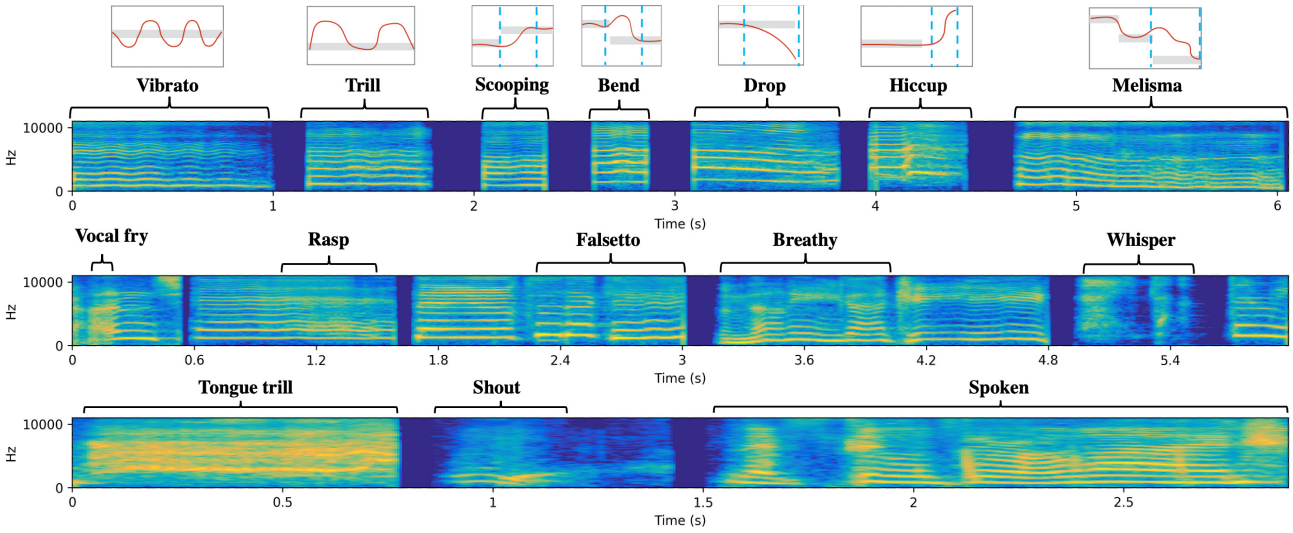

Figure 1. Singing techniques included in the dataset. (upper) Pitch techniques with a sketch of pitch contour. Gray is the target vocal note, the red line is the pitch contour and blue dotted lines are boundary of each technique. (middle) Timbral techniques. We also show the non-technique extracted from the same track. (lower) Miscellaneous techniques.

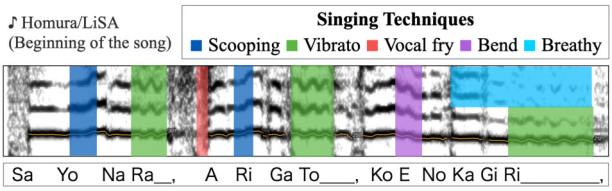

Figure 2. An overview of annotation. (middle) an excerpt spectrogram with singing technique (colored area) and vocal melody (orange curve).

3.2 Data Pre-processing

The dataset contains two types of annotations: vocal melody and singing techniques. We illustrate the annotation result in Figure 2. All annotations were conducted on isolated vocal tracks after vocal separation using Demucs v3 [24], which is a state-of-the-art model for musical source separation. During the annotation process, we confirmed that there was no dropout of vocal regions by comparing the original mixed track. Instead, we observed that the separated vocals tends to retain the back-chorus or sometimes instrumental sounds that are similar to the singer's voice (e.g., electric guitar, synthesizer).

3.3 Singing Technique Annotation

We thoroughly surveyed singing techniques based on instructional books [14, 25] and other conventional scientific research related to singing techniques [7, 13, 16, 17, 26–29] and defined the labels for the major techniques that appear in these references. We manually annotated songs using these labels.

Although many other singing techniques still exist, we considered the following 15 singing techniques in this study. The pitch contour sketches of these techniques and spectrogram examples are illustrated in Figure 1.

We note that these label names are not unique in the real-world (e.g., scooping is also called 'portamento,' 'glissando,' 'gliss-up,' 'shakuri' (in Japanese)). We made frame-level annotations on the audio tracks collected based

Table 1. The description of each singing technique appeared in the dataset.

Technique	description	type
vibrato	a periodic oscillation of pitch.	pitch
scooping	an upper continuous picth change	pitch
drop	a lower continuous picth change	pitch
bend	a short tremolo or U/inverted-U shaped pitch change	pitch
hiccup	a short hiccuping on attack/release of note	pitch
melisma	a musical arrangement in which several notes are applied to one syllable of a lyric.	pitch
trill	a continuous pitch change between two notes	pitch
falsetto	sung by falsetto register.	timbre
breathy	sung by breathy sound.	timbre
whisper	sung like whispering.	timbre
rasp	sung by a creaky voice with subharmonics.	timbre
vocal fry	sung by a creaky voice and pulse register phonation.	timbre
spoken	singing like rapping, *sprechgesang* [1], and some other styles like speaking.	misc.
shout	shouting.	misc.
tongue trill	a rolling tongue, occurred on [r] consonant.	misc.

[1] A vocal style between singing and speaking.

[2] Although discarded in the analysis, we also annotate 'unknown' labels if the region seems to represent a singing technique but is difficult to classify them into any of the techniques above and found 24 unknown techniques in total. In the techniques, there are some sounds akin to coughing and pig squeals, for example.

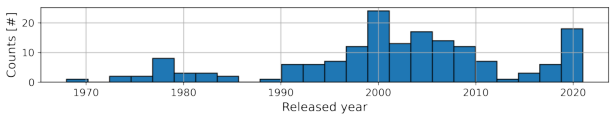

Figure 3. Distribution of song year of release.

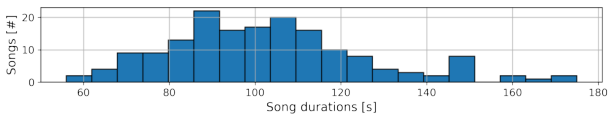

Figure 4. Distribution of song length in seconds.

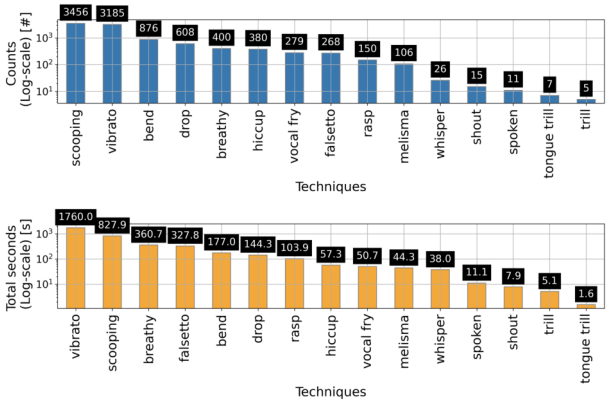

Figure 5. Statistics of the labels. upper: counts, lower: total duration.

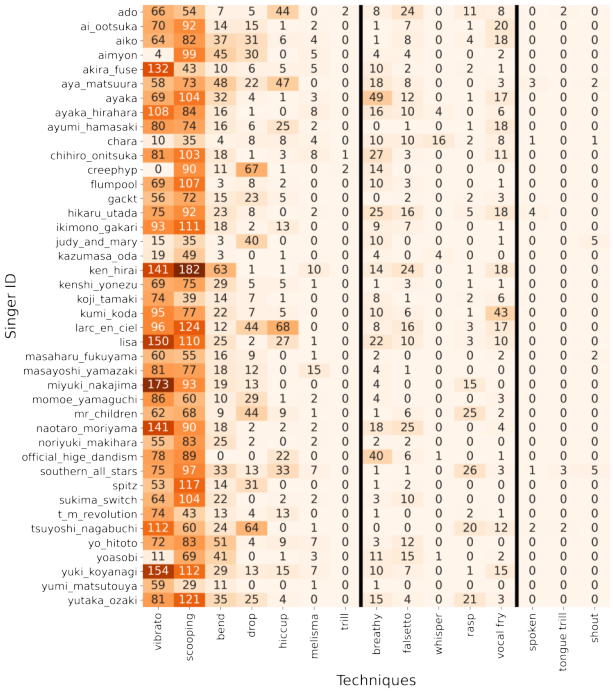

Figure 6. Occurrence distribution of singing techniques by singer. The vertical black line divides each category.

on the vocabulary. Singing techniques were carefully annotated by an experienced vocalist (i.e., the first author of the paper) with the helps of sound playback and visualizing the spectrograms and pitchgrams. The annotation process is conducted on Sonic Visualizer [30].

3.4 Melody Annotation

Since pitch is an essential component of singing technique analysis, we further annotated melodic pitch using Tony [31], followed by manual correction such as removing the unvoiced parts and reverberation tails.

4. DESCRIPTIVE STATISTICS

4.1 Song Statistics

The distribution of songs selected for the dataset based on the year of release is shown in Figure 3. Songs can be collected from various eras, ranging from 1968 to 2021. The distribution of the songs in the dataset is shown in Figure 4. The overall length was 4h 47m 39s, and the average length of a song track was 1m 43s. The ratio of technique regions per song track was 22.8%.

4.2 Label Statistics

We present the statistics for the annotated labels as a histogram at the upper side of Figure 5. The most frequent technique is 'scooping'. It is followed by 'vibrato', 'bend', and 'drop'. This indicates that such techniques are relatively common in J-POP. These techniques are also used in Japanese commercial karaoke systems for vocal assessment [32]. 'scooping', 'bend' and 'drop' are portamento, whose frequent use is sometimes considered undesirable in classical singing [25]. It can be said that this frequent use of portamento is a characteristic of J-POP.

We show the distribution of techniques by each singer in Figure 6 and find that the occurrence frequency of the techniques is different for each singer. We also confirmed that scooping appears more than 29 times for every singer, whereas several singers tend to not use vibrato frequently (e.g., 'creephyp', 'aimyon', and 'yoasobi').

The lower side of Figure 5 and Figure 7 shows the total duration and distribution of each technique using a boxenplot, respectively. Figure 7 shows that most of the techniques are relatively short (that is, the range between 0.1s and 1s.), especially for 'drops', 'scooping', 'bend', 'hiccup', and 'vocal fry'. The average length of the singing techniques was 0.4s.

5. SINGING TECHNIQUE DETECTION

In this section, we describe the detection of the singing techniques. We conducted experiments on a multi-class detection scenario, which handles classification and localization simultaneously. The problem setting is the same as that in sound event detection (SED) [33]. Figure 8 illustrates the proposed method. We investigated the performance of the CRNN model in a singing technique detec-

[3] https://seaborn.pydata.org/generated/seaborn.boxenplot.html

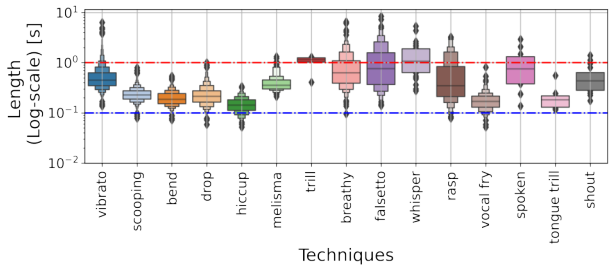

Figure 7. Distribution of each technique. Red and blue horizontal lines indicate 1 s and 0.1 s, respectively.

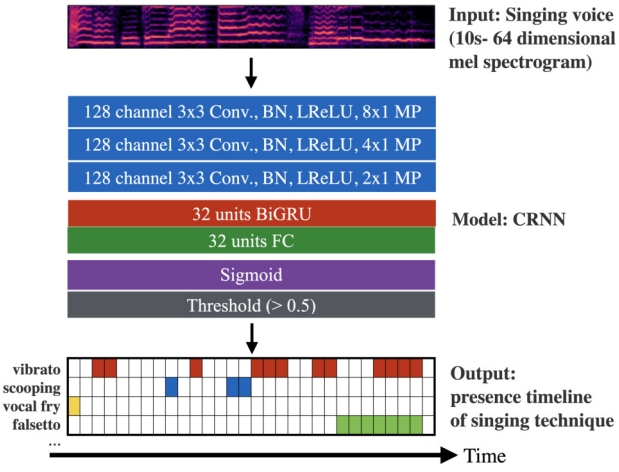

Figure 8. An overview of the singing technique detection. The configuration of the CRNN model is also shown.

tion task. In addition to running the simple setting, we also investigated how the considerations of the characteristics of the dataset improve the performance (i.e., label sparseness and pitch information). Thus, through the experiment, we attempted to answer for the following three research questions: 1) To what extent can we detect singing techniques using machine learning? 2) Can modification of the loss function treat the problem of label sparseness? 3) Can auxiliary pitch information help improve the detection performance?

5.1 Experimental Conditions

We performed singer-wise seven-fold cross validation during the experiment. We first split the singers into seven groups. Then, in each run, one group is left out for test set, another one group is for validation set , and the rest are used for training set. Owing to the label imbalances of techniques between singer, we used the 9 most common classes (i.e., 'bend,' 'breathy,' 'drop,' 'falsetto,' 'hiccup,' 'rasp,' 'scooping,' 'vibrato,' and 'vocal fry'), which appear on every fold. We divided the singer fold to balance the amount of each technique as much as possible.

We segmented the vocal tracks, which are separated by Demucs v3 [24] in advance, into 10s audio clips and non-overlapping parts at a sample rate of 44.1 kHz. Therefore, they were converted into 64-dimensional log-mel spectrograms. For experiments that involve the computation of short-time Fourier transform (STFT), we used a Hann window with 2048 samples to compute the discrete Fourier transform (DFT). The hop size was set to 10 ms in every experiment.

We used a CRNN model, which is widely used as a baseline system for SED. We adapted the settings of the model of Imoto et al. [34], which treated a similar issue (i.e, label sparseness) of ours in SED. The model configuration is shown in Figure 8. The model consists of three convolutional layers, one layer of bidirectional GRU cells, and one fully connected (FC) layers. The input was the aforementioned log-mel-spectrogram, and the output was a multi-hot vector representation of the singing technique for each time frame. All experiments were conducted using a batch size of 16, and the model parameters were optimized using an Adam optimizer with a learning rate of 10^{-4}. The model stopped training if the value of the loss function on the validation set did not improve over 10 epochs.

The model was optimized using the binary cross-entropy (BCE) loss. Performance was evaluated using segment-based recall (R), precision (P), macro-F-measure (Macro-F), and micro-F-measure (Micro-F) [35] using sed_eval. The macro-F-measure and micro-F-measure denote the class-wise and instance-wise average of the F-measure, respectively. We set the segment length to 50 ms.

5.2 Model Improvement

5.2.1 Label-frame sparsity

As mentioned in Section 4.2, most singing techniques have a short duration (i.e., shorter than 1 s). This can cause label imbalance between non-technique frames. To alleviate the problem, we investigated the effect of *focal loss* [36], which was originally proposed for image object detection. The focal loss can be represented as follows:

$$L_{fl}(p_t) = -\alpha(1-p_t)^{\gamma}\log(p_t) \quad (1)$$

where $\alpha \in [0,1]$ is a weighting factor for balancing the importance of positive and negative examples, and the term $(1-p_t)^{\gamma}$ is a modulating factor, with γ controlling the rate of dominant examples. We set $\alpha = 0.2$ and $\gamma = 2$ for all conditions in the work that used focal loss.

5.2.2 Pitch information

We demonstrated that some of the techniques, such as vibrato and scooping, have a pattern of the shape to certain extent. Therefore, it is possible that explicitly feeding the pitch helps in the detection. We further investigated the effect of the auxiliary information. Under this condition, we considered two ways of obtaining pitch in our experiment. One is the ground-truth (GT) pitch annotation mentioned in Subsection 3.4. However, when used in real-world applications, it is difficult to obtain correct pitch annotation. Hence, we also used pitch estimation predicted by CREPE [37] on a separated vocal track. As CREPE can compute pitch confidence, we adopted the pitch value

[4] https://tut-arg.github.io/sed_eval/index.html

Table 2. The results of singing technique detection.

	Macro-F	Micro-F	P	R
BCE	37.7%	56.6%	40.2%	41.7%
Focal	39.6%	57.7%	40.3%	42.6%
BCE-GT	39.1%	57.1%	41.4%	42.8%
BCE-CREPE	39.1%	57.1%	40.2%	41.7%
Focal-GT	**40.4%**	**58.3%**	**43.1%**	42.2%
Focal-CREPE	40.3%	57.9%	42.7%	**43.3%**

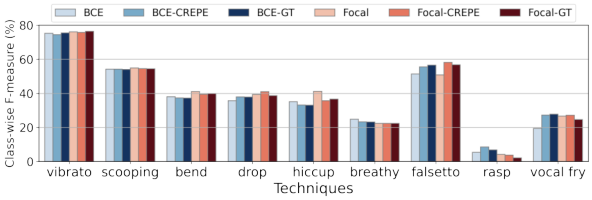

Figure 9. Class-wise F-measure

where the confidence value was higher than 0.5, whose overall accuracy was 78.0% evaluated by mir_eval [38]. We converted the pitch contour to a mel-band pitchgram that has the same frequency dimensions as the input mel-spectrogram, as in the work of singer identification [39]. Therefore, we stacked it on a mel-spectrogram along the channel axis.

5.3 Experimental Results

5.3.1 Baseline

The first row of Table 2 shows the results of the CRNN model compiled by BCE: 37.7% for Macro-F and 56.6% for Micro-F. The fact that Macro-F is much lower than Micro-F indicates that the model failed to identify minority rather than majority classes. Actually, the class-wise F-measure, as shown in Figure 9, indicates that detection of the minority techniques (e.g., 'rasp' and 'vocal fry') is more difficult than the majority techniques (e.g., 'vibrato' and 'scooping').

5.3.2 Effect of focal loss and pitch information

First, we show the results for the effect of *focal loss* in the middle part of Table 2. The detection performance improved by 1.9% in Macro-F. We further studied the class-wise F-measure, as shown in Figure 9, and confirmed that the performance of short techniques, such as 'bend' (38.0% $\rightarrow$ 41.1%), 'drop' (35.7% $\rightarrow$ 39.3%), and 'hiccup' (35.1% $\rightarrow$ 41.2%) are especially improved. This indicates that *focal loss* can adapt to the sparsity of the label.

Next, we showed the results of the effect of the auxiliary pitch information in the middle part of Table 2. A remarkable finding is that the F-measure of 'falsetto' was particularly improved (51.3% $\rightarrow$ 56.5%). We also confirmed that the recall value of 'falsetto' is raised more (64.7% $\rightarrow$ 74.1%) rather than the precision value (45.1% $\rightarrow$ 48.3%). This indicates that pitch information hint at the relationship between the sung pitch value and the occurrence of falsetto (i.e., falsetto can appear only at a higher pitch). We also confirmed that the input of CREPE pitch shows the similar tendency.

We further investigated the effects of combining these two types of auxiliary information. As shown in the bottom part of Table 2, the combination of GT pitch and focal loss is the best condition in terms of both macro- and micro-F values. We can confirm that both effects of each auxiliary information occur under this condition (red and dark-red bars in Figure 9). This indicates that the condition is robust to techniques that have a short duration or some relationships with pitch.

5.3.3 Challenges

The remaining challenges in the task are detected unnatural region lengths and singers' identity. We confirmed that one of the common mis-detection cases is from the detection of too short or frequently switching regions. It might need to consider post-processing, such as temporal filtering or probabilistic model (e.g, hidden Marcov model, hidden semi-Marcov model, etc.). As for singer identity, it can affect the performance of the timbral techniques. We labeled timbral techniques when the sung voice transformed from the ordinary voice of the singer, and it caused the confusion. For example, the 'rasp' from a singer who has clean voice and ordinary voice from a singer who has raspy voice can be confusing. Therefore, there is still an issue of disentangling singers' identity and singing techniques.

6. CONCLUSION AND FUTURE WORK

This work presented an analysis and identification of the singing technique in J-POP music as a case study. In addition, we built a new dataset consisting of 168 J-POP songs with annotation of pitch contour and singing techniques. The contributions of this study are summarized as follows: 1) we constructed a dataset with frame-level annotations of singing techniques on 168 famous J-POP vocal songs, 2) conducted a descriptive-statistical analysis of the dataset, 3) proposed a CRNN-based singing technique detection model as a baseline, and confirmed that both auxiliary pitch information and focal loss help the detection performance.

As we showed the appearance of singing techniques relates to some other factors (e.g., difference of singing technique distribution by singer, pitch information helps the detection of falsetto, etc.), we are planning further investigations of relationships or joint identification between other musical factors (e.g. musical note, lyrics, beat and tempo, etc.) or higher level information (e.g. singer, songs' emotion, mood, etc.) with making more types of annotation. Although this work focuses only on J-POP, the knowledge can be applicable to other types of singing. Similar study on other style of vocal such as K-POP, western pops or non-popular vocal and comparison between them are other interesting topics for future work.

[5] The other metrics were following; 85.9% for voice recall, 29.9% for voicing false alarms, 94.5% for raw pitch accuracy.

7. ACKNOWLEDGEMENTS

This work was supported by JST SPRING, Grant Number JPMJSP2124. Dr. Maiko Murai (Associate Prof. at the University of Tsukuba) gave advices on copyright law. Yoshihide Ishikawa, Sayori Takayama and Karolina Shi helped the preparation of the metadata. We are grateful to above all.

8. REFERENCES

[1] M. Ryynänen, *Singing Transcription*, A. Klapuri and M. Davy, Eds. Boston, MA: Springer US, 2006. [Online]. Available: https://doi.org/10.1007/0-387-32845-9_12

[2] E. J. Humphrey, S. Reddy, P. Seetharaman, A. Kumar, R. M. Bittner, A. Demetriou, S. Gulati, A. Jansson, T. Jehan, B. Lehner, A. Kruspe, and L. Yang, "An introduction to signal processing for singing-voice analysis: High notes in the effort to automate the understanding of vocals in music," *IEEE Signal Processing Magazine*, vol. 36, no. 1, pp. 82–94, 2018.

[3] H. Kenmochi and H. Ohshita, "Vocaloid - commercial singing synthesizer based on sample concatenation," in *Interspeech*, 2007.

[4] C. E. Seashore, "A musical ornament, the vibrato," *Proc. of Psychology of Music, 1938*, pp. 33–52, 1938.

[5] E. Prame, "Measurements of the vibrato rate of ten singers," *J. Acoust. Soc. Am.*, vol. 96, no. 4, pp. 1979–1984, 1994.

[6] J. Sundberg, "The perception of singing," in *The psychology of music*. Elsevier, 1999, pp. 171–214.

[7] K.-I. Sakakibara, L. Fuks, H. Imagawa, N. Tayama *et al.*, "Growl voice in ethnic and pop styles," in *Proceedings of International Symposium on Musical Acoustics*, 2004.

[8] L. Yang, S.-K. Rajab, and E. Chew, "Ava: an interactive system for visual and quantitative analyses of vibrato and portamento performance styles," in *In Proceedings of the 17th International Society for Music Information Retrieval Conference (ISMIR)*, 2016.

[9] K. Hirayama and K. Ito, "Discriminant analysis of the utterance state while singing," in *Proceedings of IEEE Symposium on Signal Processing and Information Technology (ISSPIT)*, 2012.

[10] Y. Lee, M. Oya, T. Kaburagi, S. Hidaka, and T. Nakagawa, "Differences among mixed, chest, and falsetto registers: A multiparametric study," *Journal of Voice (In Press, Corrected Proof)*, 2021.

[11] O. Nieto, "Voice transformations for extreme vocal effects," *Master thesis, Pompeu Fabra University*, 2008.

[12] N. Migita, M. Morise, and T. Nishiura, "A study of vibrato features to control singing voices," *Proceedings of International Congress on Acoustics*, pp. 23–27, 2010.

[13] Y. Yamamoto, T. Nakano, M. Goto, H. Terasawa, and Y. Hiraga, "Analysis of frequency, acoustic characteristics, and occurrence location of singing techniques using imitated j-pop singing voice," *The Special Interest Group Technical Report of IPSJ (MUS)*, no. 20, pp. 1–8, 2021, (in Japanese).

[14] AKIRA/Utana, *The Secret Singing Technique Used by Professional Singers 17 - The karaoke score improved from a 68 to a 92!-*. Tsuta-shobou, 2013, (in Japanese).

[15] P. Proutskova, C. Rhodes, T. Crawford, and G. Wiggins, "Breathy, resonant, pressed–automatic detection of phonation mode from audio recordings of singing," *Journal of New Music Research*, vol. 42, no. 2, pp. 171–186, 2013.

[16] J. Wilkins, P. Seetharaman, A. Wahl, and B. A. Pardo, "Vocalset: A singing voice dataset," in *In Proceedings of the 15th International Society for Music Information Retrieval Conference (ISMIR)*, 2018, pp. 468–474.

[17] K. L. Kim, J. Lee, S. Kum, C. L. Park, and J. Nam, "Semantic tagging of singing voices in popular music recordings," *IEEE/ACM TASLP*, vol. 28, pp. 1656–1668, 2020.

[18] V. Kalbag and A. Lerch, "Scream detection in heavy metal music," in *Proceedings of Sound and Music Computing (SMC)*, 2022.

[19] Y. Yamamoto, J. Nam, H. Terasawa, and Y. Hiraga, "Investigating time-frequency representations for audio feature extraction in singing technique classification," in *Proceedings of Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2021.

[20] J.-L. Rouas and L. Ioannidis, "Automatic classification of phonation modes in singing voice: towards singing style characterisation and application to ethnomusicological recordings," in *Interspeech*, 2016.

[21] D. Stoller, S. Dixon *et al.*, "Analysis and classification of phonation modes in singing," in *The 17th International Society for Music Information Retrieval Conference (ISMIR)*, 2016.

[22] S. S. Miryala, K. Bali, R. Bhagwan, and M. Choudhury, "Automatically identifying vocal expressions for music transcription." in *In Proceedings of the 14th International Society for Music Information Retrieval Conference (ISMIR)*, 2013.

[23] Y. Ikemiya, K. Yoshii, and K. Itoyama, "Transferring vocal expressions of a professional singer to unaccompanied singing signals," in *ISMIR-LBD 2014*, 2014.

[24] A. Défossez, "Hybrid spectrogram and waveform source separation," in *In Proceedings of the 22nd International Society for Music Information Retrieval Conference (ISMIR)*, 2021.

[25] F. Husler and Y. Rodd-Marling, *Singing: The physical nature of the vocal organ: A guide to the unlocking of the singing voice*. Random House (UK), 1976.

[26] Y. Ikemiya, K. Itoyama, and H. G. Okuno, "Transferring vocal expression of f0 contour using singing voice synthesizer," in *Proceedings of International Conference on Industrial, Engineering & Other Applications of Applied Intelligent Systems. (IEA/AIE)*, 2014, pp. 250–259.

[27] M. Panteli, R. Bittner, J. P. Bello, and S. Dixon, "Towards the characterization of singing styles in world music," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2017, pp. 636–640.

[28] M. Nakazato, "How to sing ornamented melody in j-pop: Through the analysis of kobushi in ken hirai, keisuke kuwata, chemistry and dreams come true," *Japanese Journal of Music Education Practice*, vol. 5, no. 1, pp. 32–39, 2007, (in Japanese).

[29] M. Jones, "The rhythm and blues (r&b) protest songs of the civil rights movement: Outlining the natural alignment between the foundational r&b recordings artists and the african-american church during the movement," *Master Theses, Liberty University*, 2016.

[30] C. Cannam, C. Landone, and M. Sandler, "Sonic visualiser: An open source application for viewing, analysing, and annotating music audio files," in *Proc. ACMMM 2010*, 2010, pp. 1467–1468.

[31] M. Mauch, C. Cannam, R. Bittner, J. Bello *et al.*, "Computer-aided melody note transcription using the tony software: Accuracy and eciency," in *Proceedings of the First International Conference on Technologies for Music Notation and Representation (TENOR 2015)*, 2015.

[32] D. C. LTD., "Karaoke systems," *Japan Patent 2020-166142*, 2020, (in Japanese).

[33] A. Mesaros, T. Heittola, T. Virtanen, and M. D. Plumbley, "Sound event detection: A tutorial," *IEEE Signal Processing Magazine*, vol. 38, no. 5, pp. 67–83, 2021.

[34] K. Imoto, S. Mishima, Y. Arai, and R. Kondo, "Impact of sound duration and inactive frames on sound event detection performance," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE, 2021, pp. 860–864.

[35] A. Mesaros, T. Heittola, and T. Virtanen, "Metrics for polyphonic sound event detection," *Applied Sciences*, vol. 6, no. 6, p. 162, 2016.

[36] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision (ICCV)*, 2017, pp. 2980–2988.

[37] J. W. Kim, J. Salamon, P. Li, and J. P. Bello, "Crepe: A convolutional representation for pitch estimation," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2018, pp. 161–165.

[38] C. Raffel, B. McFee, E. J. Humphrey, J. Salamon, O. Nieto, D. Liang, D. P. Ellis, and C. C. Raffel, "mir_eval: A transparent implementation of common mir metrics," in *In Proceedings of the 15th International Society for Music Information Retrieval Conference (ISMIR)*, 2014.

[39] T.-H. Hsieh, K.-H. Cheng, Z.-C. Fan, Y.-C. Yang, and Y.-H. Yang, "Addressing the confounds of accompaniments in singer identification," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE, 2020, pp. 1–5.

Papers