Self-Supervised Contrastive Learning for Singing Voices

Hiromu Yakura, Kento Watanabe, and Masataka Goto

***Abstract*—This study introduces self-supervised contrastive learning to acquire feature representations of singing voices. To acquire robust representations in an unsupervised manner, regular self-supervised contrastive learning trains neural networks to make the feature representation of a sample close to those of its computationally transformed versions. Similarly, we employ two transformations—pitch shifting and time stretching—considering the nature of singing voices. Nevertheless, we use them reversely: we train networks to push away representations of the transformed versions. The networks then attempt to discriminate changes in vocal timbres introduced by pitch shifting without time stretching and those in singing expressions introduced by time stretching without pitch shifting. Consequently, the acquired representations become attentive to vocal timbre and singing expression. This was confirmed through a singer identification task, where we trained a classifier to learn the relationship between the feature representations to the corresponding singer labels of 500 singers. As a result, the employed transformations helped the classifier improve the classification accuracy by 9.12% (top-1 accuracy: 63.08%) compared with the case where the feature representations fed to the classifier were acquired without the transformations (top-1 accuracy: 53.96%). Furthermore, the proposed approach can be extended to acquire feature representations attentive to either vocal timbre or singing expression but not to the other by changing how the transformations are incorporated. We particularly explored the characteristics of such vocal timbre- or singing expression-oriented feature representations against song genre, singer gender, and vocal technique, and confirmed that they successfully capture different aspects of singing voices.**

***Index Terms*—Singer identification, self-supervised contrastive learning, singing information processing, representation learning.**

I. Introduction

FEATURE representation learning is an important task in music information retrieval (MIR) [1], [2] because the acquired representations can be used in various applications, e.g., music recommendation [3], [4]. However, learning methods dedicated to singing voices have not been investigated extensively, although singing voices are one of the most salient components of people's musical tastes [5]. This can be attributed to the demand for a large-scale dataset from deep learning techniques and the difficulties related to preparing such a dataset, particularly when corresponding annotations are required.

This study introduces self-supervised contrastive learning to realize feature representation learning in a manner specialized to singing voices separated from polyphonic music (Fig. 1). In image and video processing, self-supervised contrastive learning [6], [7] has improved the performance of various tasks without requiring annotated datasets. Here, the core concept is training neural networks to make the feature representation of a sample close to those of its transformed versions while pushing representations from different samples away. The networks then acquire representations that are robust to such transformations. For example, existing methods in image processing frequently involve noise addition, color distortion, and flipping [7].

The proposed approach involves transformations that reflect the nature of singing voices. Importantly, we apply them backwardly, i.e., we train networks to make the feature representation of the original sample and those of its transformed versions dissimilar. Then, the networks are expected to acquire

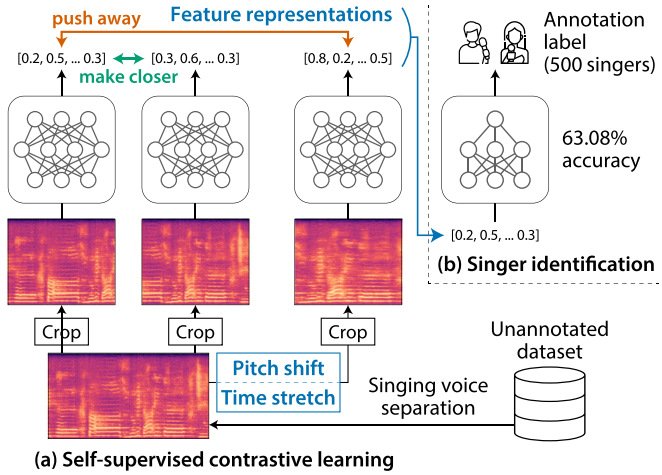

Fig. 1. (a) We introduce self-supervised contrastive learning for singing voices.[1] (b) Acquired feature representations capture the characteristics of each singer so effectively that they can enable a shallow classifier to achieve high accuracy in singer identification.

Manuscript received October 15, 2021; revised February 25, 2022; accepted April 1, 2022. Date of publication April 26, 2022; date of current version May 6, 2022. This work was supported in part by JST ACT-X under Grant JPMJAX200R, in part by JST CREST under Grant JPMJCR20D4, and in part by JSPS KAKENHI under Grants JP21J20353 and JP21H04917. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Stefan Bilbao. *(Corresponding author: Hiromu Yakura.)*

Hiromu Yakura is with the Graduate School of Science and Technology, University of Tsukuba, Ibaraki 305-8574, Japan, and also with the National Institute of Advanced Industrial Science and Technology (AIST), Ibaraki 305-8568, Japan (e-mail: hiromu.yakura@aist.go.jp).

Kento Watanabe and Masataka Goto are with the National Institute of Advanced Industrial Science and Technology (AIST), Ibaraki 305-8568, Japan (e-mail: kento.watanabe@aist.go.jp; m.goto@aist.go.jp).

Digital Object Identifier 10.1109/TASLP.2022.3169627

[1]Our source code is available at https://github.com/hiromu/contrastive-singing-voices

representations that are sensitive to specific transformations. Here, we consider two transformations, i.e., pitch shifting and time stretching. Naïve pitch shifting without time stretching, which is achieved by time-domain resampling operation followed by time-stretching operation, can change both pitch and formants [8]. Since formants are the core component of vocal timbre [9], discriminating pitch-shifted versions would yield representations attentive to vocal timbre. Additionally, time stretching without pitch shifting affects singing expression by modifying expressive articulations, e.g., vibrato rate and fundamental frequency (F0) contour in note transitions [10]. Thus, discriminating time-stretched versions would yield representations attentive to singing expression.

The proposed approach can be used not only for acquiring singer-specific feature representations that are attentive to both vocal timbre and singing expression. When we train networks to make representations of the pitch-shifted versions close to that of the original sample while pushing representations of the time-stretched versions away, we can acquire representations that are attentive to singing expression but inattentive (i.e., robust) to vocal timbre. This broadens the applicability of the acquired feature representations. For example, without using annotated data, we can retrieve singing dissimilar in terms of vocal timbre but similar in terms of singing expression, which would be difficult to realize using conventional features, such as Mel-frequency cepstral coefficients (MFCCs).

In this study, we first examined the effectiveness of the proposed approach, which is based on self-supervised contrastive learning, by applying the acquired feature representations to singer identification. We trained a classifier to learn the relationship between the acquired representations and singer labels and achieved an accuracy of 63.08% with 500 singers, 9.12% of which was attributed to the introduced transformations. Then, we explored feature representations trained to be attentive to either vocal timbre or singing expression. The results of our analysis suggest that, even without annotated data, we can retrieve singing voices that are similar in terms of only vocal timbre or singing expression. These results indicate that the proposed approach can be a powerful tool for developing new MIR applications.

II. Related Work

A. Representation Learning for Singing Voices

We share part of our motivation with singer identification because it is one of the downstream tasks that can leverage feature representations acquired by the proposed approach. Singer identification is a central topic in singing information processing [11], [12]; thus, it has been the focus of several previous studies [13]–[23]. Typically, traditional approaches [13]–[18] have trained classifiers on MFCCs or linear prediction coefficients (LPCs) to capture vocal timbre. Some methods [19]–[21] have attempted to capture singing expressions by designing handcrafted features that can capture vibrato in singing voices. In addition, recent methods [22], [23] leveraged deep learning techniques to classify singing voices from their spectrograms without relying on handcrafted features. However, these methods were explicitly designed for singer identification and not assumed for similarity computations, limiting their use in other MIR applications.

There are relatively few methods that allow similarity computation of singing voices. Early attempts [24]–[26] applied Gaussian mixture models or latent Dirichlet allocation on MFCCs or LPCs. Whereas these were unsupervised, their dependency on MFCCs or LPCs made the computed similarities only reflect vocal timbre. Some recent methods [27], [28] enable similarity computation using feature representation learning in a similar manner to us. The adoption of contrastive learning is common to the proposed approach; however, these methods were not self-supervised. They employed supervised training to make feature representations of the singing voices of the same singer close to each other while pushing those of different singers away. By contrast, the proposed approach allows us to acquire feature representations that can be used for similarity computations without depending on annotated datasets because it computationally transforms samples to be self-supervised.

Here, we acknowledge that this study can be associated with speaker identification, which has been actively researched for many years [29], [30]. Particularly, i-vector and x-vector are known to capture the characteristics of individual speakers [31], [32], and indeed, some existing methods [33], [34] have leveraged them for singer identification. On the other hand, Xia *et al.* [35] showed that their speaker identification method employing self-supervised contrastive learning outperformed i-vector and x-vector. Analogously, we can expect that the introduction of self-supervised contrastive learning for singing voices allows us to acquire feature representations that effectively capture the characteristics of each singing voice, whereas it is under researched.

B. Self-Supervised Contrastive Learning

Given this context, we explain the scheme of self-supervised contrastive learning. As discussed in Section I, self-supervised contrastive learning [36], [37] is a powerful paradigm that enables feature representation learning without annotated datasets. As shown in Fig. 2(a), it introduces both positive and negative anchors. Here, positive anchors are generated by applying computational transformations to a sample in a dataset, and negative anchors refer to other samples in the dataset. Self-supervised contrastive learning yields robust representations that can distinguish different samples in a dataset by training networks to make the feature representation of a sample similar and dissimilar to those of its positive and negative anchors, respectively. The acquired representations are known to bring performance improvement in various tasks, e.g., image classification and object recognition in the image domain and action recognition in the video domain [6], [7].

Let us assume that we are trying to acquire such feature representations by training a network $f_\theta(\cdot)$ using a dataset $X = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N\}$. We also use K^+ positive anchors $X_i^+ = \{\boldsymbol{x}_{i,1}^+, \boldsymbol{x}_{i,2}^+, \ldots, \boldsymbol{x}_{i,K^+}^+\}$ and K^- negative anchors $X_i^- =$

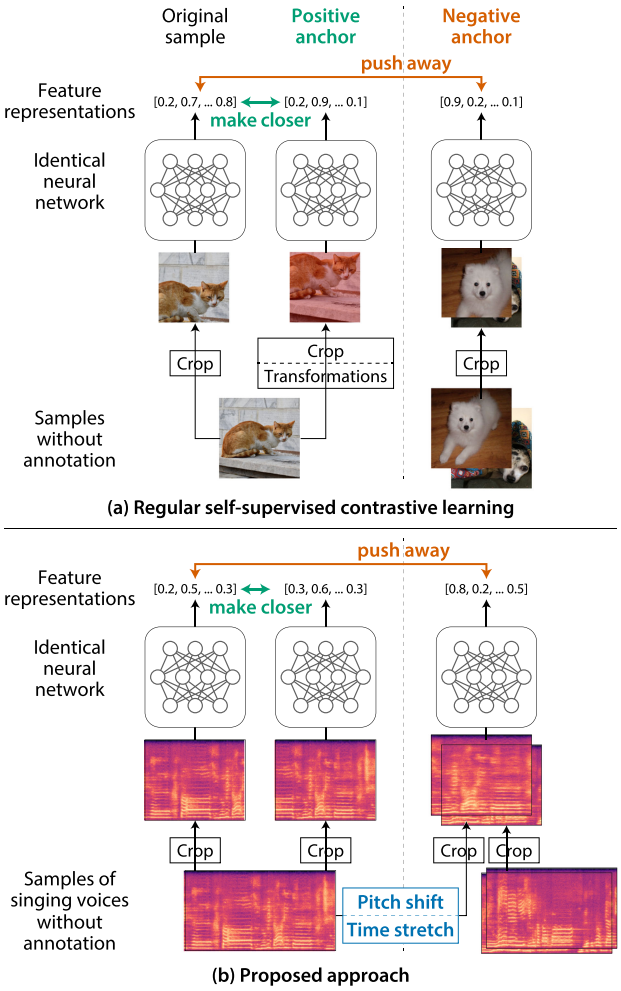

Fig. 2. (a) Regular self-supervised contrastive learning framework transforms the original sample to generate positive anchors to acquire robust feature representations. (b) The proposed approach employs pitch shifting and time stretching to generate negative anchors to make the acquired feature representations attentive.

$\{\boldsymbol{x}_{i,1}^-, \boldsymbol{x}_{i,2}^-, \ldots, \boldsymbol{x}_{i,K^-}^-\}$ prepared for each sample $\boldsymbol{x}_i$. Then, regular self-supervised contrastive learning framework trains the network via optimizing the parameter of the network θ so as to minimize the following loss function:

$$\mathcal{L}(\theta) = -\sum_{i=1}^{N}\sum_{j=1}^{K^+} \log \frac{\hat{\mathcal{L}}_\theta\left(\boldsymbol{x}_i, \boldsymbol{x}_{i,j}^+\right)}{\hat{\mathcal{L}}_\theta\left(\boldsymbol{x}_i, \boldsymbol{x}_{i,j}^+\right) + \sum_{k=1}^{K^-} \hat{\mathcal{L}}_\theta\left(\boldsymbol{x}_i, \boldsymbol{x}_{i,k}^-\right)}$$

$$\hat{\mathcal{L}}_\theta(\boldsymbol{a}, \boldsymbol{b}) = \exp\left(\frac{f_\theta(\boldsymbol{a}) \cdot f_\theta(\boldsymbol{b})}{\tau}\right) \quad (1)$$

Here, τ is a hyperparameter that controls the concentration level of the feature representations [37], [38]. As mentioned above, the positive anchors X_i^+ can be prepared by applying various computational transformations to $\boldsymbol{x}_i$, while the negative anchors X_i^- are commonly supplied with $X \setminus \{\boldsymbol{x}_i\}$ by setting $K^- = N - 1$.

Based on this formulation, various techniques have been proposed to enhance the effectiveness or efficiency of self-supervised contrastive learning. For example, SimCLR [36] generates positive anchors stochastically instead of preparing a fixed number of positive anchors a priori. Specifically, SimCLR sets $K^+ = 1$ and dynamically generates $\boldsymbol{x}_{i,1}^+$ iteratively during training through applying transformations to $\boldsymbol{x}_i$ by varying their parameters. Consequently, the diversity of the positive anchors increases, and the acquired feature representations become more robust. Then, MoCo [37] introduced a momentum encoder to improve the computational and memory efficiency of the training.

C. Self-Supervised Contrastive Learning in MIR

Despite its effectiveness, the use of self-supervised contrastive learning in MIR remains limited [39]–[41]. For example, Spijkervet and Burgoyne [41] proposed CLMR, extending SimCLR [36] to generate positive anchors by adding Gaussian noise, applying frequency filters, etc. They reported that CLMR achieved state-of-the-art performance in semantic music tagging. However, such methods did not consider the application to singing voices and were not designed to reflect multiple aspects, whereas the proposed approach considers vocal timbre- or singing expression-oriented representations.

For the purpose of acquiring feature representations of human speech, several methods leveraging self-supervised contrastive learning [42], [43] have been proposed. While they exhibited substantial performance improvement when the feature representations were applied to speech recognition, they were not designed to capture the identity of speakers. Hence, it is unclear whether they can be applied to singer identification or speaker identification. The same can be said for other methods [44]–[46] that attempt to acquire general feature representations of various auditory events, e.g., speech commands and instrumental sounds. Therefore, in this study, we explored how to exploit self-supervised contrastive learning, especially in a manner specialized to singing voices.

III. Self-Supervised Contrastive Learning for Singing Voices

A. Proposed Approach

In the proposed approach, we apply self-supervised contrastive learning, specifically MoCo [37] because of its performance and memory efficiency, by introducing transformations that reflect the nature of singing voices. As in the regular framework of self-supervised contrastive learning (Fig. 2(a)), MoCo trains networks to make the feature representation of a sample similar to those of its positive anchors, i.e., its transformed versions, in terms of their cosine similarity. In the proposed approach, we modified MoCo to use our introduced transformations to generate negative anchors (rather than positive anchors) and make their representation dissimilar to that of the original sample (Fig. 2(b)). Consequently, we can acquire feature representations sensitive to changes caused by the transformations.

A similar approach was introduced in video processing by Tao *et al.* [47], who reported that using shuffled videos, where video frames are shuffled in temporal order, as negative anchors yielded feature representations that achieve high action

recognition performance. In fact, we attempted shuffling singing voices temporally to generate negative anchors and examined the acquired representations. However, we found that the temporal shuffling produced unnatural artifacts in audio signals, e.g., noncontinuous changes, and resulted in representations that are merely sensitive to the artifacts. This taught us that we should generate negative anchors that are likely to appear in a dataset as a natural singing voice but make us feel like other singers sang them. In this sense, we introduce two transformations, i.e., pitch shifting and time stretching.

These transformations allow trained neural networks to distinguish vocal timbres and singing expressions, respectively. The vocal timbre depends on spectral envelopes and formants [9], which are altered substantially after naïve pitch shifting[2](without time stretching); hence, pitch-shifted singing voices tend to be perceived as having different vocal timbres. Therefore, training networks to distinguish pitch-shifted versions (i.e., adding pitch-shifted singing voices as negative anchors) leads to feature representations that are sensitive to the spectral changes in vocal timbres.

Time stretching yields singing voices that could sound like they have different singing expressions produced by different singers. Here, we focus on singing expressions related to short-term expressive articulations, e.g., the vibrato rate and F0 contour in note transitions. The vibrato is one of the clear expressive features in singing voices [12], and its rate is known to be fairly constant regardless of a song's tempo [49]. The steepness of the F0 contour in note transitions is also a cue to capture singing expressions because it reflects personal singing behaviors or styles [50], [51]. Such temporal expressions are altered substantially after time stretching (without pitch shifting). Thus, training networks to distinguish time-stretched versions (i.e., adding time-stretched singing voices as negative anchors) leads to feature representations sensitive to the temporal changes in singing expressions.

Here, the duration of singing voices can be varied; thus, we assume that they were cut down to the same duration before being input to networks. This assumption provides a means to prepare positive anchors; i.e., we use another cut from the same sample as a positive anchor because the cuts from the same sample are considered sung by the identical singer. Thus, their feature representations should be close to each other.

In summary, given the notation in Equation 1, we can see that the proposed approach is augmenting the negative anchors X_i^- by using the introduced transformations. Specifically, it complements X_i^-, consisting of other samples in a dataset analogously to regular self-supervised contrastive learning, with the pitch-shifted and time-stretched versions of $\boldsymbol{x}_i$. Meanwhile, the positive anchors X_i^+ were supplied with $\tilde{\boldsymbol{x}}_i$, which was cut from the same singing voice as $\boldsymbol{x}_i$. Here, as we explain later in Section III-C, the positive and negative anchors can be generated stochastically (i.e., via random sampling[3]) to increase their diversity, inheriting the design of SimCLR [36] and MoCo [37] (see Section II-B). With this scheme, the proposed approach can acquire singer-specific feature representations attentive to both vocal timbre and singing expression.

[2]In other words, for pitch shifting, we should not apply existing methods that can modify pitch while preserving formants, such as TD-PSOLA [48].

[3] Although samples (i.e., singing voices) by the same singer might be selected randomly as negative anchors, this effect is usually ignored in contrastive learning since its chance is rare.

B. Data

To realize the proposed approach, we must prepare a large dataset of singing voices. Particularly, self-supervised learning is known to demonstrate performance improvement in various tasks when a network is trained with large amounts of data [52]. Thus, inspired by previous studies [41], [53] that crawled 30 s audio previews of 240 k songs corresponding to Million Song Dataset [54] from a music streaming service, we constructed a larger dataset of 30 s previews (audio excerpts for trial listening) of songs as well as their genre labels by crawling them from a music streaming service.

The songs were collected with their artist names; however, the artist names and any other metadata were not used for the self-supervised learning of the proposed approach. Instead, using the artist names, we randomly sampled 500 artists from those with more than 50 songs in the dataset and filtered them out for the singer identification task in Section IV. As a result, we used 328,418 songs, in which all songs by the 500 sampled artists were excluded, without the artist names for our self-supervised training.

We applied Spleeter [55] to all songs to separate the singing voices from the previews (polyphonic music). We also applied singing voice detection [56] and cut consecutive silence sections whose duration was greater than 0.5 s.

C. Implementation and Training Procedure

As explained in Section III-A, the constructed dataset was used to train neural networks in an unsupervised manner. For the networks to be trained, we adopted the architecture of convolutional recurrent neural network (CRNN) [57], which has been used for singer identification in a supervised situation [23]. We implemented it using PyTorch to take a spectrogram of a 5 s audio excerpt as input and output a 256-dimensional feature representation vector. To ensure the diversity of the input, we did not cut the singing voices in the dataset into 5 s a priori. Instead, we implemented it to crop each sample of the singing voices randomly during training.

For the transformations to be applied, we prepared four types of transformation: raising the pitch by three semitones, lowering the pitch by three semitones, speeding up by $\times 1.70$, and slowing down by $\times 0.65$. As explained in Section III-A, the transformations should not be too aggressive because they would lead to feature representations that are merely sensitive to unnatural artifacts, whereas too subtle transformations would not yield effective anchors for contrastive learning. Given that, these parameters were decided on the basis of our preliminary observations to balance the naturalness of the transformed versions with the degree of distinction from the original sample.

Then, one of the four types was applied at the same probability to a cut of a singing voice cropped to generate negative anchors. Correspondingly, another cut cropped from the same sample was

used as positive anchors without applying any transformations, as mentioned in Section III-A. Here, we used the SoX utility[4] (version 14.4.2) via torchaudio[5] to implement the transformations. Specifically, we invoked SoX with the option of either `pitch +300` or `-300` for pitch shifting and either `tempo 1.70` or `0.65` for time stretching.

The training was performed for 150 epochs, which took approximately 2.5 days on a computer with four NVIDIA Tesla V100 GPUs. We also trained the same network by altering the combinations of transformations to be used (i.e., four combinations compromising both, either, or no transformation) for comparison. Adam [58] was used to optimize the network parameters with a learning rate of 0.01, which was updated by multiplying by 0.2 at 25, 50, 75, and 100 epochs. Also, we set the dictionary size of the momentum encoder, which is analogous to K^- under the scheme of MoCo [37], to 4096.[6] After this self-supervised training, the networks were frozen to be used only for extracting feature representations.

IV. Singer Identification

We first performed singer identification to evaluate the effectiveness of the feature representations acquired by the proposed approach. Here, we trained a classifier to identify the artist names from the acquired feature representations and compared its accuracy to when the same classifier trained using conventional features.

A. Data

For the singer identification task, we used the singing voices of the 500 artists we previously filtered (see Section III-B). We sampled 50 songs for each artist from the collected 30 s audio previews, yielding 25,000 songs in total. They were preprocessed by separating only the singing voices from the polyphonic music using the singing voice separation and detection technique in the same manner as Section III-B. Then, they were split into training, validation, and testing sets (40:5:5) for each artist.

B. Procedure

The singer identification task was performed in a supervised manner by training a very shallow classifier. It is common to use a shallow classifier in the evaluation of downstream tasks to examine the effectiveness of self-supervised contrastive learning, as we can see in SimCLR [36] and MoCo [37]. This is because, when even such a classifier performed better in the downstream tasks by taking the acquired feature representations as input, it can be considered that the representations capture the characteristics of samples well.

Specifically, we used a two-layer perceptron with 512 hidden units in its middle layer and softmax activation in its output layer to follow the experimental design of CLMR [41]. We trained the perceptron to learn the relationship between the acquired feature representations and the corresponding singer labels. This supervised training ran for 250 epochs, lasting approximately 2.5 h. We again used Adam [58] with a stepping learning rate, in which the learning rate is multiplied at certain epochs (see Section III-C), starting with 0.01.

TABLE I
Accuracy of Singer Identification (500 Singers)

Feature extraction	Top-1 acc.	Top-5 acc.
Conventional features (MFCC, etc.)	0.20%	1.00%
CLMR [41]	47.96%	73.64%
Proposed	53.96%	76.60%
Proposed + time stretching	61.00%	81.16%
Proposed + pitch shifting	61.28%	81.72%
Proposed + both	**63.08%**	**82.16%**

We compare the proposed approach to CLMR [41], which achieved state-of-the-art performance in music tagging by employing self-supervised contrastive learning. To facilitate a fair comparison, we first performed self-supervised training of CLMR using the same singing voices separated from the 328,418 songs used in Section III-B. The trained network of CLMR was subsequently frozen and used to extract the feature representations of the 25,000 songs (see Section IV-A). Afterward, we trained the same two-layer perceptron using the extracted representations as input and evaluated its accuracy.

Additionally, we prepared a baseline using conventional features in reference to previous studies (see Section II-A). We calculated the same audio features as those Wang and Tzanetakis [27] prepared for their baseline, i.e., 70-dimensional vectors consisting of the mean and standard deviation of chroma (12-dim), MFCC (20-dim), spectral centroid, spectral roll-off, and spectral flux (3-dim). We then trained the same two-layer perceptron using the conventional features similarly.

C. Results

The results are shown in Table I; the proposed approach of training CRNN using pitch-shifted and time-stretched samples as negative anchors obtained the best performance (i.e., 63.08% for top-1 accuracy). Here, a recent study [28] reported an accuracy of 39.3% when identifying 500 singers taken from Million Song Dataset [54]. It is difficult to directly compare it with ours because of the difference in both the training procedures employed (e.g., supervised and self-supervised) and the datasets used, but it can be said that the proposed approach demonstrated relatively high accuracy.[7]

[4]http://sox.sourceforge.net/

[5]https://pytorch.org/audio/stable/sox_effects.html

[6]We followed the parameters of Tao *et al.* [47] and did not conduct hyperparameter tuning in our study.

[7]The method proposed by Lee and Nam [28] is not compatible with ours because its design of employing supervised learning requires annotated datasets. In addition, it requires a dataset consisting of clean singing voices (i.e., not ones separated from polyphonic music), such as DAMP [59], to make the acquired feature representations robust, which hindered us from applying the method to our dataset. Thus, we did not include it in the baselines; however, merely for reference, we additionally evaluated the feature representations acquired with a pretrained network that is trained using DAMP [59] and published by Lee and Nam [28]. As a result, the same two-layer perceptron being fed their feature representations achieved 47.9% for top-1 accuracy and 71.2% for top-5 accuracy in our dataset.

The comparison with CLMR [41] further emphasizes the advantage of the proposed approach. Here, although the number of its parameters (489 k) is much less than that of CLMR (2.9 M), the proposed approach achieved comparable performance even without pitch shifting or time stretching. This is attributable to the design of CLMR, which follows regular self-supervised contrastive learning. For example, while it does not employ time stretching, CLMR applies pitch shifting to generate positive anchors in contrast to us, which may make the acquired representations inattentive to vocal timbre.

In addition, the proposed approach boosted its performance by involving the two introduced transformations. Using either pitch-shifted or time-stretched versions as negative anchors improved accuracy, and using both showed the best performance. This met our expectations given that the transformations were introduced to make the acquired representations attentive to vocal timbre and singing expression, which allow the proposed approach to outperform CLMR.

On the other hand, the two-layer perceptron trained with the conventional features failed to identify singers, as its accuracy did not differ from the expected value of the random output. This is consistent with the result of Wang and Tzanetakis [27]; that is, such conventional features achieved relatively high accuracy in identifying 20 singers from the artist20 dataset [17], [18] but are less effective when we apply them to a larger dataset. Particularly, when we inspected the outputs of the two-layer perceptron, we found that they were concentrated on a small number of singers, which led to the accuracy of the chance level. This implies that learning the relationship between the conventional features and the corresponding singer labels would be intractable for the two-layer perceptron. In other words, the conventional features could not provide enough distinction of singing voices such that the shallow classifier can distinguish 500 singers.

Conversely, the proposed approach achieved much better accuracy with the two-layer perceptron. This shows that the acquired feature representations capture the characteristics of each singer so effectively that they can enable the shallow classifier to distinguish 500 singers. Thus, we expect that replacing conventional features with the feature representations acquired by the proposed approach can improve the performance of various downstream tasks other than singer identification.

D. Effect of Self-Supervised Representation Learning

We also conducted a comparative experiment to confirm the effectiveness of the self-supervised representation learning of the proposed approach. Here, we simultaneously trained the two-layer perceptron with CRNN from scratch using the same training set comprising 20,000 songs (40 songs $\times$ 500 artists) as Section IV-B.[8] Since the architecture and number of parameters of the entire network were consistent, it is theoretically capable of achieving the accuracy presented in Section IV-C. In other words, the differences in accuracy reflected the effect of the self-supervised representation learning.

[8] It is technically possible to use all songs we collected in Section III-B to train the networks from scratch. However, the self-supervised training in Section III-C did not involve label information (i.e., artist names); thus, using the songs for this supervised training would be unfair and would yield a meaningless comparison.

TABLE II
EFFECT OF THE SELF-SUPERVISED TRAINING ON THE ACCURACY

Condition	Top-1 acc.	Top-5 acc.
Frozen after self-supervised training	63.08%	82.16%
Fully-trained from scratch	2.24%	9.00%

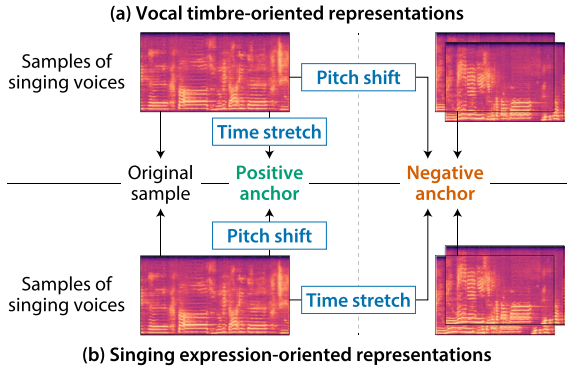

Fig. 3. (a) Vocal timbre-oriented representations were acquired using time-stretched versions as positive anchors. (b) Singing expression-oriented representations were acquired using pitch-shifted versions as positive anchors.

Table II shows the results. Although the network architecture was consistent, the accuracy was primarily improved with the help of self-supervised learning. Particularly, the accuracy of when the networks were trained from scratch was not significantly different from that obtained when using conventional features. This supports the effectiveness of the self-supervised representation learning of the proposed approach, which aligns with observations in other domains [6], [7].

V. VOCAL TIMBRE- OR SINGING EXPRESSION-ORIENTED REPRESENTATIONS

While the case of using both pitch shifting and time stretching to generate negative anchors achieved the best performance in Section IV-C, we can use either of them to generate positive anchors, as mentioned in Section I. Specifically, instead of using the pitch-shifted and time-stretched versions of $\boldsymbol{x}_i$ to augment the negative anchors X_i^-, we can consider complementing the positive anchors X_i^+ with either of them. Then, the acquired feature representations are expected to be attentive to either vocal timbre or singing expression, although the proposed approach does not require annotations to capture these aspects in its self-supervised training.

To examine this option, we newly trained neural networks by switching the usage of the pitch-shifted and time-stretched versions in such a manner and explored the acquired feature representations. Specifically, we prepared a network trained with time-stretched positive anchors and pitch-shifted negative anchors (Fig. 3(a)) and another network trained with pitch-shifted positive anchors and time-stretched negative anchors (Fig. 3(b)). Similar to Section III-C, we trained CRNN with the same dataset as Section III-B and extracted the representations of singing

TABLE III
COMPARISON OF THE VARIANCE OF FEATURE REPRESENTATIONS OF SONGS IN EACH GENRE [9]

Vocal timbre	Singing expression	Variance
Alternative	J-Pop	large
Rock	Rock	↑
Pop	Pop	
J-Pop	Alternative	
Folk	Anime	
Country	Hard Rock	
Blues	Folk	
Anime	Blues	
Metal	Reggae	
Hard Rock	Metal	
Reggae	Country	
Hip-Hop/Rap	Hip-Hop/Rap	small

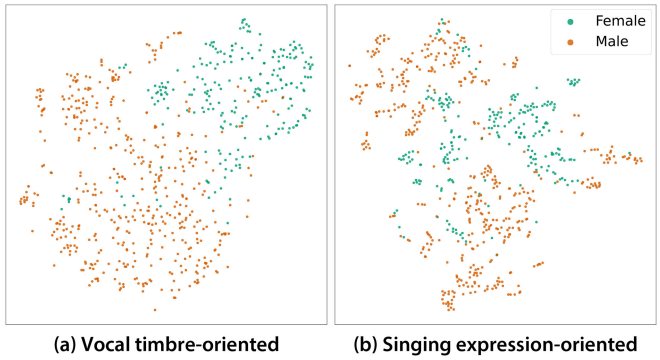

Fig. 4. Feature representations visualized by t-SNE. The plot of (a) vocal timbre-oriented representations provided a clearer distinction between male and female singers than (b) singing expression-oriented representations.

voices in the dataset using the networks. Hereafter, we refer to the representations acquired with the former network as *vocal timbre-oriented* and those acquired with the latter as *singing expression-oriented*.

A. Song Genre

We first explored the acquired representations regarding song genres because genres are often associated with singing styles from the perspective of music theory [60]. By using the genre labels of the songs in our dataset (see Section III-B), we investigated the diversity of singing voices in each genre in terms of the difference in their feature representations. Here, we calculated the variance of the representations of songs in each genre (i.e., the squared average of the L_2-distance of each representation from the mean representation of the corresponding genre) and sorted major genres in order of variance, as shown in Table III.

Hip-Hop/Rap, which is known to have unique singing styles, was ranked at the bottom in both cases, which means that *Hip-Hop/Rap* songs were considered to have similar consistent singing voices in terms of vocal timbre and singing expression. Conversely, *Alternative*, *Rock*, *Pop*, and *J-Pop*, which typically cover a wide variety of songs, were ranked at the top, which means that they were judged to have diverse vocal timbres and singing expressions. It is notable that this distinction seems to accord with the discussion in music theory regarding the nature of those genres, while it was learned in an unsupervised manner. For example, Malaway [60] emphasized the role of rhyme scheme, rather than articulation or prosody, in the individuality of singing voices, referring to the critical analysis on *Hip-Hop* and *Rap* by Krims [61].

Similarly, *Country* songs demonstrated large and small variances for their vocal timbre- and singing expression-oriented representations, respectively. This result agrees with the discussion by Malaway [60], which pointed out that the singing style of *Country* songs is "metrically regularized" based on the anthropological analysis by Feld *et al.* [62]. In addition, *Anime* songs exhibited the opposite trend, i.e., they were judged to have similar vocal timbres and diverse singing expressions. A bias to reduce the diversity of vocal timbres could exist because *Anime* songs are often sung by the voice actors of anime characters, and actors can change their expressions according to the given anime characters. However, they would have difficulty changing their timbres drastically because of the physical constraints of their body structure.

[9]Here, lower genres have less diversity in their vocal timbres or singing expressions according to the acquired vocal timbre- or singing expression-oriented representations, respectively.

B. Singer Gender

To provide another perspective, we explored the acquired representations regarding singer gender since the singing voices of male singers are expected to have similar vocal timbres when compared with those of female singers, and vice versa. Since singer gender information is unavailable for our songs, we sampled five artists for each of the 12 genres listed in Table III and manually annotated gender labels. We then visualized the feature representations of their songs by mapping them into two-dimensional space using t-SNE [63], as shown in Fig. 4.

We can see that the plots of the vocal timbre-oriented representations (Fig. 4(a)) present a clearer distinction between male and female singers than those of the singing expression-oriented representations (Fig. 4(b)), which agrees with our expectations. More specifically, we expect that the vocal timbres of male singers would be more similar to those of other male singers than those of female singers, and vice versa. Conversely, the singing expressions of male singers would not necessarily be similar to those of other male singers, in comparison with those of female singers. These results suggest that the vocal timbre- and singing expression-oriented representations have different characteristics, which can be associated with the vocal timbre and singing expression of singing voices, respectively.

C. Exploration With VocalSet

We further extracted the feature representations of singing voices in VocalSet [64] using the two networks to confirm their characteristics quantitatively. This dataset contains the singing voices of 20 singers performing various vocal techniques, e.g., vibrato and trill. Thus, it can be expected that the vocal timbre-oriented representations capture the distinction of the singers,

TABLE IV
DEGREE TO WHICH SINGING VOICES ARE CONSIDERED SIMILAR TO A QUERY SHARED SINGERS AND VOCAL TECHNIQUES WITH THE QUERY

Feature	Singer			Vocal technique		
	MRR	Prec@1	Prec@5	MRR	Prec@1	Prec@5
Vocal timbre oriented	0.826	0.747	0.601	0.756	0.640	0.525
Singing expression oriented	0.788	0.699	0.511	0.816	0.730	0.564

while the singing expression-oriented representations capture the distinction of the vocal techniques. We confirmed this by examining pairs of singing voices considered similar according to either of the representations.

Specifically, by taking each singing voice in the dataset as a query, we ranked the other singing voices based on their similarity calculated using the feature representations. We then examined whether highly ranked singing voices share their singers or vocal techniques with the query. Using the same scheme for information retrieval, we calculated the mean reciprocal ranking (*MRR*) [65], which takes a value of 1 when the representations regard a singing voice of the same singer or vocal technique to the query as the most similar. We also calculated the precision at k (*Prec@k*), which is the proportion of singing voices that share their singers or vocal techniques with the query among the top-k results. Here, we used the singing voices of nine techniques[10] performed by 20 singers; thus, random selection would result in 0.05 and 0.11 regarding singers and vocal techniques, respectively.

The results are shown in Table IV. As expected, the vocal timbre-oriented representations outperformed the singing expression-oriented representations in terms of retrieving singing voices by the same singer as a query. Conversely, the singing expression-oriented representations performed better when retrieving singing voices with the same vocal techniques. Additionally, when we conducted Student's t-test on the basis of the recommendation by Urbano *et al.* [66], we found that all metrics of *MRR*, *Prec@1*, and *Prec@5* yielded $p < 0.001$ regarding both singers and vocal techniques. In other words, the vocal timbre- and singing expression-oriented representations exhibited significantly different performances in retrieving singing voices by the same singer or those with the same vocal techniques.

These results suggest that the proposed approach of using either time stretching and pitch shifting to generate positive anchors and the other to generate negative anchors made the acquired feature representations attentive to either vocal timbre or singing expression, respectively. Notably, the feature representations used here were not trained using VocalSet. The fact that the acquired representations captured the characteristics of the singing voices of a different dataset from the one used for training implies the generalizability of the proposed approach.

[10]We excluded the singing voices labeled as *spoken* from 10 vocal techniques Wilkins *et al.* [64] used because these are speech samples (rather than singing).

VI. CONCLUSION

In this study, we have introduced self-supervised contrastive learning in a manner specialized to singing voices. Our contributions are summarized as follows:

- We enabled the acquisition of feature representations attentive to vocal timbre and singing expression in an unsupervised manner by training neural networks to discriminate pitch-shifted and time-stretched versions.
- We confirmed that the acquired representations help a classifier improve the accuracy of singer identification, as we observed an accuracy of 63.08% with 500 singers.
- We further suggested that the proposed approach can be extended to acquire feature representations to be used to retrieve singing voices that are similar in terms of either vocal timbre or singing expression.

While we have focused on the singing voice, we believe that other MIR tasks can benefit from the idea of the proposed approach that exploits self-supervised contrastive learning to capture a specific aspect of input by incorporating the domain knowledge in this field.

A. Limitations and Future Work

This study has several limitations. First, it is desirable to conduct a perceptual evaluation to examine the characteristics of the acquired feature representations in depth. Since the vocal timbre- and singing expression-oriented representations could be used in an MIR application that retrieves similar singing voices, investigating how humans perceive the retrieved singing voices would facilitate such an application.

Also, the transformation employed in the proposed approach left room for further exploration. In this study, we prepared fixed parameters for pitch shifting and time stretching based on our early observation, as explained in Section III-C. However, by conducting a comprehensive comparison that involves various parameters, we can elucidate the boundary of the effective range of the parameters. Then, we can extend the proposed approach to randomly sample the parameters within the range during the training, which would contribute to the further (in)attentiveness of the acquired feature representations, in an analogous manner to Chen *et al.* [36] (see Section II-B).

Other transformations than pitch shifting and time stretching can also be incorporated into the proposed approach. For example, instead of time stretching singing voices as a whole, we can modify a specific part of singing voices that contains vibrato in combination with existing methods for detecting vibrato from singing voices [67]. Considering that it would not affect the other parts that do not contain vibrato, incorporating this transformation to generate negative anchors would yield feature representations that are more attentive to vibrato.

Similarly, we can use singing voice conversion [68], [69] to acquire feature representations attentive to vocal timbre, substituting pitch shifting. Specifically, we can generate negative anchors that alter the vocal timbre of singing voices by using singing voice conversion techniques while maintaining other aspects, such as expressive articulations. While we did not employ it because applying singing voice conversion to all samples in

the dataset would be computationally demanding compared with pitch shifting, the acquired feature representations are expected to be specifically attentive to vocal timbre. At the same time, it can pose a possibility of yielding feature representations that are attentive to artifacts produced by singing voice conversion, which demands future research.

REFERENCES

[1] J. A. Burgoyne, I. Fujinaga, and J. S. Downie, *Music Information Retrieval*. Hoboken, NJ, USA: Wiley, 2015, ch. 15, pp. 213–228.

[2] J. Ramírez and M. J. Flores, "Machine learning for music genre: Multifaceted review and experimentation with audioset," *J. Intell. Inf. Syst.*, vol. 55, no. 3, pp. 469–499, 2020.

[3] Y. Song, S. Dixon, and M. Pearce, "A survey of music recommendation systems and future perspectives," in *Proc. Int. Symp. Comput. Music Model. Retrieval*, 2012, pp. 395–410.

[4] M. Schedl, "Deep learning in music recommendation systems," *Front. Appl. Math. Statist.*, vol. 5, p. 9, 2019.

[5] A. M. Demetriou, A. Jansson, A. Kumar, and R. M. Bittner, "Vocals in music matter: The relevance of vocals in the minds of listeners," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, 2018, pp. 514–520.

[6] L. Jing and Y. Tian, "Self-supervised visual feature learning with deep neural networks: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 11, pp. 4037–4058, Nov. 2021.

[7] A. Jaiswal, A. R. Babu, M. Z. Zadeh, D. Banerjee, and F. Makedon, "A survey on contrastive self-supervised learning," *Technologies*, vol. 9, no. 1, p. 22, 2021.

[8] R. Bristow-Johnson, "A detailed analysis of a time-domain formant-corrected pitch-shifting algorithm," *J. Audio Eng. Soc.*, vol. 43, no. 5, pp. 340–352, 1993.

[9] J. Sundberg, *The Science of the Singing Voice*. DeKalb, IL, USA: Northern Illinois Univ. Press, 1987.

[10] M. Umbert, J. Bonada, M. Goto, T. Nakano, and J. Sundberg, "Expression control in singing voice synthesis: Features, approaches, evaluation, and challenges," *IEEE Signal Process. Mag.*, vol. 32, no. 6, pp. 55–73, Nov. 2015.

[11] M. Goto, "Singing information processing," in *Proc. IEEE Int. Conf. Signal Process.*, 2014, pp. 2431–2438.

[12] E. J. Humphrey *et al.*, "An introduction to signal processing for singing-voice analysis: High notes in the effort to automate the understanding of vocals in music," *IEEE Signal Process. Mag.*, vol. 36, no. 1, pp. 82–94, Jan. 2019.

[13] T. Zhang, "Automatic singer identification," in *Proc. IEEE Int. Conf. Multimedia Expo*, 2003, pp. 33–36.

[14] H. Fujihara, T. Kitahara, M. Goto, K. Komatani, T. Ogata, and H. G. Okuno, "Singer identification based on accompaniment sound reduction and reliable frame selection," in *Proc. Int. Conf. Music Inf. Retrieval*, 2005, pp. 329–336.

[15] A. Mesaros, T. Virtanen, and A. Klapuri, "Singer identification in polyphonic music using vocal separation and pattern recognition methods," in *Proc. Int. Conf. Music Inf. Retrieval*, 2007, pp. 375–378.

[16] M. Lagrange, A. Ozerov, and E. Vincent, "Robust singer identification in polyphonic music using melody enhancement and uncertainty-based learning," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, 2012, pp. 595–600.

[17] X. Zhang, Y. Jiang, J. Deng, J. Li, M. Tian, and W. Li, "A novel singer identification method using GMM-UBM," in *Proc. Conf. Sound Music Technol.*, 2019, pp. 3–14.

[18] Y. V. S. Murthy, S. G. Koolagudi, and T. K. R. Jeshventh, "Singer identification for Indian singers using convolutional neural networks," *Int. J. Speech Technol.*, vol. 24, no. 3, pp. 781–796, 2021.

[19] T. L. Nwe and H. Li, "Exploring vibrato-motivated acoustic features for singer identification," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 2, pp. 519–530, Feb. 2007.

[20] T. L. Nwe and H. Li, "On fusion of timbre-motivated features for singing voice detection and singer identification," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2008, pp. 2225–2228.

[21] D. Y. Loni and S. Subbaraman, "Timbre-vibrato model for singer identification," in *Proc. Int. Conf. Inf. Commun. Technol. Intell. Syst.*, 2019, pp. 279–292.

[22] Z. Nasrullah and Y. Zhao, "Music artist classification with convolutional recurrent neural networks," in *Proc. Int. Joint Conf. Neural Netw.*, 2019, pp. 1–8.

[23] T. Hsieh, K. Cheng, Z. Fan, Y. Yang, and Y. Yang, "Addressing the confounds of accompaniments in singer identification," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2020, pp. 1–5.

[24] H. Fujihara and M. Goto, "A music information retrieval system based on singing voice timbre," in *Proc. Int. Conf. Music Inf. Retrieval*, 2007, pp. 467–470.

[25] H. Fujihara, M. Goto, T. Kitahara, and H. G. Okuno, "A modeling of singing voice robust to accompaniment sounds and its application to singer identification and vocal-timbre-similarity-based music information retrieval," *IEEE Trans. Speech Audio Process.*, vol. 18, no. 3, pp. 638–648, Mar. 2010.

[26] T. Nakano, K. Yoshii, and M. Goto, "Vocal timbre analysis using latent Dirichlet allocation and cross-gender vocal timbre similarity," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2014, pp. 5202–5206.

[27] C. Wang and G. Tzanetakis, "Singing style investigation by residual siamese convolutional neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2018, pp. 116–120.

[28] K. Lee and J. Nam, "Learning a joint embedding space of monophonic and mixed music signals for singing voice," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, 2019, pp. 295–302.

[29] J. H. L. Hansen and T. Hasan, "Speaker recognition by machines and humans: A tutorial review," *IEEE Signal Process. Mag.*, vol. 32, no. 6, pp. 74–99, Nov. 2015.

[30] S. S. Tirumala, S. R. Shahamiri, A. S. Garhwal, and R. Wang, "Speaker identification features extraction methods: A systematic review," *Expert Syst. Appl.*, vol. 90, pp. 250–271, 2017.

[31] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Trans. Speech Audio Process.*, vol. 19, no. 4, pp. 788–798, May 2011.

[32] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "x-vectors: Robust DNN embeddings for speaker recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2018, pp. 5329–5333.

[33] H. Eghbal-zadeh, B. Lehner, M. Schedl, and G. Widmer, "I-vectors for timbre-based music similarity and music artist classification," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, 2015, pp. 554–560.

[34] B. Sharma, B. Lehner, M. Schedl, and G. Widmer, "On the importance of audio-source separation for singer identification in polyphonic music," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2019, pp. 2020–2024.

[35] W. Xia, C. Zhang, C. Weng, M. Yu, and D. Yu, "Self-supervised text-independent speaker verification using prototypical momentum contrastive learning," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2021, pp. 6723–6727.

[36] T. Chen, S. Kornblith, M. Norouzi, and G. E. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 1597–1607.

[37] K. He, H. Fan, Y. Wu, S. Xie, and R. B. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 9726–9735.

[38] Z. Wu, Y. Xiong, S. X. Yu, and D. Lin, "Unsupervised feature learning via non-parametric instance discrimination," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3733–3742.

[39] H. Wu *et al.*, "Multi-task self-supervised pre-training for music classification," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2021, pp. 556–560.

[40] A. N. Carr, Q. Berthet, M. Blondel, O. Teboul, and N. Zeghidour, "Self-supervised learning of audio representations from permutations with differentiable ranking," *IEEE Signal Process. Lett.*, vol. 28, pp. 708–712, Mar. 2021.

[41] J. Spijkervet and J. A. Burgoyne, "Contrastive learning of musical representations," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, 2021, pp. 673–681.

[42] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Proc. Conf. Neural Inf. Process. Syst.*, 2020, pp. 12 449–12 460.

[43] W. Hsu, B. Bolte, Y. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "HuBERT: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 3451–3460, Oct. 2021.

[44] H. Al-Tahan and Y. Mohsenzadeh, "CLAR: Contrastive learning of auditory representations," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2021, pp. 2530–2538.

[45] A. Saeed, D. Grangier, and N. Zeghidour, "Contrastive learning of general-purpose audio representations," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2021, pp. 3875–3879.
[46] D. Niizumi, D. Takeuchi, Y. Ohishi, N. Harada, and K. Kashino, "BYOL for audio: Self-supervised learning for general-purpose audio representation," in *Proc. Int. Joint Conf. Neural Netw.*, 2021, pp. 1–8.
[47] L. Tao, X. Wang, T. Yamasaki, J. Chen, and S. Hicks, "Self-supervised video representation learning using inter-intra contrastive framework," in *Proc. ACM Int. Conf. Multimedia*, 2020, pp. 2193–2201.
[48] E. Moulines and F. Charpentier, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," *Speech Commun.*, vol. 9, no. 5/6, pp. 453–467, 1990.
[49] C. E. Seashore, *The Vibrato, Series Studies in the Psychology of Music*. Iowa City, IA, USA: Univ. Iowa, 1932.
[50] Y. Ohishi, H. Kameoka, K. Kashino, and K. Takeda, "Parameter estimation method of F0 control model for singing voices," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2008, pp. 139–142.
[51] T. Kako, Y. Ohishi, H. Kameoka, K. Kashino, and K. Takeda, "Automatic identification for singing style based on sung melodic contour characterized in phase plane," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, 2009, pp. 393–398.
[52] P. Goyal, D. Mahajan, A. Gupta, and I. Misra, "Scaling and benchmarking self-supervised visual representation learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 6390–6399.
[53] J. Pons, O. Nieto, M. Prockup, E. M. Schmidt, A. F. Ehmann, and X. Serra, "End-to-end learning for music audio tagging at scale," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, 2018, pp. 637–644.
[54] T. Bertin-Mahieux, D. P. W. Ellis, B. Whitman, and P. Lamere, "The million song dataset," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, 2011, pp. 591–596.
[55] R. Hennequin, A. Khlif, F. Voituret, and M. Moussallam, "Spleeter: A fast and efficient music source separation tool with pre-trained models," *J. Open Source Softw.*, vol. 5, no. 50, 2020, Art. no. 2154.
[56] S. Kum and J. Nam, "Joint detection and classification of singing voice melody using convolutional recurrent neural networks," *Appl. Sci.*, vol. 9, no. 7, 2019, Art. no. 1324.
[57] K. Choi, G. Fazekas, M. B. Sandler, and K. Cho, "Convolutional recurrent neural networks for music classification," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2017, pp. 2392–2396.
[58] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Representation*, 2015, p. 15.
[59] J. C. Smith, "Correlation analyses of encoded music performance," Ph.D. dissertation, Dept. Music, Stanford Univ., 2013.
[60] V. Malawey, *A Blaze of Light in Every Word: Analyzing the Popular Singing Voice*. Oxford, U.K.: Oxford Univ. Press, 2020.
[61] A. Krims, *Rap Music and the Poetics of Identity*. Cambridge, U.K.: Cambridge Univ. Press, 2000.
[62] S. Feld, A. A. Fox, T. Porcello, and D. Samuels, "Vocal anthropology: From the music of language to the language of song," in *A Companion to Linguistic Anthropology*, Hoboken, NJ, USA: Wiley, 2005, ch. 14, pp. 321–345.
[63] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 86, pp. 2579–2605, 2008.
[64] J. Wilkins, P. Seetharaman, A. Wahl, and B. Pardo, "VocalSet: A singing voice dataset," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, 2018, pp. 468–474.
[65] N. Craswell, "Mean reciprocal rank," in *Encyclopedia of Database Systems*. New York, NY, USA: Springer, 2009, Art. no. 1703.
[66] J. Urbano, H. Lima, and A. Hanjalic, "Statistical significance testing in information retrieval: An empirical analysis of type I, type II and type III errors," in *Proc. Annu. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2019, pp. 505–514.
[67] J. Driedger, S. Balke, S. Ewert, and M. Müller, "Template-based vibrato analysis in complex music signals," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, 2016, pp. 239–245.
[68] H. Doi, T. Toda, T. Nakano, M. Goto, and S. Nakamura, "Singing voice conversion method based on many-to-many eigenvoice conversion and training data generation using a singing-to-singing synthesis system," in *Proc. Asia Pacific Signal Inf. Process. Assoc. Annu. Summit Conf.*, 2012, pp. 1–6.
[69] A. Polyak, L. Wolf, Y. Adi, and Y. Taigman, "Unsupervised cross-domain singing voice conversion," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2020, pp. 801–805.

Hiromu Yakura received the B.E. and M.E. degrees in 2019 and 2021, respectively, from the University of Tsukuba, Tsukuba, Japan, where he is currently working toward the Ph.D. degree. His research interests include intersection of machine learning and human-computer interaction, which involves various application areas, such as creativity support, human resource development, and virtual reality. He was the recipient of Google Ph.D. Fellowship and Microsoft Research Asia Fellowship.

Kento Watanabe received the B.E. and Ph.D. degrees from Tohoku University, Sendai, Japan, in 2013 and 2018, respectively. He is currently a Researcher at the National Institute of Advanced Industrial Science and Technology (AIST), Tsukuba, Japan. His research interests include lyrics information processing, natural language processing, machine learning, and human computer interaction.

Masataka Goto received the Doctor of Engineering degree from Waseda University, Tokyo, Japan, in 1998. He is currently a Prime Senior Researcher at the National Institute of Advanced Industrial Science and Technology (AIST), Tsukuba, Japan. Over the past 30 years, he has published more than 300 papers in refereed journals and international conferences and has received 57 awards, including several best paper awards, best presentation awards, the Tenth Japan Academy Medal, and Tenth JSPS PRIZE. He has served as a committee member of over 120 scientific societies and conferences, including the General Chair of ISMIR 2009 and 2014. As the research director, he began OngaACCEL project in 2016 and RecMus project in 2021, which are five-year JST-funded research projects (ACCEL and CREST) related to music technologies.