently by different transcribers. Some examples include insertion of vowels (anaptyxis) in the Russian (Fig. 7a) and Jewish (Fig. 7b) performances, or differences in syllabic division depending on the dialect (Fig. 6b).
The only counterexample of a multi-syllable note we found was the ethnachta clause in the Jewish Romaniote tradition. The ethnachta clause is the opening of a Torah passage recitation, which is spoken rapidly before a more protracted, sung passage. It is perceived by the singer and the listeners as a marker which only makes sense as a whole, rather than a word that can be subdivided into syllables (Fig. 7c).

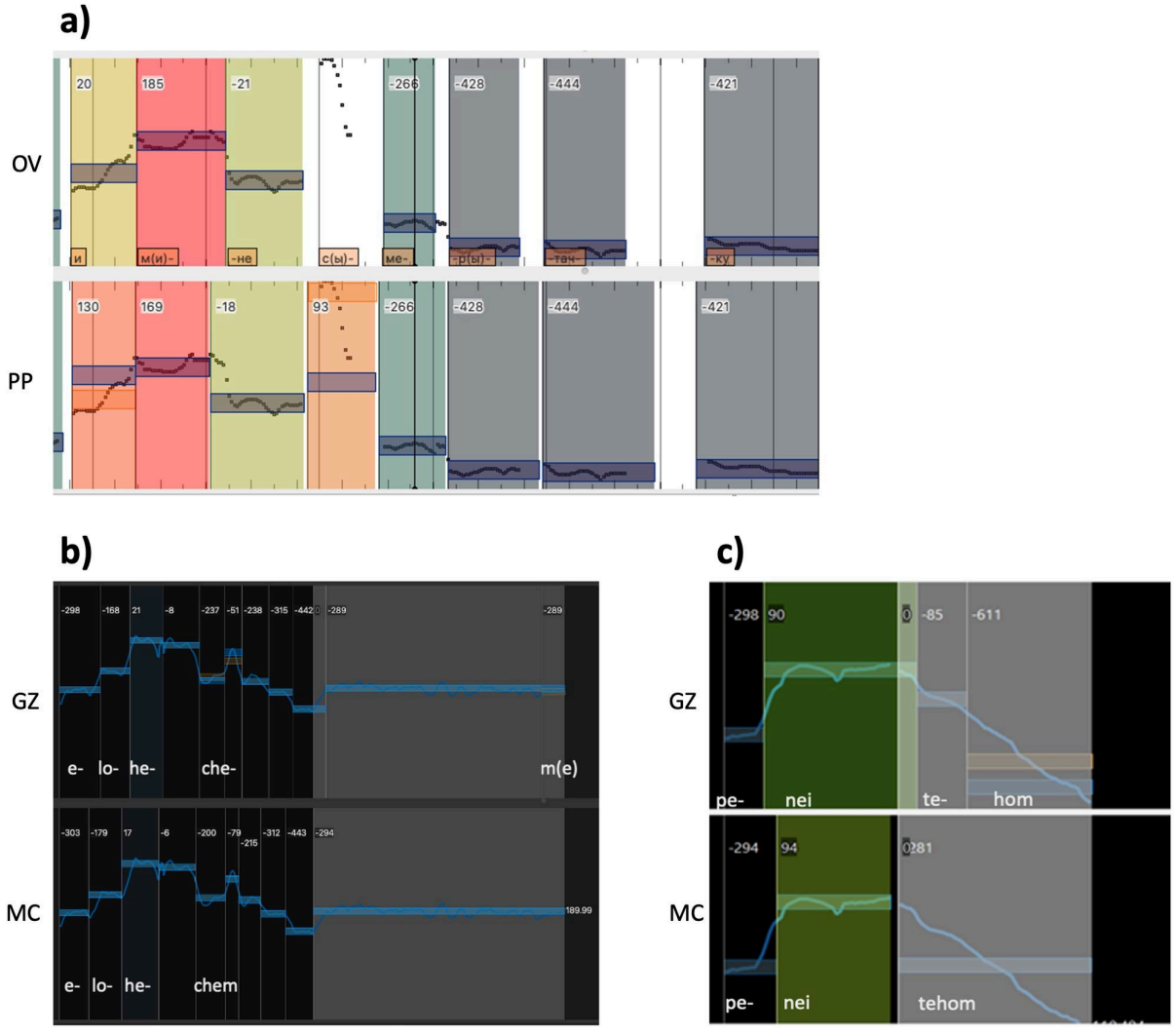

Fig. 7 Differences in syllabic interpretation: **a)** In the Russian language, syllables may include more than one consonant. In traditional village singing, insertion of vowels (anaptyxis) is common, when a tiny vowel is added to a consonant that was part of a syllable in speech, making this consonant into a syllable of its own. Here, inserted vowels are given in brackets. The sound /s(y)/ (/c(ы)/ in Cyrillic transliteration) is interpreted as a voiceless consonant (and therefore not a note) by OV, whereas PP assigns pitch to it, making it into a note. You can hear in the video how the singer creates many short syllables in her singing by inserting a vowel after each consonant. **b)** In Romaniote chanting some cantors' expressively insert vowels at phrase-final words that end in a consonant (Ø > [ə] / C_#). See the rightmost segment in CZ's transcription (red circle), which is absent in MC. (video) **c)** Multi-syllabic units: The *ethnachta* clause in the Jewish tradition was sometimes considered as one unit by the transcribers owing to its declamatory character as well as its role as a divider between two sung parts of a verse. Therefore, MC subsumed several syllables into one note segment. The example is cantor Borbolis (Gen.) chanting the words "pe-nei te-hom" (video).

Intra-syllabic Pitch Change

The most frequent disagreement across all groups involves intra-syllabic pitch change, when pitch changes within a syllable (Fig. 8,9). Examples of this include glissando and melisma (Fig. 8a-c), vibrato (Fig. 8d), intra-syllabic embellishments and runs (Fig. 8e-f), and glides (scoops) at the start and the end of the syllable (Fig. 9).

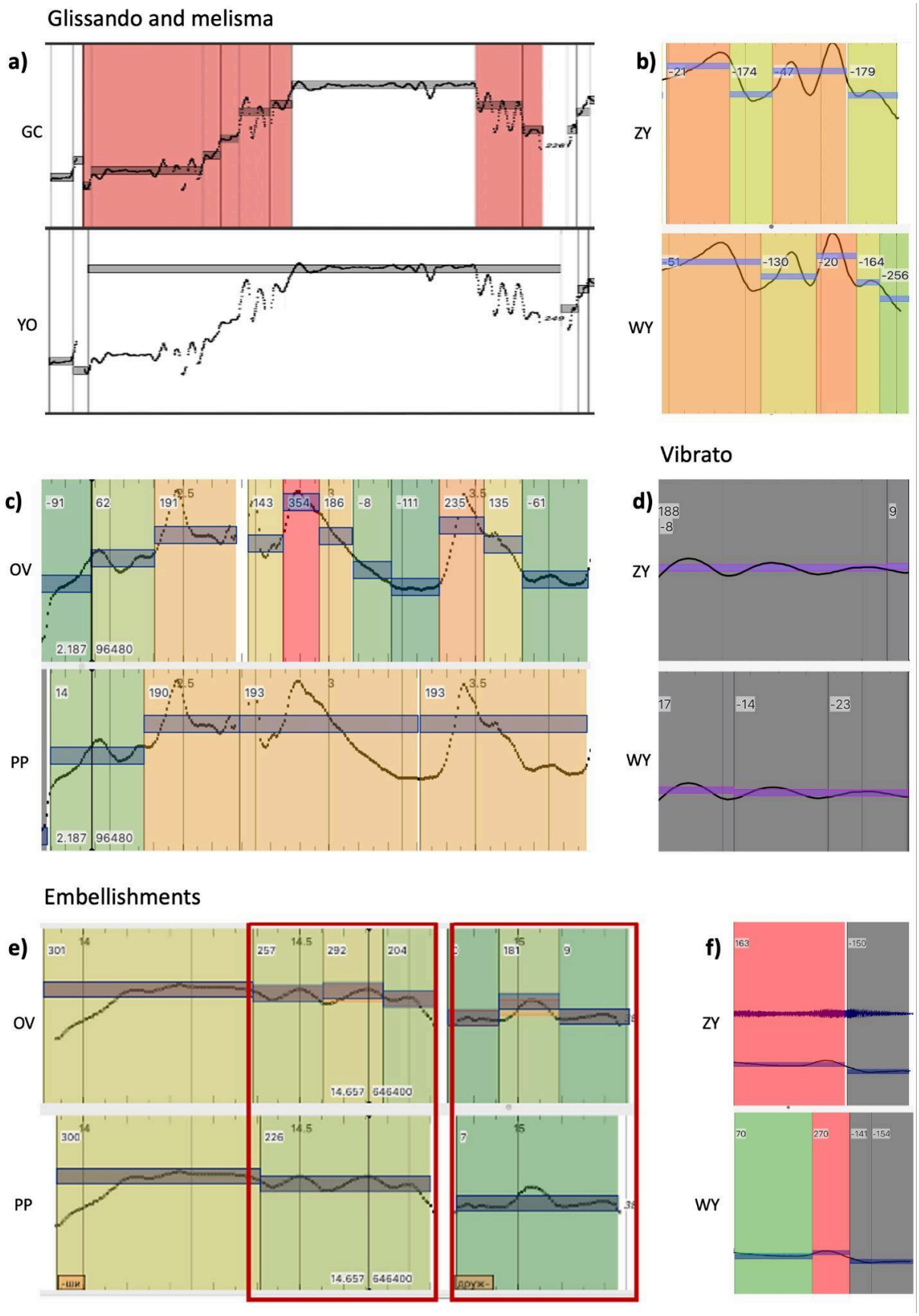

Fig. 8 Intra-syllabic pitch change: glissando and melisma (a-c), vibrato (d), embellishments (e-f). **a)** A Japanese example of a melisma (video), GC segmenting in much more detail than YO. **b)** this Chinese example shows a sung syllable (‘啊(a)’) with melisma; the two

transcriptions differ in terms of the note boundaries and number of notes (video). **c)** A Russian example of a glissando (video); PP, who segmented less, commented that her transcription was based on vocal and expressive gestures. Intra-syllabic pitch change: vibrato. **d)** A Chinese example of vibrato; ZY annotated it as one note (with vibrato), while WY segmented in three notes (video).Intra-syllabic pitch change: embellishments. **e)** In this Russian traditional song, PP considered the pitch change to be embellishments, while OV did not (video). **f)** In this Chinese example WY annotated an ornament highlighted in red with the pitch value 270 cents (video).

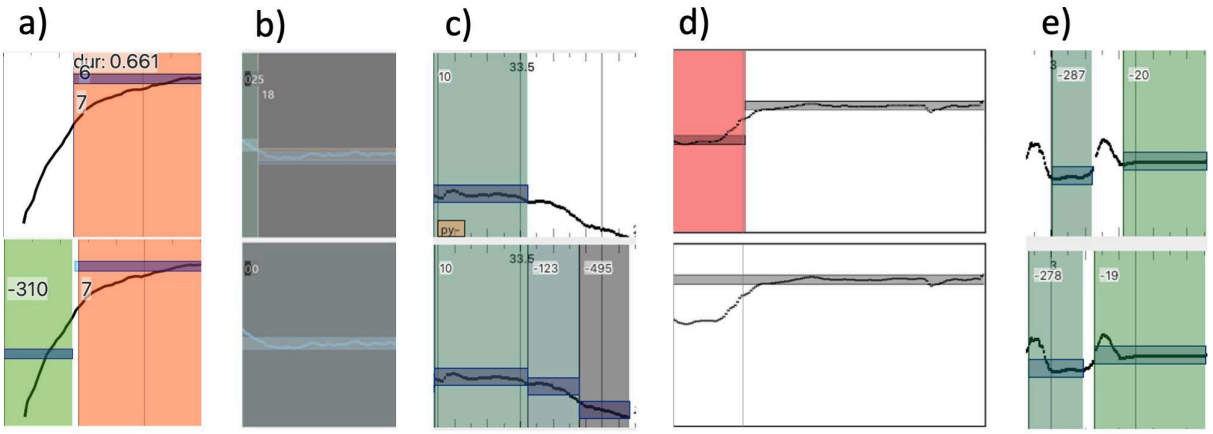

Fig. 9 Intra-syllabic pitch change: glides. Glides at the start and end of the note (sometimes called ‘scoops’) were found in each culture: **a)** Chinese, **b)** Jewish, **c)** Russian, **d)** Japanese, and **e)** Alpine.

“Voice splash”

We use the term “voice splash” to denote fast, sharp and very short changes in pitch, which were clearly audible and also visible in the f0 curve. This could be caused by a number of vocal techniques which may or may not have been intentional, and may or may not change vocal register. They can be linguistic and non-linguistic sounds, may constitute a syllable or be part of a syllable. While a canonical example is the Indian gamaka, we encountered this phenomenon in all repertoires analysed in this study (Fig. 10).

There are borderline cases where it is difficult to decide whether an intra-syllabic embellishment should be classified as voice splash, and Jewish Romaiote repertoire presented many such examples (Fig. 10e). In such cases, we relied on our musical judgement; more specifically, if the sounds seemed very expressive and were shorter than other embellishments, we classified them as voice splash.

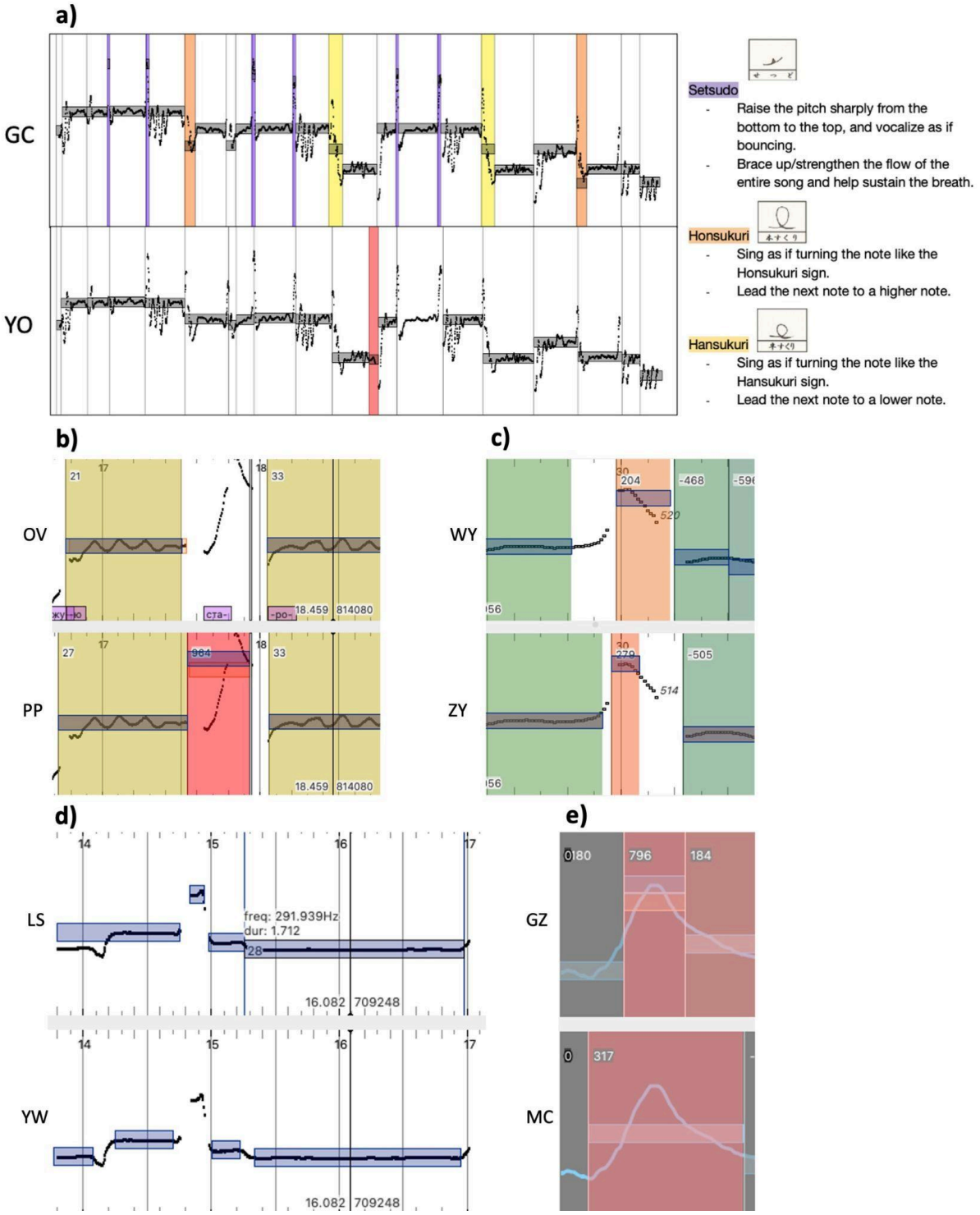

Fig. 10 “Voice splash”: **a)** In Japanese folk song, there are names for various vocal techniques which produce sharp, short changes in pitch (video). **b)** An expressive burst in a Russian lament, which incorporates sounds of crying into singing (video). **c**) In Chinese opera, a similar expressive burst is employed. **d)** A voice splash in Alpine yodel. **e**) Orange note bars represent the automatic note pitch suggestions and blue bars show the note pitch manually corrected by the transcriber. Jewish Romaniote repertoire presented many borderline voice splash cases like this one from a performance by the cantor Kofinas: while MC represented it as a part of the encompassing note, GZ segmented it and raised the pitch

considerably compared to the automatic suggestion. Due to its expressivity we decided to classify this embellishment as a voice splash.

Discussion

From these conversations, it became apparent that the differences we were finding between transcribers were not primarily perceptual, but representational. That is, discussion between the experts suggested that they had heard more or less the same pitch content, but nonetheless, they represented it differently when asked to segment that pitch content into discrete notes. Of course, the initial pitch transcriptions were machine-assisted, starting from an automatically-extracted fundamental frequency, rather than done purely by ear. As such, any potential differences in early pitch perception were controlled for – or perhaps merely hidden – by this automated step (see Methodology paper). On the other hand, a previous study that did quantify agreement among purely manual transcriptions using staff notation also found more disagreement in note segmentation than in pitch assignment (Ozaki et al. 2021), suggesting that this result is not simply an artefact of our methodology. As mentioned, however, it has been demonstrated that expert musicians tend to have finely-honed pitch perception, so purely perceptual differences may have been negligible regardless (Arndt, Schlemmer, and Van Der Meer 2020).

It is of greater interest that, despite hearing the same pitch content, equally enculturated and expert musicians still segmented it differently. For example, where one listener would represent a short pitch deviation as a separate, ornamental note, another would represent it as vibrato embellishing a long held note (see Fig 8). These differences in audio representation and categorization comprise some of the most interesting outcomes of this project, and require much further study. It is impossible in the present study, for example, to disambiguate whether these differences reflect the transcribers' internal (cognitive) representational models of the melody, their external (visual) representational models of the melody, or a complex interaction of the two. It is still striking that one can ask two experts of a musical tradition "what are the notes in this piece?" and receive such different answers. Perhaps it provides some support for Nettl's claim that "notes" – as an undifferentiated, universal concept – may not be a wholly appropriate framework for representing vocal music, and we need to have a more nuanced discussion of how non-discrete vocal gestures fit into our models (both internal and external) of music.

Although VocalNotes focused on differences between transcribers, differences in cognitive processes even within a single transcriber were also noted during discussion. For example, when transcribing the melody, the analyst could select what portion of the audio to listen to, and every analyst agreed that the length of the selection – essentially, the amount of auditory context – made a difference in how they conceptualised what they were hearing. The effect of the context was particularly striking during the note pitch correction phase, although it doubtless affected segmentation as well. A certain pitch may sound more or less correct when heard individually as opposed to in a phrase, or in a stream of multiple phrases, because the varying context changes our categorical perception, our usage of auditory memory, and our understanding of larger patterns (Tervaniemi et al. 2009). The effect of context on musical representation within a single brain, while not the focus of this project, also deserves further study.

There are many other issues of cognitive import that could be explored through this project. For example, emotional attachment to music profoundly affects the musical experience, and many of the analysts in VocalNotes have deep emotional connections to their traditions (Lamont 2012). Most analysts were also singers within the tradition themselves, and some reported using singing to determine transcription decisions such as the correct pitch for the note. Therefore, we must understand how music perception interacts with music production. Individual differences in music production abilities may be playing a large, implicit role in this task, and they vary more widely than music perception abilities, and are even more shaped by expertise (Sloboda 2000). The evidence from the Japanese team seems to point towards practitioners having more detailed segmentations compared to musically-informed and enculturated non-singers. This might relate to findings that music performers have more enhanced sensory-motor coupling than non-musicians, especially when listening to music of their own instrument (Proverbio and Bellini 2018,15-25).

In short, the VocalNotes project provides an interesting case study for examining a number of key processes in music cognition. The difficulty is in disentangling which process(es) are responsible for any given difference in the produced transcription. Only expert musicians were selected for the project, only cultural experts were included within each team, and everyone was asked to use the same tools to control for some of the variables, but in practice significant differences still arose, and therefore we suggest stricter controls for future studies (Methodology paper).

To facilitate future replication and research, we have published the VocalNotes dataset[1] containing the audio fragments and the experts' annotations in a simple, easily accessible machine readable format (f0, note onsets/durations, and note pitch as csv files, alongside extensive contextual documentation which may include the title, lyrics, performers, sub-cultural or geographical origin, genre or function, and/or recording metadata) (Proutskova et al. 2023). This cross-cultural corpus with high quality annotations can serve as a basis for further quantitative analysis of disagreements, and as ground truth for machine learning algorithms.

Aside from the final transcriptions, the transcribers discussed their process of transcription at length, providing data for qualitative analysis of different cognitive processes. This exploratory study is thus a rich example of the kind of cross-cultural, cross-disciplinary, mixed-methods approach to understanding music cognition that could produce fascinating insights in the future.

Acknowledgements

We would like to acknowledge the contribution of the teams who had to drop out of the project due to the war in Ukraine and for other reasons: Anastasiia Mazurenko of the Ukrainian team, Ieva Tihovska, Ilze Cepurniece, Zane Šmite of the Latvian team, Nana Mzhavanadze, Teona Rukhadze of the Georgian team.

[1] The annotations can be downloaded from Zenodo under the CC BY-NC-SA licence: https://zenodo.org/records/10065955; the complete VocalNotes Dataset including the audio fragments that were analysed is provided for research only via a request form: https://osf.io/4n5ry/

Author contributions

- Pilot: Polina Proutskova, Olga Velichkina
- Conceptualisation: Polina Proutskova
- Research infrastructure, admin: Polina Proutskova
- Communication: Polina Proutskova
- Translation for the Chinese team: Yukun Li
- Training in using digital tools, video tutorials: Polina Proutskova
- Methodology: led by Polina Proutskova with contributions from the teams and from Andrew Killick, John McBride and Elizabeth Phillips
- Choice of musical fragments and the purpose of transcription for each team - teams led by team leads
- Pitch curve correction, note segmentation, note pitch correction, qualitative analysis and within-team synthesis - teams
- Across-team synthesis - all teams led by Polina Proutskova, with contributions from Andrew Killick, John McBride, Elizabeth Phillips, Yukun Li
- Python routines for transcription visualisation: Polina Proutskova with contributions from John McBride, Yuto Ozaki for the Japanese team
- Data cleaning: John McBride, Polina Proutskova, Yukun Li with the help of other team members
- Dataset preparation: John McBride
- Writing: Polina Proutskova, Elizabeth Phillips, Andrew Killick
- Editing and revising: everyone

References:

Abraham, Otto, and Erich M. von Hornbostel. 1909. ‘Vorschläge Für Die Transkription Exotischer Melodien’. *Sammelbände Der Internationalen Musikgesellschaft*, 1–25.

Abraham, Otto, and Erich M Von Hornbostel. 1994. ‘Suggested Methods for the Transcription of Exotic Music’. Translated by George and Eve List. *Ethnomusicology* 38(3): 425–56.

Alekseev, Evgeny. 1990. *Notnaya Zapis’ Narodnoi Muzyki [Score Notation of Folk Music]*. Moscow: Sovetskiy Kompositor.

Arndt, Christin, Kathrin Schlemmer, and Elke Van Der Meer. 2020. ‘Same or Different Pitch? Effects of Musical Expertise, Pitch Difference, and Auditory Task on the Pitch Discrimination Ability of Musicians and Non-Musicians’. *Experimental Brain Research* 238(1): 247–58. https://doi.org/10.1007/s00221-019-05707-8.

Babacan, Onur, Thomas Drugman, Nicolas d’Alessandro, Nathalie Henrich, and Thierry Dutoit. 2013. ‘A Comparative Study of Pitch Extraction Algorithms on a Large Variety of Singing Sounds’. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 7815–19. IEEE.

Baily, John. 1988. ‘Anthropological and Psychological Approaches to the Study of Music Theory and Musical Cognition’. *Yearbook for Traditional Music* 20:114–24.

Benetos, Emmanouil, Simon Dixon, Zhiyao Duan, and Sebastian Ewert. 2019. ‘Automatic Music Transcription: An Overview’. *IEEE Signal Processing Magazine* 36(1): 20–30. https://doi.org/10.1109/MSP.2018.2869928.

Besson, Mireille, Daniele Schön, Sylvain Moreno, Andréia Santos, and Cyrille Magne. 2007. 'Influence of Musical Expertise and Musical Training on Pitch Processing in Music and Language'. *Restorative Neurology and Neuroscience* 25(3–4): 399–410.

Bittner, Rachel M, Katherine Pasalo, Juan José Bosch, Gabriel Meseguer-Brocal, and David Rubinstein. 2021. 'Vocadito: A Dataset of Solo Vocals with "F0", Note, and Lyric Annotations'. *arXiv Preprint arXiv:2110.05580*.

Brown, Steven, and Joseph Jordania. 2013. 'Universals in the World's Musics'. *Psychology of Music* 41(2): 229–48.

Cannam, Chris, Christian Landone, and Mark Sandler. 2010. 'Sonic Visualiser: An Open Source Application for Viewing, Analysing, and Annotating Music Audio Files'. In *Proceedings of the 18th ACM International Conference on Multimedia*, 1467–68.

Casey, Michael A, Remco Veltkamp, Masataka Goto, Marc Leman, Christophe Rhodes, and Malcolm Slaney. 2008. 'Content-Based Music Information Retrieval: Current Directions and Future Challenges'. *Proceedings of the IEEE* 96(4): 668–96.

Chiba, Gakuto, Yuto Ozaki, and Patrick E. Savage. 2024. 'What Is a "Note"? Agreement and Disagreement in Transcriptions of Japanese Folk Songs'. *PsyArXiv Preprint*. https://doi.org/10.31234/osf.io/nh9d2.

Choi, Soonbeom, Wonil Kim, Saebyul Park, Sangeon Yong, and Juhan Nam. 2020. 'Children's Song Dataset for Singing Voice Research'. In *ISMIR Late-Breaking Demo*.

Cottrell, Stephen. 2018. 'Big Music Data, Musicology, and the Study of Recorded Music: Three Case Studies'. *The Musical Quarterly* 101(2–3): 216–43.

De Cheveigné, Alain, and Hideki Kawahara. 2002. 'YIN, a Fundamental Frequency Estimator for Speech and Music'. *The Journal of the Acoustical Society of America* 111(4): 1917–30.

Deutsch, Diana. 2012. 'The Processing of Pitch Combinations'. In *The Psychology of Music*, 249–325. Cognition and Perception. United Kingdom: Elsevier Science.

Devaney, Johanna. 2020. 'An Empirical Evaluation of Note Segmentation and Automatic Pitch-Extraction Methods for the Singing Voice'. In *The Routledge Companion to Interdisciplinary Studies in Singing. Volume I, Development*, edited by Frank A. Russo, Beatriz Senoi Ilari, and Annabel J. Cohen, 136–48. New York, NY: Routledge.

Dowling, W Jay, and DL Harwood. 1987. 'Music Cognition'. *Psychomusicology* 7(1): 91.

Downie, J Stephen. 2003. 'Music Information Retrieval'. *Annual Review of Information Science and Technology* 37(1): 295–340.

Elhilali, Mounya, Juanjuan Xiang, Shihab A. Shamma, and Jonathan Z. Simon. 2009. 'Interaction between Attention and Bottom-Up Saliency Mediates the Representation of Foreground and Background in an Auditory Scene'. Edited by Timothy D. Griffiths. *PLoS Biology* 7(6). https://doi.org/10.1371/journal.pbio.1000129.

England, Nicholas M, Robert Garfias, Mieczyslaw Kolinski, George List, Willard Rhodes, and Charles Seeger. 1964. 'Symposium on Transcription and Analysis: A Hukwe Song with Musical Bow'. *Ethnomusicology* 8(3): 223–77.

Fu, Zih-Sing, and Li Su. 2019. 'Hierarchical Classification Networks for Singing Voice Segmentation and Transcription'. In *Proceedings of the 20th International Society for Music Information Retrieval Conference (ISMIR 2019)*, 900–907.

Goldstone, Robert L, and Andrew T Hendrickson. 2010. 'Categorical Perception'. *Wiley Interdisciplinary Reviews: Cognitive Science* 1(1): 69–78.

Gómez, Emilia, and Jordi Bonada. 2013. 'Towards Computer-Assisted Flamenco Transcription: An Experimental Comparison of Automatic Transcription Algorithms as Applied to a Cappella Singing'. *Computer Music Journal* 37(2): 73–90.

Gong, Rong, Rafael Caro Repetto, and Xavier Serra. 2017. 'Creating an a Cappella Singing Audio Dataset for Automatic Jingju Singing Evaluation Research'. In *Proceedings of the 4th International Workshop on Digital Libraries for Musicology*, 37–40.

Grannan-Rubenstein, Greta, William Grannan-Rubenstein, and Paul Thibodeau. 2014. 'Enculturation Effects of Musical Training on Pitch Discrimination'. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, 36:2985–89.

Hannon, Erin E. 2009. 'Musical Enculturation: How Young Listeners Construct Musical

Knowledge through Perceptual Experience'. In *Neoconstructivism*, edited by Scott Johnson, 1st ed., 132–56. New York: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195331059.003.0007.
Hannon, Erin E, and Laurel J Trainor. 2007. 'Music Acquisition: Effects of Enculturation and Formal Training on Development'. *Trends in Cognitive Sciences* 11(11): 466–72.
Hartmann, W.M. 1997. *Signals, Sound, and Sensation*. AIP Series in Modern Acoustics and Signal Processing. United States: AIP Press.
Herzog, Avigdor. 1964. 'Transcription and Transnotation in Ethnomusicology'. *Journal of the International Folk Music Council* 16:100–101.
Holzapfel, Andre, Emmanouil Benetos, Andrew Killick, and Richard Widdess. 2022. 'Humanities and Engineering Perspectives on Music Transcription'. *Digital Scholarship in the Humanities* 37(3): 747–64.
Hood, Mantle. 1971. *The Ethnomusicologist*. New York: McGraw-Hill.
Huron, David. 2006. *Sweet Anticipation: Music and the Psychology of Expectation*. Cambridge MA: MIT Press.
Hutchins, Sean, Catherine Roquet, and Isabelle Peretz. 2012. 'The Vocal Generosity Effect: How Bad Can Your Singing Be?' *Music Perception* 30(2): 147–59.
Jacoby, Nori, Elizabeth Hellmuth Margulis, Martin Clayton, Erin Hannon, Henkjan Honing, John Iversen, Tobias Robert Klein, et al. 2020. 'Cross-Cultural Work in Music Cognition: Challenges, Insights, and Recommendations'. *Music Perception* 37(3): 185–95. https://doi.org/10.1525/mp.2020.37.3.185.
Jairazbhoy, Nazir A. 1977. 'The" Objective" and Subjective View in Music Transcription'. *Ethnomusicology* 21(2): 263–73.
Killick, Andrew. 2020. 'Global Notation as a Tool for Cross-Cultural and Comparative Music Analysis'. *Analytical Approaches to World Music* 8(2): 235–79.
Kim, Jong Wook, Justin Salamon, Peter Li, and Juan Pablo Bello. 2018. 'CREPE: A Convolutional Representation for Pitch Estimation'. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 161–65. IEEE.
Kirsh, David. 2013. 'Embodied Cognition and the Magical Future of Interaction Design'. *ACM Transactions on Computer-Human Interaction (TOCHI)* 20(1): 1–30.
Kühnis, Jürg, Stefan Elmer, Martin Meyer, and Lutz Jäncke. 2013. 'The Encoding of Vowels and Temporal Speech Cues in the Auditory Cortex of Professional Musicians: An EEG Study'. *Neuropsychologia* 51(8): 1608–18. https://doi.org/10.1016/j.neuropsychologia.2013.04.007.
Lamont, Alexandra. 2012. 'Emotion, Engagement and Meaning in Strong Experiences of Music Performance'. *Psychology of Music* 40(5): 574–94.
Lerdahl, Fred, and Ray Jackendoff. 1983. 'An Overview of Hierarchical Structure in Music'. *Music Perception*, 229–52.
Levitin, Daniel J., Jessica A. Grahn, and Justin London. 2018. 'The Psychology of Music: Rhythm and Movement'. *Annual Review of Psychology* 69(1): 51–75. https://doi.org/10.1146/annurev-psych-122216-011740.
Li, Yukun, Emir Demirel, Polina Proutskova, and Simon Dixon. 2021. 'Phoneme-Informed Note Segmentation of Monophonic Vocal Music'. In *Proceedings of the 2nd Workshop on NLP for Music and Spoken Audio (NLP4MusA)*, 17–21.
List, George. 1963. 'The Musical Significance of Transcription (Comments on Hood, "Musical Significance")'. *Ethnomusicology* 7(3): 193–97.
———. 1974. 'The Reliability of Transcription'. *Ethnomusicology* 18(3): 353–77.
Margulis, Elizabeth Hellmuth. 2014. *On Repeat: How Music Plays the Mind*. Oxford University Press, USA.
Mauch, Matthias, Chris Cannam, Rachel Bittner, George Fazekas, Justin Salamon, Jiajie Dai, Juan Bello, and Simon Dixon. 2015. 'Computer-Aided Melody Note Transcription Using the Tony Software: Accuracy and Efficiency'. In *First International Conference on Technologies for Music Notation and Representation (TENOR 2015)*.
Mauch, Matthias, and Simon Dixon. 2014. 'pYIN: A Fundamental Frequency Estimator Using Probabilistic Threshold Distributions'. In *IEEE International Conference on Acoustics,*

Speech and Signal Processing, 659–63. IEEE.
Mauch, Matthias, Klaus Frieler, and Simon Dixon. 2014. 'Intonation in Unaccompanied Singing: Accuracy, Drift, and a Model of Reference Pitch Memory'. *The Journal of the Acoustical Society of America* 136(1): 401–11.
Meer, Vim van der. 2023. 'Praat Manual for Musicologists'. http://thoughts4ideas.eu/praat-manual-for-musicologists.
Mehr, Samuel A, Manvir Singh, Dean Knox, Daniel M Ketter, Daniel Pickens-Jones, Stephanie Atwood, Christopher Lucas, Nori Jacoby, Alena A Egner, and Erin J Hopkins. 2019. 'Universality and Diversity in Human Song'. *Science* 366(6468): eaax0868.
Meredith, D. 2015. *Computational Music Analysis*. Germany: Springer International Publishing.
Meseguer-Brocal, Gabriel, Alice Cohen-Hadria, and Geoffroy Peeters. 2018. 'DALI: A Large Dataset Of Synchronized Audio, Lyrics And Notes, Automatically Created Using Teacher-Student Machine Learning Paradigm'. In *19th International Society for Music Information Retrieval Conference*.
Metfessel, Milton. 1928. *Phonophotography in Folk Music: American Negro Songs in New Notation*. Chapel Hill: University of North Carolina Press.
Mor, Bhavya, Sunita Garhwal, and Ajay Kumar. 2020. 'A Systematic Literature Review on Computational Musicology'. *Archives of Computational Methods in Engineering* 27:923–37.
Müller, Meinard. 2015. *Fundamentals of Music Processing: Audio, Analysis, Algorithms, Applications*. Vol. 5. Springer.
Music in Motion. 2023. 'Music in Motion: The Automated Transcription for Indian Music (AUTRIM) Project by NCPA and UvA'. https://autrimncpa.wordpress.com.
Nature Editors. 2022. 'Nature Addresses Helicopter Research and Ethics Dumping'. *Nature* 606(7912): 7.
Nettl, Bruno. 2015. *The Study of Ethnomusicology: Thirty-Three Discussions*. University of Illinois Press.
Ozaki, Yuto, John McBride, Emmanouil Benetos, Peter Q Pfordresher, Joren Six, Adam T Tierney, Polina Proutskova, Emi Sakai, Haruka Kondo, and Haruno Fukatsu. 2021. 'Agreement among Human and Automated Transcriptions of Global Songs'. In *22nd International Society for Music Information Retrieval Conference (ISMIR 2021)*, 500–508. International Society for Music Information Retrieval.
Ozaki, Yuto, Adam Tierney, Peter Q. Pfordresher, John M. McBride, Emmanouil Benetos, Polina Proutskova, Gakuto Chiba, et al. 2024. 'Globally, Songs and Instrumental Melodies Are Slower and Higher and Use More Stable Pitches than Speech: A Registered Report'. *Science Advances* 10(20): eadm9797. https://doi.org/10.1126/sciadv.adm9797.
Pfordresher, Peter Q, and Steven Brown. 2017. 'Vocal Mistuning Reveals the Origin of Musical Scales'. *Journal of Cognitive Psychology* 29(1): 35–52.
Pfordresher, Peter Q, Steven Brown, Kimberly M Meier, Michel Belyk, and Mario Liotti. 2010. 'Imprecise Singing Is Widespread'. *The Journal of the Acoustical Society of America* 128(4): 2182–90.
Proutskova, Polina. 2019. 'Investigating the Singing Voice: Quantitative and Qualitative Approaches to Studying Cross-Cultural Vocal Production'. PhD Dissertation, Goldsmiths University of London.
Proutskova, Polina, John McBride, Yuto Ozaki, Gakuto Chiba, Yukun Li, Zhaoxin Yu, Wei Yue, et al. 2023. 'The VocalNotes Dataset'. In *International Society for Music Information Retrieval Conference, Late-Breaking/Demo*. Milano. https://forms.gle/W86j2koBwpkfmnBc9.
Proutskova, Polina, Olga Velichkina, John McBride, Gakuto Chiba, Miranda Crowdus, Yulia Nikolaenko, Yuto Ozaki, et al. 2024. 'VocalNotes Cross-Cultural Song Transcription Methodology: Framework, Challenges and Lessons'. *Analytical Approaches to World Music*.

Proverbio, Alice Mado, and Eleonora Bellini. 2018. ‘How the Degree of Instrumental Practice in Music Increases Perceptual Sensitivity’. *Brain Research* 1691:15–25.
Razumovskaya, E.N. 1997. *Традиционная Музыка Русского Поозерья (Traditional Music of Russian Poozer’ye)*. St.-Petersburg: Kompositor.
Reis, Katherine S, Shannon LM Heald, John P Veillette, Stephen C Van Hedger, and Howard C Nusbaum. 2021. ‘Individual Differences in Human Frequency-Following Response Predict Pitch Labeling Ability’. *Scientific Reports* 11(1): 14290.
Rosenzweig, Sebastian, Frank Scherbaum, and Meinard Muller. 2021. ‘Reliability Assessment of Singing Voice F0-Estimates Using Multiple Algorithms’. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 261–65. Toronto, ON, Canada: IEEE. https://doi.org/10.1109/ICASSP39728.2021.9413372.
Rosenzweig, Sebastian, Frank Scherbaum, David Shugliashvili, Vlora Arifi-Müller, and Meinard Müller. 2020. ‘Erkomaishvili Dataset: A Curated Corpus of Traditional Georgian Vocal Music for Computational Musicology’. *Transactions of the International Society for Music Information Retrieval* 3(1).
Ryynänen, Matti P, and Anssi P Klapuri. 2004. ‘Modelling of Note Events for Singing Transcription’. In *ISCA Tutorial and Research Workshop (ITRW) on Statistical and Perceptual Audio Processing*.
Sauvé, Sarah A, Elizabeth Phillips, Wyatt Schiefelbein, Hideo Daikoku, Shantala Hegde, and Sylvia Moore. 2023. ‘Anti-Colonial Strategies in Cross-Cultural Music Science Research’. *Music Perception* 40(4): 277–92.
Savage, P. E. 2022. *Comparative Musicology: The Science of the World’s Music.* PsyArXiv. https://doi.org/10.31234/osf.io/b36fm.
Savage, Patrick E, Steven Brown, Emi Sakai, and Thomas E Currie. 2015. ‘Statistical Universals Reveal the Structures and Functions of Human Music’. *Proceedings of the National Academy of Sciences* 112(29): 8987–92.
Scherer, Klaus R. 2003. ‘Vocal Communication of Emotion: A Review of Research Paradigms’. *Speech Communication* 40(1–2): 227–56.
Seeger, Charles. 1958. ‘Prescriptive and Descriptive Music-Writing’. *The Musical Quarterly* 44(2): 184–95.
Serra, Xavier, Michela Magas, Emmanouil Benetos, Magdalena Chudy, Simon Dixon, Arthur Flexer, Emilia Gómez Gutiérrez, Fabien Gouyon, Herrera Boyer, and Sergi Jordà Puig. 2013. ‘Roadmap for Music Information Research’. The MIReS Consortium. https://mires.eecs.qmul.ac.uk/files/MIRES_Roadmap_ver_1.0.0.pdf.
Sloboda, John A. 2000. ‘Individual Differences in Music Performance’. *Trends in Cognitive Sciences* 4(10): 397–403.
Stanyek, Jason. 2014. ‘Forum on Transcription’. *Twentieth-Century Music* 11(1): 101–61.
Steinhauer, Kimberly, Mary M McDonald, and Jo Estill. 2017. *The Estill Voice Model: Theory & Translation*. Estill Voice International.
Sturm, Bob L. T., and Arthur Flexer. 2023. ‘Validity in Music Information Research Experiments’. *arXiv Preprint arXiv:2301.01578*. https://doi.org/10.48550/ARXIV.2301.01578.
Sundberg, Johan. 1987. *The Science of the Singing Voice*. Northern Illinois University Press.
Tervaniemi, Mari, Stephanie Kruck, Wouter De Baene, Erich Schröger, Kai Alter, and Angela D Friederici. 2009. ‘Top‐down Modulation of Auditory Processing: Effects of Sound Context, Musical Expertise and Attentional Focus’. *European Journal of Neuroscience* 30(8): 1636–42.
Tilley, Leslie. 2018. ‘Analytical Ethnomusicology: How We Got Out of Analysis and How to Get Back In’. In *Springer Handbook of Systematic Musicology*, edited by Rolf Bader, 1st ed. Springer Handbooks. Germany: Springer Berlin Heidelberg.
Titze, Ingo R, and Daniel W Martin. 1998. ‘Principles of Voice Production’. *The Journal of the Acoustical Society of America* 104(3): 1148–1148.
Viitaniemi, Timo, Anssi Klapuri, and Antti Eronen. 2003. ‘A Probabilistic Model for the Transcription of Single-Voice Melodies’. In *Proceedings of the 2003 Finnish Signal*

Processing Symposium, FINSIG'03, 59–63. Tampere, Finland.
Wang, Jun-You, and Jyh-Shing Roger Jang. 2021. 'On the Preparation and Validation of a Large-Scale Dataset of Singing Transcription'. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 276–80. Toronto, ON, Canada: IEEE. https://doi.org/10.1109/ICASSP39728.2021.9414601.
Wang, Xianke, Wei Xu, Weiming Yang, and Wenqing Cheng. 2022. 'MusicYOLO: A Sight-Singing Onset/Offset Detection Framework Based on Object Detection Instead of Spectrum Frames'. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 396–400. Singapore: IEEE. https://doi.org/10.1109/ICASSP43922.2022.9746684.
Zbikowski, Lawrence Michael. 2002. *Conceptualizing Music: Cognitive Structure, Theory, and Analysis*. Oxford University Press, USA.