THE VocalNotes DATASET

Polina Proutskova
BBC/QMUL
proutskova@googlemail.com

John McBride
Centre for Soft and Living Matter
jmmcbride@protonmail.com

Yuto Ozaki
Keio University
yozaki@sfc.keio.ac.jp

Gakuto Chiba
Keio University
gane1222@sfc.keio.ac.jp

Yukun Li
Queen Mary University of London
yukun.li@qmul.ac.uk

Zhaoxin Yu
Shandong College of Arts
943469149@qq.com

Wei Yue
Shandong College of Arts
yuewei1981zgyy@126.cm

Miranda Crowdus
Concordia University
miranda.crowdus@concordia.ca

Gabriel Zuckerberg
Brown University
gabriel_zuckerberg@brown.edu

Olga Velichkina
French Society for Ethnomusicology
olga.velichkina@gmail.com

Yulia Nikolaenko
independent
mitida173@gmail.com

Yannick Wey
Lucerne University of Applied Sciences and Arts
yannick.wey@hslu.ch

Lawrence Shuster
Cornell University
lbs239@cornell.edu

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

Elizabeth Phillips
McMaster University
phille10@mcmaster.ca

Andrew Killick
University of Sheffield
a.killick@sheffield.ac.uk

ABSTRACT

The VocalNotes dataset is a collection of audio and annotations for excerpts of vocal performances from five musical traditions - Japanese Minyo, Chinese Hebei Bangzi opera, Russian traditional singing, Alpine yodel and Jewish Romaniote chant. For each tradition the dataset contains: about 10 minutes of audio; documentation for the songs from which annotated fragments originate; f0, independent onset, offset and note pitch annotations created by two or three experts; The dataset was created as part of the VocalNotes project [1]. It is released under CC-BY-NC-SA license and can be accessed by filling out a request form.

1. INTRODUCTION

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 [2], MIR-ST500 [3]), yet the resulting quality of the annotations is low and is 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 [4], Korean [5] and Chinese songs [6]. Cross-cultural datasets have been published outside MIR, e.g. by the Speech/Song project [7]. Here we expand on this by releasing a dataset of vocal performances from five different traditions – Japanese, Chinese, Russian, Alpine and Jewish – with pitch and note annotations by experts from each tradition.

Automated systems continue to fall short of human expertise in singing onset and note annotation [8].

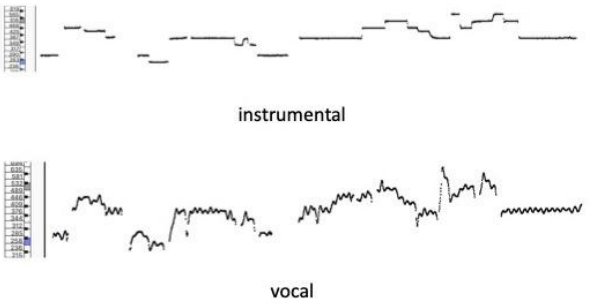

Figure 1. Fundamental frequency curve of a Russian traditional song "Vy kumushki kumitesia" from Poozerie region, played on a keyboard and sung by the traditional performer Olga Sergeeva

Typically in MIR research, one develops algorithms that aim to reproduce a 'ground truth'. One of the pitfalls is the assumption that there is one 'correct' way to transcribe music, which has been refuted by ethnomusicologists [9]–[14]. Transcriptions can differ due to cultural knowledge, or different aims as to what level of detail should be annotated. Limits of human perception can lead to uncertainties, and transcribers may have different strategies to resolve uncertainties. Singing in particular is difficult to transcribe due to inherently unstable pitch curves (Fig. 1) and vocal drift [15]. Vocal techniques such as vibrato, embellishments and glides contribute to the analytical ambiguity of singing, for humans and automated systems alike. Following [16] we provide a larger dataset with the advantage of 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.

Figure 2. The VocalNotes project participants map.

2. THE VOCALNOTES PROJECT

The VocalNotes project [1] 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 Min'yo, Chinese Bangzi opera, Russian traditional village singing, Alpine yodelling, and Romaniote Jewish chanting) each transcribed 10 minutes of vocal recordings from their culture, where transcription consisted of segmentation and note pitch correction. The experts then systematically compared their independent transcriptions, isolated and formalised contexts which led to disagreements.

Each team was led by an ethnomusicologist and included two or three transcribers deeply familiar with the tradition they were annotating. The instruction was to produce independent analytic transcriptions based on expert's perception. Yet each team had their own research question which shaped the choice of musical fragments and the transcription process:

- The Japanese team analysed 9 recordings: 3 different singers x 3 different folk songs, with one historical performance and two modern performances by the co-authors for each song.
- 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 (21 monophonic tracks from 12 songs)
- The Jewish team transcribed 10 excerpts from the few existing recordings of Romaniote Torah cantillation - a selection of the recitation of the same biblical text chanted by different Romaniote practitioners. They include speech-like declamation alongside singing.
- The Alpine team transcribed 9 solo excerpts from the central European yodelling tradition.
- The Chinese research team transcribed 11 excerpts from Hebei Bangzi, a genre of Chinese traditional opera, covering a range of tempos including slow, moderate, fast, and rubato patterns.

3. ANNOTATION METHODOLOGY

Transcription was conducted in three phases: pitch curve correction, segmentation, note pitch correction. The first two phases were conducted in Tony [17], with note pitch correction done in Sonic Visualizer [18].

The f0 curve was automatically extracted using pYIN in Tony, then manually corrected and agreed within the team. Segmentation and note pitch correction were performed independently by each transcriber. Note segmentation was done in Tony without relying on Tony's automatic segmentation suggestions. It was found that the note pitch automatically assigned by Tony to be the median of the pitch curve was not always perceived as accurate. Therefore note segments were exported to Sonic Visualiser and pitch was manually corrected where necessary. Sonic Visualiser turned out to be a much less convenient and responsive tool for note pitch correction than Tony was for note segmentation, therefore our segmentation annotations, which were performed from scratch, are of higher quality than the note pitch annotations, which were largely based on the automated suggestions.

4. THE DATASET

The dataset comprises three components: the audio files, the song documentation files and the annotations. Audio filenames are the song titles chosen by the team (usually lyrics incipits) from which the analysed fragments originated. The documentation files from each team contain information about the songs, the performers, where and when the songs were recorded, what is their social or musical function (e.g. "wedding song", or "Genesis 1:1-11"), the rights holders of the complete recordings and where these recordings can be accessed.

The annotations are in the form of .csv files, including f0, note segments with uncorrected pitch (f0 median), note segments with manually corrected pitch. The filenames are constructed of the song title, transcriber abbreviation and data type (pitches, segments or notes): `<Song_title__OV__notes.csv>`

The dataset is available under the CC BY-NC-SA 4.0 license to allow non-commercial use and facilitate replicability and further research. A version of record of the annotations is hosted on Zenodo[1]. The access to the full dataset including audio can be requested by filling out a short online form[2]. The answers will be reviewed by a team member and a link to download the dataset will be provided. This is in recognition of the fact that the audio fragments in our dataset contain sensitive religious and ritual material. We ask the users of the dataset to treat the recordings, their performers and their communities with respect and to refrain from aggregating the dataset for unrestricted online access.

The VocalNotes dataset welcomes new contributions of audio and annotations which follow the same methodology [19]. All the annotations collected during the project, the guidelines, video tutorials and software scripts are available in the VocalNotes Open Science Framework repository[3].

[1] https://doi.org/10.5281/zenodo.10065955
[2] request form at https://forms.gle/W86j2koBwpkfmnBc9
[3] https://osf.io/4n5ry/

5. REFERENCES

[1] Polina Proutskova *et al.*, 'VocalNotes – investigating perceptual differences in segmentation and pitch through transcriptions of vocal performances in five musical traditions', *Anal. Approaches World Musics J.*, in preparation 2024.

[2] G. Meseguer-Brocal, A. Cohen-Hadria, and G. Peeters, 'DALI: A large dataset of synchronized audio, lyrics and notes, automatically created using teacher-student machine learning paradigm', *ArXiv Prepr. ArXiv190610606*, 2019.

[3] J.-Y. Wang and J.-S. R. Jang, '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)*, Toronto, ON, Canada: IEEE, Jun. 2021, pp. 276–280. doi: 10.1109/ICASSP39728.2021.9414601.

[4] S. Rosenzweig, F. Scherbaum, D. Shugliashvili, V. Arifi-Müller, and M. Müller, 'Erkomaishvili Dataset: A curated corpus of traditional Georgian vocal music for computational musicology', *Trans. Int. Soc. Music Inf. Retr.*, vol. 3, no. 1, 2020.

[5] S. Choi, W. Kim, S. Park, S. Yong, and J. Nam, 'Children's song dataset for singing voice research', in *ISMIR Late-Breaking Demo*, 2020.

[6] R. Gong, R. C. Repetto, and X. Serra, 'Creating an a cappella singing audio dataset for automatic jingju singing evaluation research', presented at the Proceedings of the 4th International Workshop on Digital Libraries for Musicology, 2017, pp. 37–40.

[7] Y. Ozaki *et al.*, 'Similarities and differences in a global sample of song and speech recordings [Stage 1 Registered Report]', 2022.

[8] Y. Ozaki *et al.*, 'Agreement among human and automated transcriptions of global songs', PsyArXiv, preprint, Jul. 2021. doi: 10.31234/osf.io/jsa4u.

[9] G. List, 'The Musical Significance of Transcription (Comments on Hood," Musical Significance")', *Ethnomusicology*, vol. 7, no. 3, pp. 193–197, 1963.

[10] A. Herzog, 'Transcription and Transnotation in Ethnomusicology', *J. Int. Folk Music Counc.*, vol. 16, pp. 100–101, 1964.

[11] N. M. England, 'Symposium on Transcription and Analysis: A Hukwe Song with Musical Bow', *Ethnomusicology*, vol. 8, no. 3, pp. 223–277, 1964.

[12] G. List, 'The reliability of transcription', *Ethnomusicology*, pp. 353–377, 1974.

[13] Alekseev, E, *Notnaya zapis' narodnoi muzyki [Score notation of folk music]*. Moscow: Sovetskiy Kompositor, 1990.

[14] J. Stanyek, 'Forum on transcription', *Twent.-Century Music*, vol. 11, no. 1, pp. 101–161, 2014.

[15] M. Mauch, K. Frieler, and S. Dixon, 'Intonation in unaccompanied singing: Accuracy, drift, and a model of reference pitch memory', *J. Acoust. Soc. Am.*, vol. 136, no. 1, pp. 401–411, 2014.

[16] R. M. Bittner, K. Pasalo, J. J. Bosch, G. Meseguer-Brocal, and D. Rubinstein, 'vocadito: A dataset of solo vocals with $ f_0 $, note, and lyric annotations', *ArXiv Prepr. ArXiv211005580*, 2021.

[17] M. Mauch *et al.*, 'Computer-aided melody note transcription using the Tony software: Accuracy and efficiency', presented at the First International Conference on Technologies for Music Notation and Representation (TENOR 2015), 2015.

[18] C. Cannam, C. Landone, and M. Sandler, 'Sonic visualiser: An open source application for viewing, analysing, and annotating music audio files', presented at the Proceedings of the 18th ACM international conference on Multimedia, 2010, pp. 1467–1468.

[19] Polina Proutskova *et al.*, 'VocalNotes methodology: framework, challenges and lessons', *Anal. Approaches World Musics J.*, in preparation 2024.