Investigating Time-Frequency Representations for Audio Feature Extraction in Singing Technique Classification

Yuya Yamamoto*, Juhan Nam†, Hiroko Terasawa* and Yuzuru Hiraga*
* University of Tsukuba, Tsukuba, Japan
E-mail:{s2130507@s, terasawa@slis, hiraga@slis}.tsukuba.ac.jp
† KAIST, Daejeon, South Korea
E-mail: juhan.nam@kaist.ac.kr

***Abstract*—Singing techniques are used for expressive vocal performances by employing temporal fluctuations of timbre, pitch, and other components of the voice. In this study, we compare the performances of hand-crafted features and automatically extracted features using deep learning methods to identify different singing techniques. Hand-crafted acoustic features are based on expert knowledge of singing voice whereas the deep learning methods take low-level feature representations, such as spectrograms and raw waveforms, as inputs and learn features automatically using convolutional neural networks (CNNs). These extracted features are used as an input to the random forest classifier for comparison with the hand-crafted features for 10-class singing technique classification. We show that the CNN-based features outperform the hand-crafted features in terms of classification accuracy. Furthermore, we explore various time-frequency representations as an input to the CNNs. We show that the best performing input is multi-resolution short-time Fourier Transform (STFTs), when the CNN kernels are oblong and they slide on the frequency- and time-axis directions separately.**

I. Introduction

In a vocal performance, singers often fluctuate the pitch, loudness, and timbre of their voice to make a song more expressive. Such fluctuations are commonly called singing techniques. At the signal level, singing techniques are observed in time–frequency representations as heavy temporal modulations of harmonic frequencies such as vibrato or highly noisy components over broad frequency bands such as a whisper voice. Singing technique classification is a challenging task because dynamic changes in multiple factors such as pitch, loudness, and timbre occur simultaneously.

We investigated the time-frequency representations for singing technique classification. Traditional hand-crafted features such as Mel-frequency cepstral coefficients (MFCCs) and other representations rich in time-frequency information plugged into CNNs are compared in terms of their efficiency in the automatic classification of singing techniques.

Well-designed feature representations of a singing technique will enable an automatic discrimination among the patterns of spectro-temporal fluctuations in vocal performance. In the field of music information retrieval (MIR), hand-crafted features, which are designed based on expert knowledge, have been successful. Several hand-crafted features have been proposed to capture the characteristics of timbre and modulation. A data-driven approach based on deep neural networks (DNNs) recently outperformed conventional methods based on hand-crafted features in other MIR tasks, and we anticipate that a similar approach can also be effective in singing technique classification. In particular, we want to focus on convolutional neural networks (CNNs) as feature extractors because of their invariance to time shifts and frequency transpositions. A variety of acoustic feature representations, such as raw waveforms, time-frequency representations (e.g., STFT), or time-frequency representations using log-scaled filter banks (e.g., a Mel-spectrogram), have been employed as inputs to a CNN. However, the most suitable input representation differs depending on the type of MIR task [1]. Because of the temporal and noisy nature of singing techniques, suitable representations for singing technique classification should better capture the time-frequency properties of the audio signal than those for the other MIR tasks.

In this paper, we compare and evaluate various audio feature representations to find an effective representation for singing technique classification. Our experimental results show that a multi-resolution STFT with a CNN works best (77.8 % classification accuracy), outperforming commonly used feature representations such as an MFCC (66.9 %) and Mel-spectrogram (72.8 %).

II. Related Work

MFCCs are among the most popular hand-crafted features for timbre, containing information regarding the spectral envelopes. MFCCs are used in many singing voice-related MIR tasks such as singer identification [2], sung language identification[3], and gender identification of the singer[4].

MFCCs have also been used in combination with other acoustic features to classify singing-technique-related aspects. Stoller et al. [5] investigated a variety of acoustic features in relation to a 4-class phonation mode (i.e., normal, breathy, pressed, flow) classification performance. The authors indicated that the combination of 80-dimensional MFCCs, cepstral peak prominence, and temporal flatness is the best feature set for phonation mode classification with an accuracy of

78%. In addition, Kroher et al. [6] combined 13-dimensional MFCCs with vibrato features and statistics (e.g., the register of the singer and number of occurrences of various singing expressions) as features for singer identification. As a result, the performance reached 83.1%, which was 23.1% higher than that of MFCCs alone.

Recent studies in the area of instrument technique identification have explored representations other than MFCCs to exploit more detailed time-frequency information. There are a number of studies on instrument playing technique identification that use representations other than MFCCs, namely, a guitar playing technique using sparse coding[7], Hartley transform[8], a violin playing technique [9], piano sustain pedal detection using Mel-spectrograms [10], playing technique classification of Chinese bamboo flutes using a wavelet scattering transform [11], [12], and guqin techniques using a constant-Q transform, pitch salience, and pitch contour [13]. Finally, Lostanlen et al. [14] used a wavelet scattering transform to classify instrumental playing techniques of multiple instruments simultaneously. Using these representations, which are rich in time-frequency information for singing technique identification seem promising because they capture the time-frequency details better than MFCCs.

Several studies have used DNNs for the classification of singing voices and instrumental playing techniques. Abeßer et al. [15] investigated the efficiency of CNN-based feature extraction for pitch contour classification. The results of solving four different tasks show that a CNN with a simple structure can achieve the same discriminative performance as hand-crafted features. We used this framework for the automatic extraction of audio features.

III. Methods

A. Datasets

In this study, we use VocalSet [16], which is the only publicly available database for studies on singing technique. VocalSet is a large-scale dataset that contains singing voices by 20 different professional singers (9 female and 11 male), performing 17 different singing techniques in various contexts such as arpeggio, scale, and long tones. We selected the samples corresponding to 10 different singing techniques (belt, breathy, inhaled singing, lip trill, spoken excerpt, straight tone, trill, trillo, vibrato, and vocal fry) by all singers from VocalSet, which resulted in 915 files ranging in length from 1.7 to 21.5 s. We then split the audio signals in each file into 3-s audio clips and non-overlapping chunks at a sample rate of 44.1 kHz, resulting in 4905 samples. The details of these samples are listed in Table I.

TABLE I
SELECTED SAMPLES FROM VOCALSET.

Label name	Type	Samples #
straight	None	1241
belt	Timbre	423
breathy	Timbre	455
vocal fry	Timbre, Modulation	587
vibrato	Modulation	1034
trill	Modulation	323
trillo	Modulation	242
lip trill	Modulation	376
inhaled	Other	151
spoken	Other	73

B. Method of model comparison

To compare the hand-crafted features and other feature representations, we employ the method shown in Figure 1. Multiple feature representations are combined with a common classifier, and the classification results with each feature representation are compared. Since our focus is on the feature learning, we use a single classification algorithm for all experiments (random forest [17]). This classifier was used successfully in combination with learned features in several audio classification works.[18], [19], [20]

We trained each feature extractor (CNN) using the feature representations calculated from the training set data. In feature extractor learning, each extractor uses the relevant time-frequency representations as input and their class labels of singing techniques as targets. The details of feature extractor learning are shown in Figure 2. The output of each extractor is denoted by a feature vector. Next, we trained random forest classifier models with 50 trees using feature vectors. Finally, we evaluated the classification performance of the test set. For the evaluation, we computed multiple accuracy metrics, as described in Section III-E.

C. Hand-crafted Features

We employ a 20-dimensional MFCC and two vibrato features (vibrato extension and vibrato rate) for the hand-crafted feature set. We used Librosa [21] for the MFCC calculations. For vibrato, the pitch contour was computed using CREPE [22] and input into Essentia [23] to calculate the vibrato features. To capture various pitch modulation, the ranges of vibrato thresholds are set to 2–10 [Hz] for the vibrato rate, and 10–200 [cents] for vibrato extent (i.e., vibrato depth). Each feature was averaged over all time lengths of an audio clip. A total of 22 dimensions of the hand-crafted features (20 for MFCCs and 2 for vibrato) were used. We denote this setting of features as Hand-crafted.

D. Learning Features

Although hand-crafted features do not require a learning process, the other representations require feature extractor learning (i.e., automatic extraction of feature vectors using neural networks). Figure 2 illustrates our supervised method for feature extractor learning, which was inspired by Abeßer et al. [15]. We compared four different types of settings: a raw waveform, STFTs, Mel-spectrograms, and a wavelet-scattering transform.

1) Raw waveforms: Under this condition, we feed a raw audio waveform to the network directly. Wilkins et al. [16] used a CNN model that inputs raw waveforms for singing technique classification. We use a 1D-CNN, which has three 1D-convolution blocks. We denote this setting as a Wave.

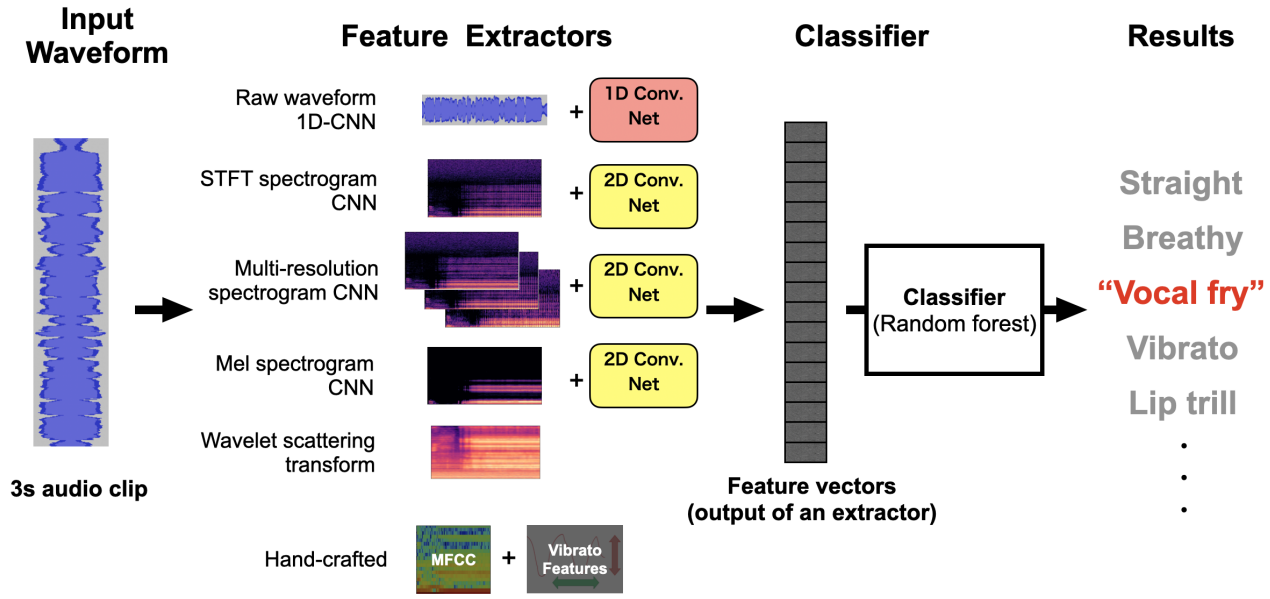

Fig. 1. Method of model comparison.

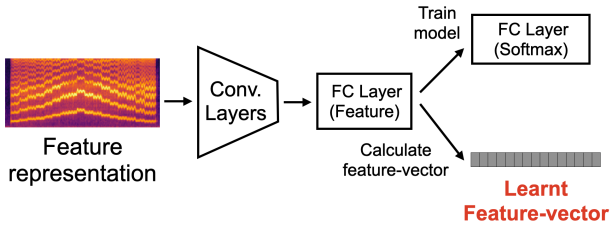

Fig. 2. Details of feature extractor learning.

TABLE II
CONFIGURATION OF STFT-BASED CNN (USED FOR BOTH STFT SPECTROGRAM CNN AND MULTI-RESOLUTION SPECTROGRAM CNN.) EACH CONVOLUTIONAL LAYER INCLUDES A BATCH NORMALIZATION, RELU ACTIVATION, AND DROPOUT (0.3).

Layer	Configuration
Conv1	Convolution (1×4), MaxPool (4×4)
Conv2	Convolution (1×16), MaxPool (4×4)
Conv3	Convolution (4×1), MaxPool (3×3)
Conv4	Convolution (16×1), MaxPool (2×2)
Flatten	
FC	512
FC (Feature)	22
FC (Softmax)	10

2) STFT magnitude spectrograms: Spectrograms using STFT are the most basic time-frequency representation. We calculated the magnitude spectrograms by applying an STFT with a Hann window with a length of 2048 and a hop size of 512. As a result, each spectrogram had 1024 frequency bins and 259 timeframes.

Takahashi et al. solved musical instrument classification using magnitude spectrograms as input for a CNN [24]. We modified their model for our spectrogram-based feature extractor to accommodate a longer signal duplication, as shown in Table II. We denote this setting as STFT. In addition, we investigated multi-resolution spectrograms [25] to capture time-frequency modulation patterns more accurately. We obtain a multi-resolution spectrogram by stacking three spectrograms with different time-frequency resolutions along the channel dimension. To maintain the same size for all spectrograms with different time-frequency resolutions, we applied zero padding while fixing the hop size. We have two conditions in this category, which we denote as Multi-1, having window sizes of (2048, 1024, 512), and Multi-2, with window sizes of (2048, 512, 128).

3) Mel spectrograms: Mel spectrograms are the most common time-frequency representations in many audio classification tasks using a DNN. Mel spectrograms were calculated by feeding magnitude spectrograms into a mel filterbank. We used 128-dimensional mel spectrograms derived from STFT-spectrograms computed with a Hann window of 2048 samples (25% overlap) at 44.1 kHz. We denote this setting as MelSpec.

4) Wavelet scattering transform: Under this condition, a wavelet scattering transform replaces the steps of "conversion into a representation" and "convolutional layers" in Figure 2. A wavelet scattering transform is a cascade of wavelet filter banks, applying a non-linearity operation (i.e., taking absolute values) after each convolution. The structure of the wavelet scattering transform is similar to that of a CNN. However, their weights are hand-crafted to encode prior knowledge of the task at hand.

TABLE III
RESULTS OF CLASSIFICATION ACCURACY.

Methods	Balanced	Accuracy	Top-2	Top-3
Hand-crafted	0.525	0.669	0.796	0.885
MelSpec	0.636	0.728	0.893	0.953
Scattering	0.668	0.754	0.894	0.947
STFT	0.713	0.770	0.920	0.946
Multi-1	0.719	0.770	**0.922**	0.964
Multi-2	**0.727**	**0.778**	0.917	**0.966**
Wave	0.589	0.684	0.849	0.927

We use Kymatio [26] for computing the wavelet scattering transform. We use first- and second-order scattering coefficients, which are the outputs of the wavelet scattering transform, as input feature representations for the FC layer. We denote this setting as Scattering. For a wavelet scattering transform only, an input signal must be a power of 2. Therefore, we set the input length to $T = 2^{17}$, which corresponds to approximately 2.97 s, which is roughly similar to 3 s for the other conditions.

E. Evaluation Metrics

We evaluated each model using four metrics: *balanced accuracy, accuracy, top-2 accuracy,* and *top-3 accuracy*. The number of samples in each class of VocalSet was imbalanced, as shown in Table I. Therefore, in addition to the normal accuracy, an evaluation using a balanced accuracy [27] was conducted. We also evaluated the *class-wise F1-score* to investigate the characteristics of each method.

For each condition, we repeat the experiment 5 times with different data splits, and calculated the mean and standard error of the above metrics. The accuracy values reported in the next section are the means of repeated measurement.

IV. EXPERIMENTS AND RESULTS

A. Experiment 1: A comparison of feature representations with fixed dimensions

First, we compared the performances of all feature representation settings under the fixed dimension size, i.e., using the feature vector of length 22. The results of Experiment 1 are shown in Table III and Figure 3. STFT-based models (STFT, Multi-1, and Multi-2) outperformed the other models. These STFT-based models performed particularly well in breathiness-related techniques such as breathing and vocal fry.

In addition, we visualized the feature vectors obtained by the hand-crafted feature and STFT-based methods (STFT and Multi-2). The number of dimensions of the feature vectors was compressed from 22 to 2 using t-distributed stochastic neighbor embedding (t-SNE) [28], and the 10 classes were visualized by highlighting them with color. Feature vectors obtained using STFT-based methods of the same class are mapped more closely to each other than those of the hand-crafted condition.

B. Experiment 2: Ablation study

We further investigated the combination of feature representation and different types of CNNs to determine the critical factors in the classification performance. To conduct this ablation study, there are two factors: the CNN architecture and time-frequency representation.

The best-performing STFT-based models have a unique architecture that differs from the standard CNN. The convolutional layers of our model are oblong, that is, the kernel length for one axis (e.g., time) is longer than that for another axis (e.g., frequency). By contrast, under the MelSpec condition, we used kernels with a square shape (3×3), which is the standard architecture for CNN-based image processing. We therefore compare all combinations of the selected input feature representations (MelSpec, STFT, and Multi-2) and CNNs. For the sake of simplicity, we denote two different types of CNN as follows: square (a CNN model in which all convolutional layers have square kernels) and oblong (a CNN model in which each convolutional layer has a length along only one axis). The configurations of these kernel shapes are listed in Table V. The results of Experiment 2 are shown in Table VI and Figure 5.

C. Experiment 3: Changing the feature vector dimension

We further investigated the performance by increasing the dimensions of the feature vectors by changing the output size of the FC layer. We examined four types of dimension sizes (i.e., 22, 44, 88, and 200) under the Multi-2 condition, which performed best in Experiment 1. The results are shown in Table VII. Increasing the size of the features does not improve the score, but instead slightly lowers the accuracy.

V. DISCUSSION

A. Singer-wise split

As a follow-up study for Experiment 1, we also tried singer-wise split, and the results are shown in Table VIII. In this split condition, the dataset was split into a training set from 15 singers and a test set 5 singers during the learning process. The results suggest that STFT-based model also outperformed the other models in the condition as well as clip-wise split.

B. Effectiveness of the proposed framework

In Experiment 1, we confirmed that STFT-based methods performed well, particularly in classifying breathiness-related singing techniques. In mel-filterbank-based representations, the contrast between the harmonic components and other noisy components becomes unclear, and the pitch contour becomes ambiguous owing to the low resolution within the frequency domain. Meanwhile, STFT-based representations maintain a clear contrast between the spectral peaks and noisy components, enabling the detection of noisy parts and a fine-scale pitch modulation. We assume that this is the reason why STFT-based representations outperformed the Mel-based representations.

In Experiment 2, we demonstrated the effectiveness of the CNN model with a convolution kernel with oblong shapes. There are many potential combinations of convolutional kernel shapes for each layer of a CNN. In fact, there are some cases in which the performance is improved by changing the shape

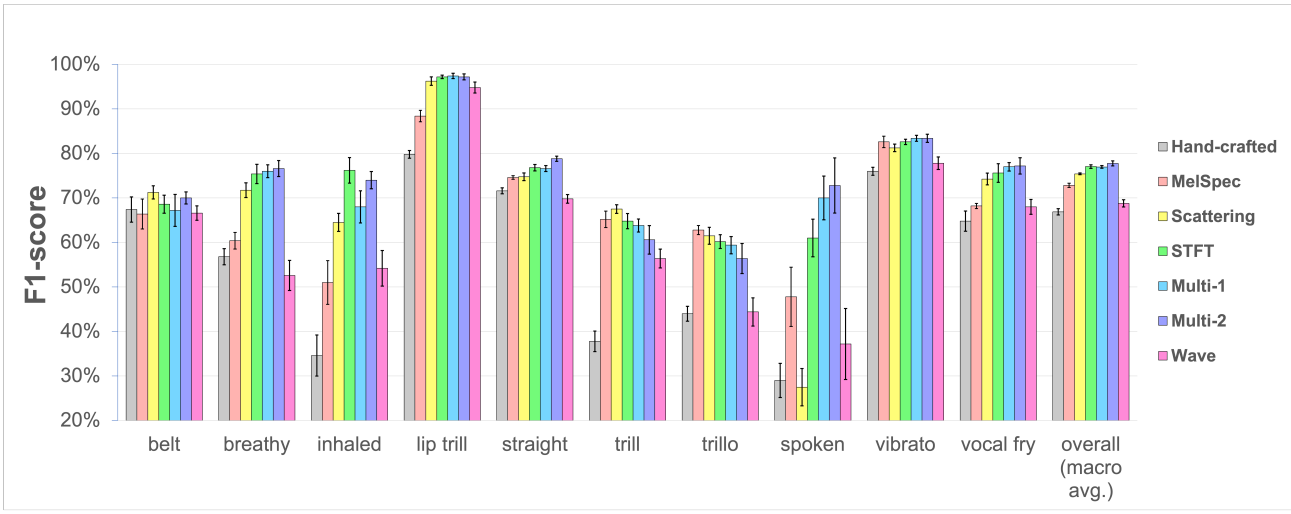

Fig. 3. Plot of class-wise F1 scores. Error bars show the standard error.

TABLE IV
CLASS-WISE F1-SCORES.(CLIP-WISE SPLIT)

Methods	Belt	Breathy	Inhaled	Lip trill	Straight	Trill	Trillo	Spoken	Vibrato	Vocal fry	Overall (macro average)
Hand-Crafted	0.674	0.568	0.346	0.798	0.716	0.378	0.440	0.290	0.760	0.648	0.669
MelSpec	0.664	0.604	0.510	0.884	0.746	0.652	**0.628**	0.478	0.826	0.682	0.728
Scattering	**0.713**	0.718	0.645	0.963	0.748	**0.675**	0.615	0.275	0.813	0.743	0.754
STFT	0.686	0.754	**0.762**	0.972	0.768	0.648	0.602	0.610	0.826	0.756	0.770
Multi-1	0.672	0.760	0.680	**0.974**	0.766	0.638	0.594	0.700	**0.834**	0.770	0.770
Multi-2	0.700	**0.766**	0.740	0.972	**0.788**	0.606	0.564	**0.728**	**0.834**	**0.772**	**0.778**
Wave	0.663	0.528	0.548	0.938	0.698	0.548	0.420	0.403	0.773	0.680	0.684

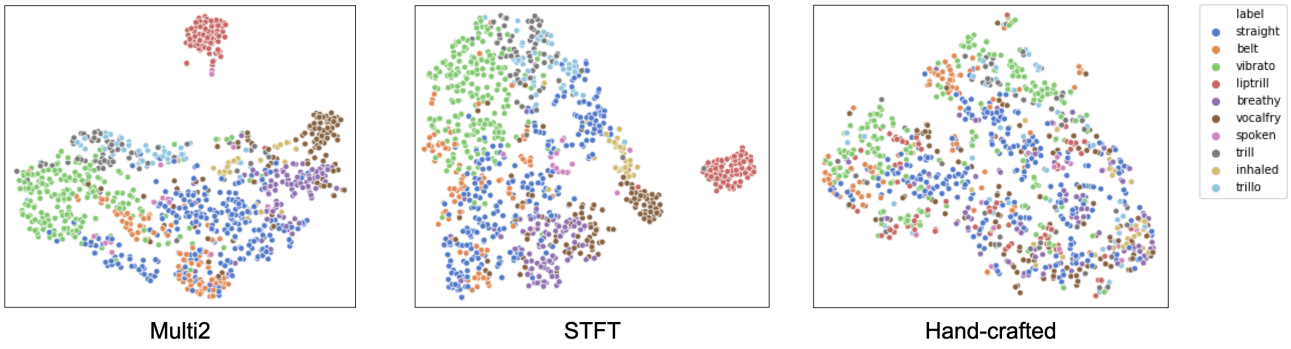

Fig. 4. Visualization of feature vector derived Multi-2 (left), STFT (center), and Hand-crafted (right).

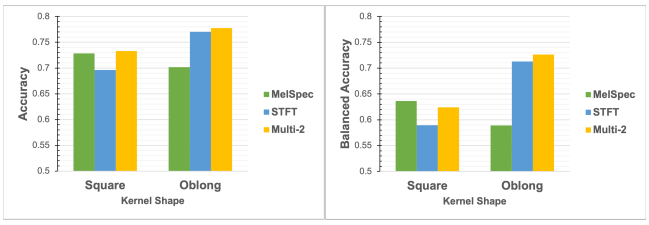

Fig. 5. Accuracy and balanced accuracy of Experiment 2.

TABLE V
SHAPE OF CONVOLUTIONAL KERNEL UNDER EACH CONDITION. THE FOUR CONVOLUTIONAL LAYERS ARE NUMBERED IN ASCENDING ORDER (CONV 1 TO 4) FROM THE INPUT LAYER.

	Conv 1	Conv 2	Conv 3	Conv 4
Square	(3×3)	(3×3)	(3×3)	(3×3)
Oblong	(1×4)	(1×16)	(4×1)	(16×1)

of the kernel [29] for automatic music tagging tasks. The shape of the kernel needs to be further studied.

TABLE VI
RESULTS OF EXPERIMENT 2.

Kernel shape	Feature	Balanced	Accuracy
Square	Multi-2	0.624	**0.733**
	STFT	0.589	0.696
	MelSpec	**0.636**	0.728
Oblong	Multi-2	**0.727**	**0.778**
	STFT	0.713	0.770
	MelSpec	0.589	0.702

TABLE VII
ACCURACY METRICS WHEN THE DIMENSION SIZE OF THE FEATURE VECTORS VARIES.

Dimension size	Balanced	Accuracy
22	**0.727**	**0.778**
44	0.713	0.770
88	0.711	0.771
200	0.716	0.773

TABLE VIII
RESULTS OF EXPERIMENT 1 UNDER THE CONDITION OF SINGER-SPLIT.

Methods	Balanced	Accuracy	Top-2	Top-3
Hand-crafted	0.377	0.513	0.706	0.803
MelSpec	0.488	0.556	0.754	0.846
Scattering	0.422	0.439	0.644	0.776
STFT	0.597	0.606	**0.803**	0.891
Multi-1	**0.605**	**0.617**	0.799	**0.898**
Multi-2	0.535	0.553	0.730	0.843
Wave	0.511	0.581	0.761	0.876

In another direction, further investigation into feature learning methods such as recurrent neural networks (RNNs), convolutional recurrent neural networks (CRNNs), or joint time–frequency scattering [12] would be interesting.

Finally, the recordings in VocalSet are monophonic, unlike many commercially distributed songs. To analyze popular songs using the methods considered in this paper, a voice separation must be applied before the singing technique classification is applied. Under such situations, singing technique classifications may decrease the accuracy. Considering that VocalSet is the only publicly available dataset that annotates singing techniques, the development of another dataset that consists of polyphonic signals (i.e., songs and accompaniments), with annotation of the singing technique, can contribute to further development of the field.

VI. CONCLUSION

This study provides an investigation into audio feature representations for singing technique classification. We compared hand-crafted features and CNN-based feature learning methods applied to various time-frequency representations. Our findings show that features learned from low-level representations, such as spectrograms, outperformed hand-crafted features based on expert knowledge. In particular, multi-resolution spectrograms performed best, with an accuracy of 77.8% in clip-wise split and 61.7% in singer-wise split. Presumably, this is due to their high ability to capture breathiness and a small modulation. We also confirmed the effectiveness of the combination of an STFT-based input feature representation and a CNN that convolves along the time and frequency axes separately. Extending this approach, we plan to further investigate the design of feature learning, e.g., changing the convolution kernel shapes of a CNN, or applying different DNN-based methods.

REFERENCES

[1] H. Purwins, B. Li, T. Virtanen, J. Schlüter, S.-Y. Chang, and T. Sainath, "Deep learning for audio signal processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 2, pp. 206–219, 2019.

[2] A. Mesaros and J. Astola, "The mel-frequency cepstral coefficients in the context of singer identification." 2005, pp. 610–613.

[3] J. Schwenninger, R. Brueckner, D. Willett, and M. E. Hennecke, "Language identification in vocal music." in *The 7th International conference for Music Information Retrieval Conference (ISMIR)*, 2006, pp. 377–379.

[4] B. Schuller, C. Kozielski, F. Weninger, F. Eyben, and G. Rigoll, "Vocalist gender recognition in recorded popular music," in *Proceedings of the 11th International Society for Music Information Retrieval Conference, ISMIR 2010 (ISMIR)*, 2010, pp. 613–618.

[5] 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.

[6] N. Kroher and E. Gómez, "Automatic singer identification for improvisational styles based on vibrato, timbre and statistical performance descriptors," in *In Proceedings of the ICMC SMC 2014 Joint Conference (ICMC SMC 2014).*, 2014.

[7] L. Su, L.-F. Yu, and Y.-H. Yang, "Sparse cepstral, phase codes for guitar playing technique classification." in *The 15th International Society for Music Information Retrieval Conference (ISMIR)*, 2014.

[8] A. B. Kruger and J. P. Jacobs, "Playing technique classification for bowed string instruments from raw audio," *Journal of New Music Research*, vol. 49, no. 4, pp. 320–333, 2020. [Online]. Available: https://doi.org/10.1080/09298215.2020.1784957

[9] J. Charles, "Playing technique and violin timbre: Detecting bad playing," Ph.D. dissertation, Ph.D.dissertation, Technological Univ. Dublin, Ireland. 2010, 2010.

[10] B. Liang, G. Fazekas, and M. Sandler, "Piano sustain-pedal detection using convolutional neural networks," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE, 2019, pp. 241–245.

[11] C. Wang, E. Benetos, V. Lostanlen, and E. Chew, "Adaptive time–frequency scattering for periodic modulation recognition in music signals," in *The 20th International Society for Music Information Retrieval Conference (ISMIR)*, 2019.

[12] C. Wang, V. Lostanlen, E. Benetos, and E. Chew, "Playing technique recognition by joint time–frequency scattering," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE, 2020, pp. 881–885.

[13] Y.-F. Huang, J.-I. Liang, I.-C. Wei, and L. Su, "Joint analysis of mode and playing technique in guqin performance with machine learning," in *The 21th International Society for Music Information Retrieval Conference (ISMIR)*, 2020.

[14] V. Lostanlen, J. Andén, and M. Lagrange, "Extended playing techniques: the next milestone in musical instrument recognition," in *Proceedings of the 5th International Conference on Digital Libraries for Musicology (DLfM)*, 2018, pp. 1–10.

[15] J. Abeßer and M. Müller, "Fundamental frequency contour classification: A comparison between hand-crafted and cnn-based features," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE, 2019, pp. 486–490.

[16] J. Wilkins, P. Seetharaman, A. Wahl, and B. A. Pardo, "Vocalset: A singing voice dataset," in *19th International Society for Music Information Retrieval Conference, (ISMIR)*. International Society for Music Information Retrieval, 2018, pp. 468–474.

[17] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[18] J. Salamon and J. P. Bello, "Unsupervised feature learning for urban sound classification," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 171–175.

[19] D. Stowell and M. D. Plumbley, "Automatic large-scale classification of bird sounds is strongly improved by unsupervised feature learning," *PeerJ*, vol. 2, p. e488, 2014.

[20] I.-Y. Jeong and K. Lee, "Learning temporal features using a deep neural network and its application to music genre classification." in *17th International Society for Music Information Retrieval Conference, (ISMIR)*, 2016, pp. 434–440.

[21] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python," in *Proceedings of the 14th python in science conference*, vol. 8, 2015, pp. 18–25.

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

[23] D. Bogdanov, N. Wack, E. Gómez Gutiérrez, S. Gulati, H. Boyer, O. Mayor, G. Roma Trepat, J. Salamon, J. R. Zapata González, X. Serra *et al.*, "Essentia: An audio analysis library for music information retrieval." International Society for Music Information Retrieval (ISMIR), 2013.

[24] T. Takahashi, S. Fukayama, and M. Goto, "Instrudive: A music visualization system based on automatically recognized instrumentation." in *19th International Society for Music Information Retrieval Conference, (ISMIR)*, 2018, pp. 561–568.

[25] K. Dressler, "Sinusoidal extraction using an efficient implementation of a multi-resolution fft," in *of the 9th International Conference on Digital Audio Effects (DAFx)*. Citeseer, 2006, p. 247.

[26] M. Andreux, T. Angles, G. Exarchakis, R. Leonarduzzi, G. Rochette, L. Thiry, J. Zarka, S. Mallat, J. Andén, E. Belilovsky *et al.*, "Kymatio: Scattering transforms in python." *Journal of Machine Learning Research*, vol. 21, no. 60, pp. 1–6, 2020.

[27] K. H. Brodersen, C. S. Ong, K. E. Stephan, and J. M. Buhmann, "The balanced accuracy and its posterior distribution," in *2010 20th international conference on pattern recognition*. IEEE, 2010, pp. 3121–3124.

[28] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." *Journal of machine learning research*, vol. 9, no. 11, 2008.

[29] J. Pons, T. Lidy, and X. Serra, "Experimenting with musically motivated convolutional neural networks," in *2016 14th International Workshop on Content-Based Multimedia Indexing (CBMI)*, 2016, pp. 1–6.