name	# unique	multi-label	description
id	1570	No	unique identifier of each music piece
instruments	32	Yes	instruments participating in performance
genres	32	Yes	music style annotations
place	81	Yes (to contain regions)	full hierarchy of the place of origin
coordinates	81	No	latitude and longitude
is-danced	2	No	binary value (0 or 1)
youtube-id	74	No	id of the episode available online
start-ts	1570	No	start timestamp of the piece
end-ts	1570	No	end timestamp of the piece

Table 1. Metadata contained in the dataset.

"end-ts" that denote the exact time (second) that a song starts and ends in the corresponding video. The duration of the music pieces sums to approximately 80 hours.

The column "youtube-id" contains the id under which the video of the full episode is available online. The count of unique values are essentially the number of episodes that were used for the creation of the dataset. A typical duration of an episode is roughly a hundred minutes. The "is-danced" binary label informs about whether a music piece is being danced by the guests of the show. The music pieces annotated with "1" are approximately 51% while in the rest of them no dance performance occurs.

The classification of Greek music in "genres" is a work that requires one to take into account a number of socio-cultural and anthropo-geographical criteria. At an abstract level, we can distinguish the music of urban centers in contrast to the music of rural areas of Greece, with the former including rebetiko, laiko, urban-folk among others, while the latter, the music of rural areas, is what we generally call traditional music. Figure 1 shows the frequencies of the genres in the dataset, with "traditional" being the dominant one constituting almost 78% of the total. Depending on the place of origin, the style of a traditional music piece varies accordingly and, thus, several sub-genres flourish, such as Epirotic for the songs originated from Epirus. The 32 unique values in this metadata category are separated into 5 distinct genres and 27 sub-genres.

From a musicological perspective, Greek traditional music can be divided into two large geographic areas, i.e., the island and the mainland Greece. Each one creates a

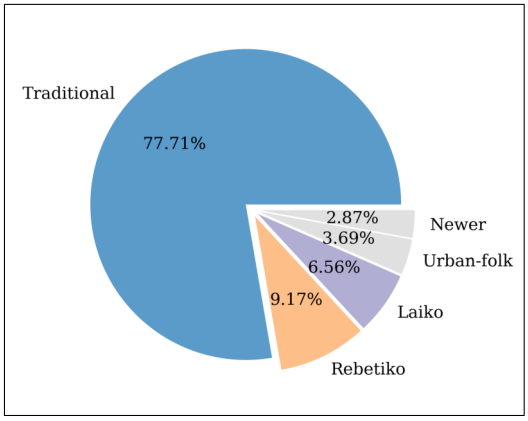

Figure 1. Relative frequencies of the music genres in the dataset.

distinctive musical feeling as there are large variations both on the rhythmic approach and the scales that are commonly used. For example, in islands we frequently come across music pieces with simple, fast rhythms while in the mainland more complex, slow rhythms are the norm.

The "place" (of origin) metadata category can be annotated with (i) a single label when the region from which a song derives is known, (ii) two labels when both region and a specific place are known and (iii) "None" denoting that there is not a specific place of origin for this piece. As an example, a music piece can be annotated with the region "Aegean sea" or with both "Aegean sea" and "Naxos", an island of the Aegean sea, if this knowledge is available. Specifically, from the 81 unique places in the dataset, 20 are regions and only half of them include the remaining 61. The most represented regions can be seen in Figure 2.

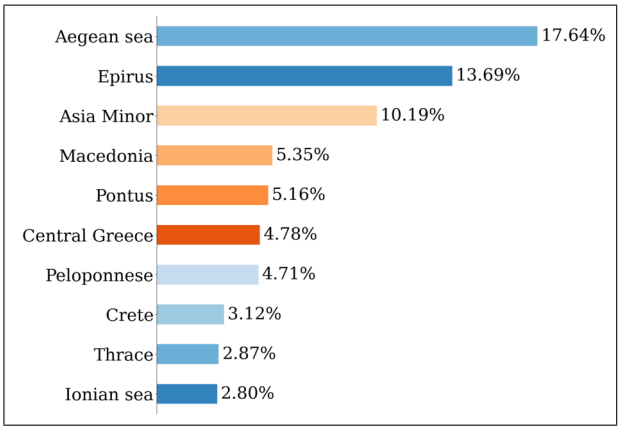

Figure 2. Relative frequencies of the most represented regions in the dataset.

The exact latitude and longitude of each place is also available at the "coordinates" column. The music pieces that do not have an explicit place of origin, such as the ones that belong to the "laiko" genre, are accounted for approximately 23% of the total. Figure 3 shows the location of the 81 places that exist in the dataset. We may notice the constant ability of music to excess the borders; places where Greek culture thrived in the past and neighboring countries

that share the same tradition, form a mosaic of people that communicated freely with each other in a musical way that has reached towards us.

Figure 3. Map of all the places of the dataset.

Analogous connections, like the ones between genres and places, one expects to be observed between places and instruments as well. Indeed, for over 100 years, the established music ensembles of Greek traditional music are generally two, namely (i) those with the violin as leading instrument (often substituted by lyra and santouri), which have a greater presence in island Greece and (ii) those with the klarino (Greek clarinet) as the leading instrument, which is dominant in the mainland.

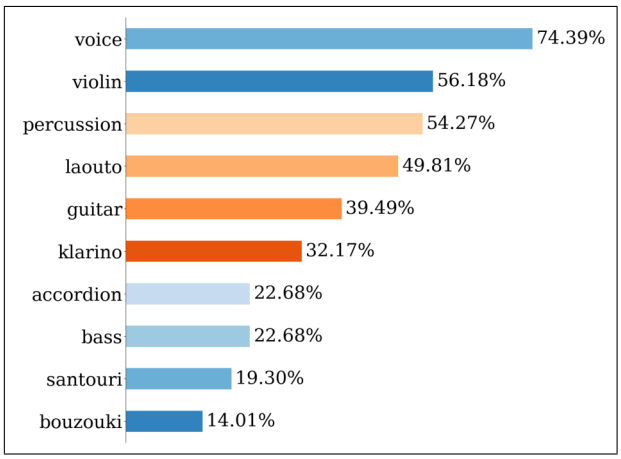

Figure 4. Relative frequencies of the most common musical instruments in the dataset.

In the popular and modern music domain, there is a great variety of instruments, but in most cases bouzouki, guitar, accordion and bass are common members of a laiko or rebetiko music ensemble. In the traditional music groups, the percussion and the laouto (Greek lute) are permanent companions of the leading instruments, offering melodic-harmonic background and rhythmic support. Of course, voice holds the main role in all kinds of performances.

In Figure 4 one can see the frequencies of the most popular instruments in the dataset. Singing voice is evident in almost 75% of it and instruments like violin, percussion and laouto, that have presence in both islands and mainland as well, are following.

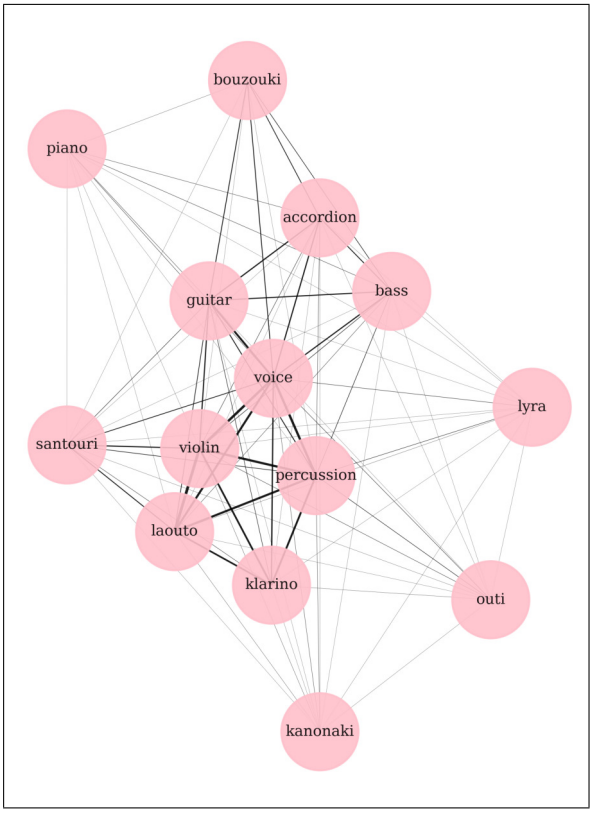

Figure 5. Graph that shows the co-occurrences of the fourteen most common musical instruments. The edge width is proportional to the number of music pieces a pair of instruments co-occur.

With regards to the music ensembles, it should be noted that 296 unique groups of instruments exist in the dataset, with the one constituted by voice, violin, percussion, laouto and klarino being by far the most popular by participating in the performance of around 12% of the music pieces. The co-occurrences of the most popular instruments can be seen in Figure 5 where the width of the graph edges is proportional to the number of pieces a pair of instruments co-occur in the dataset.

Sample rows of the dataset can be seen in Table 2. The dataset along with the baseline classification methods and the trained models are available online. [1]

The shared metadata should be considered as the version 1.0 of the dataset. In the next versions, it will be evolved towards two main directions, namely (i) the incorporation of more metadata categories such as annotations according to the content of the lyrics, the lyrics themselves as well as information about the types of the dances that occur and (ii) the addition of more music pieces by following the same process either for next episodes of the same documentary series or for other series that have a similar theme.

3. BASELINE CLASSIFICATION

The audio recording of each music piece is represented using a Mel-scaled Spectrogram (mel-spectrogram): this is

[1] https://github.com/pxaris/lyra-dataset

id	instruments	genres	place	coordinates	is-danced	youtube-id	start-ts	end-ts
alexandra	voice|violin| percussion| laouto| klarino	Traditional| Epirotic	Epirus| Zagori	39.8648| 20.9284	0	qrOwc1mLFUk	749	927
choros-tik	percussion| lyra	Traditional| Pontian	Pontus	40.9883| 39.7270	1	Aws0Y3aLaIs	1731	1886
agiothodo-ritissa	voice|violin| santouri| percussion| laouto|guitar	Rebetiko	None	None	1	0cj8BNcAhg4	2632	2853
einai-arga-poly-arga	voice|piano| guitar|bass| bouzouki| accordion	Laiko	None	None	0	zkoqg3VRVLA	2365	2614

Table 2. Sample rows of the dataset. Pipe "|" is used to separate values in a field.

a spectrogram whose frequencies are converted to the mel scale according to the equation:

$$m = 2595 \log_{10}(1 + \frac{f}{700}) = 1127 \ln(1 + \frac{f}{700}) \quad (1)$$

where **m** is the frequency in Mels and **f** is the frequency in Hz. Mel-spectrograms are calculated per fix-sized segment duration of 10 seconds. A non-overlapping window of 50 milliseconds has been applied, therefore the Mel-spectrogram size is 200 windows $\times$ 128 frequency bins.

As a baseline classification approach, each 10-second Mel-spectrogram is classified to the aforementioned tasks (genre, place and instruments) using a Convolutional Neural Network (CNN). CNNs have been widely used in general audio [15], speech [16, 17] and music [18] classification tasks. In particular, we have adopted the following architecture: 4 convolutional layers of 5 $\times$ 5 kernels, single stride, and max pooling of size 2. The number of convolutional kernels (channels) are for the first layer 32, for the second 64, for the third 128 and for the fourth 256. The final output of the convolutional layers is passed through 3 fully connected layers, with the first having an output dimension of 1024, the second 256 and the third equal to the number of classes.

Note that fix-sized duration of segments is necessary, since audio recordings do not share the same size. Adopting a much longer segment would require zero padding for several mel-spectrograms and probably more CNN parameters. In addition, splitting the song into non-overlapping segments achieves some type of data augmentation. For two of the adopted classification tasks (namely genre and place), we have trained the CNNs using a multiclass, single-label setup, while for the instrument task, which is multi-label, we have trained multiple binary CNNs, one for each instrument, which have been evaluated separately.

After the training and validation procedure of each of the aforementioned CNNs, final testing was applied on the respective test recordings. For the test set, to avoid spreading pieces from the same broadcast across data splits, we separate training and test data on an episode level. From the 74 unique episodes, we randomly split 20% of them and use all the music pieces they include, namely 330, to form the test data. Obviously, this final testing needs to be carried out on a "song-level", not a 10-sec segment level. Towards this end, a simple aggregation method was adopted, by just averaging the posteriors of the individual segment decisions of the CNNs. This aggregated estimate was used as the final prediction and evaluated in the final testing.

4. RESULTS

The performance results during the training of the baseline classifiers are shown in Table 3, computed on a validation subset that corresponds to 20% of the segments. For the multi-label task (instrument recognition), we show the F1 metric for each binary subtask separately, while for the single-label tasks (genre and place) we show the overall macro F1 for all classes. We remind that this evaluation is performed on the validation split of the 10-second data.

Classifiers type	Task	F1 (%)
Multiclass classifiers	Genre	82.3
	Place	85.5
Binary classifiers for Instruments	voice	75.8
	violin	86.3
	percussion	97.1
	laouto	96.7
	guitar	80.2
	klarino	95.0

Table 3. F1 scores of the classifiers on the 10-second segments of the training data using a 20% validation subset.

As soon as the 10-second classifiers are trained, they are applied on the whole recordings of the testing data, and a simple majority aggregation is performed to extract the final decision, as described in the previous Section. For the instrument classification task, we compute the Area Under the Curve (AUC) metric per label (binary classification subtask). The results are shown in Table 4.

Finally, the confusion matrices along with the respective F1 measures for the multiclass, single-label classification tasks of "genres" and "places" are shown in Figures 6 and 7. All genres have been taken into account for the respective classification, but not the sub-genres. On "places" task, the pieces have been classified to the 10 most common regions (including "None") plus the "other" category for the remaining. Genres classifier macro F1-score

Instrument	AUC (%)
voice	68.9
violin	85.2
percussion	95.1
laouto	93.8
guitar	73.5
klarino	90.9

Table 4. Area Under the Curve (AUC) scores of the instrument classifiers on the test data.

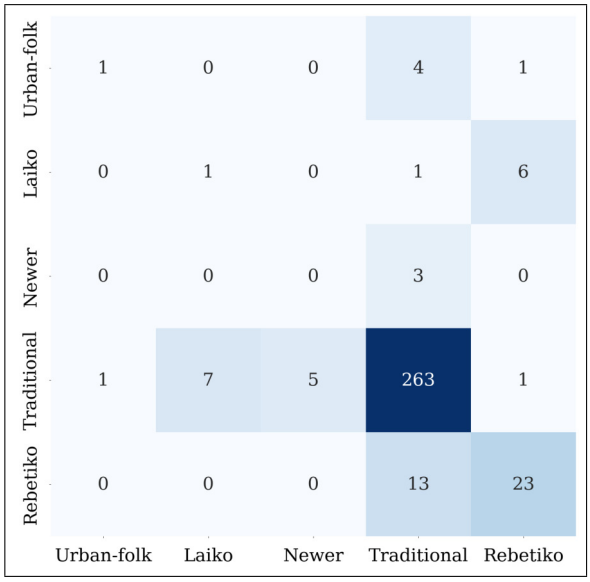

Figure 6. Confusion matrix for "genre" classes on the test data. **Macro F1-**score: **39.9**% and **Micro F1-**score: **87.2**%.

is 39.9% and micro F1-score is 87.2%, while for the places classifier the macro F1-score is 34.4% and the micro F1-score is 42.4%.

5. DISCUSSION

A reason that the "voice" classifier has lower performance compared to the other ones may be of a musicological character. Indeed, while the presence of the rest of the instruments can depend significantly on the music style of a piece, the same does not apply to "voice" as it is the dominant musical instrument in any genre. Given the fact that the binary classifier is trained to recognize an instrument (evident in a part of a music piece) in each of the 10-second segments, regardless if it is present on it or not, we expect to move towards a space with latent musical features such as the music style, where "voice" may not be as discriminative as the rest of the instruments are.

With regards to confusion matrices, the misclassifications can be either due to imbalance between classes or statistical correlations across them. Specifically, for the "places" task, the confusions between regions that are geographically near may be justifiable, while for "genres" task the imbalance between the classes seems to have significantly affected the performance of the model at the least represented ones.

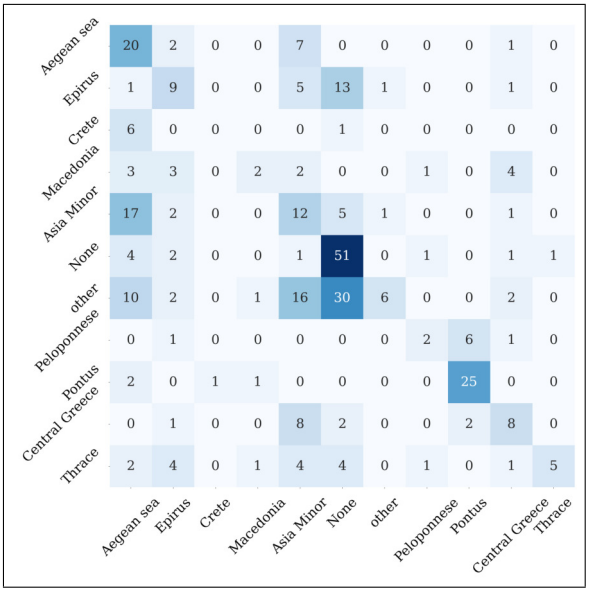

Figure 7. Confusion matrix for "place" classes on the test data. **Macro F1-**score: **34.4**% and **Micro F1-**score: **42.4**%.

We intend to improve models performance and conduct a more in-detail analysis in order to examine hypotheses such as the aforementioned ones in a follow-up study.

6. CONCLUSIONS

Greek traditional and folk music integrates components of Eastern and Western idioms, providing interesting research directions in the field of computational ethnomusicology. This paper presents "Lyra", a dataset of 1570 traditional and folk Greek music pieces that includes audio and video (timestamps and links to YouTube videos), along with annotations that describe aspects of particular interest for this dataset, including instrumentation, geographic information and labels of genre and subgenre, among others. The advantage of this dataset is that the entire content is harvested from web resources of a Greek documentary series that was produced by academics with specialization in this music and, therefore, includes high-quality and rich annotations extracted from the content of the shows. Additionally, the production of recordings and video material is professional-level, providing a common ground in terms of audio quality. Three baseline audio-based classification tasks are performed, namely instrument identification, place of origin and genre classification.

The presented results indicate that specialized tasks, that use the audio signal, can potentially provide valuable insight about several aspects of this music. The combination of video and audio signals allows possible experimentation on methods that process multimodal data. The Lyra dataset includes material that readily allows MIR tools to be employed for reaching valuable musicological results. This dataset can also, potentially, foster the expansion of the MIR methods altogether.

7. ACKNOWLEDGEMENTS

The authors would like to thank Professor Lambros Liavas, the producer, researcher and presenter of the documentary series "To Alati tis Gis - Salt of the Earth". His work not only provides a fruitful source of knowledge and art but also acts as an inspiration for new researchers. This work would not have been possible without the help of the volunteer annotators. A big thank to the students of the Department of Music Studies of National and Kapodistrian University of Athens. Finally, we would like to thank the reviewers for their valuable and constructive comments.

8. REFERENCES

[1] M. Panteli, E. Benetos, and S. Dixon, "A review of manual and computational approaches for the study of world music corpora," *Journal of New Music Research*, vol. 47, no. 2, pp. 176–189, 2018.

[2] P. Van Kranenburg, M. De Bruin, and A. Volk, "Documenting a song culture: The dutch song database as a resource for musicological research," *International Journal on Digital Libraries*, vol. 20, no. 1, pp. 13–23, 2019.

[3] A. Srinivasamurthy, G. K. Koduri, S. Gulati, V. Ishwar, and X. Serra, "Corpora for music information research in indian art music," in *Georgaki A, Kouroupetroglou G, eds. Proceedings of the 2014 International Computer Music Conference, ICMC/SMC; 2014 Sept 14-20; Athens, Greece*. Michigan Publishing, 2014.

[4] R. C. Repetto, N. Pretto, A. Chaachoo, B. Bozkurt, and X. Serra, "An open corpus for the computational research of arab-andalusian music," in *Proceedings of the 5th International Conference on Digital Libraries for Musicology*, 2018, pp. 78–86.

[5] N. Kroher, J.-M. Díaz-Báñez, J. Mora, and E. Gómez, "Corpus cofla: A research corpus for the computational study of flamenco music," *Journal on Computing and Cultural Heritage (JOCCH)*, vol. 9, no. 2, pp. 1–21, 2016.

[6] S. Abdoli, "Iranian traditional music dastgah classification." in *Proceedings of the 12th International Society for Music Information Retrieval Conference*, 2011, pp. 275–280.

[7] 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," *Transactions of the International Society for Music Information Retrieval*, vol. 3, no. 1, 2020.

[8] X. Gong, Y. Zhu, H. Zhu, and H. Wei, "Chmusic: A traditional chinese music dataset for evaluation of instrument recognition," in *2021 4th International Conference on Big Data Technologies*, 2021, pp. 184–189.

[9] M. K. Karaosmanoğlu, B. Bozkurt, A. Holzapfel, and N. D. Dişiaçık, "A symbolic dataset of turkish makam music phrases," in *Proceedings of Fourth International Workshop on Folk Music Analysis (FMA2014)*, 2014, pp. 10–14.

[10] G. B. Dzhambazov, A. Srinivasamurthy, S. Sentürk, and X. Serra, "On the use of note onsets for improved lyrics-to-audio alignment in turkish makam music," in *Devaney J, Mandel MI, Turnbull D, Tzanetakis G, editors. Proceedings of the 17th International Society for Music Information Retrieval Conference*, 2016.

[11] T. Fouloulis, A. Pikrakis, and E. Cambouropoulos, "Traditional asymmetric rhythms: A refined model of meter induction based on asymmetric meter templates," in *Proceedings of the third international workshop on folk music analysis*, 2013, pp. 28–32.

[12] D. Makris, K. L. Kermanidis, and I. Karydis, "The greek audio dataset," in *IFIP International Conference on Artificial Intelligence Applications and Innovations*. Springer, 2014, pp. 165–173.

[13] D. Makris, I. Karydis, and S. Sioutas, "The greek music dataset," in *Proceedings of the 16th International Conference on Engineering Applications of Neural Networks (INNS)*, 2015, pp. 1–7.

[14] M. Leman and P.-J. Maes, "The role of embodiment in the perception of music," *Empirical Musicology Review*, vol. 9, no. 3-4, pp. 236–246, 2015.

[15] M. Papakostas and T. Giannakopoulos, "Speech-music discrimination using deep visual feature extractors," *Expert Systems with Applications*, vol. 114, pp. 334–344, 2018.

[16] Z. Huang, M. Dong, Q. Mao, and Y. Zhan, "Speech emotion recognition using cnn," in *Proceedings of the 22nd ACM international conference on Multimedia*, 2014, pp. 801–804.

[17] K. Choi, G. Fazekas, M. Sandler, and K. Cho, "Convolutional recurrent neural networks for music classification," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 2392–2396.

[18] M. Ashraf, G. Geng, X. Wang, F. Ahmad, and F. Abid, "A globally regularized joint neural architecture for music classification," *IEEE Access*, vol. 8, pp. 220 980–220 989, 2020.

ANALYSIS AND DETECTION OF SINGING TECHNIQUES IN REPERTOIRES OF J-POP SOLO SINGERS

Yuya Yamamoto[1] **Juhan Nam** [2] **Hiroko Terasawa** [1]
[1] Doctoral Program in Informatics, University of Tsukuba, Japan
[2] Graduate School of Culture Technology, KAIST, South Korea
s2130507@s.tsukuba.ac.jp, juhan.nam@kaist.ac.kr, terasawa@slis.tsukuba.ac.jp

ABSTRACT

In this paper, we focus on singing techniques within the scope of music information retrieval research. We investigate how singers use singing techniques using real-world recordings of famous solo singers in Japanese popular music songs (J-POP). First, we built a new dataset of singing techniques. The dataset consists of 168 commercial J-POP songs, and each song is annotated using various singing techniques with timestamps and vocal pitch contours. We also present descriptive statistics of singing techniques on the dataset to clarify what and how often singing techniques appear. We further explored the difficulty of the automatic detection of singing techniques using previously proposed machine learning techniques. In the detection, we also investigate the effectiveness of auxiliary information (i.e., pitch and distribution of label duration), not only providing the baseline. The best result achieves 40.4% at macro-average F-measure on nine-way multi-class detection. We provide the annotation of the dataset and its detail on the appendix website [0] .

1. INTRODUCTION

A singing voice is one of the most essential elements of music. It provides impactful emotional expressions through melody and lyrics. In particular, in popular music, the role of the singing voice is more important, as the vocal quality and unique style of singers are critical in attracting people. Vocals mainly consist of singer individuality (i.e., vocal fold vibration and vocal tract resonance) and singing expressions (i.e., fine control of pitch, timbre, and loudness). Singing techniques, as the name suggests, are extended techniques in human singing. It characterizes singing voices as a component of expressions in vocal performances. In particular, many professional singers in popular music use singing techniques to make a performance more expressive, each in a different manner.

[0] https://yamathcy.github.io/ISMIR2022J-POP/

The singing voice is an important subject in the field of music information retrieval (MIR), and there are many prior works that extract quantitative or objective information as in the task of singing transcription [1], lyric identification [2], and singer identification [2]. Computational analysis of singing techniques based on MIR can contribute to the clarification of singing styles and can be applied to music discovery, vocal training, consumer-generated media and so on. Singing techniques have been of keen interest, especially in J-POP, both among singers and music creators; the singing style of J-POP singers has a wide variety, and their singing techniques are diverse. In addition, singing techniques are one of the evaluation criteria in Japanese commercial karaoke systems with scoring systems. Meanwhile, as for music creators, *VOCALO* songs, whose singer's voice are created by a singing voice synthesis software such as VOCALOID [3], have been established as a music genre in Japan. Many VOCALOID creators manually manipulate the voice of VOCALOID to make it more expressive, sometimes while referring to how the actual singer produce the singing voice, like how singing techniques are applied. Therefore, J-POP is an attractive research target for singing technique analysis and computational singing technique analysis may bring benefits to real-world applications.

In this regard, we present a case study of singing technique analysis in J-POP. We first introduce a new dataset, including the annotation of singing techniques for J-POP recordings. We analyze the dataset using descriptive statistics to illustrate the technique distributions and singer styles. Finally, we conduct singing technique detection based on machine learning methods, which can be beneficial to analyze the one side of the singing voice automatically. We demonstrate the detection of various singing techniques (i.e., identifying the singing technique type with its starting time and duration) and report baseline results using existing methods. Finally, we investigate the effect of the identification model with auxiliary information derived from the statistics of the dataset and pitch information.

2. RELATED WORKS

2.1 Singing Technique Analysis

Analyzing the acoustic patterns of singing expressions has been a long-standing research topic. Research on vibrato

dates to the 1930s [4]. Later, computer-based methods allowed a more quantitative analysis of vibrato parameters [5,6]. In addition to vibrato, other singing techniques, such as growling [7], portamento [8], voice registers [9,10], and extreme vocal effects [11] have also been investigated. An analysis of singing techniques was also conducted for J-POP. Migita et al. investigated the vibrato parameters of the imitated singing voices of J-POP vocalists [12]. Yamamoto et al. conducted an exploratory analysis of singing techniques that imitate the singing style of J-POP singers [13]. They found 13 types of singing techniques from 48 tracks and analyzed them with annotations. Because singing techniques in J-POP cover wide repertoires [14], such exploratory analysis is also important for the purpose of singing performance analysis. As our interests are handling and detecting multiple singing techniques rather than specifying a single technique, we annotated the label as in the manner in [13] but on the original real-world tracks of J-POP. Therefore, our new dataset consists of labels for multiple singing techniques.

2.2 Singing Technique Datasets

There are a handful of datasets that focus on the analysis of singing techniques. The Phonation Mode dataset consisted of four vocal modes (neutral, pressed, breathy, and flow) of sustained sung vowels from four singers [15]. Although the dataset includes a wide range of pitches, they are discrete, and thus lack a melodic context. VocalSet [16] handles the issue by having voices sung in contexts of scales, arpeggios, long tones, and excerpts. In addition, it covers a broader range of singing techniques, such as vibrato, trill, vocal fry, and inhaled singing. Because audio samples in the two datasets were newly recorded, they were collected under controlled conditions. Therefore, their characteristics of singing techniques might be different from those appeared in songs.

However, several datasets have been annotated for real-world vocal music. The KVT dataset was originally built for vocal-related music-tagging tasks in K-POP music [17]. It contains 70 vocal tags, of which six are related to the singing technique (whisper/quiet, vibrato, shouty, falsetto, speech-like, and non-breathy). The annotation was conducted using crowdsourcing. The MVD dataset was built to analyze screams in heavy metal music, and has four different types of screams (high fry, mid fry, low fry, and layered) [18]. The regions and types of screams in the audio files were manually annotated. Similar to these studies, we are interested in versatile singing techniques in real-world music and thus we chose commercial music to build the dataset.

2.3 Singing/Playing Technique Identification on MIR research

Studies have been conducted on the automatic identification of singing techniques beyond computational analysis. There are two scenarios for identifying a singing techniques. One is classifying sung vocal audio into singing techniques. Several studies have conducted singing technique classification on VocalSet [19] and phonation modes [20, 21]. These works identify singing techniques but do not provide time-related information, such as start time and duration. The other method is to detect the singing technique in time. Miryala et al. [22] identified the singing expressions of raga, classical music in India. They created 35 recordings in eight ragas sung by six singers and showed a classification accuracy of 84.7% using a rule-based classifier. Yang et al. [8] proposed AVA, an interface for analyzing vibrato and portamento based on the filter diagonalization method (FDM) and hidden Markov model (HMM), respectively. They also provided a case study of the analysis of vibrato and portamento in the Beijing opera. Ikemiya et al. [23] provided rule-based parameterization of four singing expressions (vibrato, kobushi, gliss-up, gliss-down), extracted them from recorded vocal tracks to transfer to other vocal tracks, and demonstrated them on two sung excerpts. Kalbag et al. [18] provided scream detection for heavy metal music. They also built the MVD dataset and demonstrated classification, including non-scream singing and non-vocal music. Since singing techniques appear locally on sung vocals, we adapted these temporal detection strategies for identification.

3. DATASET

In this section, we describe the organization of our new dataset used for the analysis of this study. Due to the limit of the space, more details are described in the appendix site [0] .

3.1 Song Selection

To cover a wide range of singing techniques, the dataset should include various types of vocalists, considering gender, genre, tempo, and mood. We first listed 42 solo singers (21 males and 21 females), and four famous hit songs were selected from each singer to be as different as possible. Each song was performed as solo performance in Japanese. We collected audio tracks from commercial CD recordings of the J-POP songs. We trimmed the collected audio tracks and annotated only the first consecutive section (i.e., verse-A, verse-B and Chorus). Since we prioritized including various songs in our dataset rather than fully annotating multiple verses in a single song, we could collect a diverse set of singing styles.

When using popular commercial songs for academic research, copyright is always an issue. The works by Kim [17] on K-POP and Kalbag [18] on heavy metal music are good examples of developing datasets using commercial music, which distributes only labels about the performed music but not the audio signals or score information. We followed their solutions and made our dataset that consists of singing technique and pitch (frequency) contour labels.