Deep Learning Approaches in Topics of Singing Information Processing

Chitralekha Gupta, *Member, IEEE*, Haizhou Li, *Fellow, IEEE*, and Masataka Goto

(Overview Article)

***Abstract*—Singing, the vocal productionof musical tones, is one of the most important elements of music. Addressing the needs of real-world applications, the study of technologies related to singing voices has become an increasingly active area of research. In this paper, we provide a comprehensive overview of the recent developments in the field of singing information processing, specifically in the topics of singing skill evaluation, singing voice synthesis, singing voice separation, and lyrics synchronization and transcription. We will especially focus on deep learning approaches including modern representation learning techniques for singing voices. We will also provide an overview of contributions in public datasets for singing voice research.**

***Index Terms*—Singing information processing, singing voice, singing skill evaluation, singing voice synthesis, singing voice separation, lyrics synchronization, lyrics transcription.**

I. Introduction

SINGING, the vocal production of musical tones, is so fundamental to humans that today, we rarely wonder about its origins. Voice is presumed to be the oldest musical instrument, and there is evidence of singing being universally present in human culture since antiquity [1]. Scientists have argued that the ability to produce something melodic, such as humming and mother-infant vocalizations, may have preceded the ability to form the consonants and vowels to make meaningful speech [2]. But what is the purpose of singing? Before written language, stories were passed down to generations through songs, as songs are often more memorable. Chants and hymns were part of religious rituals, and tales of history and heroics were often in the form of ballads and epics. From a broad evolutionary perspective, there are several theories about why singing was beneficial for humans[1].

Although the origins of singing are still debated, what cannot be denied is its ability to evoke emotions and the role it plays in everyone's lives. We listen to emotional singing to change our mood, or hum to our favorite song in the shower. The immense popularity of singing idol shows, music channels, radio stations, online music services, and karaoke applications shows how we are so surrounded by singing in our daily lives today. Singing has educational [3], entertainment [2], and therapeutic [4], [5] value, which prompts academia and industry to investigate methods to characterize different aspects of singing voice, for applications such as singing skill evaluation and singing synthesis. A research field of such broad studies related to singing technologies is named *singing information processing* [6], [7], and the previous overview article on singing information processing [8] covered the topics of singing synthesis, lyrics transcription and synchronization, vocal timbre analysis, music information retrieval based on singing voices, and singing skill evaluation from the perspective of the traditional methods of handcrafted features and parametric statistical modeling.

Singing and speech have commonality since they share the same underlying voice organ that consists of three units [9]: the breathing apparatus, the vocal folds, and the vocal tract. Due to their commonality, methods developed for solving problems in the speech domain can serve as a foundation for similar problems in the context of singing voice. In this paper, we focus on singing information processing topics that analyze and characterize singing voice, while also being inspired by the commonality between singing and speech. However, the wide variation in the units of the voice organ manifests itself as differences between singing and speech voices. For example, controlled manipulation of the vocal fold vibrations results in larger pitch variation in singing than in speech. In singing, vowels are often stretched in time to sustain musical notes, whereas in speech, the duration of vowels is comparatively small and less varying. Moreover, singing voice often contains embellishments in pitch such as vibrato. Due to those differences, the methods developed for speech voice have not always been directly applicable to singing

Manuscript received 2 December 2021; revised 25 April 2022 and 11 June 2022; accepted 11 June 2022. Date of publication 13 July 2022; date of current version 29 July 2022. The work of Chitralekha Gupta was supported by the Academic Research Council, Ministry of Education (ARC, MOE), Singapore under Grant MOE2018-T2-2-127. The work of Haizhou Li was supported by the Guangdong Provincial Key Laboratory of Big Data Computing, The Chinese University of Hong Kong, Shenzhen, China under Grant B10120210117-KP02. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Stefan Bilbao. *(Corresponding author: Chitralekha Gupta.)*

Chitralekha Gupta is with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore 117583 (e-mail: chitralekha@u.nus.edu).

Haizhou Li is with the Chinese University of Hong Kong, Shenzhen 518172, China, also with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore 117583, and also with the Kriston AI Lab, Xiamen, China (e-mail: haizhouli@cuhk.edu.cn).

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

Digital Object Identifier 10.1109/TASLP.2022.3190732

[1]https://www.economist.com/christmas-specials/2008/12/18/why-music

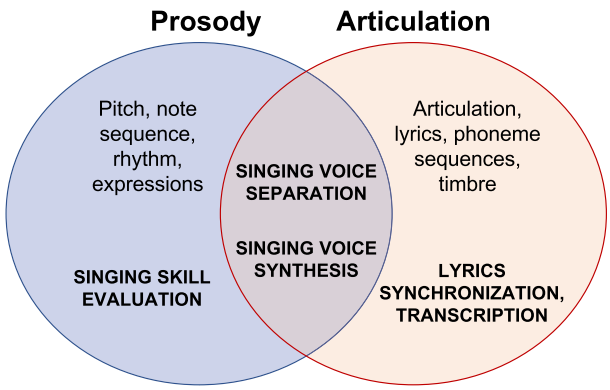

Fig. 1. Overview of the research topics according to the predominant singing voice attributes, i.e. prosody and articulation.

voice [10], [11], and various methods specific to the singing voice have been developed.

With the advent of deep learning and modern representation learning techniques, in this paper, we provide an overview of the recent advances of singing information processing in four important topics: singing skill evaluation, singing voice synthesis, singing voice separation, and lyrics synchronization and transcription. Singing voice can be characterized broadly and independently using two main aspects: *prosody* and *articulation*. Prosodic attributes such as pitch, rhythm, and expressive elements such as vibrato [12], [13] represent the way the musical notes of a song are sung by a singer. Articulation-related attributes represent the way the lyrical contents of the song, consisting of words or phonetic sequences, are uttered. Statistical modeling of the prosodic aspect of singing voice is important for problems such as singing skill evaluation, whereas modeling the articulation aspect of singing voice is needed for tasks such as lyrics synchronization and transcription. Singing voice synthesis and singing voice separation involve modeling both the prosody and articulation aspects of singing voice. Fig. 1 provides an overview of the topics elaborated in this paper according to the singing voice attributes predominantly modeled.

This paper is organized as follows. In Section II, we give a short description of basic terms and concepts from music theory. In Section III, we formulate and discuss various techniques for singing skill evaluation. In Section IV, we discuss the developments in the area of singing voice synthesis. Section V will provide an overview of the techniques of singing voice separation from polyphonic music. In Section VI, the topics of lyrics-to-audio alignment and lyrics transcription will be discussed. In Section VII, we will summarize the datasets that have been made publicly available for singing research in the music information retrieval (MIR) community. We conclude in Section VIII.

II. Music Fundamentals

In this section, we briefly describe some fundamental concepts and terms of music and singing voice that will be referred to in the rest of the paper.

A. Pitch and F0

Pitch represents the perceived fundamental frequency of a sound [14]. The fundamental frequency, which is the rate of vibration of the vocal folds, is referred to as *F0* [15]. Vocal fold vibration results in puffs of air which in turn result in pressure variations, which reach our ears as sound [9]. F0 is measured in terms of frequencies (in cycles per second, or Hertz). Strictly speaking, the pitch is a perceptual attribute, though the F0 is a physical attribute, but the term pitch is often used to refer to F0.

A sung vowel, like any other periodic signal, has a spectrum with energy primarily at integer multiples of F0; these separable signal components are called the harmonics of F0. The F0 range of singing voice is much larger than that of speech. For males, the F0 in singing voice can typically vary from 70 to 500 Hz and for females, from 150 to 700 Hz [9].

Pitch is a perceptual attribute of a sound, perceived on a real-valued scale. In signals with clear harmonic structure, like singing, the perceived pitch of a sound is almost perfectly predicted by its F0. Human pitch perception is approximately logarithmic with respect to F0, i.e., constant pitch changes in music refers to a constant ratio of F0s. The perceived distance between the pitches 220 Hz and 440 Hz is the same as the perceived distance between the pitches 440 Hz and 880 Hz. Each of those distances corresponds to an octave, and twelve-tone equal temperament divides an octave into 12 intervals equally spaced on a logarithmic scale, called semitones.

B. Musical Note and Melody

A *musical note* typically represents the pitch and the duration of a sound in musical notation (score). A note can also represent a pitch class. A *pitch class* is all pitch values related to each other by an octave, which, in Western music, is also referred to as *chroma* [16]. Assuming the equal-tempered scale, there are twelve chroma values that consist of the twelve pitch spelling attributes as used in the Western music notation. For example, the pitch class C consists of the Cs in all octaves. Pitch class is derived from the fact that human pitch-perception is quasi-periodic and pitches belonging to the same pitch class are perceived as having a similar tonal quality. One main property of chroma features is that they capture harmonic and melodic characteristics of music, while being robust to changes in timbre and instrumentation.

Melody is the temporal sequence of musical notes [17], i.e., a linear succession of musical notes that the listener perceives as a single entity. For singing voice, the pitch contour is not just made up of discrete horizontal lines corresponding to distinct steady notes, but is a continuously evolving curve that has macro and micro-tonal pitch movements and expressive musical elements.

C. Lyrics

Lyrics are words that make up a song, usually consisting of *verses* and *choruses*. In popular music, when two or more sections of the song have almost identical music accompaniment but different lyrics, each section is considered as a verse. On the other hand, chorus refers to the repeated sections of the song

having the same or similar set of lyrics. Chorus may contrast with the verse melodically, rhythmically, and harmonically, may assume a higher level of dynamics and activity, and is often with added background instruments.

The notion of rhythm occurs in the lyrics, which is the measured flow of words and phrases as determined by the relation of long and short, or stressed and unstressed syllables [18], [19]. In music, prosody is the way the composer sets the lyrics of a vocal composition in the assignment of syllables to notes in the melody. One syllable of a word in the lyrics is generally assigned to one musical note [20], [21]. As observed in [22], the frequency of one note corresponding to one syllable is much higher than more than one note corresponding to one syllable or one note corresponding to more than one syllables. Thus, syllable duration is closely related to the musical note duration, i.e., one steady note duration is likely to correspond to one syllable duration [20].

III. Singing Skill Evaluation

Singing has been a popular medium of entertainment and a desirable skill. It is also used for rehabilitation and therapy for treating speech disorders such as aphasia [4], [5], [23]. This prompts researchers to study computer-assisted singing learning [24]–[26]. Recently, karaoke singing apps and online music/video sharing services have provided a platform for people to practice, to learn and to showcase their talent. Therefore, automatic singing quality assessment has become an active research area to provide meaningful feedback to singers or to aid music therapy for speech rehabilitation [27].

Singing skill evaluation or singing quality assessment often refers to the degree to which a particular vocal production meets professional standards of excellence. For reliable assessment, it is important to identify vocal attributes that relate to human ratings and objectively define singing excellence. Past studies have identified several perceptual singing-voice parameters that play a significant role in subjective evaluation of singing skill. One study described twelve generally accepted criteria used in the evaluation of Western classical singing by expert music teachers [28], which are: *appropriate vibrato, resonance/ring, color/warmth, intensity, dynamic range, efficient breath management, evenness of registration, flexibility, freedom throughout vocal range, intonation accuracy, legato line,* and *diction*. Oates *et al.* [29] proposed an auditory-perceptual rating scale for operatic singing voice, which consisted of five perceptual parameters, *appropriate vibrato, ring, pitch accuracy, evenness throughout the range,* and *strain*, and these parameters were proven to be unambiguous and covered all aspects of operatic voice. Nakano *et al.* [30] and Goto [8] also summarize acoustic parameters related to singing skill. However, those parameters may not be suitable for evaluating a non-trained or a novice singer. For example, as first studied by Sundberg *et al.* [31], the presence of singing formant, which is an additional vocal resonance, is typically observed in operatic style of singing. Such an operatic style is a specific way of singing that, one may argue, can be unsuitable and undesirable for singing lessons or karaoke performances, especially for beginners [32].

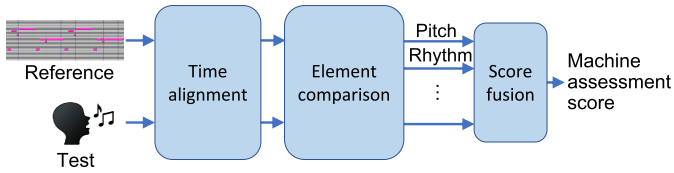

Fig. 2. Diagram of a typical reference-dependent singing skill evaluation system that includes a time-alignment algorithm, an element comparison module, and a score fusion module.

Cao *et al.* [33] summarized five perceptual evaluation criteria for assessing non-trained singers. Those perceptual parameters were: *intonation accuracy*, described as singing in tune, where suitable key transposition is allowed; *rhythm consistency*, described as singing with appropriate tempo speed, where slight tempo variation is allowed; *timbre brightness*, described as brilliance of tone, a sensation of brightness of spectrum; *vocal clarity*, described as vocal vibrations of a clear, well-produced tone; and *overall performance*, described as the overall evaluation integrating all perceptual parameters. It is important to note that the perceptual parameters and criteria for assessment could differ between different singing styles and genres such as traditional Indian music [34], [35] or Jingju music of Beijing Opera [36], [37]. While music performance research has been inclusive of various musical genres, the vast majority of studies have been concerned with Western music [38]. Therefore, the remainder of this section focuses primarily on Western popular music.

The aim of computational techniques for evaluating singing skill is to derive objective metrics that measure the quality of a singing rendition in the same way as music experts do. Automatic singing skill (or quality) evaluation methods seek to provide quantitative measurement of the quality of a singing rendition on the basis of all or a subset of the perceptual criteria that humans use, to provide meaningful feedback to the singers. There have been broadly two approaches for automatic singing skill evaluation [39]–[41]: reference-dependent and reference-independent. In the following, we will introduce both approaches, but since deep learning approaches have not yet been popular for reference-dependent singing skill evaluation, we will focus more on reference-independent evaluation based on deep learning.

A. Reference-Dependent Techniques

Conventionally, the earliest methods for automatic singing skill evaluation relied only on the comparison of a test singing rendition against an ideal reference, such as the MIDI (Musical Instrument Digital Interface) notes/score of the song or a professional singing rendition of the song, as shown in Fig. 2. Such techniques have already been widely adopted for automatic evaluation of karaoke singing, where the reference and test sequences are of the same length and synchronized thanks to the same background music. However, when they are not of the same length and/or not time-synchronized, a time-alignment algorithm such as dynamic time warping (DTW) is required. After comparing the input test singing with the reference, these

methods for singing skill evaluation produce scores on musical elements, such as pitch, rhythm, and volume-related features.

1) Intonation Accuracy: Intonation accuracy or pitch accuracy evaluation has been the most common method for singing quality assessment, primarily because studies have shown that intonation accuracy is one of the most important perceptual parameters when music experts assess singing quality [42], [43]. In one of the earliest studies, Lal [44] proposed a pitch-based similarity measure to compare a test singing clip to the reference singing clip. In another study, Tsai and Lee [32] proposed an automatic evaluation system for karaoke singing in which they computed the Euclidean distance of the note sequence of test singing voice from that of the intended reference song, time-aligned with each other using DTW, to compute the pitch accuracy rating. Other reference-dependent techniques [13], [33], [45] have also been proposed.

Although MIDI notes used as the reference approximately represent the sequence of sung notes, they are unable to represent human voice since singing voice comprises of pitch transitions, modulations, and different voice timbres. Therefore, studies for singing quality assessment have explored comparison with an ideal reference singing rendition [43], [46], [47], instead of the MIDI notes of the song. The drawback of this approach is that the choice of an ideal singing rendition is subjective. Moreover, such a gold standard reference for comparison limits the scope of creative deviations of a singer.

2) Rhythm Consistency: Rhythm consistency is another important feature for singing evaluation. Tsai and Lee [32] evaluated rhythm by comparing the note-onset strength of the background accompaniment of karaoke to that of the test singing. Molina *et al.* [48] and Lin *et al.* [49] designed a method to evaluate rhythm in the absence of background accompaniment by aligning the test pitch contour with the reference pitch contour using DTW, and obtained the rhythm score by computing the deviation of the optimal path in the DTW cost matrix from its straight line regression fit. Here, a straight line with an angle different from 45 degrees represents a good rhythmic performance but at a different tempo from the reference, which is not penalized by this measure. So the deviation of the optimal path from the regression line (which may not be at 45 degrees) serves as an indicator of the rhythm accuracy. As an extension to this idea, Gupta *et al.* [47] used the sequence of MFCC vectors, instead of the error-prone pitch contours, to compute the DTW alignment between the reference and test singing renditions. The assumption was that if the sequences of phonemes and words are uttered correctly, MFCC would capture the spectral characteristics of the uttered words, thus making this rhythm measure independent of inaccurate pitch estimation.

3) Perceptual Quality: According to music psychology studies of human perception, humans first convert the perceived singing audio into a weighted representation of the identified perceptual parameters [50], and then make a judgment of overall quality in a holistic manner [28], [29], [43]. For example, Nakano *et al.* [12] fuse features derived from intonation and vibrato to classify recordings of professional singers as either 'good' or 'bad' and achieve an accuracy of up to 87%, depending on the gender of the singer, though it is reference-independent. Gupta *et al.* [47] presented a reference-dependent measure of evaluation called perceptual evaluation of singing quality (PESnQ) where localized errors in time and frequency (e.g. certain phrases sung with bad pitch or bad rhythm) have a greater subjective impact than distributed error. Furthermore, Gupta *et al.* [43] explored early and late fusion of pitch, rhythm, timbre, and vibrato related features to map those objective features to the overall singing quality judgment scores by human music experts. They found that the late fusion method achieved a higher correlation with human judgment than early fusion.

4) Discussion: An advantage of reference-dependent singing skill evaluation is its ability to provide detailed feedback both in terms of different perceptual parameters as well as the temporal location at which errors occur. A drawback is that this approach is inevitably constrained either by the need for a reference singer/singing or the availability of digital sheet music for a song.

All of the above techniques introduced for reference-dependent singing skill evaluation are not based on deep learning; deep learning has not yet been used for reference-dependent singing skill evaluation. There have been some recent studies on deep learning approaches for assessing piano performances [51] and flute performances [52]. In both of these works, the reference and the test are encoded to a latent space through a stack of CNNs which are then compared to predict the human assessment ratings. Such approaches could be possibly explored for singing performances in future.

B. Reference-Independent Techniques

Studies have shown that music experts can evaluate singers with a high level of consensus even when the song is new to them [30], which implies that there are underlying inherent characteristics of singing quality that differentiate between good and poor singing. This motivates the investigation of singing quality assessment without a reference [12]. While reference-independent techniques are much desired in practical applications, there are not many studies in the literature. A summary of recent studies is shown in Table I. In general, the reference-independent techniques can be grouped into two categories: characterization of singing techniques, and data-driven learning approaches.

1) Characterization of Singing Techniques: This group of methods characterize the features extracted from a singing rendition on the basis of music theoretical rules and train a classification or regression model with human assessment scores as ground-truth to predict the overall singing quality score, as shown in Fig. 3. Two types of singing characterization methods are employed: *absolute* that involves characterizing only the test singing vocal input and *relative* that characterizes the test singing vocal input in comparison to other singing vocal inputs.

In an early study, Nakano *et al.* [12] designed an evaluation scheme based on pitch interval accuracy and vibrato, which are regarded as features that function independently from the individual characteristics of singer or melody. They used pitch interval accuracy to measure the averaged amount of the offset

TABLE I
SUMMARY OF RECENT REFERENCE-INDEPENDENT APPROACHES FOR SINGING SKILL EVALUATION

Paper	Parameters	Framework	Dataset/Codebase link	Performance
[54] (2017)	Intonation, Voice Quality, Dynamics	60 Low-Level Descriptors (LLD) and their derivatives, total of 6,373 features extracted using OpenSmile. Classifiers: Support Vector Machine (SVM), Naïve Bayes, K-Nearest Neighbors (KNN), Decision Tree	110 singing home recorded renditions crawled from YouTube. Ratings from crowd-sourcing platforms, three rating classes: poor, fair, good	Unweighted Average Recall: 59.30%
[55] (2018)	Intonation	Pitch histogram modeling with Gaussian Mixture Model (GMM), aggregate rank-ordering	A mix of NUS48E and DAMP datasets, 5 songs, 10 singers per song. Human ratings from crowd-sourcing.	Spearman Correlation: 0.86
[56] (2019)	Overall	Spectrogram input, convolutional Bidense network to predict overall singing quality class	19,478 singing renditions using karaoke application Kwai. 10 human raters, 2 classes: good or poor	Classification Accuracy: 89.55%
[39] (2020)	Intonation, Rhythm, Overall	Inter-singer distances, pitch histogram and DTW alignment based. MLP	400 singing renditions in total across 4 distinct songs, each song sung by 100 unique singers. Curated subset of DAMP dataset. Human ratings from crowd-sourcing.	Spearman Correlation: 0.71
[57] (2020)	Overall rank order	Mel-spectrogram, pitch histogram. Twin CRNN with comparative loss conditioned on pitch histogram	Same as in [39]	Spearman Correlation: 0.73
[58] (2020)	Overall	Mel-spectrogram, CQT, Chromagram, pitch histogram. CRNN conditioned on pitch histogram	Same as in [39]	Spearman Correlation: 0.76
[59] (2021)	Intonation and Overall	CQT, pitch histogram. CRNN conditioned on pitch histogram	Same as in [39] plus artificially augmented data with pitch variations. 3 pitch classes good medium poor†	Pearson Correlation: Overall: 0.77; Pitch classification: 96%
[22] (2021)	Rhythm	CQT, rhythm histogram. CRNN conditioned on rhythm histogram	100 Chinese language singing renditions of pop songs, artificially augmented with duration variations, 5 artificial rhythm ratings between -1 to 1‡	Pearson Correlation for Rhythm Quality Score: 0.75

†[59] Code: https://github.com/AME430/Towards-Training-Explainable-Singing-Quality-Assessment-Network-with-Augmented-Data.git.
‡[22] Code: https://github.com/AME430/TOWARDS-REFERENCE-INDEPENDENT-RHYTHM-ASSESSMENT-OF-SOLO-SINGING.git.
Note that we include the performance indicators as reported in the papers, which may not be comparable across rows as the test datasets are different.

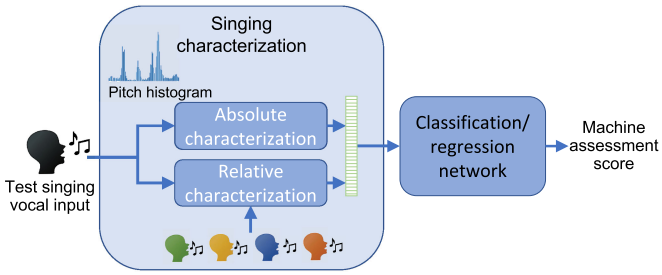

Fig. 3. Diagram of a typical reference-independent singing skill evaluation system that employs singing characterization method for quality assessment.

of the logarithmic-scale F0 values of the singing within a musical semitone grid (corresponding to equal temperament in the Western music tradition). Before computing the pitch interval accuracy, the tuning of the semitone grid is first adjusted by shifting it along the frequency axis so that the grid can be best matched with the F0 values. If the F0 values tend to have small offsets from the semitone grid throughout the song, then the pitch interval accuracy becomes high and the singing is considered to be of good quality. There were also studies that primarily measure the quality of pitch histogram [53]–[55].

With the immense amount of online uploads on singing platforms, Gupta *et al.* [39] leveraged the comparative statistics between singers as well as music theory to derive a leaderboard of singers, where the singers are rank-ordered according to their singing quality relative to each other. They designed inter-singer relative measures based on the hypothesis that given a song that has a particular sequence of notes and a rhythm, it can be sung correctly in one or a few consistent ways, but incorrectly in many different, and dissimilar ways. The good singers would share many characteristics such as the frequently hit notes, the sequence of notes, and the overall consistency in the rhythm of the song, but different bad singers will deviate from the intended song in different amounts and ways. They proposed a framework to combine these inter-singer relative distance measures with the pitch histogram-based intonation measures to provide a comprehensive singing quality assessment without relying on a reference. A linear combination of the rank ordering provided by each of these measures showed a Spearman rank correlation of 0.71 with human judgments.

2) Data-Driven Learning Approaches: With the advent of deep learning, computational neural network solutions have been explored for reference-independent singing skill evaluation that does not depend on handcrafted features. The success of deep learning methods relies on the design of network architecture as well as feature representation. Zhang *et al.* [56] proposed a convolutional neural network (CNN) architecture named Bi-DenseNet that used magnitude spectrograms as input representation and was trained in a supervised way to discriminate good singing renditions from poor quality singing renditions.

This network consisted of input convolutional layers, followed by bi-dense blocks, where each bi-dense block consisted of two parallel convolution filter banks, one with horizontal filters, and the other with vertical filters. This framework accounted for the multi-scale temporal and spectral features of singing voices through the proposed bi-dense blocks. Wang and Tzanetakis [60] utilized CNNs in a Siamese architecture trained on both Mel-spectrograms and constant-Q transforms (CQT) to investigate singing style. CNNs convolve the input with learnable kernels and are efficient at learning localized features.

To improve the encoding of long-term temporal dependencies, recurrent neural networks (RNNs) are studied. RNNs calculate the output of a time step from both the input of this time step and the hidden state of the previous step. Moreover, RNNs can process the output of a CNN to form a convolutional recurrent neural network (CRNN). In this case, the initial convolutional layers capture local information, and the recurrent layer summarizes it along time. Pati *et al.* [61] trained a fully CNN on pitch contours and a CRNN model on Mel-scaled spectrograms (Mel-spectrograms) to assess music performances of pitched wind instruments.

Huang *et al.* [58] adopted the CRNN architecture to learn features from the input and predict evaluation scores of singers in a supervised training setup, thus, using CRNN for absolute characterization of singing vocalization as in Fig. 3. Mel-spectrogram, CQT, and chromagram input features were compared. CQT, which uses geometrically spaced frequency bins to ensure that the Q factors (i.e., the ratio of the center frequencies to bandwidths) of all bins are constant, was found to have the best performance in predicting singing quality score. CQT, in general, is found to be well suited for representing music data, since because of the constant Q factor, it can capture essential audio information from both low and high frequencies with sufficient resolution. In addition, Huang *et al.* [58] incorporated pitch histogram as a conditioning vector appended to the embedding from the CRNN. This hybrid framework of spectral features and pitch histogram (CPH-CRNN) showed the best performance, with a Spearman rank correlation of 0.76 with human judgments, for a large test set of unseen singers. The pitch histogram helped in capturing information related to intonation accuracy, while other rhythm and timbre related parameters are captured through the spectral features. However, for unseen songs, although the hybrid CPH-CRNN framework performed better than CQT-CRNN, the Spearman rank correlation with human judgment dropped to 0.56. One reason was that only four unique songs (each sung by 100 singers) were present in the training data. Such a small number of unique songs made the model not general enough across different songs.

Gupta *et al.* [57] incorporated a twin-neural network consisting of a CRNN framework for each branch where the inputs are Mel-spectrograms, along with pitch histograms as a conditioning vector, similar to [58]. They used a comparative loss, instead of a contrastive loss, to train this twin network. This loss was designed in such a way that the network learns which of the two input singing voices is more preferable in terms of singing quality. Many such comparisons lead to a rank-order of singing voices. The advantage of such a comparative network is that the two arms of the network would be able to project each singing voice input to a compressed latent space that only represents the discriminatory singing quality properties independent of the song or the singer. Indeed, the rank correlation of the output of this framework was 0.65 when compared to human judgment on unseen songs test set. However, the drawback of this comparative framework is that at the time of inference, a given test singing audio needs to be compared against all other existing singing renditions in the database to find its right rank position.

Although these data-driven methods have achieved good performance, they depend on a dataset annotated by music experts, which is not easily scalable. The lack of annotated datasets has been a major hurdle in singing skill evaluation research. Li *et al.* [59] used pitch shifting as a data augmentation technique to create *negative examples* of singing quality from a dataset that consisted of only professional grade singing voices singing a rich variety of songs (86 distinct songs). Upon training the same network as in [58] with this augmented dataset, the experiment with unseen songs and singers showed better performance than [58]. Similarly, Gupta *et al.* [22] used time-scaling of phonemes as a data augmentation technique to create negative examples of rhythm quality of solo-singing voices from the professional grade dataset.

In the recent data-driven methods [57], [58], [62], neural network frameworks have been employed to learn implicit features from the time-frequency representations of the singing voice, allowing the model to learn the inherent characteristics of singing voice through supervised learning while not depending on a reference singing rendition. However, they are trained to only give an overall assessment score, as such a score is the only human annotation practically available for large datasets. Therefore, such systems are unable to provide detailed feedback to the singers about their singing quality in terms of musical parameters such as pitch accuracy and rhythm correctness. Li *et al.* [59] used an augmented dataset that contained artificial or pseudo ground-truths of pitch accuracy score to train a multi-task CRNN framework to simultaneously predict pitch accuracy score and overall singing quality score. This was the first step towards an explainable singing-quality-evaluation neural network in a reference-independent setting. Exploring the correlation between syllable duration and note duration, Gupta *et al.* [22] proposed a rhythm representation based on syllable duration, i.e., syllable duration histogram as an indicator of rhythm quality in a solo singing rendition when the musical score information is not available. They incorporated this rhythm representation as a conditioning feature in CRNN framework.

C. Comparison of Reference-Dependent and Independent Methods

Reference-dependent methods have the advantage of providing detailed feedback to the singers about their singing quality in terms of musical parameters such as pitch accuracy and rhythm correctness, and also providing an assessment for every short duration of singing, making them a popular technique for use

in karaoke applications with real-time feedback. They, however, require a reference and assume that the reference is a perfect model that should be reproduced by a singer. On the other hand, the existing reference-independent methods have the advantage of not needing a reference for every song, and rather allowing the singer to be different from a reference while following suitable singing practices. Data-driven reference-independent methods based on deep learning, however, require a dataset of overall assessment scores annotated by music experts for training data, and cannot provide detailed feedback to the singers since they typically provide only an overall score. Although some methods [22], [59] have attempted to provide an explainable score using data augmentation techniques, much needs to be done to build standard training and test datasets with human labels that go beyond an overall assessment score. As reported in [39], the longer the singing input, the better the prediction accuracy. So, these methods are not yet suitable for real-time feedback.

D. Future Directions

Research in singing skill evaluation has mainly focused on the evaluation of the fundamental qualities of singing vocals, such as pitch and rhythm. Deeper levels of singing quality assessment, such as expressions, emotions, flexibility, creativity, and personality, require further studies. While a lot of work has been done for popular karaoke-style singing, objective measures across different genres and styles of singing need further investigation. For example, the criteria of evaluation of a rap singing will be different from that of a jazz singing, or a Chinese opera singing from a Western classical singing.

Since reference-dependent and independent methods have complementary properties, their possible combinations would also be an interesting direction of research in the future since such combinations could benefit from their relative advantages.

IV. Singing Voice Synthesis

Singing voice synthesis (SVS) has been an active research area for a long time [63], [64]. It is common to synthesize the singing voice from written lyrics that follow the musical score or MIDI notes. This approach is the basis for many commercial products. This type of singing voice synthesis was later termed as *text-to-singing (or lyrics-to-singing) synthesis*, just like text-to-speech (TTS) synthesis for speech synthesis. In 2007, *speech-to-singing synthesis*, a novel approach of synthesising singing voice from speaking voice (speech or spoken lyrics), was coined by Saitou *et al.* [65]. Furthermore, in 2009, *singing-to-singing synthesis* was coined by Nakano *et al.* [66], which synthesises singing voice from another singing voice (e.g. singing voice from a different singer or with bad quality). Since then, each of the three tasks, namely text-to-singing synthesis [67], speech-to-singing synthesis [65], [68], and singing-to-singing synthesis [66], [69] has become an extensive area of research. Even *singing-to-speech synthesis* was proposed by Aso *et al.* [70] in 2010, though speech synthesis is beyond the scope of this paper.

The most important aspect of these tasks is to synthesise appropriate prosody of singing voice while retaining the linguistic content and the intended singer's identity. Since the earliest works, singing synthesis methods have been heavily based on speech synthesis methods [63]. Traditionally, the techniques involved in text-to-singing synthesis were based on linear predictive coding [71], formant synthesis [72], concatenative unit selection technology such as VOCALOID [73], and statistical parametric modeling such as Sinsy [74]. Speech-to-singing and singing-to-singing synthesis methods have explicitly tried to control and modify singing specific features such as the F0 contour, vibrato, and phoneme duration through alignment between source and target [65], [68]. With the advent of deep neural network (DNN) approaches, the control over voice is more implicit, resulting in improved naturalness and quality of the synthesized singing voice [67], [75], [76].

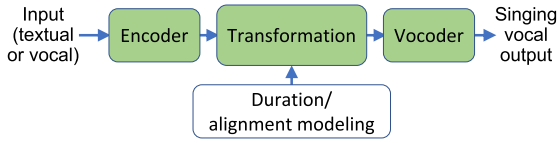

Fig. 4. Diagram of a singing voice synthesis system that takes either textural (lyrics and musical scores) or vocal (speech or singing voice) as input and generates a singing vocal output.

The workflow of singing voice synthesis can be summarized in Fig. 4. The framework consists of an encoder that converts the input into an embedding. The input to the encoder could be textual such as lyrics and musical scores, or vocal such as speech or singing voice. A duration modeling unit time-aligns the encoded input representation with the encoded output representation. The input and the output are not of equal temporal dimensions, as textual inputs could be of a much shorter length than the length of audio output frames or waveform samples, therefore duration modeling helps in temporally aligning the input and the output. The transformation unit maps the encoded input embedding to the encoded output acoustic features conditioned on the duration model. Finally, a vocoder is employed to convert the output acoustic features into an audio waveform.

Among the three singing voice synthesis tasks, the architecture of encoder and transformation units may vary with the type of input. However, they all require a vocoder for waveform generation. Parametric vocoders such as STRAIGHT [77] and WORLD [78] decompose the signal into phonetic and pitch components, so that the pitch can be modified easily to match any target melody. The analysis part of these vocoders estimates multiple acoustic features while the synthesis part converts these features into a time domain waveform. For example, the WORLD vocoder consists of analysis algorithms to estimate F0, spectral envelope, and aperiodicity, while a synthesis algorithm based on the minimum-phase response incorporates these parameters to synthesize the waveform. In many recent studies, a neural network architecture is employed to model the acoustic features, which are then fed as inputs to the parametric vocoder to synthesize the audio waveform. Blaauw *et al.* [79] proposed a neural parametric vocoder based on WaveNet, instead of the parametric vocoders. WaveNet is a probabilistic autoregressive model consisting of dilated causal convolutions that are used to synthesize the audio waveform sample-by-sample conditioned

TABLE II
SUMMARY OF SINGING VOICE SYNTHESIS TECHNIQUES WITH TEXTUAL INPUT (TEXT-TO-SINGING)

Paper	Architecture	Dataset	Demo/Samples	MCD	MOS
Sinsy-DNN [75], [95] (2018)	DNN	70 Japanese children's songs	https://www.sinsy.jp/	5.06	3.74
[76] (2020)	CNN-FCN	55 Japanese children's songs and 55 J-POP songs	https://www.techno-speech.com/news-20181214a-en	NA	4.23
NPSS [79] (2017)	Autoregressive	31 Japanese children's songs	http://www.dtic.upf.edu/~mblaauw/NPSS/	5.54	3.43
DAR [93] (2019)	Autoregressive-prenet multi-head attention	100 Chinese songs	http://home.ustc.edu.cn/~yiyh/interspeech2019/	3.51	NA
XiaoiceSing [81] (2020)	Autoregressive with joint modeling	2,297 Mandarin pop songs	https://xiaoicesing.github.io/	5.42	3.61
FFT-NPSS [67] (2020)	E2E transformer-based	35 English songs	https://mtg.github.io/singing-synthesis-demos/transformer/	NA	2.87
ByteSing [96] (2021)	E2E encoder-decoder framework	90 Chinese songs	https://bytesings.github.io/paper1.html	5.76	NA
WGAN-Sing [80] (2019)	Wasserstein deep convolutional GAN	Multi-singer, 48 English songs [97]	https://pc2752.github.io/sing_synth_examples/†	5.36	NA
BEGANSing [98] (2020)	Autoregressive conditional GAN	50 Korean children's songs	https://soonbeomchoi.github.io/saebyulgan-blog/‡	NA	3.12
[99] (2020)	Multi-singer; GAN with multiple random window discriminators	770 Chinese pop songs from one female singer, 200 Chinese pop songs from 6 singers	https://jiewu-demo.github.io/INTERSPEECH2020/	NA	4.12

†[80] Code: https://github.com/MTG/WGANSing.
‡[98] Code: https://github.com/SoonbeomChoi/BEGANSing.
The performance metrics are not directly comparable across different rows as the test data may be different. MCD: Mel cepstral distortion (dB) (lower is better), MOS: Mean opinion score (five-point naturalness score) (higher is better).

on the estimated acoustic features. The neural network solutions are known to outperform the parametric vocoders in terms of voice quality, if the target singing voice is sufficiently similar to singing voices in training datasets. However, they are typically computationally more expensive, which calls for network architecture of low computational cost, such as WaveRNN.

In this section, we will summarize the recent advances in singing voice synthesis, categorized into two types of input, textual or vocal. A summary of the recent methods of singing voice synthesis from textual input is provided in Table II, and those from vocal inputs (singing voice and/or speech) is provided in Table III.

A. *Singing Voice Synthesis From Textual Input*

Singing voice synthesis (SVS) and text-to-speech (TTS) synthesis are related but distinct research fields. While both fields try to generate signals mimicking the human voice, singing voice synthesis models a higher range of pitch values and vowel durations [80]. Moreover, while speech synthesis is controlled primarily by words or syllables, singing voice synthesis is additionally controlled by the musical score, which puts constraints on the pitch and timing. These constraints and differences have resulted in SVS being its own active field of research, separate from TTS.

Singing voice synthesis is a task of synthesizing singing voice from textual input that includes lyrics and a musical score. A typical SVS system consists of an acoustic model (transformation model) that generates acoustic features (e.g., Mel-spectrogram) conditioned on encoded lyrics, musical score, and duration. A vocoder is then used to convert these generated acoustic features into a waveform, as shown in the inference phase of Fig. 5. At the time of training, the acoustic characteristics of singing vocals are learnt from a database of singing vocals of a singer, and the temporal relationship between musical score and phonetic duration is learnt through duration modeling, as shown in the training phase in Fig. 5.

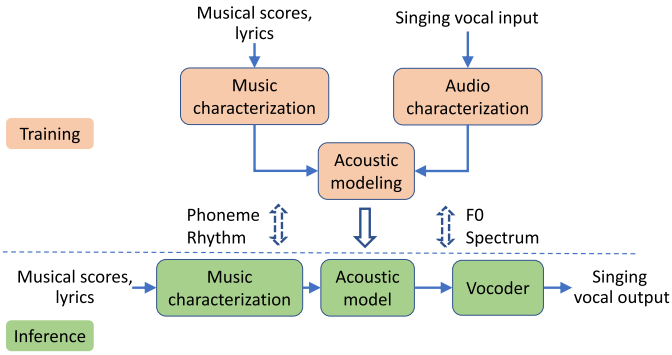

Fig. 5. Overview of the training and inference phases of singing voice synthesis systems with textual inputs.

In order to synthesize an expressive and rhythmic singing voice of high quality, three aspects are generally taken into consideration [81]:

- An effective F0 synthesizing model to establish the complex patterns in the F0 contour of singing,
- A spectrum model that predicts spectral features for phonetic articulation with adequate naturalness of sound quality, and
- A duration model that can learn the correspondence between note duration and phonetic duration, and in turn adhere to the rhythmic constraints of singing.

TABLE III
SUMMARY OF SINGING VOICE SYNTHESIS TECHNIQUES WITH VOCAL INPUT (SINGING VOICE CONVERSION AND SPEECH-TO-SINGING)

Paper	Input Type	Architecture	Dataset	Demo	MOS_{sim}	$MOS_{natural}$
[110] (2019)	Speech, Singing	Autoencoder	Smule DAMP [115], NUS48E [97]	https://enk100.github.io/Unsupervised_Singing_Voice_Conversion/	3.69	4.10
PitchNet [111] (2020)	Singing	Autoencoder with pitch regression network	NUS48E [97]	https://tencent-ailab.github.io/pitch-net/	3.64	3.75
[113] (2020)	Singing	GAN	NUS48E [97]	https://singing-conversion.github.io/	NA	4.04
[112] (2021)	Singing	ASR as content encoder, GAN as conversion model	16 singers, Mandarin songs	https://lzh1.github.io/singVC/†	3.57	3.75
FastSVC [114] (2021)	Singing	ASR as content encoder, waveform generator as conversion model	NUS-48E [97], Chinese singing dataset	https://nobody996.github.io/FastSVC/	3.58	4.00
[116] (2020)	Singing	Conditional encoder-decoder	255 songs, 15 singers	https://juheo.github.io/DTS/	NA	NA
[117] (2019)	Singing	RNN-deep bidirectional LSTM	4 singers from MIR 1k [118]	https://sites.google.com/site/singingvoiceconversion2018/	3.40	3.50
[119] (2020)	Singing	Variational autoencoder (VAE)	VocalSet [62]	https://ismir19-217.github.io/icassp20-audio-sample/index.html	NA	NA
VAW-GAN [120] (2020)	Singing	Variational autoencoding Wasserstein GAN	NUS48E [97]	https://kunzhou9646.github.io/singvaw-gan/	NA	3.03
[121] (2019)	Singing, Speech	Speaker adaptation with autoregressive	NUS48E [97]	https://mtg.github.io/singing-synthesis-demos/voice-cloning/	NA	NA
[122] (2021)	Singing	Semi-supervised encoder-decoder	41 English pop songs	https://mtg.github.io/singing-synthesis-demos/semisupervised/	NA	3.42
DeepSinger [123] 2020	Singing	Transformer-based FastSpeech model	Chinese, Cantonese, English songs crawled from the web	https://speechresearch.github.io/deepsinger/	NA	3.78
Zero-shot[124] (2020)	Singing	Speaker embedding network with encoder-decoder	7 hours singing voice data	https://sites.google.com/izotope.com/ismir2020-audio-demo	3.05	2.67
[125] (2019)	Speech	Aligner and DB-LSTM conditioned on i-vectors	(Parallel) NUS48E [97], NHSS [126]	https://xiaoxue1117.github.io/sample/	NA	3.95
DurIAN-SC [127] (2020)	Speech	Encoder, aligner, autoregressive decoder	Non-parallel Mandarin speech and singing datasets	https://tencent-ailab.github.io/learning_singing_from_speech/	3.74	3.71
[128] (2020)	Speech	Encoder-decoder conditioned on pitch contour	NUS48E [97]	https://jayneelparekh.github.io/icassp20/‡	NA	NA

†[112] Code: https://github.com/google/REAPER.
‡[128] Code: https://github.com/jayneelparekh/sp2si-code.
The performance metrics are not directly comparable across different rows as the test data may be different. Mean opinion score (MOS) on five-point naturalness score ($MOS_{natural}$) and voice similarity score (MOS_{sim}) (higher is better).

In the following sub-sections, we briefly discuss the traditional methods for singing voice synthesis, and discuss in detail various deep learning frameworks explored for this task. This includes feed-forward DNN-based approaches, autoregressive prediction models, approaches to overcome over-smoothness of the generated vocals, and approaches explored to control pitch expressiveness and fidelity.

1) Traditional Methods: The earliest works in singing synthesis involved physical speech synthesis systems that were also capable of singing synthesis, such as the acoustic tube model of Kelly and Lochbaum [82]. However, these were computationally expensive and not commercially viable. Another early voice model was Rodet's formant wave function (FOF) [83] which is a time-domain waveform model of the impulse response of individual formants, where the control parameters define the center frequency and bandwidth of the formant being modeled, and the rate at which the FOFs are generated and added determines the base frequency of the voice. Singing synthesis using formant models has been extensively used and studied in MUSSE (MUsic and Singing Synthesis Equipment) synthesizer [72] for studying music performance through synthesis-by-rule [84], and has been adapted for real-time control in performance [72].

More recent studies in singing voice synthesis generated sounds using unit concatenation [73], [85] or HMM-based statistical parametric synthesis [74], [86] methods. A typical unit concatenation synthesis system receives the score and lyric information, selects the necessary phonetic samples from a large corpus of singing recordings, concatenates them, and "smooths" the pitch and timbre around the junction of samples in frequency domain. Such systems have been deployed in various commercial singing voice synthesis products, such as VOCALOID [73], because they can provide good sound quality and naturalness in certain settings. They usually require large databases of singing voice recordings from professional singers, as well as manual labeling and segmentation effort. Since the units may not always connect smoothly, developers of those products often make careful efforts to ensure that singing voices are recorded with clear note boundaries, and that segmentation labels correctly annotate these boundary times.

A statistical parametric synthesis system consists of training and synthesis components. During training, the spectrum, excitation, and vibrato parts are extracted from a singing voice database and then modeled by context-dependent HMMs, which also include state duration modeling. Pitch adaptive training [87] is used to generate singing voices in any pitch. The singing voice waveform is synthesized from the acoustic parameters predicted by a trained HMM, thereby requiring less data to construct a system compared to unit-selection systems. However, HMM-based methods tend to have several limitations, such as excessive averaging (oversmoothing) and an overly static sound and noticeable state transitions in long sustained vowels.

2) Early Deep Learning Frameworks: With the rapid evolution of deep learning, several SVS systems based on deep neural networks have been proposed in the past few years that have demonstrated their superiority over traditional HMM-based ones. The steps for training an SVS model using DNN [88] are usually as follows: first, use a pre-trained HMM to align frame-by-frame the musical score feature sequence with the acoustic feature sequence of the corresponding singing audio from the database, and then, exploit a DNN to learn the mapping relationship between the musical score features and acoustic features. During synthesis, an arbitrarily given musical score including lyrics to be synthesized is first converted to a label sequence, which is mapped to an acoustic feature sequence by the trained DNN using forward propagation. Then, the spectrum and F0 parameters are generated by a speech parameter generation algorithm [89] from the acoustic feature sequence that provides a parameter trajectory corresponding to natural singing voice. Finally, a singing voice is synthesized from the generated spectrum and F0 parameters by using a neural vocoder [79]. Hono *et al.* [75] introduced a DNN-based SVS system named Sinsy that provided an online website for SVS. Trajectory training, a vibrato model, and a time-lag model were introduced into the system to synthesize high quality singing voices. Experimental results showed that the DNN-based methods are better than the HMM-based methods which were the state-of-the-art before then. Although there are correlations between neighbouring frames in singing data, the feed-forward DNN-based approach assumes that each frame is generated independently, resulting in a one-to-one mapping between linguistic/musical and acoustic features frame-by-frame, resulting in a discontinuous output. As a solution, long short-term memory recurrent neural networks (LSTM-RNN) [90] provide an elegant way to model sequential data that take into account short- and long-term correlations between neighbouring frames, i.e., previous input features can be used to predict the output features at each frame. Experimental results showed that LSTM-RNN performed better than DNN for singing voice synthesis. Later, Nakamura *et al.* [76] proposed a framework based on CNNs combined with fully connected networks to account for long-term dependencies of singing voices. They showed that the proposed framework can generate natural trajectories without the use of the speech parameter generation algorithm [89]. Moreover, the training of CNNs and the generation of acoustic features are fast, because there is no recurrent structure in this architecture.

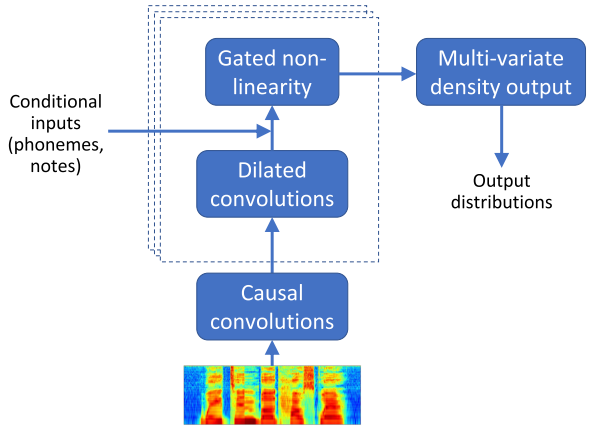

Fig. 6. Autoregressive models for singing voice synthesis [92].

The SVS methods are heavily inspired from existing TTS methods. However, the differences between singing voice and speech must be considered when designing SVS methods. First, in singing voice, there are dynamic movements in the prosody-related acoustic features, independent of the linguistic features. For example, there are dynamic and expressive movements in the F0 contour of singing voice, such as vibrato, overshoot, and fine-fluctuations [64], [68], [91]. The spectral features of singing voice are also affected by these kinds of F0 movements. However, it is difficult to model these local dynamic characteristics of acoustic features using conventional DNNs or LSTM-RNNs directly. Second, the predicted F0 contours should be consistent with the input musical notes, which cannot be guaranteed by these acoustic models for SVS. The synthetic voice may be perceived as out of tune if the predicted F0 contours deviate too much from the pitch determined by musical notes.

3) Autoregressive Prediction Models: Autoregressive architectures make predictions based on predicted past acoustic features, allowing them to better model rapid modulations in singing voice. Blaauw *et al.* [92] proposed a modified version of Wavenet for SVS called Neural Parametric Singing Synthesis (NPSS) which modeled the features produced by a parametric vocoder, rather than the raw waveform. A parametric vocoder decomposes the signal into phonetic and pitch components through which it becomes convenient to match any target melody with any lyrics, requiring less training data to sufficiently cover the entire pitch-timbre space. NPSS consisted of a neural network that takes a window of past acoustic features as input and predicts a probability distribution of acoustic features corresponding to the current time step, as shown in Fig. 6. Experimental results demonstrated that this method outperformed the HMM-based statistical parametric models and the concatenative method. The system proposed in [92] only produced timbre-related features, but did not produce features related to music expression, such as F0 and phonetic timing. In [79], an extension of the system was presented that also included F0 and phonetic timing prediction. This autoregressive approach improved reproduction of consonants and a more natural variation of predicted parameters over time, compared to statistical parametric and concatenative

systems. Moreover, this approach offered greater flexibility and more robustness to small misalignments between phonetic and acoustic features in the training data.

Yi *et al.* [93] proposed a modified deep autoregressive (DAR) model to better describe the dependencies among the acoustic features of consecutive frames. A DAR model follows the idea of feeding the target data of previous frames as additional input to a uni-directional recurrent layer. Yi *et al.* [93] extended this DAR model to predict spectral parameters with a *prenet* module consisting of multi-head self-attention layers [94] that computes a weighted combination of the historical frames and extracts high-level representations. This approach of concatenating the output of attention layers, rather than the past feature vectors, was better at predicting continuous spectral features, compared to the more discrete F0 feature. Finally, a WaveRNN vocoder was built to synthesize the waveforms of singing voice from the predicted F0 and spectral features. Subjective and objective evaluation showed that the proposed DAR method for acoustic modeling was significantly preferred over an RNN-based baseline method because of the advantages of DARs at modeling the temporal dependency of acoustic features across frames. Although the autoregressive model can achieve high quality, it suffers from exposure bias and time-consuming inference due to the forward dependency.

In these systems, the duration, F0, and spectrum models are trained independently, which usually leads to the neglect of consistency between them in the resultant singing voice. Lu *et al.* [81] proposed an SVS system called XiaoiceSing, which employed an integrated network to jointly model the spectrum, F0, and duration, assuming that the correlation between them can be expressed inside the neural network. The system was based on the TTS model FastSpeech [100] and added singing-specific design to suit the SVS task. Specifically, they added a residual connection between note pitch and the predicted F0 to reliably model the F0 of the singing voice. Moreover, during training, they added a syllable duration loss, along with the phoneme duration loss, which resulted in rhythm enhancement.

4) End-to-End (E2E) Frameworks: Many of these systems aim to train an acoustic model to predict the acoustic feature inputs to a parametric vocoder, e.g. F0 and phoneme duration. This pipeline cannot exceed the upper bound of the vocoder performance. Moreover, the need for pre-aligned training data, separate phonetic transcription, or a separate duration model is a major constraint in many of these systems, as the existing automatic alignment tools (e.g. forced alignment with an HMM) do not yield sufficiently accurate results on expressive singing, often requiring manual correction. To overcome these issues, there have been efforts to design end-to-end frameworks that directly generate a spectrogram.

Blaauw *et al.* [67] proposed a sequence-to-sequence singing synthesizer based on feed-forward Transformer, which avoids the need for training data with pre-aligned phonetic and acoustic features. The attention sub-layer of this network learns to implicitly time-align the inputs score embeddings to output spectrogram. Moreover, this framework has the advantage of faster inference time and less exposure bias than autoregressive models.

Angelini *et al.* [101] proposed UTACO, an attention-based sequence-to-sequence (AS2S) architecture, inspired from TTS Tacotron [102] framework, to implicitly model singing voice. The system has a front-end that takes a musical score as input and outputs the note embeddings (consisting of the phoneme sequence, note sequence, and duration) that are sent to an attention encoder. The output of the attention encoder is sent to the AS2S architecture, and finally the decoder produces Mel-spectrograms. The spectrograms are finally synthesized with a WaveNet vocoder. This system required considerably less explicit modeling of voice features such as F0 patterns, vibrato, and note and phoneme durations, than previous models in the literature. In a similar approach, ByteSing [96] employed a Tacotron-like encoder-decoder structure as the acoustic model, and an auxiliary phoneme duration prediction model is utilized to expand the input sequence, which can enhance the model controllable capacity, model stability, and tempo prediction accuracy. WaveRNN vocoder is adopted as the neural vocoder to further improve the voice quality of synthesized songs.

Another end-to-end sequence-to-sequence model proposed by Lee *et al.* [103] is based on the deep convolutional TTS model [104] that is known for efficient end-to-end TTS modeling. The proposed model uses content-based encoder-decoder attention with an autoregressive decoder. There is an initial alignment of the input states to the output time steps. Since Korean syllable structure has at most one onset and one coda consonant, they proposed a phonetic enhancement masking method where the first and last frames of the note are assigned to each consonant respectively, and the remaining frames are assigned to the vowel. This method produced more accurate pronunciation. Moreover, they proposed a conditional adversarial training method for the generation of more realistic singing voices.

In order to achieve high performance, the sequence-to-sequence end-to-end singing voice synthesizer involves increased complexity of the model that requires a large amount of training data from a singer to generalize well, which is difficult and expensive to collect in specific application scenarios.

5) Overcoming Over-Smoothness: DNN-based acoustic models are generally trained on a single loss criterion such as mean square error (MSE). However, the distribution of acoustic features is multimodal, as humans can sing the same lyrics in many different ways. The conventional training approaches of neural networks cannot learn to model more complex distributions of acoustic features than a unimodal Gaussian distribution. Hence, the estimated parameters tend to be over-smoothed, which leads to deterioration of the naturalness of the synthesized singing voice. Hono *et al.* [105] introduced Generative Adversarial Network (GAN) to the DNN-based singing synthesis system. GANs have achieved great success in modeling the distributions of complex data because GAN-based training is equivalent to minimizing the divergence between true data distribution and generated data distribution. Different from the vanilla GAN, [105] used the music score feature sequence as the generator input, instead of random noise. Moreover, the generator is used to model frame-wise vocoder features, which models the inter-feature dependencies within a frame of the output. They also proposed a conditional GAN (CGAN) where

the discriminator is conditioned by the musical score features. The training method based on GAN and CGAN alleviated the over-smoothing, and improved the naturalness of synthesized singing voice compared with the DNN-based method, and CGAN performed better than GAN.

Inspired by the Deep Convolutional Generative Adversarial Networks (DCGAN) architecture, Chandna *et al.* [80] proposed a novel block-wise generation network for SVS, and optimized it with Wasserstein-GAN algorithm to handle the training instability issues of GAN, called WGANSing. The network takes a block of consecutive frame-wise linguistic and F0 features, along with global singer identity as input and outputs vocoder features, and allows modeling temporal dependencies between features within each block. Temporal dependency is further modeled via autoregression through a neural vocoder [79].

Choi *et al.* [98] proposed an autoregressive conditional GAN for a Korean SVS system which uses spectrogram in a previous time step as input to produce spectrogram in the current time step. It takes advantage of the autoregressive technique to generate continuous spectrogram without abrupt temporal discontinuities. Moreover, the network employed a boundary equilibrium GAN (BEGAN) objective to generate spectrogram that uses an autoencoder for discriminator. While the original GAN matches the distributions between real and generated samples directly, BEGAN balances discriminator and generator using the autoencoder loss that allows more stable training.

All the above-mentioned GAN-based SVS systems only adopted a single discriminator directly operating on the whole sample sequence, resulting in a lack of diversity in the sample distribution assessment. To overcome this, Wu *et al.* [99] introduced multiple random window discriminators (MRWDs) into a multi-singer singing voice model. MRWDs is an ensemble of discriminators that operate on randomly sized segments of samples by using different sizes of windows, rather than the whole sample sequences. This method has a data augmentation effect that is helpful since the training data is limited for each singer. Also, Wu *et al.* [99] only focused on spectrum modeling and assumed that F0 is known. Additionally, they incorporated an adversarial loss of a singer identity classifier. Therefore, this makes the encoder independent of singer identity and lets it focus on learning a latent representation of acoustic feature.

6) Pitch Contour Fidelity: One challenge that SVS systems face is learning expressive pitch from the model itself. This is because the amount of singing voice data recorded with the corresponding musical score is limited, whereas there exist many combinations of musical factors (such as melody, note, and accent) that make pitch variation complicated. This sparseness of pitch context in a database is a challenge, i.e., the pitch of the synthesized singing voice must accurately follow the note pitch of the musical score even if the note pitch to be synthesized is outside the range of the training data.

A pitch normalization technique was proposed for F0 modeling in DNN-based SVS [88]. In this technique, the differences between the log F0 sequence extracted from waveforms and the note pitch are modeled. Some studies [81] introduced a residual connection between note pitch and the predicted F0, which is similar to pitch normalization. These techniques have the advantage of being robust to rare or unseen data and avoiding out-of-tune generation. Some systems [79], [106] utilize a data augmentation technique by pitch-shifting the training data. However, this technique requires more training time due to the increased amount of training data, and it is difficult to reproduce the voice characteristics and singing styles that change according to the pitch. A post-processing strategy has also been proposed [93]. For this strategy, F0 should be modified for each voiced segment, which may generate a discontinuous F0 contour at the edge of the voiced segment.

In [95], a non-autoregressive neural vocoder called PeriodNet [107] is adopted, which is a non-autoregressive GAN-based neural vocoder that is shown to be more robust for generating accurate pitch. Moreover, an automatic pitch correction technique is incorporated that ensures accurate pitch in the synthesized singing voices.

Another unique characteristic of singing voices is that F0 includes periodic fluctuations due to vibrato. In [75], [95], the vibrato was separated from the original F0 sequence in advance and modeled with sinusoidal parameters. The advantage of this approach is that it provides direct control over the vibrato intensity and frequency in the synthesis stage. On the other hand, the autoregressive or end-to-end frameworks, which do not use the decomposition approach, directly model the F0 sequence with the vibrato component using the neural network. In these frameworks, the direct parametric control over vibrato would not be possible in the synthesis stage, but their advantage is that they can reproduce more complex vibrato shapes that are otherwise difficult to represent by sinusoidal parameters.

B. Singing Voice Synthesis From Vocal Input

While most state-of-the-art singing voice synthesis systems generate only a fixed voice or a set of fixed voices with the lyrics and musical score as inputs, the methods for singing voice synthesis from vocal input aim to generate a personalized singing voice [68]. These include singing-to-singing and speech-to-singing synthesis methods. Originally speaking, singing-to-singing synthesis [66] means a method that automatically controls parameters of text-to-singing synthesis to imitate the vocal input. In this paper, we use its term in a broader sense, and singing-to-singing synthesis can mean a method that takes a singing voice as input and outputs a singing voice with another personality. Given this broad definition, singing-to-singing synthesis can cover singing voice conversion. Singing voice conversion is usually not regarded as singing synthesis, but this paper dares to introduce it from the viewpoint of singing synthesis.

Personalized singing renditions by singing-to-singing and speech-to-singing synthesis can serve as a reference singing for singing learners. Moreover, it can be employed to beautify the singing renditions of amateur singers, which has a strong commercial application [68]. Furthermore, this direction of research serves as a bridge between speech and singing voice analysis, providing valuable insights into the production and perception of speech and singing voices. In the following sub-sections, we summarize the techniques for singing voice

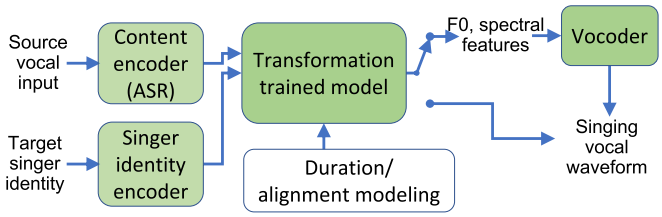

Fig. 7. Overview of singing voice conversion approaches.

conversion and speech-to-singing conversion. We discuss the singing voice conversion techniques from the perspectives of the ability of the system to map from one singer to another or many-to-many using parallel or non-parallel datasets in fully supervised or semi/unsupervised manners. The speech-to-singing conversion techniques are discussed broadly from the perspectives of template-based and model-based approaches.

1) Singing Voice Conversion: A singing voice conversion (SVC) system involves converting a source singing to a target singing by changing the timbre of the singer. SVC shares similar motivations with the conventional speech voice conversion where the person-dependent traits are transformed from source to target while the person-independent content is carried over [108]. The task of SVC, however, differs from speech voice conversion because in speech, the prosody that includes pitch, dynamics, and duration of words describes the speaker's manner of speaking or identity, whereas in singing, the prosody (i.e., musical notes) is primarily determined by the song. Therefore, the prosody-related features in speech are considered as person-dependent traits and are transformed from the source to the target speaker [109]. On the other hand, in SVC, the manner of singing is considered as rather person-independent, though this is not accurate since some singers actually have person-dependent prosodies, and only the characteristics of singer identity, such as the timbre, tend to be considered as the person-dependent traits to be converted. Hence, most of the work in SVC focuses on spectrum conversion.

a) Use of parallel datasets: The traditional methods in SVC rely on parallel training datasets, that is, different singers are required to sing the same song. With parallel training data of two singers, the spectral mapping between source and target is straightforward. However, such a setup requires large amounts of parallel singing data recordings as well as accurate time-alignment between these parallel singing voices, which are both time-consuming and expensive tasks. In addition, these SVC methods can only achieve one-to-one conversion, and the generalization ability is weak.

b) Non-parallel data and multi-singer approaches: Most recent SVC systems train a content encoder to extract content features from a source singing signal and a conversion model to transform content features to either acoustic features or waveforms, as shown in Fig. 7. One class of SVC approaches jointly trains the content encoder and the conversion model as an autoencoder model [110], [111]. Another class of SVC approaches separately trains those models [112]–[114]. These approaches train an automatic speech recognition (ASR) model as the content encoder. The conversion model can be a GAN [69], [113], [114], which directly generates waveform from the content features; or a regression model, which transforms content features to spectral features (e.g., Mel-spectrograms) and adopts an additionally trained neural vocoder to generate a waveform.

Singing voice conversion can also be viewed as a multi-singer singing voice synthesis system that should not only produce accurate singing prosody but also suitably reflect the identity of a given singer. To achieve this, methods for adding conditional inputs reflecting the singer's identity to the network have been proposed. For example, Chandna *et al.* [80] proposed a method of expressing each singer's identity by a one-hot vector. This method is straightforward and simple, but has the limitation of requiring re-training every time a new singer needs to be added.

Lee *et al.* [116] proposed a method to directly map a singing query into an embedding that defines singer's identity, and separately conditioned the timbre (linguistic content) and pitch (singing style) decoders with the encoded singer identity, while treating timbre and singing style as two independent factors, each being influenced by the singer identity.

Chen *et al.* [117] proposed a many-to-one SVC method trained on non-parallel data. They viewed the phonetic posteriors generated by a robust ASR as the singer-independent content. Then, RNN was used to model the mapping from the source phonetic posteriors to the acoustic features of the target singer. Features such as F0 and aperiodicity were extracted from the source singing voice together with the target acoustic features to reconstruct the target singing voice through vocoder.

Variational autoencoder (VAE) [119], variational autoencoding Wasserstein GAN (VAW-GAN) [120], and phonetic posteriorgram (PPG) models [112] are also investigated for non-parallel SVC. The combination of a PPG model and a waveform generator achieved promising SVC performance [113], [114]. In this approach, different models need to be trained for different target singers, and the model performance depends on the quality of an ASR engine used to extract the phonetic content.

Hu *et al.* [129] proposed a cycle-consistent generative adversarial learning model, MSVC-GAN, which can use non-parallel data to realize many-to-many SVC. Through adversarial loss, MSVC-GAN learns the acoustic feature distribution of different singers, and establishes the forward and reverse mapping among the acoustic features of different singers, that is, the model can learn a unique function for all many-to-many mappings. Moreover, Sisman *et al.* [108] proposed the CycleGAN-based conversion framework for non-parallel singing voice conversion. A caveat in these systems is that the training data needs to have sufficient number of singing examples of all the intended singing voices to appropriately model the singer identity.

c) Semi-supervised and unsupervised speaker adaptation approaches: In this context, the ability to create a new voice from a small amount of recordings, e.g., 2 min, is desirable. In such cases, a technique of voice cloning, also known as voice fitting or speaker adaptation, is used to leverage data from many speakers combined with a small amount of adaptation data from the target speaker to create a voice model that outperforms a model trained on just the adaptation target data from scratch [121]. Blaauw *et al.* [121] adopted a speaker adaptation

technique in the baseline autoregressive SVS [79] by fine-tuning speaker embedding and model weights for a few iterations on a small amount of the new target speaker's data, which is the advantage of this technique. It involves assigning a random embedding to the unseen voice, and resuming model training on data of the unseen voice so as to update this embedding and perform any necessary refinements to the model.

Bonada *et al.* [122] proposed a semi-supervised singing synthesizer which effectively disentangles singer identity from the linguistic content, thus the network learns new voices from audio data only, without any annotations such as phonetic segmentation. It consists of an encoder-decoder model with two encoders, linguistic and acoustic. The system is first trained in a supervised manner, using a labeled multi-singer dataset. Then, to learn a new voice in an unsupervised manner, the pre-trained acoustic encoder is used to fine-tune the decoder for the target singer. Finally, at inference, the pre-trained linguistic encoder is used together with the decoder of the new voice, to produce acoustic features from the linguistic input.

Nachmani *et al.* [110] proposed a WaveNet autoencoder for unsupervised singing voice conversion. The model neither requires parallel training data between singers, nor does it need to transcribe audio into text or musical notes. This approach adopts an adversarial speaker classifier to disentangle singer information from the encoder output. To further improve this method, an additional pitch adversarial mechanism is added to remove pitch information from the encoder output [111].

Ren *et al.* [123] built a complete pipeline for training a multi-lingual and multi-singer SVC system from scratch, called DeepSinger, using singing training data mined from music websites. The pipeline consisted of data crawling, singing voice separation, lyrics-to-singing alignment, data filtration based on confidence of alignment step, and singing modeling using Transformer-based FastSpeech model [100] to directly generate linear-spectrogram from lyrics, instead of the traditional acoustic features in parametric synthesis. Additionally, a reference encoder is designed as part of the singing model to capture the timbre of a singer from noisy singing data. DeepSinger thus synthesized high-quality singing voices in multiple languages and multiple singers using data directly mined from the web, without any high quality singing data. A reference audio (singing or speech) of any target singer should also be provided to extract the voice characteristics such as timbre for the synthesized singing voice. A drawback of this pipeline is that the pitch of this mined data may not always be correct as it may contain out-of-tune data, which would impact the quality of a singing voice.

Nercesian [124] proposed to use a speaker embedding network together with the WORLD vocoder to perform zero-shot SVC, and designed two encoder-decoder architectures to realize it. The speaker embedding network can adapt to the new voice on the fly, train the unlabeled data model, and conduct the initial training on a widely-available large speech dataset, followed by model adaptation on a smaller singing voice dataset.

2) Speech-to-Singing Conversion: Although data-efficient singing voice synthesis methods [121] and unsupervised singing voice conversion methods [110] can effectively generate new sounds, they still need to obtain a small number of singing voice samples from the target speaker, which limits the application of singing voice synthesis to relatively limited scenarios, that is, the singing voice of the target speaker must be available. However, speech samples are much easier to collect than singing voices. Thus, a speech-to-singing conversion system is designed to take the spoken utterances from a speaker as input, manipulate the parameters of the speech to match with the prosody of the song, and finally the transformed parameters are used to generate the output singing voices while retaining the linguistic content and the speaker's identity [65].

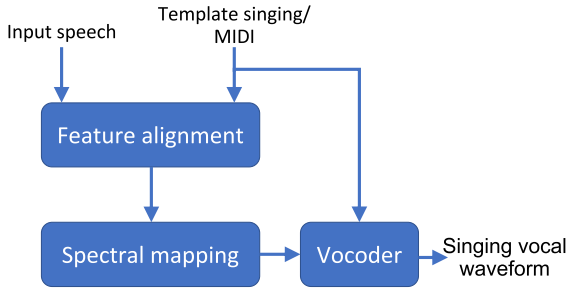

Fig. 8. Overview of speech-to-singing conversion systems.

An overview of the speech-to-singing conversion systems is shown in Fig. 8. They are broadly classified into two categories, the template-based conversion [125] and the model-based conversion [65], depending on how the reference prosody of singing is generated. The template-based approach uses reference prosody as a template that is extracted from high-quality singing. The model-based approach, on the other hand, generates singing prosody by imposing prosody control with the help of the digital score of the song (MIDI). Both approaches need temporal alignment between the reference prosody (good singing voice or MIDI) and the input speech, in terms of musical rhythm and melody. The synchronization information obtained from the temporal alignment is used for the subsequent frame-level parameter conversion. The parameter conversion method should transform the prosodic and spectral characteristics of the input speech into that of singing voice, while retaining the phonetic content of the speech and its speaker's identity. The temporal alignment between the input speech and the reference becomes a crucial step for accurate speech-to-singing conversion since the alignment errors are perceived as distortions in the synthesized singing [68].

For the template-based approach, various alignment schemes have been explored to time-align the input speech to reference singing rendition. For example, in the dual alignment scheme [130], [131], instead of directly aligning the input speech to the target singing, a target speech is first aligned to the target singing, and then the input speech to the target speech. This approach splits the burden of mismatch between input and target speaker identity, and between speech and singing, into two stages. However, such a multi-step approach accumulates errors from different steps. For the model-based approach, on the other hand, the lyrical words are already aligned to the musical notes. Therefore, once the input speech is aligned with the lyrical words

TABLE IV
SUMMARY OF SINGING VOICE SEPARATION TECHNIQUES SINCE 2017

Architecture	Approach	Prediction	Codebase	Extra Training Data?	Vocals SDR
DeepConvNet (2017) [143]	Data-Driven	Spectrogram	https://github.com/MTG/DeepConvSep/	No	2.37
Wave-U-Net (2018) [151]	Data-Driven	Waveform	https://github.com/f90/Wave-U-Net-Pytorch	No	3.25
Wavenet (2018) [152]	Data-Driven	Waveform	https://github.com/francesclluis/source-separation-wavenet/	No	3.35
Spect-U-Net (2018) [153]	Data-Driven	Spectrogram	https://github.com/s603122001/Music-Source-Separation	No	5.74
MMDenseLSTM (2018) [154]	Data-Driven	Spectrogram	-	No	6.60
				804 songs	7.24
Open-Unmix (2019) [136]	Data-Driven	Spectrogram	https://github.com/sigsep/open-unmix-pytorch	No	6.32
Spleeter (2020) [135]	Data-Driven	Spectrogram	https://github.com/deezer/spleeter	25k songs†	6.86
Meta-Tasnet (2020) [155]	Content-Informed	Waveform	https://github.com/pfnet-research/meta-tasnet	No	6.40
DPRNN (2020) [156]	Data-Driven	Waveform	https://github.com/facebookresearch/svoice	No	6.92
Meseguer-Brocal et al. [157] (2020)	Content-Informed	Spectrogram	https://github.com/gabolsgabs/vunet	NA	NA
Schulze-Forster et al. (2019) [158]	Content-Informed	Spectrogram	https://github.com/schufo/wiass	NA	NA
D3Net (2021) [159]	Data-Driven	Spectrogram	https://github.com/sony/ai-research-code/tree/master/d3net	No	7.24
				1.5k songs	7.80
Conv-Tasnet (2021) [160]	Data-Driven	Waveform	https://github.com/facebookresearch/demucs/tree/v2	No	6.43
				150 songs	6.74
Demucs (2021) [160]	Data-Driven	Waveform	https://github.com/facebookresearch/demucs	No	6.84
				150 songs	7.29
LaSAFT-GPoCM (2021) [161]	Content-Informed	Spectrogram	https://github.com/ws-choi/Conditioned-Source-Separation-LaSAFT	No	7.33
X-UMX (2021) [162]	Data-Driven	Spectrogram	https://github.com/sony/ai-research-code/tree/master/x-umx	No	6.61
ByteMSS (2021) [163]	Data-Driven	Spectrogram	https://github.com/bytedance/music_source_separation	No	8.98
Mimilakis et al. (2021) [164]	Data-Driven, Unsupervised	Spectrogram	https://github.com/Js-Mim/rl_singing_voice	NA	NA

†Each track is 30 seconds.
The separation performance of each architecture is indicated in terms of signal-to-distortion ratio (SDR) (dB) on the test set of MUSDB18 dataset [150]. If training is done on extra data, other than the train set of MUSDB18, then that is indicated under the column "Extra Training Data?". If the test data is not MUSDB18 dataset or the reported result is not comparable, the column "Vocals SDR" is NA.

through forced alignment of an ASR system, it is also aligned with the musical notes [65].

Parameter conversion in both template-based and model-based approaches is interpreted in terms of the source-filter model [132], where the *source* characteristics represent the prosody of the song, while the *filter* characteristics represent the speaker's identity and phonetic content. The F0 contour, representing excitation source parameters, is extracted from the template or generated by the model, while the spectral envelope, representing vocal tract filter parameters, is extracted from the input speech to preserve the speaker identity in synthesized singing, except that these spectral characteristics have to be modified to resemble those of singing voices.

The template-based approach benefits from the natural prosody obtained from the reference singing voice, whereas the model-based approach lacks the expressiveness of natural singing. On the other hand, the model-based approach is more scalable than the template-based approach because recording high quality singing for every possible song is tedious, while preparing digital scores for songs is relatively easy.

Recently, Zhang *et al.* [127] proposed an algorithm to directly synthesize high-quality target speaker singing voice by learning speech features from normal speech samples. The key was to integrate both speech and singing voice synthesis into a unified framework and learn the universal speaker embeddings that can be shared between both speech and singing voice synthesis tasks. They proposed DurIAN-4S, a speech and singing voice synthesis system based on the previously proposed autoregressive generation model DurIAN. The entire model is trained together with the learnable speaker embedding as the conditional input of the model. By selecting a different speaker embedding in the process of singing voice generation, the model can generate different singing voices. Parekh *et al.* [128] explored a method to achieve speech-to-singing conversion by employing minimal additional information over the melody contour.

C. Future Directions

Although the quality of synthesized singing voices has improved in recent years, the generated voices are still far from reflecting emotions in singing voices as talented professional singers do. Research on singing voice synthesis is thus not mature at all and there is much room for improvement in many directions. Multilingual singing voice synthesis and conversion is another interesting direction.

Currently, much emphasis is given to the generation of natural synthesized voices. However, the potential of singing synthesis technologies themselves is not limited to the naturalness of the generated sound. Digital sound synthesizers have been shown to be capable of synthesizing various novel but attractive musical sounds, which have already been widely used and accepted in music cultures. Such creative uses of singing synthesizers are

yet to be explored, though many creators using the VOCALOID singing synthesizer (e.g. *Hatsune Miku*) have shown creative uses in their songs since 2007, such as songs having extremely high speed/frequency singing.

In addition, there is a need for creating a standard benchmark and open test datasets as well as a need for investigating objective evaluation metrics. For example, the existing objective metrics such as MCD were originally designed for speech synthesis tasks, and may not capture all essential aspects of singing voice.

V. Singing Voice Separation

Music recordings consist of mixtures of multiple sound objects, such as lead vocal and different accompanying instruments. Therefore, manipulation of individual sound objects requires the separation of the audio mixture into several tracks (sound sources). This general task is called audio source separation or music source separation. This section focuses on singing voice separation, a more focused task of extracting the singing voice signal from a music recording. An overview paper by Rafii *et al.* [133] in 2018 provides a detailed and comprehensive review on the last 50 years of research on the topic of lead and accompaniment separation in music. In this section, we provide an overview of recent data-driven methods of singing voice separation based on deep learning.

The data-driven methods exploit large databases of audio examples where both the isolated lead and accompaniment signals are available, and leverage on machine learning methods to learn how to separate the singing voice from the mixture. Typical spectrogram-based methods estimate a mapping between the audio mixture and either the time-frequency (TF) mask for separating the sources, or the corresponding spectrograms. They learn to predict a power spectrogram for each source and reuse the phase from the input mixture to synthesise individual waveforms. These methods decompose music into its constituent components using multiple non-linear layers to learn the optimal hidden representations or masks from data in a supervised setting, either in a purely data-driven approach [134]–[136] or in combination with music information [137], [138].

In recent years, a level of maturity has been achieved in the task of singing voice separation as seen by the performance of methods in the Signal Separation Evaluation Campaign (SiSEC) 2018 [139] and the wide use of open source music source separation implementations such as Open-Unmix [136] and Spleeter [135]. This progress has activated other closely related MIR tasks that use these methods as a pre-processing step, such as singing pitch estimation, singing melody transcription, singer identification, cover song detection, and lyrics synchronization and transcription, though most of them have already been tackled by using traditional singing voice separation [8].

The data-driven methods are classified into pure data-driven methods and content-informed data-driven methods. Table IV shows a summary of those methods introduced in the followings.

A. Pure Data-Driven Methods

We first review various purely data-driven methods from the perspectives of estimating the time-frequency (TF) masks in spectrogram-based methods, estimating the waveform directly, exploring methods to handle the lack of training data, exploring adversarial networks for source separation, estimating multi-scale and multi-resolution features, and involving phase estimation in spectrogram-based methods. An overview block diagram of data-driven singing voice separation methods is shown in Fig. 9.

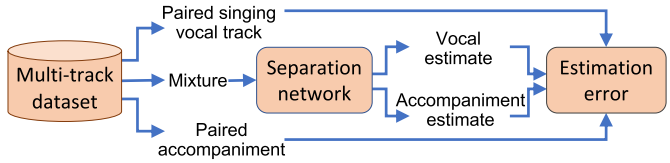

Fig. 9. Overview of supervised singing voice separation.

1) Estimating Time-Frequency Masks: The most adopted solution for audio source separation is the estimation of a time-varying and source-dependent filter, i.e., the usage of time-frequency (TF) masking, which is applied to the mixture. One of the earliest works that proposed deep neural networks for singing voice separation was by Huang *et al.* [140], [141], in which the objective was to separate singing voice and accompaniment music simultaneously by optimizing TF-masks with a network which models all the sources directly. In order to model the temporal structure and long-term dependencies of the sources while estimating the TF-masks, various neural network architectures have since been explored, such as fully connected feedforward network (FFN) [142], CNN [143], RNN [140], [141], LSTM [144], and GRU [145].

The most widely-used strategy to achieve the estimation of individual sources employs time-frequency (TF) filters or masks in the STFT domain, which are derived from rational models which incorporate prior information about the spectral representation of each source in the mixture [146]–[148]. In contrast, another approach is to estimate the TF-mask within the learning process on the basis of prior knowledge of the target source. Huang *et al.* [141] proposed a bidirectional GRU architecture along with skip-filtering connections to separate the lead vocal from the background music, which was also adopted by Mimilakis *et al.* [145]. Such skip connections allow the input spectrogram to propagate through the network to operate on intermediate representations within the network, and force the network to learn the TF-mask through optimization, based on the prior knowledge of the magnitude spectrogram of the target source and not on the prior knowledge of an ideal TF-mask. Jansson *et al.* [134] proposed a U-Net architecture, and Park *et al.* [149] proposed a stacked hourglass network in which the input spectrogram is compressed by the encoder into a bottleneck layer to obtain a lower dimensional descriptor and then the descriptor is re-expanded to the size of the target spectrogram by the decoder, as shown in Fig. 10. In addition, with the help of additional skip connections (or similar structures) between layers at the same hierarchical level in the encoder and decoder, these networks allow low-level spectral information to flow directly from the high resolution input to the high-resolution output which helps the reconstruction by providing finer details in the decoding

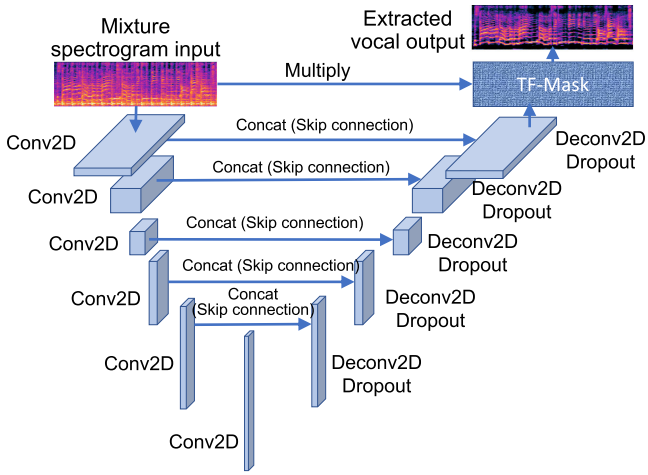

Fig. 10. Example of the U-Net architecture for singing voice separation [134].

directly from the different layers of the encoder (which would be otherwise progressively lost during encoding). The output of the last layer is responsible for masking the input mixture spectrogram.

A reproducible spectrogram-based baseline called Open-Unmix has been released by Stöter *et al.* [136]. It consists of a bidirectional LSTM with skip connections to predict the TF-mask, and shows competitive performance in SiSEC 2018 evaluation campaign [139] amongst systems trained only on the 100 songs open dataset called MUSDB18 [150]. MM-DenseLSTM, a multi-band dense net with LSTMs at different scales of the encoder and decoder, proposed by Takahashi *et al.* [154] was trained on 807 proprietary songs and held the absolute record of signal-to-distortion ratio (SDR) in the SiSEC 2018 campaign [139]. Hennequin *et al.* [135] proposed another U-Net architecture for spectrogram masking, called Spleeter, that is trained on the unreleased Bean dataset [165] composed of short excerpts from nearly 25,000 songs. Spleeter and Open-Unmix provide comparable and strong performance and have been widely adopted by the digital music industry. More recently, Takahashi and Mitsufuji [159] improved the spectrogram-domain state-of-the-art performance with D3Net that uses dilated convolutions with dense connection.

In many cases, the masking process is not a learnable function or is not encapsulated into the deep learning optimization. Consequently, most of the existing spectrogram-based methods rely on a post processing step using the generalized multi-channel Wiener filtering. Nugraha *et al.* [166] showed that Wiener filtering is an efficient post-processing step for spectrogram-based models and is commonly used by all top-performing systems of this category. On the other hand, Mimilakis *et al.* [167] proposed a method that learns and optimizes a source-dependent mask at the time of training and does not need the aforementioned post processing step. Lin *et al.* [168] proposed an inclusion of an ideal binary mask as the target label along with a cross-entropy training objective in a CNN architecture. This target mask learning step again allows eliminating the Wiener filter post-processing step.

2) Direct Estimation of Waveform: Stoller *et al.* [151] proposed Wave-U-Net, an end-to-end network that uses a similar topology as the U-Net with skip connections [134] but works directly on the time-domain audio waveform, therefore avoiding the problems related to reconstruction of the audio signal from the magnitude spectrogram without the phase information. Similarly, Defossez *et al.* [160] introduced Demucs, a U-Net architecture for music source separation in the waveform domain, that consists of a convolutional autoencoder with skip connections, similar to Wave-U-Net, but they used transposed convolutions rather than linear interpolation followed by a convolution with a stride of 1 because for the same increase in the receptive field, transposed convolutions require 4 times less operations and memory. Thus, this architecture enabled simultaneous estimation of multiple sources without running out of memory and outperformed other waveform-based methods.

3) Handling the Lack of Data: The success of U-Net and Wave-U-Net models depends on the availability of a large amount of proprietary training data. Cohen-Hadria *et al.* [169] investigated data augmentation techniques such as pitch-shifting and time-scaling on publicly available smaller datasets and compared the performance of U-Net and Wave-U-Net models. In order to avoid overfitting, besides using data augmentation, Perez-Lapillo *et al.* [170] employed an additional regularization term in the loss function, called the minimum hyperspherical energy for the Wave-U-Net architecture, where the diversity of neurons is promoted by minimizing the hyperspherical energy of the neurons in each layer. Since convolutional encoder-decoder frameworks are sensitive to the sound level of the input, Lin *et al.* [171] showed that a combination of data augmentation, frame normalization, and zero-mean convolution makes the network sound-level invariant.

Mimilakis *et al.* [164] successfully used unpaired vocal and accompaniment data to learn non-negative, smooth representations with a denoising autoencoder with the help of an unsupervised objective. Seetharaman *et al.* [172] proposed a stage-wise teacher-student algorithm where a clustering-based labeler (teacher) assigns time-frequency bin labels with a confidence measure, and a separator network (student) is trained on these labels afterward. Recently, Wang *et al.* [173] presented a semi-supervised singing voice separation method that uses a noisy self-training framework. Self-training is a semi-supervised framework in which a pre-trained teacher network assigns pseudo-labels for unlabeled data. Then a larger student network is trained with the self-labeled dataset combined with the dataset with labeled ground-truth.

Michelashvili *et al.* [174] proposed a semi-supervised singing voice separation network, in which the training data contains a set of samples of mixed music (singing voice and instrumental) and an unmatched set of instrumental music. Therefore, neither matched singing voice nor accompaniment are available during training. Their solution was to learn a function that maps a mixture to the instrumental component, such that at run-time, the vocal audio is computed by subtracting the estimated instrumental component from the mixture. During training, synthetic mixed samples were created by mixing reconstructed singing voices with random instrumental samples. Additionally, two

discriminators were trained: one for enforcing the distribution of the estimated samples from instrumental domain to match the distribution of the instrumental training set, and the other for enforcing the distribution of the mixed synthetic samples to match the distribution of mixture training set.

Kang *et al.* [175] proposed a singing voice separation approach based on the curriculum learning framework, in which learning is started with only easy examples and then the task difficulty is gradually increased. They define easy examples as the ones in which one source is obviously dominant over the other, where the dominance factor depends on the relative intensity of vocals and instruments.

4) Using Adversarial Networks: Fan *et al.* [176] introduced GANs for singing voice separation. They proposed a strategy to first estimate the initial weights of the generator by a supervised training of the generator that takes a mixture spectrogram as input and predicts the separated vocal and background music spectrograms. The generator is further trained as an unsupervised GAN where the discriminator detects if the combination of the predicted vocal and background music tracks is real or fake compared to the original mixture. Moreover, the authors included an additional trainable layer in the generator that does the TF-masking process implicitly and predicts the resultant spectrogram, instead of predicting the mask itself. Stoller *et al.* [177] proposed a semi-supervised GAN that makes use of a large amount of available unlabeled music tracks as well as datasets of solo source instrument recordings. One discriminator network per source is trained to distinguish separator estimates made on the unlabeled music from real samples taken from the respective source dataset. The separator aims to output more realistic sources as judged by the discriminators, in addition to minimizing the supervised loss on multi-track data. With only a few multi-track datasets available, often extensive data augmentation is used to combat overfitting [144], [178]. Mixing random tracks, however, can reduce separation performance in a supervised setting as instruments in real music are strongly correlated with the leading vocals. Stoller *et al.* [177] showed that using unpaired source and mixture recordings in a GAN training setting benefits from data augmentation without the drawbacks of creating unrealistic music mixtures.

5) Multi-Scale, Multi-Resolution Features: Music relies heavily on its structural repetitions to build logical structure and meaning in a song. These repetitions are made of recurring elements at different time scales, from small duration basic elements such as individual notes, to larger structures, such as chords [179]. These multi-scale repetitions effectively distinguish the musical accompaniment from the vocals which are less redundant and mostly harmonic [133]. Therefore, effectively modeling the repetitive structures in the mixture signal would be a promising solution for DNN-based singing voice separation. Since the repetitive structures in music can be observed as the similarities between different regions in the TF representations, the separation network needs to attend to the different TF regions in order to capture the dependencies across different frequencies in the mixture. Multi-resolution CNNs, which can capture multi-resolution features by constructing various sized receptive fields allowing a large context to model, both in time and frequency [134], [143], [151]. However, the convolution operator in the convolutional encoder-decoder networks [134], [143], [149], [151], which has a local receptive field, can only model the repeating patterns locally. Even cascaded convolutional layers with skip connections cannot effectively capture the dependencies across multiple layers for repetitive structures [180]. Yuan *et al.* [181] proposed a skip-attention mechanism wherein, instead of directly feeding features from different layers from encoder to decoder through skip connections, an inter-attention layer is used to make this connection. This attention layer accounts for a longer context, therefore modeling the repetitive structures in the musical source. It was noted that the optimization algorithm becomes less effective in capturing the dependencies across multiple layers in the cascaded structures and it requires too many layers to cover a sufficiently large input field to model global information, making the network training too difficult [180].

Another kind of multi-resolution CNN directly implements multi-resolution receptive fields in the same layer using multiple sets of various-sized convolutional operators, such as multi-resolution convolutional autoencoder [182] and multi-resolution fully CNN [183]; thereby extracting multi-resolution features without deepening the cascade structure, unlike the U-Net [134] or stacked hourglass networks [149]. The advantage of this kind of multi-resolution CNN over cascaded structures is that the problem of the optimization algorithm not being effective in capturing the dependencies across multiple layers does not affect. On the other hand, a minor linear shift in TF representations could cause significant distortions in vocal and music perception which affects these types of multi-resolution CNNs, while the cascaded structures use skip connections to transmit low-level information between different layers [134], [151], [181] to overcome this problem.

D3Net by Takahashi and Mitsufuji [159] uses multi-dilated convolution that has different dilation factors in a single layer to model different resolutions, i.e., local and global information simultaneously. Yuan *et al.* [184] proposed a solution that incorporates multi-resolution CNN while decreasing the computational complexity of the model. They implemented a multi-resolution pooling CNN, which uses various-sized pooling operators to extract multi-resolution features, and then used Neural Architecture Search to use an evolutionary mechanism to allow the network to automatically search for effective multi-resolution pooling CNN structures. They explored optimizing this network in terms of a single objective taking into account only separation performance as well as multiple objectives taking into account both separation performance and model complexity.

6) Involving Phase Estimates: The spectrogram-based separation methods often suffer from incorrect phase reconstruction which degrades the performance. Kong *et al.* [163] showed the effectiveness of estimating the phase by estimating complex ideal ratio masks where they decouple the estimation of these masks into magnitude and phase estimations. Moreover, they proposed a deep residual U-Net architecture with up to 143 layers.

B. Content-Informed Data-Driven Methods

There has been a growing interest in including additional information in the separation network for making it more robust to challenging conditions because articulation (e.g. lyrics) as well as prosody (e.g. musical score) are closely related to different components of a music audio mixture. In the following sub-sections, we discuss various music source separation networks that use additional side information under two categories, lyric-informed and other musical-attribute-informed (musical score, instrument label, or vocal characteristics).

1) Lyrics-Informed: The lyrics of a song have a promising possibility to be used as additional information for singing voice separation since lyrics are closely related to singing voice. Chandna *et al.* [185] trained an encoder via knowledge distillation to learn a content embedding that contained implicit phonetic information. From this embedding, a decoder estimated features which, along with F0 estimates, were then resynthesized into a time domain voice signal estimate by a vocoder. The results show that the intelligibility of synthesized vocals is improved through phonetic features, but the overall subjective audio quality is lower than purely data-driven methods. Takahashi *et al.* [186] used linguistic features extracted from an end-to-end ASR model as additional information for source separation networks. It resulted in robust performance in noisy conditions. An advantage of these approaches is that alignment between audio and text is not required because the phonetic information is extracted directly from the mixtures. However, this phonetic information is only implicit and the mixture remains the only source of information.

Jeon *et al.* [187] conditioned singing voice separation on lyrics that were manually aligned at syllable level. They use a deep text encoder and evaluated on a private dataset of Korean amateur solo singing recordings that were mixed with unrelated accompaniments.

Meseguer-Brocal *et al.* [157] used lyrics transcripts aligned at word level to condition U-Net based singing voice separation network, and improvements over the classic U-Net are reported. Parameters are successfully estimated from the words to transform deep features in the U-Net encoder. However, it is not clear whether the improvement is caused by the higher number of parameters in the conditioned U-Net, or the voice activity information inherent in aligned text, or by the phonetic information.

Schulze-Forster *et al.* [158] adapted the attention mechanism for allowing the use of weakly-labeled side information by learning lyrics alignment as a byproduct. Schulze-Forster *et al.* [188] further considered aligned phoneme sequences from lyrics transcripts, in contrast to word sequences in [157], as an additional input to the separation model. They proposed a sequential framework where a phoneme alignment network is followed by a modified Open-Unmix architecture for singing voice separation such that it takes the aligned phoneme information as an additional input. They show that this text information helps in preserving spectral properties of the articulated phonemes in the separated voice signals.

2) Musical-Attribute-Informed: Chandna *et al.* [189] trained a neural network to estimate vocal-specific vocoder parameters (instead of directly estimating the waveform or spectrogram) from a mixture signal. Those parameters can be used to synthesize the singing voices using a vocoder. Since a vocal-specific synthesis approach is used, the estimated vocal track has no direct interference from the backing track, leading to a lower distortion factor. However, since it is a resynthesis of the voice signal instead of a TF-masking approach, the perceived quality and intelligibility are not as high as that one can achieve from the TF-masks. Similarly, Swaminathan *et al.* [190] used frame level vocal activity information as an augmented feature input to a singing voice separation network, and found that the network informed with vocal activity learns to differentiate between vocal and non-vocal regions and reduces interference and artifacts.

Singing voice separation and vocal F0 estimation are tightly related tasks. The outputs of singing voice separation systems have been used as inputs to vocal F0 estimation systems; conversely, vocal F0 has been used as side information to improve singing voice separation. Nakano *et al.* [137] proposed a multi-task learning approach to jointly perform F0 estimation and separation of singing voices from music signals. It consists of a deep convolutional neural network for vocal F0 saliency estimation and a U-Net with an encoder shared by two decoders specialized for separating vocal and accompaniment parts. Between these two networks, a differentiable layer that converts an F0 saliency spectrogram into harmonic masks indicating the locations of harmonic partials of a singing voice is introduced. The harmonic masks derived from the F0 estimates are further refined through the U-Net and the whole network is trained jointly. Similarly, Jansson *et al.* [138] proposed an architecture for jointly separating vocals and estimating the F0, and showed that joint learning is advantageous.

Petermann *et al.* [191] conditioned the U-Net separation network with pitch for separating choir ensemble. Choral ensembles consist of simultaneous and harmonic singing which leads to high structural similarity and overlap between the spectral components of the sources in a choral mixture, making source separation for choirs a harder task. They showed that the network with the additional pitch information surpasses pitch-agnostic network.

Music source separation networks generally use supervised learning where the mixture signals and the isolated sources (which typically include vocals, bass, drums, and others) are available for training, and build dedicated models for each task to isolate. Recently, there have been some approaches to build unified music source separation model, where one can choose to separate a particular instrument or singing voice with a conditioning vector. Meseguer-Brocal *et al.* [192] proposed a Conditioned-U-Net (CU-Net) which adds a control mechanism to the standard U-Net that decides the music source (vocals, bass, drums, or others) to be isolated, according to a one-hot-encoding input vector. A feature-wise linear modulation (FiLM) layer is then used to modulate the intermediate representations with the conditioning vector, where the parameters of the FiLM

layer are jointly learnt with the main network. Thus, with CU-Net, vocal separation, along with other instrument separations, could happen with a single separation model achieving comparable performances as the dedicated ones at a lower cost.

Choi *et al.* [161] incorporated the idea of source-based conditioning [192] in the time-distributed U-Net framework [193] such that a source-attentive frequency transformation block can capture the source-dependent frequency patterns. They also proposed a Gated Point-wise Convolutional Modulation layer that extends the concept of FiLM layer [192] by incorporating inter-channel interactions. This CU-Net extension is shown to outperform the other existing methods on singing vocals extraction task on MUSDB18 dataset.

C. Future Directions

The singing voice separation systems have been mostly evaluated on pop and rock songs. Their performances are likely to vary for other music genres. Further investigations are required for different genres and singing styles. Moreover, some songs contain multiple simultaneous singing voices, such as duets and unison singing. Separating each of those singing voices with singer identification is challenging and needs more investigation.

Singing voices may be recorded in different kinds of background conditions, for example, through karaoke singing apps, where traffic noise, other irrelevant music, other people's voices, etc. are present. Singing voice separation for such widely different recording conditions will bring interesting new challenges and applications.

VI. Lyrics Synchronization and Transcription

Lyrics are an important component of singing voice, and various research directions have been explored to give machines the ability to understand lyrics in different ways [194]. Lyrics synchronization, also known as lyrics-to-audio alignment, is the task of finding time boundaries of the lyrical units (words, phonemes) of given lyrics with a given polyphonic or solo singing audio. Lyrics transcription, also known as lyrics recognition, is the task of recognizing the sung lyrics from audio. Both of these tasks are closely related. In fact, lyrics synchronization can be considered as a sub-task of a lyrics transcription system, where the sequence of words in the output of the transcription system is fixed to the given lyrics, and the job of the recognizer is to find the optimal path between the audio and the given lyrics, i.e., the time stamps of the lyrical units. Lyrics transcription, on the other hand, needs to transcribe the lyrics of the singing voices in a song given the audio alone. Thus, lyrics synchronization is easier than lyrics transcription.

In the last decade, there has been considerable interest in digital music services and applications with lyrics-related features, such as displaying lyrics of songs synchronized with their audio [10], [195] and searching music using lyrics or their topics [196]. Although some well-known hit songs already have lyric time stamps within existing karaoke databases, new or not-so-popular songs, amateur-based cover songs, and live recordings may not. An automatic lyrics synchronization system could reduce the huge amount of time and labor required to manually construct such time stamps. Additionally, lyrics synchronization as well as lyrics transcription could also be used as a front-end for many content-based MIR applications [8] such as karaoke lyrics display, query-by-singing, and keyword spotting, as well as applications involving lyrical rhymes and emotions.

In this overview, we will provide a historical perspective of techniques explored for lyrics synchronization and transcription, and the recent developments in lyrics modeling in solo-singing and polyphonic music, where lyrics modeling includes both lyrics synchronization and transcription. A summary of recent techniques is presented in Table V.

A. Historical Perspective

A straightforward adaptation of automatic speech recognition (ASR) techniques for singing voice is difficult [10], [197] since singing has a higher degree of variation in pronunciation and prosody than speech [198] and is typically accompanied with interfering background music. Lyrics transcription of singing voice is thus more challenging than ASR, and since it is sometimes too difficult, research on lyrics-to-audio alignment has been much pursued given the text of lyrics as prior information.

When precise word-level lyrics-to-audio alignment was not necessary, early studies exploited the knowledge of the musical structure of the song to align the lyrics [199], [200]. LyricAlly [200] used the structural information of popular songs to align lines or sections of lyrics to the music audio.

To achieve finer word-level or phoneme-level lyrics-to-audio alignment in clean or separated singing voices, early lyrics-to-audio alignment systems used hidden Markov model (HMM) based acoustic models which are utilized to extract frame-level phoneme posterior probabilities. Then a forward-pass decoding algorithm called forced alignment was applied on these posteriograms, obtaining word and phoneme alignments [10], [11], [197], [201]–[203]. In forced alignment, vocal audio and its corresponding lyrics text are automatically aligned at the word and phoneme levels, with the help of the acoustic model and a pronunciation lexicon that converts graphemes to phonemes. Due to lack of large annotated datasets, existing ASR acoustic models trained on speech voices were usually applied to singing voices. In using such acoustic models, it was important to reduce the mismatch between speech and singing signals by adapting the speech acoustic models with a small amount of singing data using maximum a posterior (MAP) or maximum likelihood linear regression (MLLR) [10], [197], [201], [203].

B. Lyrics Modeling in Solo-Singing

In 2017, Smule made their solo-singing karaoke dataset called Digital Archive of Mobile Performances (DAMP)[2] available for research purposes. This dataset is collected via a karaoke app, and therefore has inconsistent recording conditions, bleeding of

[2]https://ccrma.stanford.edu/damp/

TABLE V
SUMMARY OF LYRICS SYNCHRONIZATION AND TRANSCRIPTION METHODS

Paper	Alignment/ Transcription	Approach	Codebase	Performance	
				Solo	Poly
Kruspe [204][†1] (2017)	Alignment	DNN-HMM trained on publicly available solo-singing data, no vocal extraction during testing	NA	A: 2.87s	A: 9.03s
Dzhambazov [205][†1] (2017)	Alignment	Extraction-transcription, HMM-based, trained on solo-singing	NA	A: 4.46s	A: 11.64s
Gupta [203][†2] (2018)	Alignment	GMM-HMM trained on publicly available solo-singing data, no vocal extraction during testing	https://github.com/chitralekha18/lyrics-aligned-solo-singing-dataset	A: 2.66s	A: 7.01s
Wang [206][†2] (2018)	Alignment	GMM-HMM, direct modeling on proprietary polyphonic music data	NA	A: 0.35s	A: 4.12s
Stoller [207][†3] (2019)	Both	Direct modeling on proprietary polyphonic data, end-to-end framework	NA	A: 0.21s	A: 0.26s, T: 70.9%
Gupta [208][†3] (2019)	Both	Direct modeling on publicly available polyphonic data, TDNN-F based	https://github.com/chitralekha18/AutoLyrixAlign Demo: https://autolyrixalign.hltnus.org/	A: 0.13s	A: 0.19s, T: 44.0%
Dabike [209] (2019)	Transcription	TDNN-F based, trained and tested on publicly available solo-singing data	https://github.com/groadabike/Kaldi-Dsing-task	NA	NA
Zhang[210][†4] (2020)	Alignment	Extraction-transcription, trained on proprietary polyphonic data, DNN-HMM models	NA	A: 0.11s	A: 0.41s
Demirel [211] (2020)	Transcription	Convolutional time delay neural network with self attention, trained and tested on publicly available solo-singing data	https://github.com/emirdemirel/AutomaticLyricsTranscription-with-Self-Attention	NA	NA
Demirel [212][†4] (2020)	Both	Extraction-transcription, trained on publicly available solo-singing data, TDNN-F with recursive search method	NA	A: 0.93s, T: 37.02%	A: 0.61s, T: 79.39%
Gao [213][†4] (2020)	Both	Direct modeling on publicly available polyphonic data, TDNN-F based	NA	A: 0.09s, T: 41.86%	A: 0.19s, T: 47.25%
Vaglio [214] (2020)	Alignment	Extraction-transcription, trained on end-to-end framework for multi-lingual lyrics alignment	https://github.com/deezer/MultilingualLyricsToAudioAlignment	NA	A: 0.22s
Demirel [215] (2021)	Transcription	Direct modeling with a publicly available solo-singing and polyphonic data, multi-stream TDNN architecture	https://github.com/emirdemirel/ALTA	NA	T: 37.33%

[†1] https://www.music-ir.org/mirex/wiki/2017:Automatic_Lyrics-to-Audio_Alignment_Results.
[†2] https://www.music-ir.org/mirex/wiki/2018:Automatic_Lyrics-to-Audio_Alignment_Results.
[†3] https://www.music-ir.org/mirex/wiki/2019:Automatic_Lyrics-to-Audio_Alignment_Results.
[†4] https://www.music-ir.org/mirex/wiki/2020:Automatic_Lyrics-to-Audio_Alignment_Results.
Reported performance on Hansen's a capella test dataset (solo) and Mauch's polyphonic test dataset (poly) (that are used in MIREX challenges), if available, are indicated. The performance indicator for lyrics synchronization (A) is average boundary error (in seconds) and for lyrics transcription (T) is WER (in %).

background music in the solo singing tracks, out-of-vocabulary words, and incorrectly pronounced words because of unfamiliar lyrics [203]. The DAMP dataset consisted of 3-4 minute songs with lyrics text and is useful for research, but its first version did not have lyric time stamps.

Kruspe [204] and Dzhambazov [216] presented systems for the lyrics alignment challenge in MIREX 2017. The acoustic models in [204] were trained using 6,000 songs from the DAMP dataset. The average error in word boundaries (i.e., absolute deviation of the onset boundary of a word with respect to the ground truth) for an a capella test dataset in the best performing system by Kruspe was 2.87 seconds. The lack of availability of lyric time stamps in the training data was one of the reasons for the poor performance.

Gupta *et al.* [203] designed a semi-supervised method to automatically obtain weak line-level (short duration) lyrics annotation of a subset of approximately 50 hours of solo singing from the DAMP data. They used an iterative method to gradually split the full song into shorter segments of a few seconds with reliable lyrics. They adapted a DNN-HMM speech acoustic model to singing voice with this data. It showed 36.32% word error rate (WER)[3] in a lyric recognition experiment on short solo-singing test phrases from the same dataset. In [55], these singing-adapted models were further enhanced to capture long duration vowels with a duration-based lexicon modification, which further reduced the WER to 29.65%.

In the next release of the DAMP dataset, the lyric prompts were made available, which gave approximate start and end time stamps of shorter utterances of singing voices. Dabike *et al.* [209] used them to train a factorized time delay neural network (TDNN-F) acoustic model using Kaldi Speech Recognition Toolkit[4], and obtained the best WER of 19.60%, which was the new state-of-the-art baseline recognition result for solo-singing data. TDNN-F models are widely used in ASR systems due to their capability of successfully modeling long-term context and their ability to be parallelized, unlike RNNs. Demirel *et al.* [211] extended this framework to include six 2-D convolutional layers

[3]WER is a common metric of performance of speech recognition systems defined as the percentage of the total number of errors, i.e., substitutions, deletions, insertions, with respect to the total number of words in the ground-truth transcription.

[4]https://kaldi-asr.org

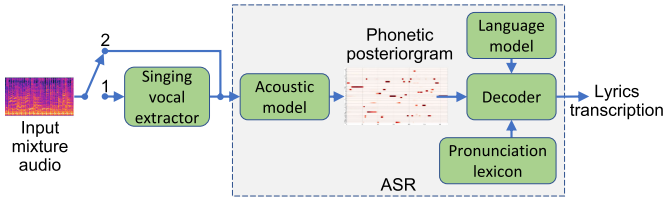

Fig. 11. Overview of lyrics transcription system for polyphonic music with traditional automatic speech recognition (ASR) framework, showing the two typical approaches, 1. extraction-transcription and 2. direct modeling.

at the front-end followed by the TDNN-F architecture, and finally a time-restricted self-attention layer. This architecture further improved the WER performance on the same solo singing test data to 14.96%, attributed to further context awareness through the convolutional and self-attention layers. Demirel *et al.* [215] further modified the TDNN-F architecture to include multiple streams of TDNN layers trained in parallel, where each stream has a unique time dilation rate capturing different temporal resolutions. This system improved the lyrics transcription performance on Mauch's polyphonic test dataset [202] with a WER of 37.33%.

C. Lyrics Modeling in Polyphonic Music

Past studies have developed acoustic models with only solo-singing recordings [55], [203], [204], [209], [217]–[220], and those models perform well for lyrics transcription on solo-singing test data [203], [209]. Acoustic models trained on solo-singing data, however, perform poorly when directly applied to polyphonic music test data [221], [222], and as seen in the GSLW2 result of lyrics-to-audio alignment task of MIREX 2018[5], they resulted in a significant drop in performance when applied to singing voices in the presence of background music.

Singing voices are often highly correlated with the corresponding background music, resulting in overlapping frequency components [223]. The varied range of voice quality of artists combined with different types of musical instruments makes the problem of lyrics synchronization and transcription highly challenging in polyphonic music. As shown in Fig. 11, the solutions can be grouped into two broad strategies, 1) *extraction-transcription* approach which extracts singing voices first through a source separation model and then transcribes the extracted vocals, and 2) *direct modeling* approach which directly models the polyphonic audio, i.e., singing voices mixed with background music. Moreover, in both of these approaches, there have been content-informed techniques used to supplement the capabilities of these data-driven modeling methods.

1) Extraction-Transcription: In the extraction-transcription approach, a singing voice separation method is employed as a pre-processing step [10], [197], [201], [211], [216], [224], [225].

Demirel *et al.* [225] dealt with the lyrics transcription problem in polyphonic music by using a pre-trained solo singing acoustic model [211] to recognize lyrics in the singing voices extracted by the pre-trained singing voice separation methods, Spleeter [135] and Demucs [160]. However, this approach involved a mismatch in quality between clean singing voices and the extracted singing voices. Since singing voice separation methods are not perfect, the extracted vocals often contain distortions and synthesis artifacts.

Gupta *et al.* [226] used domain adaptation to adapt an acoustic model trained on solo singing data with two kinds of in-domain data: direct polyphonic audio and singing voices extracted by the singing voice separation method, Wave-U-Net [151]. The in-domain adaptation of acoustic models using polyphonic data was found to perform better than the acoustic models adapted with extracted vocals. This suggested that using polyphonic data for adaptation, instead of imperfect extracted vocals, helps in capturing the spectro-temporal variations useful for lyrics modeling. However, with a different setting, Basak *et al.* [227] showed that the in-domain adaptation using the extracted vocals was found to outperform that using polyphonic music. They proposed a data augmentation method of using a vocoder-based speech-to-singing conversion technique for generating artificial singing examples for training lyrics transcription model. For the purpose of in-domain fine-tuning of a solo-singing end-to-end lyric recognition system, they compared direct polyphonic audio with singing vocals extracted by the state-of-the-art D3Net [159], which has a superior separation performance compared to the previously used Wave-U-Net, and obtained the above-mentioned result.

The performance of lyrics synchronization and transcription is negatively affected by the imperfect singing voice extraction front-end [208], [224], as the separation artifacts often make the words unrecognizable. Moreover, the extraction-transcription approach usually requires a separate training setup for the singing voice separation method. Training the voice extraction and transcription modules separately may lead to possible mismatch between training and testing, though this approach has the advantage of being able to use the results of state-of-the-art singing voice separation research that has been actively investigated.

2) Direct Modeling: Many recent works have explored data intensive direct modeling approaches to lyrics-to-audio alignment. The direct modeling approach does not remove the background music, but rather makes use of it. In MIREX 2018, Wang [206] presented a system that achieved a mean alignment error (AE)[6] of 4.12 seconds on a standard test dataset (Mauch's polyphonic dataset [202]) for word alignment evaluation. They used 7,300 annotated proprietary English songs to train GMM-HMM models. Stoller *et al.* [207] presented an end-to-end system based on the Wave-U-Net architecture that predicts character probabilities directly from raw audio. The system was trained on more than 44,000 proprietary songs with line-level lyrics annotations. They achieved an impressive 0.35

[5]https://www.music-ir.org/mirex/wiki/2018:Automatic_Lyrics-to-Audio_Alignment_Results

[6]AE measures the absolute time deviation (in seconds) between the actual time stamp and the estimated time stamp at the beginning of each lyrical unit (word). The error is averaged over all words of a song, and then over all songs in a test dataset.

TABLE VI
A SUMMARY OF DATASETS OPENLY AVAILABLE FOR SINGING VOICE RESEARCH

Dataset	Content	Link	Applications
RWC Music Database [230], [231]	Popular Music Database (100 songs) (Japanese and English), Royalty-Free Music Database (15 songs), etc.	https://staff.aist.go.jp/m.goto/RWC-MDB/	Lyrics transcription, melody transcription
VocalSet [62]	10.1 hours of monophonic recorded audio of professional singers, 20 singers, 17 vocal techniques on 5 vowels, a total of 3560 recordings	https://zenodo.org/record/1442513	Singing skill assessment
DAMP Sing! (Smule Inc.)	300x30x2; 1100 hours solo singing and lyrics two singers (male and female), from the 300 most popular songs, from 30 countries, mobile app recording quality	https://zenodo.org/record/2747436	Lyrics transcription in solo singing [209], [211], singer identification and query by singing [232], singing style and intonation pattern analysis [60], [233]
DSing (DAMP Sing! Lyrics Curated) [209]	150 hours curated English songs data from the DAMP dataset; removed noisy data	https://github.com/groadabike/Kaldi-Dsing-task	Lyrics transcription in solo singing [209], [211], [215]
DAMP-VSEP [234]	11,494 compositions (155 countries, 36 languages, 6456 artists) with backing tracks, one or more isolated vocals, and a mixture of the two	https://zenodo.org/record/3553059	Singing voice separation [173]
DAMP Aligned [203]	50 hours training data, 2.3 hours test; lyrics aligned and short segments	https://github.com/chitralekha18/lyrics-aligned-solo-singing-dataset	Lyrics transcription in solo singing [203], [226], [224]
DALI [228]	134 hours English polyphonic song utterances with aligned lyrics	https://github.com/gabolsgabs/DALI	Lyrics transcription in polyphonic music
NUS48E [97]	2.8 hours recordings of the sung and spoken lyrics of 48 (20 unique) English songs by 12 subjects and transcriptions and duration annotations at the phone-level	https://smcnus.comp.nus.edu.sg/nus-48e-sung-and-spoken-lyrics-corpus/	Speech-singing conversion [235], singing synthesis, pronunciation evaluation [236], phoneme alignment in solo singing
NHSS [126]	100 songs sung and spoken by 10 singers, resulting in total of 7 hours audio data	https://hltnus.github.io/NHSSDatabase/index.html	Speech-singing conversion, singing synthesis, lyrics alignment in solo singing
NUS48E+ SingEval [43]	2 songs, 20 singers; music experts labels on pitch, rhythm, etc.	https://github.com/chitralekha18/PESnQ_APSIPA2017	Singing skill evaluation [47], [43], [22], [59]
DAMP SingEval [39]	400 renditions (4 songs, 100 singers per song), each rated by humans on the basis of singing quality	https://github.com/chitralekha18/SingEval.git	Singing skill evaluation [39], [58], [57], [22], [59]
Nitech[79]	31 studio quality recordings of a female singer singing Japanese children songs, phoneme and musical note annotations	http://hts.sp.nitech.ac.jp/archives/2.3/HTS-demo_NIT-SONG070-F001.tar.bz2	Singing voice synthesis [79], [88], [75]
Kara1k [237]	2,000 songs (1,000 studio recorded cover songs, 1,000 original artist songs), pure singing voice and instrumental tracks, title, genre, singer gender, lyrics language (mostly English, some French, German, and Spanish)	http://yannbayle.fr/karamir/index.php	Cover song identification, singer gender identification, singing voice separation, lyrics transcription
MIR1k [118]	1,000 Chinese song clips with the music accompaniment and the singing voice recorded as left and right channels, respectively; manual annotations of pitch contours in semitone, indices and types for unvoiced frames, lyrics, and vocal/non-vocal segments; speech recordings of the lyrics by the same singer	https://sites.google.com/site/unvoicedsoundseparation/mir-1k	Music source separation, lyrics transcription
MedleyDB [238]	70 songs, with individually processed stems and raw audio for each track, contains tags for genre, f0 melody time stamps and instrument activations	http://medleydb.weebly.com/downloads.html	Music source separation
iKALA [239]	252 excerpts of 30-seconds; separate vocal and instrument tracks, songs recorded by six different singers, pitch labels and lyrics (with time stamps)	http://mac.citi.sinica.edu.tw/ikala/index.html	Music source separation
DSD100 [240]	100 tracks of professionally mixed songs, separate tracks for vocal and instrument sources	https://sigsep.github.io/datasets/dsd100.html	Music source separation
MUSDB18 [150]	100 tracks from DSD100, 46 tracks from MedleyDB, and 4 tracks from other sources	https://zenodo.org/record/1117372	Music source separation
CCMixter [241]	50 songs, vocal and background music tracks	https://members.loria.fr/ALiutkus/kam/	Music source separation
Phonation Modes [242]	900 samples of sustained sung vowels in four phonation modes, breathy, neutral, flow and pressed, by one female singer	https://osf.io/pa3ha/wiki/home/	Singing voice analysis

seconds mean AE on the Mauch's dataset. However, end-to-end systems require a large amount of annotated training data to perform well as seen in [207], while publicly available acoustic resources for polyphonic music are limited. Moreover, while this system performed well for the task of lyrics-to-audio alignment, it showed a high word error rate (WER) for the task of lyrics transcription.

Gupta *et al.* [208] trained the TDNN pipeline of Kaldi that consists of separate acoustic model, language model, and pronunciation lexicon, with a publicly available polyphonic dataset called DALI [228] that contains about 3,000 songs with line-level lyrics alignment. This system outperformed the lyrics-to-audio alignment as well as transcription results of Stoller *et al.* [207], showing that when the data is limited,

the standard acoustic modeling pipeline performs better than end-to-end pipeline.

3) Content-Informed Lyrics Modeling: In the direct modeling approach, there have been works in which additional content information has been incorporated as side information to the data-driven modeling. Gupta *et al.* [208] trained a music-informed acoustic model that incorporated music genre-specific information into acoustic models. The underlying assumption was that lyrics intelligibility depends on the background instrumentation and syllable rates, which are correlated with the genre of the music [229]. They categorized music genres broadly into pop, hip hop, and metal, and used this tag at the time of training through genre-specific lexicon. This model outperformed the acoustic model without any additional content information for both the tasks of lyrics-to-audio alignment as well as lyrics transcription, which suggests that lyrics acoustic models can benefit from musical information, e.g., music genre, available through the meta-data of the songs. This approach [213] showed the best AE of 0.19 seconds for the task of lyrics synchronization and the best WER of 47.25% for the task of lyrics transcription on Mauch's dataset in MIREX 2020[7].

Mauch *et al.* [202] exploited the correlation between chords and lyrics for the purpose of lyrics-to-audio alignment in polyphonic music. As seen in lyrics text in song books, chord annotations are sometimes partially available on lyrics. To leverage those chord annotations, a lyrics-to-audio alignment HMM was extended to incorporate chords and chroma features; each hidden state has two emissions, one for the phoneme feature (MFCC) for lyrics and the other for chroma feature for chords. Missing chord information was also recovered by locating phrase-level boundaries based on the partially given chords. They showed that the additional chord information boosted the lyric-to-audio alignment performance.

Most of the above methods explored lyrics modeling in a single language that was English. Vaglio *et al.* [214] employed an end-to-end approach for lyrics-to-audio alignment applied in a multilingual context using a universal set of phonemes as an additional intermediate representation, and showed that this approach can help in bridging between different languages.

D. Future Directions

Despite the recent advances with large public datasets and the application of successful speech recognition techniques, the performance of the current lyrics transcription systems remains far from perfect, thus the problem of lyrics transcription for polyphonic music is still unsolved. For example, metal and rap songs are seen to be particularly difficult for the transcription systems. Further studies are needed that analyze the error patterns of the current systems, and architectural modifications specific to singing voices and lyrics need to be designed.

Most of the current systems of both lyrics synchronization and transcription are heavily language dependent for acoustic modeling. This limits the scalability of such systems across different regions of the world because of the lack of data in different languages. Therefore, it is important to study and design a unified lyrics transcription model for multiple languages through a unified phoneme dictionary and techniques for low-resource language modeling.

There are professional musicians who are capable of transcribing multiple aspects of the audio mixture of a song, such as lyrics, chords, musical notes, and drum beats, while taking their relationship into account. Most of the current automatic systems focus on just one of the tasks. Multi-task systems that leverage auxiliary tasks to boost the performance of the main task of lyrics transcription would be worth exploring in the future.

VII. Datasets for Singing Voice Research

Due to music copyright restrictions, one of the main hindrances for research in singing voice has been the lack of appropriately annotated publicly available datasets. In recent years, companies such as Smule Inc. and KaraFun have contributed datasets to the research community, and researchers have come together to prepare annotated datasets for the community such as NHSS [126], DALI [228], and NUS48E [97]. These datasets are used or can potentially be used for multiple tasks. For example, NHSS is a speech and singing parallel dataset along with manual word boundaries, which can be used for different singing information processing tasks such as lyrics-to-audio alignment, study of rhythm in sung lyrics, and singing skill evaluation. A regularly updated list of data resources for MIR applications is published by audiocontentanalysis.org.[8] We collate and summarize the content and potential applications of the recently published datasets related to the tasks of singing information processing in Table VI.

VIII. Conclusion

We have provided a comprehensive overview of the recent developments in various topics of singing information processing: singing skill evaluation, singing voice synthesis, singing voice separation, and lyrics synchronization and transcription. This paper has especially focused on the latest deep learning techniques explored in each of these topics. We have also provided summary tables that contain a collection of papers along with the available resources, such as codebases, datasets, and some indicative performances. Furthermore, we have listed publicly available datasets for singing voice research along with their possible applications. We hope this paper can serve as an easy-to-access resource to stimulate further advances in singing information processing research.

References

[1] J. Koopman, *A Brief History of Singing*. Lawrence University, Appleton, WI, USA, 1994. Accessed on: Jul. 19, 2022. [Online]. Available: http://www2.lawrence.edu/fast/KOOPMAJO/brief.html

[2] J. Montagu, "How music and instruments began: A brief overview of the origin and entire development of music, from its earliest stages," *Front. Sociol.*, vol. 2, 2017, Art. no. 8.

[7] https://www.music-ir.org/mirex/wiki/2020:MIREX2020_Results

[8] http://www.audiocontentanalysis.org/data-sets/

[3] A. J. Cohen, "Research on singing: Development, education and well-being–introduction to the special volume on "singing and psychomusicology"," *Psychomusicol.: Music, Mind Brain*, vol. 21, no. 1-2, 2011, Art. no. 1.

[4] A. Norton, L. Zipse, S. Marchina, and G. Schlaug, "Melodic intonation therapy," *Ann. New York Acad. Sci.*, vol. 1169, no. 1, pp. 431–436, 2009.

[5] M. L. Albert, R. W. Sparks, and N. A. Helm, "Melodic intonation therapy for aphasia," *Arch. Neurol.*, vol. 29, no. 2, pp. 130–131, 1973.

[6] M. Goto, T. Saitou, T. Nakano, and H. Fujihara, "Singing information processing based on singing voice modeling," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2010, pp. 5506–5509.

[7] 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.

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

[9] J. Sundberg and T. D. Rossing, "The science of singing voice," *Acoust. Soc. Amer.*, vol. 87, no. 1, pp. 462–463, 1990.

[10] H. Fujihara, M. Goto, J. Ogata, and H. G. Okuno, "LyricSynchronizer: Automatic synchronization system between musical audio signals and lyrics," *IEEE J. Sel. Topics Signal Process.*, vol. 5, no. 6, pp. 1252–1261, Oct. 2011.

[11] M. McVicar, D. P. Ellis, and M. Goto, "Leveraging repetition for improved automatic lyric transcription in popular music," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2014, pp. 3117–3121.

[12] T. Nakano, M. Goto, and Y. Hiraga, "An automatic singing skill evaluation method for unknown melodies using pitch interval accuracy and vibrato features," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2006, pp. 1706–1709.

[13] O. Mayor, J. Bonada, and A. Loscos, "Performance analysis and scoring of the singing voice," in *Proc. Int. Conf. Audio Games, Audio Eng. Soc.*, 2009, pp. 1–7.

[14] A. Klapuri, "Introduction to music transcription," in *Signal Processing Methods for Music Transcription*. Berlin, Germany: Springer, 2006, pp. 3–20.

[15] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1978, vol. 100.

[16] D. Tymoczko, *A Geometry of Music: Harmony and Counterpoint in the Extended Common Practice*. London, U.K.: Oxford Univ. Press, 2010.

[17] V. Kliewer, "Melody: Linear aspects of twentieth-century music," *Aspects of Twentieth-Century Music*, pp. 270–321, 1975.

[18] D. W. Harding, D. C. Harding, and D. W. Harding, *Words Into Rhythm: English Speech Rhythm in Verse and Prose*. Cambridge, U.K.: Cambridge Univ. Press, 1976.

[19] G. Fant, A. Kruckenberg, and L. Nord, "Stress patterns and rhythm in the reading of prose and poetry with analogies to music performance," in *Music, Language, Speech and Brain*. Berlin, Germany: Springer, 1991, pp. 380–407.

[20] E. Nichols, D. Morris, S. Basu, and C. Raphael, "Relationships between lyrics and melody in popular music," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, 2009, pp. 471–476.

[21] J. Pons Puig, R. Gong, and X. Serra, "Score-informed syllable segmentation for a cappella singing voice with convolutional neural networks," in *Proc. Int. Soc. Music Inf. Retrieval*, 2017, pp. 483–489.

[22] C. Gupta, J. Li, and H. Li, "Towards reference-independent rhythm assessment of solo singing," in *Proc. Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf.*, 2021, pp. 912–919.

[23] I. Van Der Meulen, V. De Sandt-Koenderman, W. Mieke, M. H. Heijenbrok, E. Visch-Brink, and G. M. Ribbers, "Melodic intonation therapy in chronic aphasia: Evidence from a pilot randomized controlled trial," *Front. Hum. Neurosci.*, vol. 10, 2016, Art. no. 533.

[24] D. Hoppe, M. Sadakata, and P. Desain, "Development of real-time visual feedback assistance in singing training: A review," *J. Comput. Assist. Learn.*, vol. 22, no. 4, pp. 308–316, 2006.

[25] G. F. Welch, D. M. Howard, and C. Rush, "Real-time visual feedback in the development of vocal pitch accuracy in singing," *Psychol. Music*, vol. 17, no. 2, pp. 146–157, 1989.

[26] D. M. Howard and G. F. Welch, "Microcomputer-based singing ability assessment and development," *Appl. Acoust.*, vol. 27, no. 2, pp. 89–102, 1989.

[27] Y. Li, "Technologies and music therapy from the perspective of music therapists," in *Proc. 4th Int. Conf. Biol. Inf. Biomed. Eng.*, 2020, pp. 1–5.

[28] J. Wapnick and E. Ekholm, "Expert consensus in solo voice performance evaluation," *J. Voice*, vol. 11, no. 4, pp. 429–436, 1997.

[29] J. M. Oates, B. Bain, P. Davis, J. Chapman, and D. Kenny, "Development of an auditory-perceptual rating instrument for the operatic singing voice," *J. Voice*, vol. 20, no. 1, pp. 71–81, 2006.

[30] T. Nakano, M. Goto, and Y. Hiraga, "Subjective evaluation of common singing skills using the rank ordering method," in *Proc. Int. Conf. Music Percep. Cogn.*, 2006, pp. 1507–1512.

[31] J. Sundberg, "The level of the 'singing formant' and the source spectra of professional bass singers," *Speech Transmiss. Lab. Quart. Prog. Status Rep.*, vol. 4, pp. 21–39, 1970.

[32] W.-H. Tsai and H.-C. Lee, "Automatic evaluation of Karaoke singing based on pitch, volume, and rhythm features," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 4, pp. 1233–1243, May 2012.

[33] C. Cao, M. Li, X. Wu, H. Suo, J. Liu, and Y. Yan, "Automatic singing performance evaluation for untrained singers," *IEICE Trans. Inf. Syst.*, vol. E 92.D, no. 8, pp. 1596–1600, 2009.

[34] C. Gupta and P. Rao, "Objective assessment of ornamentation in Indian classical singing," in *Speech, Sound and Music Processing: Embracing Research in India*. Berlin, Germany: Springer, 2012, pp. 1–25.

[35] M. Clayton, *Time in Indian Music: Rhythm, Metre, and Form in North Indian Rag Performance*. London, U.K.: Oxford Univ. Press, 2008.

[36] R. Gong, "Automatic assessment of singing voice pronunciation: A case study with Jingju music," Ph.D. dissertation, Universitat Pompeu Fabra, Barcelona, Spain, 2018.

[37] S. Zhang, R. Caro Repetto, and X. Serra, "Understanding the expressive functions of Jingju metrical patterns through lyrics text mining," in *Proc. Int. Soc. Music Inf. Retrieval*, 2017, pp. 397–403.

[38] A. Lerch, C. Arthur, A. Pati, and S. Gururani, "Music performance analysis: A survey," in *Proc. Int. Soc. Music Inf. Retrieval*, 2019, pp. 33–43.

[39] C. Gupta, H. Li, and Y. Wang, "Automatic leaderboard: Evaluation of singing quality without a standard reference," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 13–26, 2020.

[40] S. Gururani, K. A. Pati, C.-W. Wu, and A. Lerch, "Analysis of objective descriptors for music performance assessment," in *Proc. Int. Conf. Music Percep. Cogn.*, 2018.

[41] A. Vidwans, S. Gururani, C.-W. Wu, V. Subramanian, R. V. Swaminathan, and A. Lerch, "Objective descriptors for the assessment of student music performances," in *Proc. Int. Conf. Semantic Audio. Audio Eng. Soc*, 2017.

[42] C. Cao, M. Li, J. Liu, and Y. Yan, "A study on singing performance evaluation criteria for untrained singers," in *Proc. Int. Conf. Signal Process.*, 2008, pp. 1475–1478.

[43] C. Gupta, H. Li, and Y. Wang, "A technical framework for automatic perceptual evaluation of singing quality," *APSIPA Trans. Signal Inf. Process.*, vol. 7, 2018, Art. no. e10.

[44] P. Lal, "A comparison of singing evaluation algorithms," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2006.

[45] J. Abeßer, J. Hasselhorn, C. Dittmar, A. Lehmann, and S. Grollmisch, "Automatic quality assessment of vocal and instrumental performances of ninth-grade and tenth-grade pupils," in *Proc. Int. Symp. Comput. Music Multidisciplinary Res.*, 2013.

[46] T. Nakano, M. Goto, and Y. Hiraga, "MiruSinger: A singing skill visualization interface using real-time feedback and music CD recordings as referential data," in *Proc. Int. Symp. Multimedia Workshops (Demonstrations)*, 2007, pp. 75–76.

[47] C. Gupta, H. Li, and Y. Wang, "Perceptual evaluation of singing quality," in *Proc. Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf.*, 2017, pp. 12–15.

[48] E. Molina, I. Barbancho, E. Gómez, A. M. Barbancho, and L. J. Tardón, "Fundamental frequency alignment vs. note-based melodic similarity for singing voice assessment," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2013, pp. 744–748.

[49] C.-H. Lin, Y.-S. Lee, M.-Y. Chen, and J.-C. Wang, "Automatic singing evaluating system based on acoustic features and rhythm," in *Proc. Int. Conf. Orange Technol.*, 2014, pp. 165–168.

[50] S. Hutchins and S. Moreno, "The linked dual representation model of vocal perception and production," *Front. Psychol.*, vol. 4, 2013, Art. no. 825.

[51] W. Wang, J. Pan, H. Yi, Z. Song, and M. Li, "Audio-based piano performance evaluation for beginners with convolutional neural network and attention mechanism," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 1119–1133, 2021.

[52] J. Huang, Y.-N. Hung, A. Pati, S. K. Gururani, and A. Lerch, "Score-informed networks for music performance assessment," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, 2020.

[53] E. Nichols, C. DuHadway, H. Aradhye, and R. F. Lyon, "Automatically discovering talented musicians with acoustic analysis of youtube videos," in *Proc. Int. Conf. Data Mining*, 2012, pp. 559–565.

[54] J. Böhm, F. Eyben, M. Schmitt, H. Kosch, and B. Schuller, "Seeking the superstar: Automatic assessment of perceived singing quality," in *Proc. Int. Joint Conf. Neural Netw.*, 2017, pp. 1560–1569.

[55] C. Gupta, H. Li, and Y. Wang, "Automatic evaluation of singing quality without a reference," in *Proc. Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf.*, 2018, pp. 990–997.

[56] N. Zhang, T. Jiang, F. Deng, and Y. Li, "Automatic singing evaluation without reference melody using bi-dense neural network," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2019, pp. 466–470.

[57] C. Gupta, L. Huang, and H. Li, "Automatic rank-ordering of singing vocals with twin-neural network," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, 2020, pp. 416–423.

[58] L. Huang, C. Gupta, and H. Li, "Spectral features and pitch histogram for automatic singing quality evaluation with CRNN," in *Proc. Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf.*, 2020, pp. 492–499.

[59] J. Li, C. Gupta, and H. Li, "Training explainable singing quality assessment network with augmented data," in *Proc. Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf.*, 2021, pp. 904–911.

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

[61] K. A. Pati, S. Gururani, and A. Lerch, "Assessment of student music performances using deep neural networks," *Appl. Sci.*, vol. 8, no. 4, 2018, Art. no. 507.

[62] 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.

[63] P. R. Cook, "Singing voice synthesis: History, current work, and future directions," *Comput. Music J.*, vol. 20, no. 3, pp. 38–46, 1996.

[64] 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.

[65] T. Saitou, M. Goto, M. Unoki, and M. Akagi, "Speech-to-singing synthesis: Converting speaking voices to singing voices by controlling acoustic features unique to singing voices," in *Proc. Workshop Appl. Signal Process. Audio Acoust.*, 2007, pp. 215–218.

[66] T. Nakano and M. Goto, "VocaListener: A singing-to-singing synthesis system based on iterative parameter estimation," in *Proc. Sound Music Comput. Conf.*, 2009, pp. 343–348.

[67] M. Blaauw and J. Bonada, "Sequence-to-sequence singing synthesis using the feed-forward transformer," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2020, pp. 7229–7233.

[68] K. Vijayan, H. Li, and T. Toda, "Speech-to-singing voice conversion: The challenges and strategies for improving vocal conversion processes," *IEEE Signal Process. Mag.*, vol. 36, no. 1, pp. 95–102, Jan. 2019.

[69] B. Sisman, K. Vijayan, M. Dong, and H. Li, "SINGAN: Singing voice conversion with generative adversarial networks," in *Proc. Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf.*, 2019, pp. 112–118.

[70] S. Aso *et al.*, "SpeakBySinging: Converting singing voices to speaking voices while retaining voice timbre," in *Proc. Int. Conf. Digit. Audio Effects*, 2010, pp. 114–12.

[71] J. A. Moorer, "The use of linear prediction of speech in computer music applications," *J. Audio Eng. Soc.*, vol. 27, no. 3, pp. 134–140, 1979.

[72] G. Carlsson, S. Termstrom, J. Sundberg, and T. Ungvary, "A new digital system for singing synthesis allowing expressive control," in *Proc. Int. Comput. Music Conf., Int. Comput. Music Assoc.*, 1991, pp. 315–315.

[73] H. Kenmochi and H. Ohshita, "VOCALOID - Commercial singing synthesizer based on sample concatenation," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2007, pp. 4009–4010.

[74] K. Oura, A. Mase, T. Yamada, S. Muto, Y. Nankaku, and K. Tokuda, "Recent development of the HMM-based singing voice synthesis system – Sinsy," in *Proc. ISCA Workshop Speech Synth.*, 2010.

[75] Y. Hono *et al.*, "Recent development of the DNN-based singing voice synthesis system – Sinsy," in *Proc. Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf.*, 2018, pp. 1003–1009.

[76] K. Nakamura, S. Takaki, K. Hashimoto, K. Oura, Y. Nankaku, and K. Tokuda, "Fast and high-quality singing voice synthesis system based on convolutional neural networks," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2020, pp. 7239–7243.

[77] H. Kawahara, I. Masuda-Katsuse, and A. De Cheveigne, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based f0 extraction: Possible role of a repetitive structure in sounds," *Speech Commun.*, vol. 27, no. 3/4, pp. 187–207, 1999.

[78] M. Morise, F. Yokomori, and K. Ozawa, "WORLD: A vocoder-based high-quality speech synthesis system for real-time applications," *IEICE Trans. Inf. Syst.*, vol. 99, no. 7, pp. 1877–1884, 2016.

[79] M. Blaauw and J. Bonada, "A neural parametric singing synthesizer modeling timbre and expression from natural songs," *Appl. Sci.*, vol. 7, no. 12, 2017, Art. no. 1313.

[80] P. Chandna, M. Blaauw, J. Bonada, and E. Gómez, "WGANSing: A multi-voice singing voice synthesizer based on the Wasserstein-GAN," in *Proc. Eur. Signal Process. Conf.*, 2019, pp. 1–5.

[81] P. Lu, J. Wu, J. Luan, X. Tan, and L. Zhou, "XiaoiceSing: A high-quality and integrated singing voice synthesis system," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2020, pp. 1306–1310.

[82] J. Kelly and C. Lochbaum, "Speech synthesis," in *Proc. 4th Int. Congr. Acoust.*, 1962, pp. 1–4.

[83] X. Rodet, "Time-domain formant-wave-function synthesis," in *Spoken Language Generation and Understanding*. Berlin, Germany: Springer, 1980, pp. 429–441.

[84] J. Sundberg, "Synthesis of singing by rule," in *Current Directions in Computer Music Research*. Cambridge, MA, USA: MIT Press, 1989, pp. 45–55.

[85] M. Macon, L. Jensen-Link, E. B. George, J. Oliverio, and M. Clements, "Concatenation-based midi-to-singing voice synthesis," in *Proc. Audio Eng. Soc. Conv. 103, Audio Eng. Soc.*, 1997.

[86] K. Saino, H. Zen, Y. Nankaku, A. Lee, and K. Tokuda, "An HMM-based singing voice synthesis system," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2006, pp. 2274–2277.

[87] K. Oura, A. Mase, Y. Nankaku, and K. Tokuda, "Pitch adaptive training for HMM-based singing voice synthesis," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2012, pp. 5377–5380.

[88] M. Nishimura, K. Hashimoto, K. Oura, Y. Nankaku, and K. Tokuda, "Singing voice synthesis based on deep neural networks," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2016, pp. 2478–2482.

[89] K. Tokuda, T. Kobayashi, and S. Imai, "Speech parameter generation from HMM using dynamic features," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 1995, pp. 660–663.

[90] J. Kim, H. Choi, J. Park, M. Hahn, S. Kim, and J.-J. Kim, "Korean singing voice synthesis system based on an LSTM recurrent neural network," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2018, pp. 1551–1555.

[91] T. Saitou, M. Unoki, and M. Akagi, "Extraction of f0 dynamic characteristics and development of f0 control model in singing voice," in *Proc. Int. Conf. Auditory Display*, 2002, pp. 1–4.

[92] M. Blaauw and J. Bonada, "A neural parametric singing synthesizer," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2017, pp. 4001–4005.

[93] Y.-H. Yi, Y. Ai, Z.-H. Ling, and L.-R. Dai, "Singing voice synthesis using deep autoregressive neural networks for acoustic modeling," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2019, pp. 2593–2597.

[94] A. Vaswani *et al.*, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[95] Y. Hono, K. Hashimoto, K. Oura, Y. Nankaku, and K. Tokuda, "Sinsy: A deep neural network-based singing voice synthesis system," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 2803–2815, 2021.

[96] Y. Gu *et al.*, "ByteSing: A. Chinese singing voice synthesis system using duration allocated encoder-decoder acoustic models and WaveRNN vocoders," in *Proc. Int. Symp. Chin. Spoken Lang. Process.*, 2021, pp. 1–5.

[97] Z. Duan, H. Fang, B. Li, K. C. Sim, and Y. Wang, "The NUS sung and spoken lyrics corpus: A quantitative comparison of singing and speech," in *Proc. Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf.*, 2013, pp. 1–9.

[98] S. Choi, W. Kim, S. Park, S. Yong, and J. Nam, "Korean singing voice synthesis based on auto-regressive boundary equilibrium GAN," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2020, pp. 7234–7238.

[99] J. Wu and J. Luan, "Adversarially trained multi-singer sequence-to-sequence singing synthesizer," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2020, pp. 1296–1300.

[100] Y. Ren *et al.*, "FastSpeech: Fast, robust and controllable text to speech," in *Proc. Adv. Int. Conf. Neural Inf. Process. Syst.*, 2019, pp. 3171–3180.

[101] O. Angelini, A. Moinet, K. Yanagisawa, and T. Drugman, "Singing synthesis: With a little help from my attention," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2020.

[102] Y. Wang *et al.*, "Tacotron: Towards end-to-end speech synthesis," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2017, pp. 4006–4010.

[103] J. Lee, H.-S. Choi, C.-B. Jeon, J. Koo, and K. Lee, "Adversarially trained end-to-end Korean singing voice synthesis system," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2019, pp. 2588–2592.

[104] H. Tachibana, K. Uenoyama, and S. Aihara, "Efficiently trainable text-to-speech system based on deep convolutional networks with guided attention," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2018, pp. 4784–4788.

[105] Y. Hono, K. Hashimoto, K. Oura, Y. Nankaku, and K. Tokuda, "Singing voice synthesis based on generative adversarial networks," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2019, pp. 6955–6959.

[106] A. Mase, K. Oura, Y. Nankaku, and K. Tokuda, "HMM-based singing voice synthesis system using pitch-shifted pseudo training data," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2010, pp. 845–848.

[107] Y. Hono, S. Takaki, K. Hashimoto, K. Oura, Y. Nankaku, and K. Tokuda, "PeriodNet: A non-autoregressive waveform generation model with a structure separating periodic and aperiodic components," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2021, pp. 6049–6053.

[108] B. Sisman and H. Li, "Generative adversarial networks for singing voice conversion with and without parallel data," in *Proc. Speaker Lang. Recognit. Workshop*, 2020, pp. 238–244.

[109] B. Sisman, M. Zhang, and H. Li, "Group sparse representation with WaveNet vocoder adaptation for spectrum and prosody conversion," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 6, pp. 1085–1097, Jun. 2019.

[110] E. Nachmani and L. Wolf, "Unsupervised singing voice conversion," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2019, pp. 2583–2587.

[111] C. Deng, C. Yu, H. Lu, C. Weng, and D. Yu, "PitchNet: Unsupervised singing voice conversion with pitch adversarial network," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2020, pp. 7749–7753.

[112] Z. Li *et al.*, "PPG-based singing voice conversion with adversarial representation learning," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2021, pp. 7073–7077.

[113] 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.

[114] S. Liu, Y. Cao, N. Hu, D. Su, and H. Meng, "FastSVC: Fast cross-domain singing voice conversion with feature-wise linear modulation," in *Proc. IEEE Int. Conf. Multimedia Expo*, 2021, pp. 1–6.

[115] S. Sing!, "Smule.digital archive mobile performances(damp)," Accessed on: Nov. 15, 2021. [Online]. Available: https://ccrma.stanford.edu/damp/

[116] J. Lee, H.-S. Choi, J. Koo, and K. Lee, "Disentangling timbre and singing style with multi-singer singing synthesis system," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2020, pp. 7224–7228.

[117] X. Chen, W. Chu, J. Guo, and N. Xu, "Singing voice conversion with non-parallel data," in *Proc. Conf. Multimedia Inf. Process. Retrieval*, 2019, pp. 292–296.

[118] C.-L. Hsu and J.-S. R. Jang, "On the improvement of singing voice separation for monaural recordings using the MIR-1 K dataset," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 2, pp. 310–319, Feb. 2010.

[119] Y.-J. Luo, C.-C. Hsu, K. Agres, and D. Herremans, "Singing voice conversion with disentangled representations of singer and vocal technique using variational autoencoders," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2020, pp. 3277–3281.

[120] J. Lu, K. Zhou, B. Sisman, and H. Li, "VAW-GAN for singing voice conversion with non-parallel training data," in *Proc. Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf.*, 2020, pp. 514–519.

[121] M. Blaauw, J. Bonada, and R. Daido, "Data efficient voice cloning for neural singing synthesis," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2019, pp. 6840–6844.

[122] J. Bonada and M. Blaauw, "Semi-supervised learning for singing synthesis timbre," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2021, pp. 7083–7087.

[123] Y. Ren, X. Tan, T. Qin, J. Luan, Z. Zhao, and T.-Y. Liu, "DeepSinger: Singing voice synthesis with data mined from the web," in *Proc. SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2020, pp. 1979–1989.

[124] S. Nercessian, "Zero-shot singing voice conversion," in *Proc. Int. Soc. Music Inf. Retrieval*, 2020, pp. 70–76.

[125] X. Gao, X. Tian, R. K. Das, Y. Zhou, and H. Li, "Speaker-independent spectral mapping for speech-to-singing conversion," in *Proc. Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf.*, 2019, pp. 159–164.

[126] B. Sharma, X. Gao, K. Vijayan, X. Tian, and H. Li, "NHSS: A speech and singing parallel database," *Speech Commun.*, vol. 133, pp. 9–22, 2021.

[127] L. Zhang *et al.*, "DurIAN-SC: Duration informed attention network based singing voice conversion system," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2020, pp. 1231–1235.

[128] J. Parekh, P. Rao, and Y.-H. Yang, "Speech-to-singing conversion in an encoder-decoder framework," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2020, pp. 261–265.

[129] J. Hu, C. Yu, and F. Guan, "Non-parallel many-to-many singing voice conversion by adversarial learning," in *Proc. Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf.*, 2019, pp. 125–132.

[130] L. Cen, M. Dong, and P. Chan, "Template-based personalized singing voice synthesis," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2012, pp. 4509–4512.

[131] K. Vijayan, M. Dong, and H. Li, "A dual alignment scheme for improved speech-to-singing voice conversion," in *Proc. Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf.*, 2017, pp. 1547–1555.

[132] B. Lindblom and J. Sundberg, "The human voice in speech and singing," in *Springer Handbook of Acoustics*. Berlin, Germany: Springer, 2014, pp. 703–746.

[133] Z. Rafii, A. Liutkus, F.-R. Stöter, S. I. Mimilakis, D. FitzGerald, and B. Pardo, "An overview of lead and accompaniment separation in music," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 8, pp. 1307–1335, Aug. 2018.

[134] A. Jansson, E. J. Humphrey, N. Montecchio, R. M. Bittner, A. Kumar, and T. Weyde, "Singing voice separation with deep U-net convolutional networks," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, 2017, pp. 745–751.

[135] 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.

[136] F.-R. Stöter, S. Uhlich, A. Liutkus, and Y. Mitsufuji, "Open-unmix - A reference implementation for music source separation," *J. Open Source Softw.*, vol. 4, no. 41, 2019, Art. no. 1667.

[137] T. Nakano, K. Yoshii, Y. Wu, R. Nishikimi, K. W. Edward Lin, and M. Goto, "Joint singing pitch estimation and voice separation based on a neural harmonic structure renderer," in *Proc. Workshop Appl. Signal Process. Audio Acoust.*, 2019, pp. 160–164.

[138] A. Jansson, R. M. Bittner, S. Ewert, and T. Weyde, "Joint singing voice separation and f0 estimation with deep U-Net architectures," in *Proc. Eur. Signal Process. Conf.*, 2019, pp. 1–5.

[139] F.-R. Stöter, A. Liutkus, and N. Ito, "The 2018 signal separation evaluation campaign," in *Proc. Int. Conf. Latent Variable Anal. Signal Separation*, 2018, pp. 293–305.

[140] P.-S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Singing-voice separation from monaural recordings using deep recurrent neural networks," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, 2014, pp. 477–482.

[141] P.-S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Joint optimization of masks and deep recurrent neural networks for monaural source separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 12, pp. 2136–2147, Dec. 2015.

[142] S. Uhlich, F. Giron, and Y. Mitsufuji, "Deep neural network based instrument extraction from music," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2015, pp. 2135–2139.

[143] P. Chandna, M. Miron, J. Janer, and E. Gómez, "Monoaural audio source separation using deep convolutional neural networks," in *Proc. Int. Conf. Latent Variable Anal. Signal Separation*, 2017, pp. 258–266.

[144] S. Uhlich *et al.*, "Improving music source separation based on deep neural networks through data augmentation and network blending," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2017, pp. 261–265.

[145] S. I. Mimilakis, K. Drossos, T. Virtanen, and G. Schuller, "A recurrent encoder-decoder approach with skip-filtering connections for monaural singing voice separation," in *Proc. Int. Workshop Mach. Learn. Signal Process.*, 2017, pp. 1–6.

[146] A. Liutkus and R. Badeau, "Generalized wiener filtering with fractional power spectrograms," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2015, pp. 266–270.

[147] H. Erdogan, J. R. Hershey, S. Watanabe, and J. Le Roux, "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2015, pp. 708–712.

[148] G. Puy, A. Ozerov, N. Q. Duong, and P. Pérez, "Informed source separation via compressive graph signal sampling," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2017, pp. 1–5.

[149] S. Park, T. Kim, K. Lee, and N. Kwak, "Music source separation using stacked hourglass networks," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, 2018, pp. 289–296.

[150] Z. Rafii, A. Liutkus, F.-R. Stöter, S. I. Mimilakis, and R. Bittner, "The MUSDB18 corpus for music separation," Dec. 2017. [Online]. Available: https://doi.org/10.5281/zenodo.1117372

[151] D. Stoller, S. Ewert, and S. Dixon, "Wave-U-Net: A multi-scale neural network for end-to-end audio source separation," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, 2018, pp. 334–340.

[152] F. Lluís, J. Pons, and X. Serra, "End-to-end music source separation: Is it possible in the waveform domain?," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2019, pp. 4619–4623.

[153] J.-Y. Liu and Y.-H. Yang, "Denoising auto-encoder with recurrent skip connections and residual regression for music source separation," in *Proc. Int. Conf. Mach. Learn. Appl.*, 2018, pp. 773–778.

[154] N. Takahashi, N. Goswami, and Y. Mitsufuji, "MMDenseLSTM: An efficient combination of convolutional and recurrent neural networks for audio source separation," in *Proc. Int. Workshop Acoust. Signal Enhancement*, 2018, pp. 106–110.

[155] D. Samuel, A. Ganeshan, and J. Naradowsky, "Meta-learning extractors for music source separation," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2020, pp. 816–820.

[156] E. Nachmani, Y. Adi, and L. Wolf, "Voice separation with an unknown number of multiple speakers," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 7164–7175.

[157] G. Meseguer-Brocal and G. Peeters, "Content-based singing voice source separation via strong conditioning using aligned phonemes," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, 2020, pp. 819–827.

[158] K. Schulze-Forster, C. Doire, G. Richard, and R. Badeau, "Weakly informed audio source separation," in *Proc. Workshop Appl. Signal Process. Audio Acoust.*, 2019, pp. 273–277.

[159] N. Takahashi and Y. Mitsufuji, "Densely connected multidilated convolutional networks for dense prediction tasks," in *Proc. Comput. Vis. Pattern Recognit.*, 2021, pp. 993–1002.

[160] A. Défossez, N. Usunier, L. Bottou, and F. Bach, "Music source separation in the waveform domain," 2021, *arXiv:1911.13254*.

[161] W. Choi, M. Kim, J. Chung, and S. Jung, "LaSAFT: Latent source attentive frequency transformation for conditioned source separation," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2021, pp. 171–175.

[162] R. Sawata, S. Uhlich, S. Takahashi, and Y. Mitsufuji, "All for one and one for all: Improving music separation by bridging networks," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2021, pp. 51–55.

[163] Q. Kong, Y. Cao, H. Liu, K. Choi, and Y. Wang, "Decoupling magnitude and phase estimation with deep ResUNet for music source separation," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, 2021, pp. 342–349.

[164] S. I. Mimilakis, K. Drossos, and G. Schuller, "Unsupervised interpretable representation learning for singing voice separation," in *Proc. Eur. Signal Process. Conf.*, 2021, pp. 1412–1416.

[165] L. Prétet, R. Hennequin, J. Royo-Letelier, and A. Vaglio, "Singing voice separation: A study on training data," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2019, pp. 506–510.

[166] A. A. Nugraha, A. Liutkus, and E. Vincent, "Multichannel music separation with deep neural networks," in *Proc. Eur. Signal Process. Conf.*, 2016, pp. 1748–1752.

[167] S. I. Mimilakis, K. Drossos, J. F. Santos, G. Schuller, T. Virtanen, and Y. Bengio, "Monaural singing voice separation with skip-filtering connections and recurrent inference of time-frequency mask," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2018, pp. 721–725.

[168] K. W. E. Lin, B. Balamurali, E. Koh, S. Lui, and D. Herremans, "Singing voice separation using a deep convolutional neural network trained by ideal binary mask and cross entropy," *Neural Comput. Appl.*, vol. 32, no. 4, pp. 1037–1050, 2020.

[169] A. Cohen-Hadria, A. Roebel, and G. Peeters, "Improving singing voice separation using deep U-Net and Wave-U-Net with data augmentation," in *Proc. Eur. Signal Process. Conf.*, 2019, pp. 1–5.

[170] J. Perez-Lapillo, O. Galkin, and T. Weyde, "Improving singing voice separation with the Wave-U-net using minimum hyperspherical energy," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2020, pp. 3272–3276.

[171] K. W. E. Lin and M. Goto, "Zero-mean convolutional network with data augmentation for sound level invariant singing voice separation," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2019, pp. 251–255.

[172] P. Seetharaman, G. Wichern, J. L. Roux, and B. Pardo, "Bootstrapping deep music separation from primitive auditory grouping principles," in *Proc. Int. Conf. Mach. Learn. Workshop Self-Supervision Audio Speech*, 2020.

[173] Z. Wang, R. Giri, U. Isik, J.-M. Valin, and A. Krishnaswamy, "Semi-supervised singing voice separation with noisy self-training," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2021, pp. 31–35.

[174] M. Michelashvili, S. Benaim, and L. Wolf, "Semi-supervised monaural singing voice separation with a masking network trained on synthetic mixtures," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2019, pp. 291–295.

[175] S. Kang, J.-S. Park, and G.-J. Jang, "Improving singing voice separation using curriculum learning on recurrent neural networks," *Appl. Sci.*, vol. 10, no. 7, 2020, Art. no. 2465.

[176] Z.-C. Fan, Y.-L. Lai, and J.-S. R. Jang, "SVSGAN: Singing voice separation via generative adversarial network," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2018, pp. 726–730.

[177] D. Stoller, S. Ewert, and S. Dixon, "Adversarial semi-supervised audio source separation applied to singing voice extraction," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2018, pp. 2391–2395.

[178] M. Miron, J. Janer Mestres, and E. Gómez Gutiérrez, "Generating data to train convolutional neural networks for classical music source separation," in *Proc. Sound Music Comput. Conf.*, 2017, pp. 227–233.

[179] J. Paulus, M. Müller, and A. Klapuri, "State of the art report: Audio-based music structure analysis," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, 2010, pp. 625–636.

[180] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-attention generative adversarial networks," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 7354–7363.

[181] W. Yuan, S. Wang, X. Li, M. Unoki, and W. Wang, "A skip attention mechanism for monaural singing voice separation," *IEEE Signal Process. Lett.*, vol. 26, no. 10, pp. 1481–1485, Oct. 2019.

[182] E. M. Grais, H. Wierstorf, D. Ward, and M. D. Plumbley, "Multi-resolution fully convolutional neural networks for monaural audio source separation," in *Proc. Int. Conf. Latent Variable Anal. Signal Separation*, 2018, pp. 340–350.

[183] E. M. Grais, D. Ward, and M. D. Plumbley, "Raw multi-channel audio source separation using multi-resolution convolutional auto-encoders," in *Proc. Eur. Signal Process. Conf.*, 2018, pp. 1577–1581.

[184] W. Yuan, B. Dong, S. Wang, M. Unoki, and W. Wang, "Evolving multi-resolution pooling CNN for monaural singing voice separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 807–822, 2021.

[185] P. Chandna, M. Blaauw, J. Bonada, and E. Gómez, "Content based singing voice extraction from a musical mixture," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2020, pp. 781–785.

[186] N. Takahashi, M. K. Singh, S. Basak, P. Sudarsanam, S. Ganapathy, and Y. Mitsufuji, "Improving voice separation by incorporating end-to-end speech recognition," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2020, pp. 41–45.

[187] C.-B. Jeon, H.-S. Choi, and K. Lee, "Exploring aligned lyrics-informed singing voice separation," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, 2020.

[188] K. Schulze-Forster, C. S. J. Doire, G. Richard, and R. Badeau, "Phoneme level lyrics alignment and text-informed singing voice separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 2382–2395, 2021.

[189] P. Chandna, M. Blaauw, J. Bonada, and E. Gomez, "A vocoder based method for singing voice extraction," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2019, pp. 990–994.

[190] R. V. Swaminathan and A. Lerch, "Improving singing voice separation using attribute-aware deep network," in *Proc. Int. Workshop Multilayer Music Representation Process.*, 2019, pp. 60–65.

[191] D. Petermann, P. Chandna, H. Cuesta, J. Bonada, and E. Gómez, "Deep learning based source separation applied to choir ensembles," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, 2020, pp. 733–739.

[192] G. Meseguer-Brocal and G. Peeters, "Conditioned-u-Net: Introducing a control mechanism in the U-Net for multiple source separations," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, 2019, pp. 159–165.

[193] W. Choi, M. Kim, J. Chung, D. Lee, and S. Jung, "Investigating U-Nets with various intermediate blocks for spectrogram-based singing voice separation," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, 2020.

[194] K. Watanabe and M. Goto, "Lyrics information processing: Analysis, generation, and applications," in *Proc. 1st Workshop NLP Music Audio*, 2020, pp. 6–12.

[195] J. Kato, T. Nakano, and M. Goto, "TextAlive: Integrated design environment for kinetic typography," in *Proc. ACM SIGCHI Conf. Hum. Factors Comput. Syst.*, 2015, pp. 3403–3412.

[196] K. Tsukuda, K. Ishida, and M. Goto, "Lyric jumper: A lyrics-based music exploratory web service by modeling lyrics generative process," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, 2017, pp. 544–551.

[197] H. Fujihara, M. Goto, J. Ogata, K. Komatani, T. Ogata, and H. G. Okuno, "Automatic synchronization between lyrics and music CD recordings based on viterbi alignment of segregated vocal signals," in *Proc. IEEE Int. Symp. Multimedia*, 2006, pp. 257–264.

[198] H. Fujihara and M. Goto, "Lyrics-to-audio alignment and its application," *Multimodal Music Process.*, Dagstuhl Follow-UPS. Dagstuhl, Germany: Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, vol. 3, 2012, pp. 23–36, doi: 10.4230/DFU.Vol3.11041.23.

[199] K. Lee and M. Cremer, "Segmentation-based lyrics-audio alignment using dynamic programming," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, 2008, pp. 395–400.

[200] M.-Y. Kan, Y. Wang, D. Iskandar, T. L. Nwe, and A. Shenoy, "Lyrically: Automatic synchronization of textual lyrics to acoustic music signals," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 2, pp. 338–349, Feb. 2008.

[201] A. Mesaros and T. Virtanen, "Automatic recognition of lyrics in singing," *EURASIP J. Audio, Speech, Music Process.*, vol. 2010, no. 1, 2010, Art. no. 546047.

[202] M. Mauch, H. Fujihara, and M. Goto, "Integrating additional chord information into HMM-based lyrics-to-audio alignment," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 1, pp. 200–210, Jan. 2012.

[203] C. Gupta, R. Tong, H. Li, and Y. Wang, "Semi-supervised lyrics and solo-singing alignment," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, 2018, pp. 600–607.

[204] A. M. Kruspe, "Bootstrapping a system for phoneme recognition and keyword spotting in unaccompanied singing," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, 2016, pp. 358–364.

[205] G. Dzhambazov, "Knowledge-based probabilistic modeling for tracking lyrics in music audio signals," Ph.D. dissertation, Universitat Pompeu Fabra, Barcelona, Spain, 2017.

[206] C.-C. Wang, "MIREX2018: Lyrics-to-audio alignment for instrument accompanied singings," in *Music Inf. Retrieval Eval. eXchange Audio-Lyrics Alignment Challenge*, 2018. Accessed on: Jul. 19, 2022. [Online]. Available: https://www.music-ir.org/mirex/abstracts/2018/CW2.pdf

[207] D. Stoller, S. Durand, and S. Ewert, "End-to-end lyrics alignment for polyphonic music using an audio-to-character recognition model," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2019, pp. 181–185.

[208] C. Gupta, E. Yılmaz, and H. Li, "Automatic lyrics alignment and transcription in polyphonic music: Does background music help?," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2020, pp. 496–500.

[209] G. R. Dabike and J. Barker, "Automatic lyric transcription from Karaoke vocal tracks: Resources and a baseline system," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2019, pp. 579–583.

[210] B. Zhang, W. Wang, E. Zhao, and S. Lui, "Lyrics-to-audio alignment for dynamic lyric generation," in *Music Inf. Retrieval Eval. eXchange Audio-Lyrics Alignment Challenge*, 2020. Accessed on: Jul. 19, 2022. [Online]. Available: https://music-ir.org/mirex/abstracts/2020/ZWZL1.pdf

[211] E. Demirel, S. Ahlbäck, and S. Dixon, "Automatic lyrics transcription using dilated convolutional neural networks with self-attention," in *Proc. Int. Joint Conf. Neural Netw.*, 2020, pp. 1–8.

[212] E. Demirel, S. Ahlback, and S. Dixon, "A recursive search method for lyrics alignment," in *Music Inf. Retrieval Eval. eXchange Audio-Lyrics Alignment Challenge*, 2020. Accessed on: Jul. 19, 2022. [Online]. Available: https://www.music-ir.org/mirex/abstracts/2020/DDA3.pdf

[213] X. Gao, C. Gupta, and H. Li, "Lyrics transcription and lyrics-to-audio alignment with music-informed acoustic models," in *Music Inf. Retrieval Eval. eXchange Audio-Lyrics Alignment and Lyrics Transcription Challenges*, 2020. Accessed on: Jul. 19, 2022. [Online]. Available: https://www.music-ir.org/mirex/abstracts/2020/GL1.pdf

[214] A. Vaglio, R. Hennequin, M. Moussallam, G. Richard, and F. d'Alché Buc, "Multilingual lyrics-to-audio alignment," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, 2020, pp. 512–519.

[215] E. Demirel, S. Ahlbäck, and S. Dixon, "MSTRE-Net: Multistreaming acoustic modeling for automatic lyrics transcription," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, 2021, pp. 151–158.

[216] G. B. Dzhambazov and X. Serra, "Modeling of phoneme durations for alignment between polyphonic audio and lyrics," in *Proc. Sound Music Comput. Conf.*, 2015, pp. 281–286.

[217] D. Kawai, K. Yamamoto, and S. Nakagawa, "Speech analysis of sung-speech and lyric recognition in monophonic singing," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2016, pp. 271–275.

[218] A. Mesaros and T. Virtanen, "Adaptation of a speech recognizer for singing voice," in *Proc. Eur. Signal Process. Conf.*, 2009, pp. 1779–1783.

[219] A. Mesaros, "Singing voice identification and lyrics transcription for music information retrieval," in *Proc. Conf. Speech Technol. Hum.-Comput. Dialogue*, 2013, pp. 1–10.

[220] C.-P. Tsai, Y.-L. Tuan, and L.-S. Lee, "Transcribing lyrics from commercial song audio: The first step towards singing content processing," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2018, pp. 5749–5753.

[221] C. Gupta, B. Sharma, H. Li, and Y. Wang, "Lyrics-to-audio alignment using singing-adapted acoustic models and non-vocal suppression," in *Music Inf. Retrieval Eval. eXchange Audio-Lyrics Alignment Challenge*, 2018. Accessed on: Jul. 19, 2022. [Online]. Available: https://www.music-ir.org/mirex/abstracts/2018/GSLW1.pdf

[222] G. R. Dabike and J. Barker, "The sheffield university system for the MIREX 2020: Lyrics transcription task," in *Music Inf. Retrieval Eval. eXchange Lyrics Transcription Challenge*, 2020. Accessed on: Jul. 19, 2022. [Online]. Available: https://www.music-ir.org/mirex/abstracts/2020/RB1.pdf

[223] M. Ramona, G. Richard, and B. David, "Vocal detection in music with support vector machines," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2008, pp. 1885–1888.

[224] B. Sharma, C. Gupta, H. Li, and Y. Wang, "Automatic lyrics-to-audio alignment on polyphonic music using singing-adapted acoustic models," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2019, pp. 396–400.

[225] E. Demirel, S. Ahlbäck, and S. Dixon, "Low resource audio-to-lyrics alignment from polyphonic music recordings," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2021, pp. 586–590.

[226] C. Gupta, E. Yılmaz, and H. Li, "Acoustic modeling for automatic lyrics-to-audio alignment," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2019, pp. 2040–2044.

[227] S. Basak, S. Agarwal, S. Ganapathy, and N. Takahashi, "End-to-end lyrics recognition with voice to singing style transfer," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2021, pp. 266–270.

[228] G. Meseguer-Brocal, A. Cohen-Hadria, and G. Peeters, "DALI: A large dataset of synchronized audio, lyrics and notes, automatically created using teacher-student machine learning paradigm," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, 2019, pp. 431–437.

[229] N. Condit-Schultz and D. Huron, "Catching the lyrics: Intelligibility in twelve song genres," *Music Percep.: Interdiscipl. J.*, vol. 32, no. 5, pp. 470–483. 2015.

[230] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, "RWC music database: Popular, classical, and jazz music databases," in *Proc. Int. Conf. Music Inf. Retrieval*, 2002, pp. 287–288.

[231] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, "RWC music database: Music genre database and musical instrument sound database," in *Proc. Int. Conf. Music Inf. Retrieval*, 2003, pp. 229–230.

[232] 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.

[233] S. Wager *et al.*, "Intonation: A dataset of quality vocal performances refined by spectral clustering on pitch congruence," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2019, pp. 476–480.

[234] Smule, "DAMP-VSEP: Smule digital archive of mobile performances – vocal separation," Oct. 2019. [Online]. Available: https://doi.org/10.5281/zenodo.3553059

[235] D. Ayllón, F. Villavicencio, and P. Lanchantin, "A strategy for improved phone-level lyrics-to-audio alignment for speech-to-singing synthesis," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2019, pp. 2603–2607.

[236] C. Gupta, D. Grunberg, P. Rao, and Y. Wang, "Towards automatic mispronunciation detection in singing," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, 2017, pp. 390–396.

[237] Y. Bayle *et al.*, "Kara1k: A karaoke dataset for cover song identification and singing voice analysis," in *Proc. Int. Symp. Multimedia*, 2017, pp. 177–184.

[238] R. M. Bittner, J. Salamon, M. Tierney, M. Mauch, C. Cannam, and J. P. Bello, "MedleyDB: A multitrack dataset for annotation-intensive mir research," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, 2014, vol. 14, pp. 155–160.

[239] T.-S. Chan *et al.*, "Vocal activity informed singing voice separation with the Ikala dataset," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2015, pp. 718–722.

[240] A. Liutkus *et al.*, "The 2016 signal separation evaluation campaign," in *Proc. Int. Conf. Latent Variable Anal. Signal Separation*, 2017, pp. 323–332.

[241] A. Liutkus, D. Fitzgerald, Z. Rafii, B. Pardo, and L. Daudet, "Kernel additive models for source separation," *IEEE Trans. Signal Process.*, vol. 62, no. 16, pp. 4298–4310, Aug. 2014.

[242] P. Proutskova, C. Rhodes, G. Wiggins, and T. Crawford, "Breathy or resonant - A controlled and curated dataset for phonation mode detection in singing," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, 2012, pp. 589–594.

Chitralekha Gupta (Member, IEEE) received the Bachelor's degree in engineering from Maharaja Sayajirao University, Vadodara, India, in 2008, the Master's degree in engineering from the Indian Institute of Technology Bombay (IIT-B), Mumbai, India, in 2011, and the Ph.D. degree from the National University of Singapore (NUS), Singapore, in 2019. She is currently a Postdoctoral Research Fellow with NUS. Her research interests include singing voice analysis, applications of ASR in music, and neural audio synthesis. She was awarded a start-up grant from the Graduate Research Innovation Program of NUS and has founded a music tech company MuSigPro Pte. Ltd. Singapore, in 2019. She was the recipient of the NUS Dean's Graduate Research Achievement Award 2018, and the Best Student Paper Award in APSIPA 2017. She was a co-captain at MIREX 2020 for the tasks lyrics-to-audio alignment and lyrics transcription, and has played an active role in the organizing committees of international conferences, such as ISMIR 2022 and 2017, ICASSP 2022, and ASRU 2019.

Haizhou Li (Fellow, IEEE) received the B.Sc., M.Sc., and Ph.D. degrees in electrical and electronic engineering from the South China University of Technology, Guangzhou, China, in 1984, 1987, and 1990, respectively. He is currently a Presidential Chair Professor with the School of Data Science, The Chinese University of Hong Kong, Shenzhen, China, and the Department of Electrical and Computer Engineering, National University of Singapore, Singapore. His research interests include automatic speech recognition, speech information processing, natural language processing, and neuromorphic computing. He was the Editor-in-Chief of IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH AND LANGUAGE PROCESSING (during 2015–2018), a Member of the Editorial Board of Computer Speech and Language since 2012, a Member of IEEE Speech and Language Processing Technical Committee (during 2013–2015), the President of the International Speech Communication Association (during 2015–2017), the President of Asia Pacific Signal and Information Processing Association (during 2015–2016), and the President of Asian Federation of Natural Language Processing (during 2017–2018). He was the General Chair of ACL 2012, INTERSPEECH 2014, ASRU 2019, and ICASSP 2022. Dr Li is a Fellow of the ISCA, and a Fellow of Academy of Engineering Singapore. He was the recipient of the National Infocomm Award 2002, and the President's Technology Award 2013 in Singapore. He was named Nokia Visiting Professors in 2009 by the Nokia Foundation, and Bremen Excellence Chair Professor in 2019 by the University of Bremen, Germany.

Masataka Goto received the Doctor of Engineering degree from Waseda University, Tokyo, Japan, in 1998. He is currently a Prime Senior Researcher with the National Institute of Advanced Industrial Science and Technology (AIST), Tsukuba, Japan. In 1992, he was one of the first to start working on automatic music understanding and has since been at the forefront of research in music technologies and music interfaces based on those technologies. Over the past 30 years, he has authored or coauthored 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 was a committee member of more than 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.