An Overview of Voice Conversion and Its Challenges: From Statistical Modeling to Deep Learning

Berrak Sisman, *Member, IEEE*, Junichi Yamagishi, *Senior Member, IEEE*, Simon King, *Fellow, IEEE*, and Haizhou Li, *Fellow, IEEE*

(Overview Article)

***Abstract*—Speaker identity is one of the important characteristics of human speech. In voice conversion, we change the speaker identity from one to another, while keeping the linguistic content unchanged. Voice conversion involves multiple speech processing techniques, such as speech analysis, spectral conversion, prosody conversion, speaker characterization, and vocoding. With the recent advances in theory and practice, we are now able to produce human-like voice quality with high speaker similarity. In this article, we provide a comprehensive overview of the state-of-the-art of voice conversion techniques and their performance evaluation methods from the statistical approaches to deep learning, and discuss their promise and limitations. We will also report the recent Voice Conversion Challenges (VCC), the performance of the current state of technology, and provide a summary of the available resources for voice conversion research.**

***Index Terms*—Voice conversion, speech analysis, speaker characterization, vocoding, voice conversion evaluation, voice conversion challenges.**

I. Introduction

VOICE conversion (VC) is a significant aspect of artificial intelligence. It is the study of how to convert one's voice to sound like that of another without changing the linguistic content. Voice conversion belongs to a general technical field of speech synthesis, which converts text to speech or changes the properties of speech, for example, voice identity, emotion, and accents. Stewart, a pioneer in speech synthesis, commented in 1922 [1], the really difficult problem involved in the the artificial production of speech-sounds is not the making of a device which shall produce speech, but in the manipulation of the apparatus. As voice conversion is focused on the manipulation of voice identity in speech, it represents one of the challenging research problems in speech processing.

There has been a continuous effort in quest for effective manipulation of speech properties since the debut of computer-based speech synthesis in the 1950s. The rapid development of digital signal processing in the 1970s greatly facilitated the control of the parameters for speech manipulation. While the original motivation of voice conversion could be simply novelty and curiosity, the technological advancements from statistical modeling to deep learning have made a major impact on many real-life applications, and benefited the consumers, such as personalized speech synthesis [2], [3], communication aids for the speech-impaired [4], speaker de-identification [5], voice mimicry [6] and disguise [7], and voice dubbing for movies.

In general, a speaker can be characterized by three factors that are 1) linguistic factors that are reflected in sentence structure, lexical choice, and idiolect; 2) supra-segmental factors such as the prosodic characteristics of a speech signal, and 3) segmental factors that are related to short term features, such as spectrum and formants. When the linguistic content is fixed, the supra-segment and the segmental factors are the relevant factors concerning speaker individuality. An effective voice conversion technique is expected to convert both the supra-segment and the segmental factors. Despite much progress, voice conversion is still far from perfect. In this paper, we celebrate the technological advances, at the same time we expose their limitations. We will discuss the state-of-the-art technology from historical and technological perspectives.

A typical voice conversion pipeline includes a speech analysis, mapping, and reconstruction modules as illustrated in Fig. 1, that is referred to as analysis-mapping-reconstruction pipeline. The speech analyzer decomposes the speech signals of a source speaker into features that represent supra-segmental and segmental information, and the mapping module changes them towards the target speaker, finally the reconstruction module re-synthesizes time-domain speech signals. The mapping module has taken the centre stage in many of the studies. These techniques can be categorized in different ways, for example, based

Manuscript received August 3, 2020; revised October 24, 2020; accepted October 31, 2020. Date of publication November 17, 2020; date of current version December 7, 2020. The work of Berrak Sisman was supported by SUTD Start-up Grant Artificial Intelligence for Human Voice Conversion under Grant SRG ISTD 2020 158 and SUTD AI Grant titled: "The Understanding and Synthesis of Expressive Speech by AI" (PIE-SGP-AI-2020-02). The work of Haizhou Li was supported by the National Research Foundation, Singapore under its AI Singapore Programme under AISG Award No: AISG-GC-2019-002 and AISG-100E-2018-006. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Heiga Zen. *(Corresponding author: Berrak Sisman.)*

Berrak Sisman is with the Information Systems Technology and Design (ISTD) Pillar of Singapore University of Technology and Design (SUTD), 487372, Singapore (e-mail: berraksisman@u.nus.edu).

Junichi Yamagishi is with the National Institute of Informatics, Japan and University of Edinburgh, Edinburgh 101-8430, U.K. (e-mail: jyamagis@nii.ac.jp).

Simon King is with the University of Edinburgh, Edinburgh EH8 9AB, U.K. (e-mail: Simon.King@ed.ac.uk).

Haizhou Li is with the Department of Electrical and Computer Engineering, National University of Singapore 117583, Singapore (e-mail: haizhou.li@nus.edu.sg).

Digital Object Identifier 10.1109/TASLP.2020.3038524

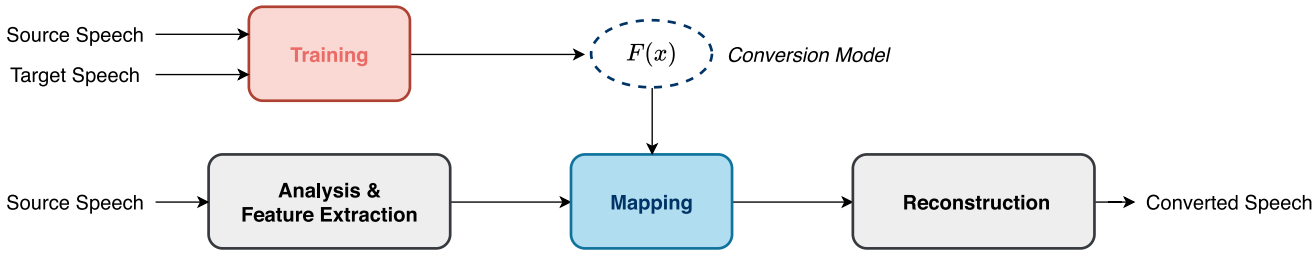

Fig. 1. Typical flow of a voice conversion system. The pink box represents the training of the mapping function, while the blue box applies the mapping function at run-time, in a 3-step pipeline process $\mathcal{Y} = (\mathsf{R} \circ \mathsf{F} \circ \mathsf{A})(\mathcal{X})$.

on the use of training data - parallel vs non-parallel, the type of statistical modeling technique - parametric vs non-parametric, the scope of optimization - frame level vs utterance level, and the workflow of conversion - direct mapping vs inter-lingual. Let's first give an account from the perspective of the use of training data. The early studies of voice conversion were focused on spectrum mapping using parallel training data, where speech of the same linguistic content is available from both the source and target speaker, for example, vector quantization (VQ) [8] and fuzzy vector quantization [9]. With parallel data, one can align the two utterances using Dynamic Time Warping [10]. The statistical parametric approaches can benefit from more training data for improved performance, just to name a few, Gaussian mixture model [11]–[13], partial least square regression [14] and dynamic kernel partial least squares regression (DKPLS) [15].

One of the successful statistical non-parametric techniques is based on non-negative matrix factorization (NMF) [16] and it is known as the exemplar-based sparse representation technique [17]–[20]. It requires a smaller amount of training data than the parametric techniques, and addresses well the over-smoothing problem. We note that the muffled sound effect occurs when the spectra are smoothed. The family of sparse representation techniques include phonetic sparse representation, group sparsity implementation [21], [22], that greatly improved the voice quality on small parallel training dataset.

The studies on voice conversion towards non-parallel training data [23]–[28] open up the opportunities for new applications. The challenge is how to establish the mapping between non-parallel source and target utterances. The INCA alignment technique by Erro *et al.* [27] represent one of the solutions to the non-parallel data alignment problem [29]. With the alignment techniques, one is able to extend the voice conversion techniques from parallel data to non-parallel data, such as the extension to DKPLS [30] and speaker model alignment method [31]. Phonetic Posteriograms, or PPG-based [32], approach represents another direction of research towards non-parallel training data. While the alignment technique doesn't use external resources, the PPG-based approach makes use of automatic speech recognizer to generate intermediate phonetic representation [33], [34] as the inter-lingual between the speakers. Successful applications include Phonetic Sparse Representation [22].

Wu and Li [6], and Mohammadi and Kain [35] provided an overview of voice conversion systems from the perspective of time alignment of speech features followed by feature mapping, that represents the statistical modeling school of thought. The advent of deep learning techniques represents an important technology milestone in the voice conversion research [36]. It has not only greatly advanced the state-of-the-art, but also transformed the way we formulate the voice conversion research problems. It also opens up a new direction of research beyond the parallel and non-parallel data paradigm. Nonetheless, the studies on statistical modeling approaches have provided profound insights into many aspects of the research problems that serve as the foundation work of today's deep learning methodology. In this paper, we will give an overview of voice conversion research by providing a perspective that reveals the underlying design principles from statistical modeling to deep learning.

Deep learning's contributions to voice conversion can be summarized in three areas. Firstly, it allows the mapping module to learn from a large amount of speech data, therefore, tremendously improves voice quality and similarity to target speaker. With neural networks, we see the mapping module as a nonlinear transformation function [37], that is trained from data [38], [39]. LSTM represents a successful implementation with parallel training data [40]. Deep learning made a great impact on non-parallel data techniques. The joint use of DBLSTM and i-vector [41], KL divergence and DNN-based approach [42], variational auto-encoder [43], average modeling [44], DBLSTM based Recurrent Neural Networks [32], [45] and end-to-end Blow model [46] bring the voice quality to a new height. More recently, Generative Adversarial Networks such as VAW-GAN [47], CycleGAN [48]–[50], and many-to-many mapping with StarGAN [51] further advance the state-of-the-art.

Secondly, deep learning has created a profound impact on vocoding technology. Speech analysis and reconstruction modules are typically implemented using a traditional parametric vocoder [11]–[13], [52]. The parameters of such vocoders are manually tuned according to some over-simplified assumptions in signal processing. As a result, the parametric vocoders offer a suboptimal solution. Neural vocoder is a neural network that learns to reconstruct an audio waveform from acoustic features [53]. For the first time, neural vocoder becomes trainable and data-driven. WaveNet vocoder [54] represents one of the popular neural vocoders, that directly estimates waveform samples from the input feature vectors. It has been studied intensively, for example, speaker dependent and independent WaveNet vocoder [54], [55], quasi-periodic WaveNet vocoder [56], [57], adaptive WaveNet vocoder with GANs [58], factorized WaveNet vocoder [59], and refined WaveNet vocoder with VAEs [60]

that are known for their natural sounding voice quality. WaveNet vocoder has been widely adopted in traditional voice conversion pipelines, such as GMM [55], sparse representation [61], [62] systems. Other successful neural vocoders include WaveRNN vocoder [63], WaveGlow [64] and FloWaveNet [65] that are excellent vocoders in their own right.

Thirdly, deep learning represents a departure from the traditional analysis-mapping-reconstruction pipeline. All the above techniques largely follow the voice conversion pipeline as in Fig. 1. As neural vocoder is trainable, it can be trained jointly with mapping module [58] and even with analysis module to become end-to-end solution [66].

Voice conversion research used to be a niche area in speech synthesis. However, it has become a major topic in recent years. In the 45th International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2020), voice conversion papers represent more than one-third of the papers under the speech synthesis category. The growth of research community was accelerated by collaborative activities across academia and industry, such as voice conversion challenge (VCC) 2016, which was first launched [67]–[69] at INTERSPEECH 2016. VCC 2016 is focused on the most basic voice conversion task, that is voice conversion for parallel training data recorded in acoustic studio. It establishes the evaluation methodology and protocol for performance benchmarking, that are adopted widely in the community. VCC 2018 [70]–[72] proposes a non-parallel training data challenge, and also connects voice conversion with anti-spoofing of speaker verification studies. VCC 2020 puts forward a cross-lingual voice conversion challenge for the first time. We will provide an overview of the series of challenges and the publicly available resources in this paper.

This paper is organized as follows: In Section II, we present the typical flow of voice conversion that includes feature extraction, feature mapping and waveform generation. In Section III, we study the statistical modeling for voice conversion with parallel training data. In Section VIII, we study statistical modeling for voice conversion without parallel training data. In Section IX, we study the deep learning approaches for voice conversion with parallel training data, and beyond parallel training data. In Section VI, we explain the evaluation techniques for voice conversion. In Section VII and VIII, we summarize the series of voice conversion challenges, and publicly available research resources for voice conversion. We conclude in Section IX.

II. Typical Flow of Voice Conversion

The goal of voice conversion is to modify a source speaker's voice to sound as if it is produced by a target speaker. In other words, a voice conversion system only modifies the speaker-dependent characteristics of speech, such as formants, fundamental frequency (F0), intonation, intensity and duration, while carrying over the speaker-independent speech content.

The core module of a voice conversion system performs the conversion function. Let's denote the source and target speech signals as $\mathcal{X}$ and $\mathcal{Y}$ respectively. As will be discussed later, voice conversion is typically applied to some intermediate representation of speech, or speech feature, that characterizes a speech frame. Let's denote the source and target speech features as $\mathbf{x}$ and $\mathbf{y}$. The conversion function can be formulated as follows,

$$\mathbf{y} = \mathsf{F}(\mathbf{x}) \tag{1}$$

where $\mathsf{F}(\cdot)$ is also called frame-wise mapping function in rest of this paper. As illustrated in Fig. 1, a typical voice conversion framework is implemented in three steps: 1) speech analysis, 2) feature mapping, and 3) speech reconstruction, that we call the analysis-mapping-reconstruction pipeline. We discuss in detail next.

A. Speech Analysis and Reconstruction

The speech analysis and reconstruction are two crucial processes in the 3-step pipeline. The goal of speech analysis is to decompose speech signals into some form of intermediate representation for effective manipulation or modification with respect to the acoustic properties of speech. There have been many useful intermediate representation techniques that were initially studied for speech communication, and speech synthesis. They become handy for voice conversion. In general, the techniques can be categorized into model-based representations, and signal-based representations.

In model-based representation, we assume that speech signal is generated according to a underlying physical model, such as source-filter model, and express a frame of speech signal as a set of model parameters. By modifying the parameters, we manipulate the input speech. In signal-based representation, we don't assume any models, but rather represent speech as a composition of controllable elements in time domain or frequency domain. Let's denote the intermediate representation for source speaker as $\mathbf{x}$, speech analysis can be described by a function,

$$\mathbf{x} = \mathsf{A}(\mathcal{X}) \tag{2}$$

Speech reconstruction can be seen as an inverse function of the speech analysis, that operates on the modified parameters and generates an audible speech signal. It works with speech analysis in tandem. For example, A vocoder [52] is used to express a speech frame with a set of controllable parameters that can be converted back into a speech waveform. A Griffin-Lim algorithm is used to reconstruct a speech signal from a modified short-time Fourier transform after amplitude modification [73]. As the output speech quality is affected by the speech reconstruction process, speech reconstruction is also one of the important topics in voice conversion research. Let's denote the modified intermediate representation and the reconstructed speech signal for target speaker as $\mathbf{y}$ and $\mathcal{Y} = \mathsf{R}(\mathbf{y})$, voice conversion can be described by a composition of three functions,

$$\begin{aligned} \mathcal{Y} &= (\mathsf{R} \circ \mathsf{F} \circ \mathsf{A})(\mathcal{X}) \\ &= \mathsf{C}(\mathcal{X}) \end{aligned} \tag{3}$$

that represents the typical flow of a voice conversion system as a 3-step pipeline. As the mapping is applied frame-by-frame, the number of converted speech features $\mathbf{y}$ is the same as that of the source speech features $\mathbf{x}$ if speech duration is not modified in the process.

While speech analysis and reconstruction make possible voice conversion, just like other signal processing techniques, they inevitably also introduce artifacts. Many studies were devoted to minimize such artifacts. We next discuss the most commonly used speech analysis and reconstruction techniques in voice conversion.

1) Signal-Based Representation: Pitch Synchronous Over-Lap and Add (PSOLA) is an example of signal-based representation techniques. It decomposes a speech signal into overlapping speech segments [74], each of which represents one of the successive pitch periods of the speech signal. By overlap-and-adding these speech segments with a different pitch periods, we can reconstruct the speech signal of a different intonation. As PSOLA operates directly on the time-domain speech signal [74], the analysis and reconstruction do not introduce significant artifacts. While PSOLA technique is effective for modification of fundamental frequency of speech signals, it suffers from several inherent limitations [75], [76]. For example, unvoiced speech signal is not periodic, and the manipulation of time-domain signal not straightforward.

Harmonic plus Noise Model (HNM) represents another signal-based representation approach. It works under the assumption that a speech signal can be represented as a harmonic component plus a noise component that is delimited by the so-called maximum voiced frequency [77]. The harmonic component is modeled as the sum of harmonic sinusoids up to the maximum voiced frequency, while the noise component is modeled as Gaussian noise filtered by a time-varying autoregressive filter. As HNM decomposition is represented by some controllable parameters, it allows for easy modification speech [78], [79].

2) Model-Based Representation: The model-based technique assumes that the input signal can be mathematically represented by a model whose parameters vary with time. A typical example is the source-filter model that represents a speech signal as the outcome of an excitation of the larynx (source) modulated by a transfer (filter) function determined by the shape of the supralaryngeal vocal tract. A vocoder, a short form of voice coder, was initially developed to minimize the amount of data that are transmitted for voice communication. It encodes speech into slowly changing control parameters, such as linear predictive coding and mel-log spectrum approximation [80], that describe the filter, and re-synthesizes the speech signal with the source information at the receiving end. In voice conversion, we convert the speech signals from a source speaker to mimic the target speaker by modifying the controllable parameters.

The majority of vocoders are designed based on some form of the source-filter model of speech production, such as mixed excitation with a spectral envelope, and glottal vocoders [81]. STRAIGHT or "Speech Transformation and Representation using Adaptive Interpolation of weiGHTed spectrum" is one of the popular vocoders in speech synthesis and voice conversion [82]. It decomposes a speech signal into: 1) a smooth spectrogram which is free from periodicity in time and frequency; 2) a fundamental frequency (F0) contour which is estimated using a fixed-point algorithm; and 3) a time-frequency periodicity map which captures the spectral shape of the noise and its temporal envelope. STRAIGHT is widely used in voice conversion because its parametric representation facilitates the statistical modeling of speech, that allows for easy manipulation of speech [11], [83], [84].

Parametric vocoders are widely adopted for analysis and reconstruction of speech in voice conversion studies [8], [9], [11], [12], [47], [48], [85], [86], and continue to play a major role today [17], [21], [22]. The traditional parametric vocoders are designed to approximate the complex mechanics of the human speech production under certain simplified assumptions. For example, the interaction between F0 and formant structure is ignored, and the original phase structure is discarded [87]. The assumption of stationary process in the short-time window, and time-invariant linear filter, also give rise to "robotic" and "buzzy" voice. Such problems become more serious in voice conversion as we modify both F0 and the formant structure of speech among others at the same time. We believe that vocoding can be improved by considering the interaction between the parameters.

3) WaveNet Vocoder: Deep learning offers a solution to some of the inherent problems of parametric vocoders. WaveNet [66] is a deep neural network that learns to generate high quality time-domain waveform. As it doesn't assume any mathematical model, it is a data-driven solution that requires a large amount of training data.

The joint probability of a waveform $\mathcal{X} = x_1, x_2, ..., x_N$ can be factorized as a product of conditional probabilities.

$$p(\mathcal{X}) = \prod_{n=1}^{N} p(x_n | x_1, x_2, ..., x_{n-1}) \tag{4}$$

A WaveNet is constructed with many residual blocks, each of which consists of 2×1 dilated causal convolutions, a gated activation function and 1×1 convolutions. With additional auxiliary features h, WaveNet can also model conditional distribution $p(x|h)$ [66]. Eq. (4) can then be written as follows:

$$p(\mathcal{X}|h) = \prod_{n=1}^{N} p(x_n | x_1, x_2, ..., x_{n-1}, h) \tag{5}$$

A typical parametric vocoder performs both analysis and reconstruction of speech. However, most of today's WaveNet vocoders only cover the function of speech reconstruction. It takes some intermediate representations of speech as the input auxiliary features, and generate speech waveform as the output. WaveNet vocoder [54] outperforms remarkably the traditional parametric vocoders in terms of sound quality. Not only can it learn the relationship between input features and output waveform, but also it learns the interaction among the input features. It has been successfully adopted as part of the state-of-the-art speech synthesis [3], [88]–[91] and voice conversion [54], [55], [57], [60]–[62], [88], [92]–[99] systems.

There have been promising studies on using vocoding parameters as the intermediate representations in WaveNet vocoding. A speaker independent WaveNet vocoder [54] is studied by utilizing the STRAIGHT vocoding parameters, such as F0, aperiodicity, and spectrum as the inputs of WaveNet. In this way, WaveNet learns a sample-by-sample correspondence between

the time-domain waveform and the input vocoding parameters. When such a WaveNet vocoder is trained on speech signals from a large speaker population, we obtain a speaker independent vocoder [54]. By adapting the speaker independent WaveNet vocoder with speaker specific data, we obtain a speaker dependent vocoder that generates personalized voice output [58], [60]. The study on WaveNet vocoder also opens up opportunities for the use of other non-vocoding parameters as the input. For example, a recent study adopts phonetic posteriogram (PPG) in WaveNet vocoding with promising results in voice conversion with non-parallel training data [96]–[99]. Another study adopts latent code of autoencoder and speaker embedding as the speech representation for WaveNet vocoder [100].

4) Recent Progress on Neural Vocoders: More recently, speaker independent WaveRNN-based neural vocoder [63] became popular as it can generate human-like voices from both in-domain and out-of-domain spectrogram [101]–[103]. Another well-known neural vocoder that achieves high-quality synthesis performance is WaveGlow [64]. WaveGlow is a flow-based network capable of generating high quality speech from mel-spectrogram [104]. WaveGlow benefits from the best of Glow and WaveNet so as to provide fast, efficient and high-quality audio synthesis, without the need for auto-regression. We note that WaveGlow is implemented using only a single network with a single cost function, that is to maximize the likelihood of the training data, which makes the training procedure simple and stable [105].

WaveNet [66] uses an auto-regressive (AR) approach to model the distribution of waveform sampling points, that incurs a high computational cost. As an alternative to auto-regression, a neural source-filter (NSF) waveform modeling framework is proposed [106], [107]. We note that NSF is straightforward to train and fast to generate waveform. It is reported 100 times faster than WaveNet vocoder, and yet achieving comparable voice quality on a large speech corpus [108].

More recently, Parallel WaveGAN [109] has also been proposed to generate high-quality voice using a generative adversarial network. Parallel WaveGAN is a distillation-free and fast waveform generation method, where a non-autoregressive WaveNet is trained by jointly optimizing multi-resolution spectrogram and adversarial loss functions. We note that Parallel WaveGAN is able to generate high-fidelity speech even with its compact architecture. We note that generating coherent raw audio waveforms with GANs is challenging. Another GAN method for generating high quality audio waveform is known as MelGAN [110]. MelGAN shows the effectiveness of GAN-based approaches for high quality mel-spectrogram inversion in speech synthesis, music domain translation and unconditional music synthesis.

B. Feature Extraction

With speech analysis, we derive vocoding parameters that usually contains spectral and prosodic components to represent the input speech. The vocoding parameters characterize the speech in a way that we can reconstruct the speech signal later on after transmission. This is particularly important in speech communication. However, such vocoding parameters may not be the best for transformation of voice identity. More often, the vocoding parameters are further transformed into speech features, that we call feature extraction in Fig. 1, for more effective modification of the acoustic properties in voice conversion.

For the spectral component, feature extraction aims to derive low-dimensional representations from the high-dimensional raw spectra. Generally speaking, the spectral features are be able to represent the speaker individuality well. The feature not only fit the spectral envelope well, but also be able to be converted back to spectral envelope. They should have good interpolation properties that allow for flexible modification.

The magnitude spectrum can be warped to Mel or Bark frequency scale that are perceptually meaningful for voice conversion. It can also be transformed into cepstral domain using a finite number of coefficients using the Discrete Cosine Transform of log-magnitude. Cepstral coefficients are less correlated. In this way, high dimension magnitude spectrum is transformed to lower dimension feature representation. The commonly used speech features include Mel-cepstral coefficients (MCEP), linear predictive cepstral coefficients (LPCC), and line spectral frequencies (LSF). Typically, a speech frame is represented by a feature vector.

Short-time analysis has been the most practical way of speech analysis. Unfortunately it inherently ignores the temporal context of speech, that is crucial in voice conversion. Many studies have shown that multiple frames [18], [111], dynamic features [62], and phonetic segments serve as effective features in feature mapping.

For the prosodic component, feature extraction can be used to decompose prosodic signal, such as fundamental frequency (F0), aperiodicity (AP), and energy contours, into speaker dependent and independent parameters [84]. In this way, we can carry over the speaker independent prosodic patterns, while converting speaker dependent ones during the feature mapping.

C. Feature Mapping

In the typical flow of voice conversion, feature mapping performs the modification of speech features from source to target speaker. Spectral mapping seeks to change the voice timbre, while prosody conversion seeks to modify the prosody features, such as fundamental frequency, intonation and duration. So far, spectral mapping remains the center of many voice conversion studies.

During training, we learn the mapping function, $\mathsf{F}(\cdot)$ in Eq.(1), from training data. At run time inference, the mapping function transforms the acoustic features. A large part of this paper is devoted to the study of the mapping function. In Section III, we will discuss the traditional statistical modeling techniques with parallel training data. In Section VIII, we will review the statistical modeling techniques that do not require parallel training data. In Section IX, we will introduce a number of deep learning approaches, which includes 1) parallel training data of paired speakers; and 2) beyond parallel data of paired speakers.

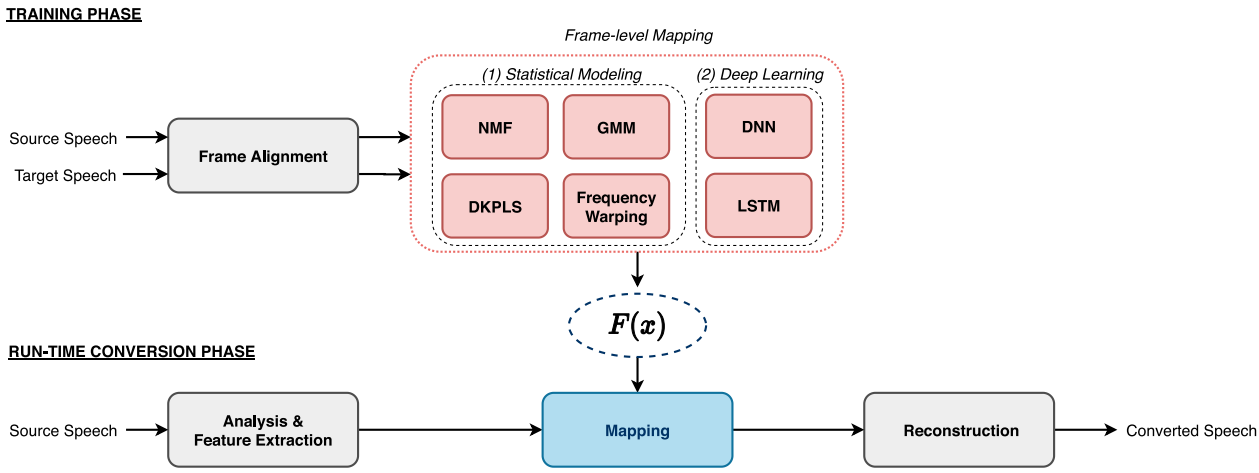

Fig. 2. Training and run-time inference of voice conversion with parallel training data under the frame-level mapping paradigm. The pink boxes represent the training algorithms of the models that result in the mapping function $F(x)$ in blue box for run-time inference. Dotted box (1) includes examples of statistical approaches, and (2) includes examples of deep learning approaches.

III. Statistical Modeling for Voice Conversion With Parallel Training Data

Most of the traditional voice conversion techniques assume availability of parallel training data. In other words, the mapping function is trained on paired utterances of the same linguistic content spoken by source and target speaker. Voice conversion studies started with statistical approaches [112] in late 1980s, that can be grouped into parametric and non-parametric mapping techniques. Parametric techniques makes assumptions about the underlying statistical distributions of speech features and their mapping. Non-parametric ones make fewer assumptions about the data, but seek to fit the training data with the best mapping function, while maintaining some ability to generalize to unseen data.

Parametric techniques, such as Gaussian mixture model (GMM) [113], Dynamic Kernel Partial Least Square Regression, PSOLA mapping technique [75], represent a great success in the recent past. The vector quantization approach to voice conversion is a typical non-parametric technique. It maps codewords between source and target codebooks [8]. In this method, a source feature vector is approximated by the nearest codeword in the source codebook, and mappped to the corresponding codeword in the target codebook. To reduce the quantization error, fuzzy vector quantization was studied [9], [114], where continuous weights for individual clusters are determined at each frame according to the source feature vector. The converted feature vector is defined as a weighted sum of the centroid vectors of the mapping codebook. Recently, the non-negative factorization approach marks a successful non-parametric implementation.

We will discuss a typical frame-level mapping paradigm under the assumption of parallel training data, as illustrated in Fig. 2. During the training phase, given parallel training data from a source speaker $\mathbf{x}$ and a target speaker $\mathbf{y}$, frame alignment is performed to align the source speech vectors and target speech vectors to obtain the paired speech feature vector $\mathbf{z} = \{\mathbf{x}, \mathbf{y}\}$. Dynamic time warping is feature-based alignment technique that is commonly used. Speech recognizer, that is equipped with phonetic knowledge, can also be used to perform model-based alignment. Frame alignment has been well studied in speech processing. In voice conversion, a large body of literature has been devoted to the design of frame-level mapping function.

A. Gaussian Mixture Models

In Gaussian mixture modeling (GMM) approach to voice conversion [113], we represent the relationship between two sets of spectral envelopes, from source and target speakers, using a Gaussian mixture model. The Gaussian mixture model is a continuous parametric function, that is trained to model the spectral mapping. In [113], harmonic plus noise (HNM) features are used in the feature mapping, which allows for high-quality modifications of speech signals. The GMM approach is seen as an extension to the vector quantization approach [8], [9], that results in improved voice quality. However, the speech quality is affected by some factors, e.g., spectral movement with inappropriate dynamic characteristics caused by the frame-by-frame conversion process, and excessive smoothing of converted spectra [115]–[117].

To address the frame-by-frame conversion issue, a maximum likelihood estimation technique was studied to model the spectral parameter trajectory [11]. This technique aims to estimate an appropriate spectrum sequence using dynamic acoustic features. To address the over-smoothing issue, or the muffled effect, joint density Gaussian mixture model (JD-GMM) was studied [2], [11] to jointly model the sequences of spectral features and their variances using maximum likelihood estimation, that increases the global variance of the spectral features. The JD-GMM method involves two phases: off-line training and run-time conversion phases. During the training phase, Gaussian mixture model (GMM) is adopted to model the joint probability density $p(\mathbf{z})$ of the paired feature vector sequence $\mathbf{z} = \{\mathbf{x}, \mathbf{y}\}$, which represents the joint distribution of source speech $\mathbf{x}$ and target

speech $\mathbf{y}$:

$$p(\mathbf{z}) = \sum_{k=1}^{K} w_k^{(z)} \mathcal{N}\left(\mathbf{z}|\mu_k^z, \Sigma_k^{(z)}\right)$$
$$\mu_k^z = \begin{bmatrix} \mu_k^x \\ \mu_k^y \end{bmatrix}, \Sigma_k^{(z)} = \begin{bmatrix} \Sigma_k^{(xx)} & \Sigma_k^{(xy)} \\ \Sigma_k^{(yx)} & \Sigma_k^{(yy)} \end{bmatrix} \quad (6)$$

where K is the number of Gaussian components, w_k is the weight of each Gaussian, μ_k^z and $\Sigma_k^{(z)}$ are the mean vector and the covariance matrix of the kth Gaussian component $\mathcal{N}(\mathbf{z}|\mu_k^z, \Sigma_k^{(z)})$, respectively. To estimate the model parameters of the JD-GMM, expectation-maximization (EM) algorithm [118]–[121] is used to maximize likelihood on the training data. During the run-time conversion phase, JD-GMM model parameters are used to estimate the conversion function. We note that JD-GMM training method provides estimates of the model parameters robustly, especially when the amount of training data is limited.

A post-filter based on modulation spectrum modification is found useful to address the inherent over-smoothing issue in statistical modeling [122], such as GMM approach, which effectively compensates the global variance. The GMM approach is a parametric solution [123]–[127]. It represents a successful statistical modeling technique that works well with parallel training data.

B. Dynamic Kernel Partial Least Squares

The family of parametric techniques also include linear [75], [76] or non-linear mapping functions. With the local mapping functions, each frame of speech is typically transformed independently from the neighboring frames, which causes temporal discontinuities to the output [76].

To take into account the time-dependency between speech features, a dynamic kernel partial least squares (DKPLS) technique was studied [15]. This method is based on a kernel transformation of the source features to allow non-linear modeling, and concatenation adjacent frames to model the dynamics. The non-linear transformation takes advantage of the global properties of the data that GMM approach doesn't. It was reported that DKPLS outperforms GMM approach [113] in terms of voice quality. This method is simple and efficient, and does not require massive tuning. More recently, DKPLS-based approaches are studied to overcome the over-fitting and over-smoothing problems by feature combination strategy [128].

While statistical modeling for the mapping of spectral features has been well studied, conversion of prosody is often achieved by simply shifting and scaling F0, which is not sufficient for high-quality voice conversion. Hierarchical modeling of prosody, for different linguistic units at several distinct temporal scales, represents an advanced technique for prosody conversion [84], [129]–[131]. DKPLS has created a platform for multi-scale prosody conversion through wavelet transform [132] that shows significant improvement in naturalness over the F0 shifting and scaling technique.

C. Frequency Warping

Parametric techniques, such as GMM [113] and DKPLS [15], usually suffer from over-smoothing because they use the minimum mean square error [83] or the maximum likelihood [11] function as the optimization criterion. As a result, the system produces acoustic features that represent statistical average, and fails to capture the desired details of temporal and spectral dynamics.

Additionally, parametric techniques generally employ low-dimensional features, as discussed in Section II.B, such as the Mel-cepstral coefficients (MCEP) or line spectral frequencies (LSF) to avoid the curse of dimensionality. The low dimensional features, however, are doomed to lose spectral details because they have low-resolution. Statistical averaging and low-resolution features both lead to the muffled effect of output speech [133].

To preserve the necessary spectral details during conversion, a number of frequency warping-based methods were introduced. The frequency warping technique directly transforms the high resolution source spectrum to that of the target speaker through a frequency warping function. In recent literature, the warping function is either realized by a single parameter, such as VTLN-based approaches [26], [134]–[137], or represented as a piecewise linear function [75], [133], [138], which has become a mainstream solution.

The goal of piecewise linear warping function is to align a set of frequencies between the source and target spectrum by minimizing the spectral distance or maximizing the correlation between the converted and target spectrum. More recently, the parametric frequency warping technique was incorporated with a non-parametric exemplar-based technique, that achieves good performance [111].

D. Non-Negative Matrix Factorization

Non-negative matrix factorization (NMF) [139] is an effective data mining technique that has been widely used, especially for reconstruction of high quality signals, such as in speech enhancement [140], [141], speech de-noising [142], [143], noise and speech estimation [144]. It factorizes a matrix into two matrices, a dictionary and an activation matrix, with the property that all three matrices have no negative elements. The NMF-based techniques are shown effective in voice conversion with very limited training data. It marks a major progress of non-parametric approach to voice conversion since vector quantization technique was introduced. Successful implementation includes non-negative spectrogram deconvolution [145], locally linear embedding (LLE) [146], and unit selection [20]. In NMF-based approaches, a target spectrogram is constructed as a linear combination of exemplars. Therefore, over-smoothing problem can also arise. To overcome the over-smoothing problem, several effective techniques were developed, that we summarize next.

1) Sparse Representation: One effective way to alleviate the over-smoothing problem is to apply sparsity constraint to the activation matrix, referred to as exemplar-based sparse representation. As illustrated in Fig. 3, a pair of dictionaries $\mathbf{A}$ and $\mathbf{B}$ are first constructed from speech feature vectors, that we

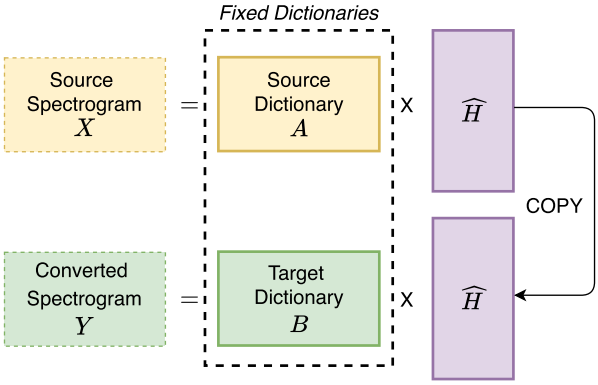

Fig. 3. Illustration of non-negative matrix factorization for exemplar-based sparse representation.

call aligned exemplars, from source and target. $[\mathbf{A}; \mathbf{B}]$ is also called the coupled dictionary. At run-time, let's consider a speech utterance as a sequence of speech feature vectors, that form a spectrogram matrix. The matrix of a source utterance $\mathbf{X}$ can be represented as,

$$\mathbf{X} \approx \mathbf{A}\hat{\mathbf{H}} \tag{7}$$

Due to the non-negative nature of spectrogram, NMF technique is employed to estimate the source activation matrix $\hat{\mathbf{H}}$, which is constrained to be sparse. Mathematically, we estimate $\hat{\mathbf{H}}$ by minimizing an objective function,

$$\hat{\mathbf{H}} = \underset{\mathbf{H} \geq \mathbf{0}}{\operatorname{argmin}}\, \mathbf{d}\left(\mathbf{X}, \mathbf{AH}\right) + \lambda ||\mathbf{H}|| \tag{8}$$

where λ is the sparsity penalty factor. To estimate activation matrix $\hat{\mathbf{H}}$, a generalised Kullback-Leibler (KL) divergence is used. It is assumed that source and target dictionaries $\mathbf{A}$ and $\mathbf{B}$ can share the same source activation matrix $\hat{\mathbf{H}}$. Therefore, the converted spectrogram for the target speaker can be written as,

$$\hat{\mathbf{Y}} = \mathbf{B}\hat{\mathbf{H}}. \tag{9}$$

where the activation matrix $\hat{\mathbf{H}}$ serves as the pivot to transfer source utterance $\mathbf{X}$ to target utterance $\mathbf{Y}$. The sparse representation framework continues to attract much attention in voice conversion. The recent studies include its extension to discriminative graph-embedded NMF approach [19], phonetic sparse representation for spectrum conversion [22], and its application to timbre and prosody conversion [147], [148].

2) Phonetic Sparse Representation: As the frame-level mapping is done at acoustic feature level, the coupled dictionary $[\mathbf{A}; \mathbf{B}]$ is therefore called acoustic dictionary. With the scripts of the training data and a general purpose speech recognizer, we are able to obtain phonetic labels and their boundaries. Studies have shown that the strategy of dictionary construction plays an important role in voice conversion [149]. The idea of selecting sub-dictionary according to the run-time speech content shows improved performance [21]. Phonetic sparse representation [22] is an extension to sparse representation for voice conversion. It is built on the idea of phonetic sub-dictionaries, and dictionary selection at run-time. The study shows that multiple phonetic sub-dictionaries consistently outperform single dictionary in exemplar-based sparse representation voice conversion [21], [22]. However, the phonetic sparse representation relies on a speech recognizer at run-time to help select the sub-dictionary.

3) Group Sparse Representation: Sisman *et al.* [62] proposed group sparse representation to formulate both exemplar-based sparse representation [145], and phonetic sparse representation [22] under a unified mathematical framework. With the group sparsity regularization, only the phonetic sub-dictionary that is relevant to the input features is likely to be activated at run-time inference. Unlike phonetic sparse representation that relies on a speech recognizer for both training and run-time inference, group sparse representation only requires the speech recognizer during training when we build the phonetic dictionary. It was reported that group sparse representation provides similar performance to that of phonetic sparse representation when performing both spectrum and prosody conversion [62].

IV. Statistical Modeling for Voice Conversion With Non-Parallel Training Data

It is easy to understand that it is more straightforward to train a mapping function from parallel than non-parallel training data. However, parallel training data are not always available. In real-world applications, there are situations where only non-parallel data are available. Intuitively, if we can derive the equivalents of speech frames or segments between speakers from non-parallel data, we are able to establish or to refine the mapping function using the conventional linear transformation parameter training, such as GMM, DKPLS or frequency warping.

There were a number of attempts to do so. For example, one idea is to find source-target mapping between unsupervised feature clusters [150]. Another is to use a speech recognizer to index the target training data so that we can retrieve similar frames from target database for a unknown source frame at run-time [151]. Unfortunately, each of the steps may produce errors that accumulate and may lead to a poor parameter estimation [150]. There was also a study to use a hidden Markov model (HMM) that is trained for the target speaker, then the parameters of GMM-based linear transformation function are estimated in such a way that the converted source vectors exhibit maximum likelihood with respect to the target HMM [152]. This method shows comparable performance with methods of parallel data. However, it requires that the orthography of the training utterances be known, that limits its use.

Next we will discuss three clusters of studies and their representative work, 1) INCA algorithm, 2) unit selection algorithm, and 3) speaker modeling algorithm.

A. INCA Algorithm

INCA refers to an Iterative combination of a Nearest Neighbor search step and a Conversion step Alignment method [27]. It learns a mapping function by finding the nearest neighbor of each source vector in the target acoustic space. It is based on a hypothesis that an iterative refinement of the basic nearest neighbour method, in tandem with the voice conversion system, would lead to a progressive alignment improvement. The main idea is that the intermediate voice, $\mathbf{x}_s^k$, obtained after the previous nearest neighbour alignment can be used as the source voice

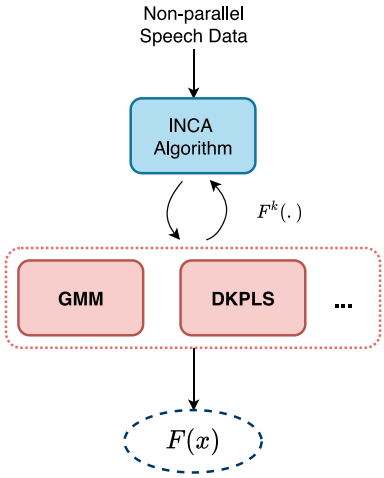

Fig. 4. Training of a frame-wise mapping function is an iterative process between the nearest neighbor search step (INCA alignment) and the conversion step (a parametric mapping function).

during the next iteration.

$$\mathbf{x}_s^{k+1} = \mathsf{F}^k(\mathbf{x}_s^k) \tag{10}$$

During training, the optimization process is repeated until the current intermediate voice, $\mathbf{x}_s^k$, is close enough to target voice, $\mathbf{y}_t$. INCA represents a successful framework for the non-parallel training data problem, where the nearest neighbor search step (INCA alignment) and the conversion step (a parametric mapping function) iterates to optimize the mapping function, as illustrated in Fig. 4.

INCA was first implemented with GMM approach [113] for voice conversion to estimate a linear mapping function. As INCA does not require any phonetic or linguistic information, it not only works for non-parallel training data, but also works for cross-lingual voice conversion. Experiments show that the INCA implementation of a cross-lingual system achieves similar performance to its intra-lingual counterpart that is trained on parallel data [27].

INCA was further implemented with DKPLS approach [15] that was discussed in Section III.B for parallel training data. The idea [30] is to use the INCA alignment algorithm [27] to find the corresponding frames from the source and target datasets, that allows the DKPLS regression to find a non-linear mapping between the aligned datasets. It was reported [30] that the INCA-DKPLS implementation produces high-quality voice that is comparable to implementation with parallel training data on the same amount of training data.

B. Unit Selection Algorithm

Unit selection algorithms have been widely used to generate natural-sounding speech in speech synthesis. It is known to produce high speaker similarity and voice quality [77], [153], [154] because the synthesized waveform is formed of sound units directly from the target speaker [155]. The unit selection algorithm optimizes the unit selection from a voice inventory of a target speaker. It was suggested [156] to make use of unit selection synthesis system to generate parallel versions of the

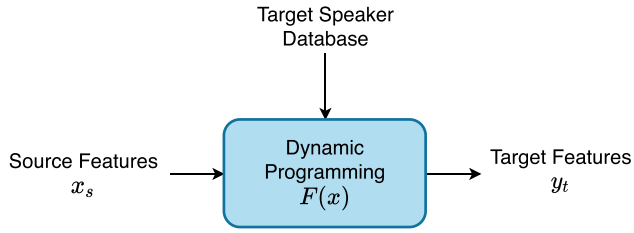

Fig. 5. Run-time inference of unit selection algorithm that doesn't model a mapping function with parameters, but rather searches for output feature sequence directly from target speaker database, and optimizes the output at utterance level.

training sentences from non-parallel data. With the resulting pseudo-parallel data, the statistical modeling techniques for parallel training data, that we discuss in Section III, can be readily applied. While this approach produces satisfactory voice quality [156], it requires a large speech database to develop the the voice inventory, that is not always practical in reality.

Another idea is to follow what we do in unit selection speech synthesis by defining a speech feature vector as a unit [24]. Given an utterance of M speech feature vectors $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_M\}$ from the source speaker, a dynamic programming is applied to find the sequence of feature vectors $\mathbf{y}_i$ from the target speaker, that minimizes a cost function,

$$\mathbf{Y} = \arg\min_{y} \left(\alpha \sum_{i=1}^{M} d_1(\mathbf{x}_i, \mathbf{y}_i) + (1-\alpha) \sum_{i=2}^{M} d_2(\mathbf{y}_i, \mathbf{y}_{i-1}) \right) \tag{11}$$

where $d_1(\cdot)$ represents the acoustic distance between a source and a target feature vector, while $d_2(\cdot)$ is the concatenative cost between two target feature vectors. With the acoustic distance, we make sure that the retrieved speech features from the target speakers are close to those of the source; with the concatenative cost, we encourage the consecutive speech frames from the target speaker database to be retrieved together in a multi-frame segment. As illustrated in Fig. 5, unit selection algorithm is a non-parametric solution because we don't model the conversion with parameters. It optimizes the output by applying a dynamic programming to find the best feature vector sequence from the target speaker database. The mapping function $\mathbf{Y} = \mathsf{F}(\mathbf{X})$ is defined by the cost function Eq. (11) itself, and optimized at the utterance level.

C. Speaker Modeling Algorithm

The techniques for text-independent speaker characterization are readily available for non-parallel training data, where a speaker can be modeled by a set of parameters, such as a GMM or i-vector. One is possible to make use such speaker models to perform voice conversion.

Mouchtaris *et al.* [157] used a GMM-based technique to model relationship between reference speakers in advance and apply the relationship for a new speaker. Toda *et al.* [158] proposed an eigenvoice approach that performs two mappings, one to map from the source speaker to an eigenvoice (or average voice) trained from reference speakers, and another from the

eigenvoice to the target speaker. These approaches don't require parallel training data, they do require parallel data from some reference speakers.

In speaker verification, the joint factor analysis method [159] decomposes a supervector into speaker independent, speaker dependent and channel dependent components, each of which is represented by a low-dimensional set of factors. This aims to disentangle speaker from other speech content for effective speaker verification. Inspired by this idea, we argue [160] that similar decomposition would be useful in voice conversion, where we would like to separate speaker information from the linguistic content, and apply factor analysis on the speaker specific component.

With factor analysis, the speaker specific component can be represented by a low-dimensional set of latent variables via the factor loadings. One of the ideas [160] is to estimate the phonetic component and factor loadings from non-parallel prior data. In this way, during the training process, we only estimate a low-dimensional set of speaker identity factors and a tied covariance matrix instead of a full conversion function from the source-target parallel utterances. Even though parallel utterances are still required for estimating the conversion function, the use of prior data allows us to obtain a reliable model from much fewer training samples than those required by conventional JD-GMM [161].

Another idea is to perform the voice conversion in i-vector [159] speaker space, where i-vector is used to disentangle a speaker from the linguistic content. The primary motivation is that an i-vector can be extracted in an unsupervised manner regardless of speaker or speech content, which opens up new possibilities especially for non-parallel data scenarios where source and target speech is of different content or even in different languages [28], [45], [162]. Kinnunen *et al.* [163] study a way to shift the acoustic features of input speech towards target speech in the i-vector space. The idea is to learn a function that maps the i-vector of the source utterance to that of the target. With the mapping function, we are able to convert the source speech frame-by-frame to the target. This technique is free of any parallel data, and text transcription.

V. Deep Learning for Voice Conversion

Voice conversion is typically a research problem with scarce training data. Deep learning techniques are typically data driven, that rely on big data. However, this is actually the strength of deep learning in voice conversion. Deep learning opens up many possibilities to benefit from abundantly available training data, so that the voice conversion task can focus more on learning the mapping of speaker characteristics. For example, it shouldn't be the job of voice conversion task to infer low level detail during speech reconstruction, a neural vocoder can learn from large database to do so [100]. It shouldn't be a task of voice conversion to learn how to represent an entire phonetic system of a spoken language, a general purpose acoustic model of neural ASR [164] or TTS [165] system can learn from a large database to do so. By leveraging the large database, we free up the conversion network from using its capacity to represent low level detail and general information, but instead, to focus on the high level semantics necessary for speaker identity conversion.

Deep learning techniques also transform the way we implement the analysis-mapping-reconstruction pipeline. For effective mapping, we need to derive adequate intermediate representation of speech, that was discussed in Section II. The concept of *embedding* in deep learning provides a new way of deriving the intermediate representation, for example, latent code for linguistic content, and speaker embedding for speaker identity. It also makes the disentanglement of speaker from speech content much easier.

In this section, we will summarize how deep learning helps address existing research problems, such as parallel and non-parallel data voice conversion. We will also review how deep learning breaks new ground in voice conversion research.

A. Deep Learning for Frame-Aligned Parallel Data

The study on deep learning approaches for voice conversion started with parallel training data, where we use a neural network as an improved regression function to approximate the frame-wise mapping function $\mathbf{y} = \mathsf{F}(\mathbf{x})$ under the frame-level mapping paradigm in Fig. 2.

1) DNN Mapping Function: The early studies on DNN-based voice conversion methods are focused on spectral transformation. DNN mapping function, $\mathbf{y} = \mathsf{F}(\mathbf{x})$, has some clear advantage over other statistical models, such as GMM, and DKPLS. For instance, it allows for non-linear mapping between source and target features, and there is little restriction to the dimension of features to be modeled. We note that conversion on other acoustic features, such as fundamental frequency and energy contour, can also be done similarly [166]. Desai *et al.* [83] proposed a DNN to map a low-dimensional spectral representation, such as mel-cepstral coefficients (MCEP), from source to target speaker. Nakashika *et al.* [167] proposed to use Deep Belief Nets (DBNs) to extract latent features from source and target cepstrum coefficients, and use a neural network with one hidden layer to perform conversion between latent features. Mohammadi *et al.* [168] furthered the idea by studying a deep autoencoder from multiple speakers to derive a compact representations of speech spectral feature. High-dimensional representation of spectrum has also been used in a more recent work [169] for spectral mapping, together with dynamic features and a parameter generation algorithm [170]. Chen *et al.* [171] proposed to model the distributions of spectral envelopes of source and target speakers respectively through a layer-wise generative training. Generally speaking, DNN for spectrum and/or prosody transformation requires a large amount of parallel training data from paired speakers, which is not always feasible. But it opens up opportunities for us to make use of speech data from multiple speakers beyond source and target, to better model the source and the target speakers, and to discover better feature representations for feature mapping.

2) LSTM Mapping Function: To model the temporal correlation across speech frames in voice conversion, Nakashika *et al.* [172] explore the use of Recurrent Temporal Restricted Boltzmann Machines (RTRBM), a type of recurrent neural networks.

The success of Long-Short Term Memory (LSTM) [173], [174] in sequence to sequence modeling inspires the study of LSTM in voice conversion, which leads to an improvement of naturalness and continuity of the speech output.

The LSTM network architecture consists of a set of memory blocks and gates, that support the storage and access to long-range contextual information [175]. LSTM can learn the optimal amount of contextual information for voice conversion. A bidirectional LSTM (BLSTM) network is expected to capture sequential information and maintain long-range contextual features from both forward sequence and backward sequence [45].

Sun *et al.* [40] and Ming *et al.* [176] proposed a deep bidirectional LSTM network (DBLSTM) by stacking multiple hidden layers of BLSTM network architecture, that is shown to outperform DNN voice conversion even without using dynamic features. While DBLSTM-based voice conversion approach generates high-quality synthesized voice, it typically requires a large speech corpus from source and target speakers for training, that limits the scope of the applications in practice [40].

Just like GMM approach, DNN and LSTM techniques rely on external frame aligner during training data preparation, as illustrated in Fig. 2. At run-time, the conversion process follows the typical flow of 3-step pipeline, and doesn't change the speech duration during the conversion.

B. Encoder-Decoder With Attention for Parallel Data

The research problems of voice conversion are centered around alignment and mapping, which are interrelated both during training and at run-time inference, as illustrated in Fig. 2. During training, more accurate alignment helps build better mapping function, that explains why we prefer parallel training data. At run-time inference, the frame-level mapping paradigm doesn't change the duration of the speech during the conversion. While it is possible to model and predict the duration for voice conversion output, it is not straightforward to incorporate duration model and mapping model in a systematic manner. Deep learning provides a new solution to this research problem.

The attention mechanism [177], [178] in encoder-decoder structure neural network brings about a paradigm change. The idea of attention was first successfully used in machine translation [177], speech recognition [179], and sequence-to-sequence speech synthesis [88], [180]–[182], that led to many parallel studies in voice conversion [183], [184]. With the attention mechanism, the neural network learns the feature mapping and alignment at the same time during training. At run-time inference, the network automatically decides the output duration according to what it has learnt. In other words, the frame-aligner in Fig. 2 is no longer required.

There are several variations based on recurrent neural networks, such as sequence-to-sequence conversion network (SCENT) [183], and AttS2S-VC [184]. They follow the widely-used architecture of encoder-decoder with attention [185], [186]. Suppose that we have a source speech $\mathbf{x} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_{T_s}\}$. The encoder network first transforms the input feature sequences into hidden representations, $\mathbf{h} = \{\mathbf{h}_1, \mathbf{h}_2, ..., \mathbf{h}_{T_h}\}$ at a lower frame rate with $T_h < T_s$, which are suitable for the decoder to deal with. At each decoder time step, the attention module aggregates the encoder outputs by attention probabilities and produces a context vector. Then, the decoder predicts output acoustic features frame by frame using context vectors. Furthermore, a post-filtering network is designed to enhance the accuracy of the converted acoustic features to generate the converted speech $\mathbf{y} = \{\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_{T_y}\}$. During training, the attention mechanism learns the mapping dynamics between source sequence and target sequence. At run-time inference, the decoder and the attention mechanism interacts to perform the mapping and alignment at the same time. The overall architecture is illustrated in Fig. 6.

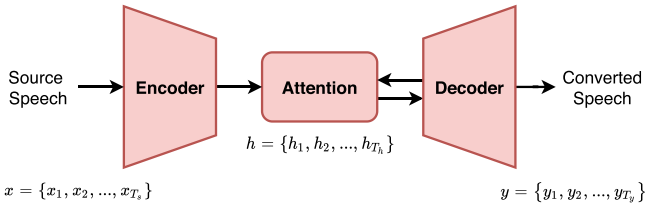

Fig. 6. Encoder-decoder mechanism with attention for voice conversion.

While recurrent neural networks represent an effective implementation for sequence-to-sequence conversion, recent studies have shown that convolutional neural networks also learn well the long-term dependencies [66], [187]. It employs an attention mechanism that effectively makes possible parallel computations for encoding and decoding. During decoding, the causal convolution design allows the model to generate an output sequence in an autoregressive manner. Kameoka *et al.* proposed a convolutional neural networks implementation for voice conversion [188], that is called ConvS2S-VC. Recent studies show that ConvS2S-VC outperforms its recurrent neural network counterparts in both pairwise and many-to-many voice conversion [184]. The encoder-decoder structure with attention marks a departure from the frame-level mapping paradigm. The attention doesn't perform the mapping frame-by-frame, but rather allows the decoder to attend to multiple speech frames and uses the soft combination to predict an output frame in the decoding process. With the attention mechanism, the duration of the converted speech T_y is typically different from that of the source speech T_s to reflect the differences of speaking style between source and target. This represents a way to handle both spectral and prosody conversion at the same time. The studies have attributed the improvement of voice quality to the effective attention mechanism. The attention mechanism also represents the first step towards relaxing the rigid requirement of parallel data in voice conversion.

C. Beyond Parallel Data of Paired Speakers

In Section III and IV, we study statistical modeling for voice conversion with parallel training data and non-parallel training data. The advent of deep learning has broken new ground for voice conversion research. We now go beyond the paradigm of parallel and non-parallel training data. Traditionally, *nonparallel training data* refers to the case where nonparallel utterances

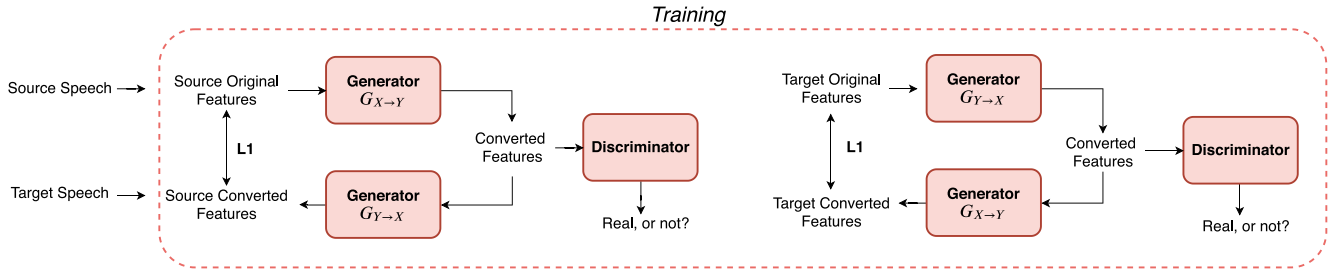

Fig. 7. Training a CycleGAN with cycle-consistency loss of L1 norm for voice conversion with non-parallel training data of paired speakers. L1 norm represents the least absolute errors.

from source-target speaker pair are required. However, the recent studies show that, deep learning has enabled many voice conversion scenarios without the need of parallel data. In this section, we summarize the studies into four scenarios,

1) Non-parallel data of paired speakers,
2) Leveraging TTS systems,
3) Leveraging ASR systems, and
4) Disentangling speaker from linguistic content.

1) Non-Parallel Data of Paired Speakers: Voice conversion with non-parallel training data is a task similar to image-to-image translation [189]–[193], which is to find a mapping from a source domain to a target domain without the need of parallel training data. Let's draw a parallel between image-to-image translation and voice conversion. In image translation, we would like to translate a horse to a zebra, where we preserve the structure of horse and change the coat of horse to that of zebra [194]–[198], in voice conversion, we would like to transform one voice to that of another, while preserving the linguistic, and prosodic content.

CycleGAN is based on the concept of adversarial learning [199], which is to train a generative model to find a solution in a min-max game between two neural networks, called as generator (G) and discriminator (D). It is known to achieve remarkable results [194] on several tasks where paired training data does not exist, such as image manipulation and synthesis [194], [196], [200]–[204], speech enhancement [205], speech recognition [206], speech synthesis [207], [208], and music translation [209].

As the speech data are non-parallel, alignment is not easily achieved. Kaneko and Kameoka first studied a CycleGAN [48], [49], [210], [211] that incorporates three loss functions: adversarial loss, cycle-consistency loss, and identity-mapping loss, to learn forward and inverse mapping between source and target speakers. The adversarial loss measures how distinguishable between the data distribution of converted features and source features $\mathbf{x}$ or target features $\mathbf{y}$. For the forward mapping, it is defined as follows:

$$L_{ADV}(G_{X\to Y}, D_Y, X, Y) = \mathbb{E}_{y\sim P(y)}[D_Y(\mathbf{y})] + \mathbb{E}_{x\sim P(x)}[log(1 - D_Y(G_{X\to Y}(\mathbf{x}))] \quad (12)$$

The closer the distribution of converted data with that of target data, the smaller the loss becomes.

The adversarial loss only tells us whether $G_{X\to Y}$ follows the distribution of target data and does not ensure that the contextual information, that represents the general sentence structure we would like to carry over from source to target, is preserved. To ensure that we maintain the consistent contextual information between $\mathbf{x}$ and $G_{X\to Y}(\mathbf{x})$, the cycle-consistency loss, that is presented in Fig. 7, is introduced,

$$L_{CYC}(G_{X\to Y}, G_{Y\to X}) = \mathbb{E}_{\mathbf{x}\sim P(\mathbf{x})}[\|G_{Y\to X}(G_{X\to Y}(\mathbf{x})) - \mathbf{x}\|_1] + \mathbb{E}_{\mathbf{y}\sim P(y)}[\|G_{X\to Y}(G_{Y\to X}(\mathbf{y})) - \mathbf{y}\|_1] \quad (13)$$

where $\|\cdot\|_1$ refers to a L1 norm function, or least absolute errors, that is known to produce sharper spectral features. This loss encourages $G_{X\to Y}$ and $G_{Y\to X}$ to find an optimal pseudo pair of $(\mathbf{x}, \mathbf{y})$ through circular conversion. To encourage the generator to find the mapping that preserves underlying linguistic content between the input and output [212], an identity mapping loss is introduced as follows,

$$L_{ID}(G_{X\to Y}, G_{Y\to X}) = \mathbb{E}_{\mathbf{x}\sim P(x)}[\|G_{Y\to X}(\mathbf{x}) - \mathbf{x}\|] + \mathbb{E}_{\mathbf{y}\sim P(y)}[\|G_{X\to Y}(\mathbf{y}) - \mathbf{y}\|] \quad (14)$$

Combining these three loss functions, we can obtain the overall loss function of CycleGAN [48], [210].

CycleGAN represents a successful deep learning implementation to find an optimal pseudo pair from non-parallel data of paired speakers. It doesn't require any frame alignment mechanism such as dynamic time warping or attention. Experimental results show that, with non-parallel training data, CycleGAN achieves comparable performance to that of GMM-based system that is trained on twice amount of parallel data [48]. Moreover, with the adversarial training, it effectively overcomes the over-smoothing problem, which is known to be one of the main factors leading to speech-quality degradation. We note that more recently, CycleGAN-VC2, an improved version of CycleGAN-VC has been studied [211], that further improves CycleGAN by incorporating three new techniques: an improved objective (two-step adversarial losses), improved generator (2-1-2D CNN), and improved discriminator (PatchGAN). CycleGAN has been successfully applied in mono-lingual [49], [213], cross-lingual voice conversion [214], emotional voice conversion [215], [216] and rhythm-flexible voice conversion [217].

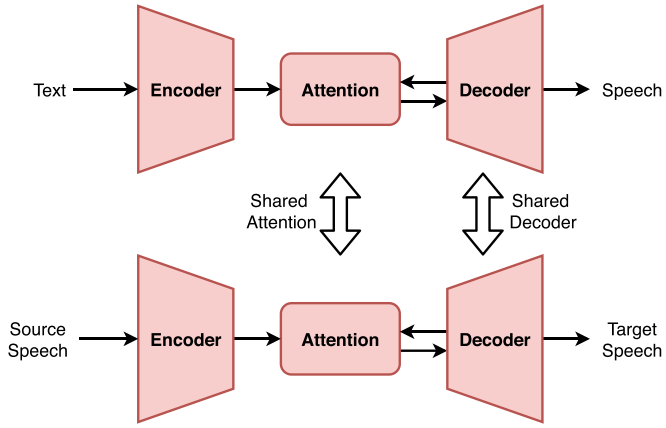

Fig. 8. Upper panel is a TTS flow, and the lower panel is a voice conversion flow. Both follow similar encoder-decoder with attention architecture. The voice conversion network leverages the TTS network, that is linguistically informed.

Unlike the encoder-decoder structure, CycleGAN follows a generative modeling architecture that doesn't explicitly model some internal representations to support flexible manipulation, such as voice identity, duration of speech, and emotion. Therefore, it is more suitable for voice conversion between a specific source and target pair. Nonetheless, it represents an important milestone towards non-parallel data voice conversion.

2) Leveraging TTS Systems: We have discussed the deep learning architectures for voice conversion that do not involve text. One of the important aspects of voice conversion is to carry forward the linguistic content from source to target. Voice conversion and TTS systems are similar in the sense that they both aim to generate high quality speech with the appropriate linguistic content. A TTS system provides a mechanism for the speech to adhere to the linguistic content. The ideas to leverage TTS mechanism can be motivated in different ways. Firstly, a TTS system is trained on a large speech database that offers a high quality speech re-construction mechanism given the linguistic content; secondly, a TTS system is equipped with a quality attention mechanism that is needed by voice conversion.

Encoder-decoder models with attention have recently shown considerable success in modeling a variety of complex sequence-to-sequence problems. Tacotron [89], [180], [218]–[220] represents one of the successful text-to-speech (TTS) implementations, that has been extended to voice conversion [3], [183], [221]. The strategy to leverage TTS knowledge is built on the ideas of shared attention knowledge and/or shared decoder architecture as illustrated in Fig. 8. Zhang *et al.* [221] proposed a transfer learning technique for voice conversion network to learn from the phonetic context vectors derived from a TTS attention mechanism, and to share the decoder with a TTS system, that represents a typical example of such leverage.

Zhang *et al.* proposed a joint training system architecture for both text-to-speech and voice conversion [3] by extending the model architecture of Tacotron, which features a multi-source sequence-to-sequence model with a dual input, and dual attention mechanism. By taking only text as input, the system performs speech synthesis. The system can also take either voice alone, or both text and voice, denoted as hybrid TTS & VC, as input for voice conversion. The multi-source encoder-decoder model is trained with a decoder that is linguistically informed via the TTS joint training, as illustrated as shared decoder in Fig. 8. Experiments show that the joint training has improved the voice conversion task with or without text input at run-time inference.

Park *et al.* proposed a voice conversion system, known as Cotatron, that is built on top of a multi-speaker Tacotron TTS architecture [165]. At run-time inference, the pre-trained TTS system is used to derive speaker-independent linguistic features of the source speech. This process is guided by the transcription of the input speech, as such, text transcription of source speech is required at run-time inference. The system uses the TTS encoder to extract speaker-independent linguistic features, or disentangle the speaker identity. The decoder then takes the attention-aligned speaker-independent linguistic features as the input, and the target speaker identity as the condition, to generate a target speaker's voice. In this way, voice conversion leverage the attention mechanism or *shared attention* from TTS, as shown in Fig. 8. Cotatron is designed to perform one-to-many voice conversion. A study [222], that shares similar motivation with [165] but is based on the Transformer instead of Tacotron, suggests transferring knowledge from a learned TTS model to benefit from large-scale, easily accessible TTS corpora.

Zhang *et al.* [223] proposed to improve the sequence-to-sequence model [183] by using text supervision during training. A multi-task learning structure is designed which adds auxiliary classifiers to the middle layers of the sequence-to-sequence model to predict linguistic labels as a secondary task. The linguistic labels can be obtained either manually or automatically with alignment tools. With the linguistic label objective, the encoder and decoder are expected to generate meaningful intermediate representations which are linguistically informed. The text transcripts are only required during training. Experiments show that the multi-task learning with linguistic labels effectively improves the alignment quality of the model, thus alleviates issues such as mispronunciation.

The neural representation of deep learning has facilitated the interaction between TTS and voice conversion. By leveraging TTS systems, we hope to improve the training and run-time inference of voice conversion with by adhering to linguistic content. However, such techniques usually require a large training corpus. Recent studies introduced a framework for creating limited-data VC system [222], [224], [225] by bootstrapping from a speaker-adaptive TTS model. It deserves future studies as to how voice conversion can benefit from TTS systems without involving large training data.

3) Leveraging ASR Systems: Deep learning approaches for voice conversion typically require a large parallel corpus for training. This is partly because we would like to learn the latent representations that describe the phonetic systems. The requirement of training data has limited the scope of potential applications. We know that most ASR systems are already trained with a large corpus. They already describe well the phonetic systems in different ways. The question is how to leverage the latent representations in ASR systems for voice conversion.

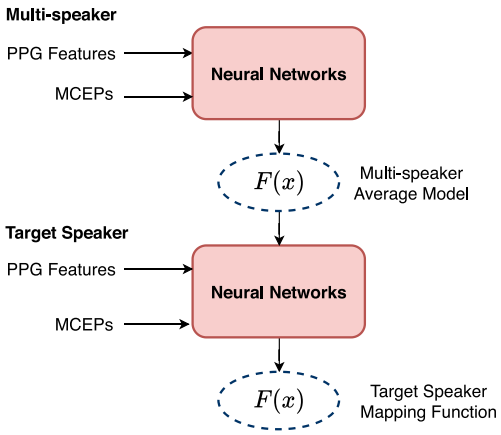

Fig. 9. Training phase of the average modeling approach that maps PPG features to MCEP features for voice conversion [44].

One of the ideas is to use the context posterior probability sequence produced by the ASR model with sequence to sequence learning to generate a target speech feature sequence [164]. In this model, the system has an encoder-decoder structure similar to Fig. 6, except that it uses a speech recognizer as the encoder, and a speech synthesizer as the decoder. Another study is to guide a sequence to sequence voice conversion model by an ASR system, which augments inputs with bottleneck features [183]. Recently, an end-to-end speech-to-speech sequence transducer, Parrotron [226], was studied. Parrotron learns to convert speech spectrogram of any speakers, with multiple accents and imperfections, to the voice of a single predefined target speaker. Parrotron accomplishes this by using an auxiliary ASR decoder to predict the transcript of the output speech, conditioned on the encoder latent representation. The multi-task training of Parrotron optimizes the decoder to generate the target voice, at the same time, constrains the latent representation to retain linguistic information only. The ASR decoder aims to disentangle the speaker's identity from the speech. The above techniques adopt the encoder-decoder with attention architecture.

It is another way to look at voice conversion that speech consists of two components, speaker dependent component and speaker independent component. If we are able to decompose speech signals into the two components, we can carry over the former, and only convert the latter to achieve voice conversion. The average modeling technique represents one of the successful implementations [41], where we build a mapping function to convert phonetic posteriogram (PPG) [32] to acoustic features. The PPG features are derived from an ASR system, that can be considered as speaker independent. We train the mapping function from multi-speaker, non-parallel speech data. In this way, one doesn't need to train a full conversion model for each target speaker. The average model can be adapted towards the target with a small amount of target speech. The training and adaptation of the average model are illustrated in Fig. 9.

There were several follow-up studies along this direction, for example, Tian *et al.* propose a PPG to waveform conversion [96], and a average model with speaker identity [159] as a condition [44]. Zhou *et al.* propose to use PPG as the linguistic features for cross-lingual voice conversion [162]. Liu *et al.* propose to use PPG for emotional voice conversion [227]. Zhang *et al.* also show that the average model framework can benefit from a small amount of parallel training data using an error reduction network [228].

4) Disentangling Speaker From Linguistic Content: In the context of voice conversion, speech can be considered as a composition of speaker voice identity and linguistic content. If we are able to disentangle speaker from the linguistic content, we can change the speaker identity independently of the linguistic content. Auto-encoder [229] represents one of the common techniques for speech disentanglement, and reconstruction. There are other techniques such as instance normalization [230] and vector quantization [231], [232], that are effective in disentangling speaker from the content.

An auto-encoder learns to reproduce its input as its output. Therefore, parallel training data is not required. An encoder learns to represent the input with a latent code, and a decoder learns to reconstruct the original input from the latent code. The latent code can be seen as an information bottleneck which, on one hand, lets pass information necessary, e.g. speaker independent linguistic content, for perfect reconstruction, and on the other hand, forces some information to be discarded, e.g. speaker, noise and channel information [85]. Variational auto-encoder (VAE) [233] is the stochastic version of auto-encoder, in which the encoder produces distributions over latent representations, rather than deterministic latent codes, while the decoder is trained on samples from these distributions. Variational auto-encoder is more suitable than deterministic auto-encoder in synthesizing new samples.

Chorowski *et al.* [100] provide a comparison of three auto-encoding neural networks by studying how they learn a representation from speech data to separate speaker identity from the linguistic content. It was shown that discrete representation, that is the latent code obtained from VQ-VAE [234], [235], preserves the most linguistic content while also being the most speaker-invariant. Recently, a group latent embedding technique for VQ-VAE is studied to improve the encoding process, which divides the embedding dictionary into groups and uses the weighted average of atoms in the nearest group as the latent embedding [236].

The concept of a VAE-based voice conversion framework [43] can be illustrated in Fig. 10. The decoder reconstructs the utterance by conditioning on the latent code extracted by the encoder, and separately on a speaker code, which could be an one-hot vector [43], [237] for a close set of speakers, or an i-vector [159], bottleneck speaker representation [238], or d-vector [239] for an open set of speakers. By explicitly conditioning the decoder on speaker identity, the encoder is forced to capture speaker-independent information in the latent code from a multi-speaker database.

Just like other auto-encoder, VAE decoder tends to generate over-smoothed speech. This can be problematic for voice conversion because the network may generate poor quality buzzy-sounding speech. Generative adversarial networks (GANs) [240] were proposed as one of the solutions to the over-smoothing problem [241]. GANs offer a general

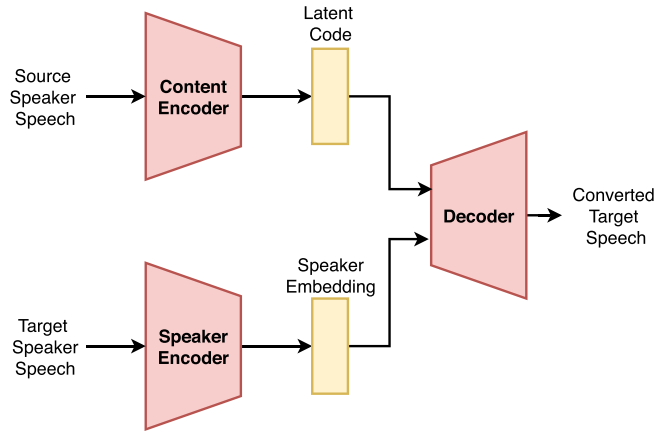

Fig. 10. Typical auto-encoding network for voice conversion, where the encoders and decoder learn to disentangle speaker from linguistic content. At run-time, the linguistic content of the source speech represented by latent code and speaker embedding of a target speaker are combined to generate target speech.

framework for training a data generator in such a way that it can deceive a real/fake discriminator that attempts to distinguish real data and fake data produced by the generator [242]–[244]. By incorporating the GAN concept into VAE, VAE-GAN was studied for voice conversion with non-parallel training data [47] and in cross-lingual voice conversion [214]. It was shown that VAE-GAN [240] produces more natural sounding speech than the standard VAE method [43], [238].

A recent study on sequence-to-sequence non-parallel voice conversion [245] shows that it is possible to explicitly model the transfer of other aspects of speech, such as source rhythm, speaking style, and emotion to the target speech.

VI. Evaluation of Voice Conversion

Effective quality assessment of voice quality is required to validate the algorithms, to measure the technological progress, and to benchmark a system against the state-of-the-art. Typically, we report the results in terms of objective and subjective measurements.

To provide an objective evaluation, a reference speech is required. The common objective evaluation metrics include Mel-cepstral distortion (MCD) [246] for spectrum, and PCC [247] and RMSE [248]–[250] for prosody. We note that, such metrics are not always correlated with human perception partly because they measure the distortion of acoustic features rather than the waveform that humans actually listen to.

Subjective evaluation metrics, such as the mean opinion score (MOS) [2], [251]–[253], preference tests [18], [254] and best-worst scaling [255] could represent the intrinsic naturalness and similarity to the target. We note that, for subjective evaluation to be meaningful, a large number of listeners are required, that is not always possible in practice.

A. Objective Evaluation

1) Spectrum Conversion: To provide an objective evaluation, first of all, we need a reference utterance spoken by the target speaker. Ideally the converted speech is very close to the reference speech. We can measure the differences between them by comparing their spectral distances. However, there is no guarantee that the converted speech and the reference speech is of the same length. In this case, a frame aligner is required to establish the frame-level mapping.

Mel-cepstral distortion (MCD) [246] is commonly used to measure the difference between two spectral features [62], [67], [256], [257]. It is calculated between the converted and target Mel-cepstral coefficients, or MCEPs, [258], [259], $\hat{\mathbf{y}}$ and $\mathbf{y}$. Suppose that each MCEP vector consists of 24 coefficients, we have $\hat{\mathbf{y}} = \{m^c_{k,i}\}$ and $\mathbf{y} = \{m^t_{k,i}\}$ at frame k, where i denotes the ith coefficient in the converted and target MCEPs.

$$MCD[dB] = \frac{10}{\ln 10}\sqrt{2\sum_{i=1}^{24}(m^t_{k,i} - m^c_{k,i})^2} \tag{15}$$

We note that a lower MCD indicates better performance. However, MCD value is not always correlated with human perception. Therefore, subjective evaluations, such as MOS and similarity score, are also conducted.

2) Prosody Conversion: Speech prosody of an utterance is characterized by phonetic duration, energy contour, and pitch contour. To effectively measure how close the prosody patterns of converted speech is to the reference speech, we need to provide measurements for the three aspects. The alignment between the converted speech and the reference speech provides the information about how much the phonetic duration differs one another. We can derive the number of frames that deviate from the ideal diagonal path on average, such as frame disturbance [260], to report the differences of phonetic duration. Pearson Correlation Coefficient (PCC) [62], [215] and Root Mean Squared Error (RMSE) have been widely used as the evaluation metrics to measure the linear dependence of prosody contours or energy contours between two speech utterances. We next take the measurement of two prosody contours as an example. PCC between the aligned pair of converted and target F0 sequences is given as follows,

$$\rho(F0^c, F0^t) = \frac{cov(F0^c, F0^t)}{\sigma_{F0^c}\sigma_{F0^t}} \tag{16}$$

where σ_{F0^c} and σ_{F0^t} are the standard deviations of the converted F0 sequences ($F0^c$) and the target F0 sequences ($F0^t$), respectively. We note that a higher PCC value represents better F0 conversion performance.

The RMSE between the converted F0 and the corresponding target F0 is defined as,

$$RMSE = \sqrt{\frac{1}{K}\sum_{k=1}^{K}(F0^c_k - F0^t_k)^2} \tag{17}$$

where $F0^c_k$ and $F0^t_k$ denote the converted and target F0 features, respectively. K is the length of $F0$ sequence, or the total number of frames. We note that a lower RMSE value represents better $F0$ conversion performance. The same measurement applies to energy contours as well.

Other generally-accepted metrics for prosody transfer include F0 Frame Error (FFE) [261] and Gross Pitch Error (GPE) [262]. We note that GPE reports the percentage of voiced frames whose

pitch values are more than 20% different from the reference, while FFE reports the percentage of frames that either contain a 20% pitch error or a voicing decision error [263].

B. Subjective Evaluation

Mean Opinion Score (MOS) has been widely used in listening tests [13], [40], [61], [62], [264]–[268]. In MOS experiments, listeners rate the quality of the converted voice using a 5-point scale: "5" for excellent, "4" for good, "3" for fair, "2" for poor, and "1" for bad. There are several evaluation methods that are similar to MOS, for example: 1) DMOS [269]–[271], which is a "degradation" or "differential" MOS test, requiring listeners to rate the sample with respect to this reference, and 2) MUSHRA [272]–[274], which stands for MUltiple Stimuli with Hidden Reference and Anchor, and requires fewer participants than MOS to obtain statistically significant results.

Another popular subjective evaluation is preference test, also denoted as AB/ABX test [2], [40], [275] or XAB test [11], [276]. In AB tests, listeners are presented with two speech samples and asked to indicate which one has more of a certain property; for example in terms of naturalness, or similarity. In ABX test, similar to that of AB, two samples are given but an extra reference sample is also given. Listeners need to judge if A or B more like X in terms of naturalness, similarity, or even emotional quality [215]. In XAB test, listeners are presented the original target speech sample first, and then a pair of converted voices randomly. We note that it is not practical to use AB, ABX or XAB test for the comparison of many VC systems at the same time. MUSHRA is another type of voice quality test in telecommunication [277], where the reference natural speech and several other converted samples of the same content are presented to the listeners in a random order. The listeners are asked to rate the speech quality of each sample between 0 and 100.

It is known that people are good at picking the extremes but their preferences for anything in between might be fuzzy and inaccurate when presented with a long list of options. Best-Worst Scaling (BWS) [255] is proposed for voice conversion quality assessment [22], where listeners are presented only with a few randomly selected options each time. With many such BWS decisions, Best-Worst Scaling can handle a long list of options and generates more discriminating results, such voice quality ranking, than MOS and preference tests.

We note that subjective measures can represent the intrinsic naturalness and similarity of a voice conversion system. However, such evaluation can be time-consuming and expensive as they involve a large number of listeners.

C. Evaluation With Deep Learning Approaches

The study of perceptual quality evaluation seeks to approximate human judgement with computational models of psychoacoustic motivation. It provides insights into how humans perceive speech quality in listening tests, and suggests assessment metrics, that are required in speech communication, speech enhancement, speech synthesis, voice conversion and any other speech production or transmission applications. Perceptual Evaluation of Speech Quality (PESQ) [278] is an ITU-T recommendation that is widely used as industry standard. It provides objective speech quality evaluation that predicts the human-perceived speech quality.

However, the PESQ formulation requires the presence of reference speech, that considerably restricts its use in voice conversion applications, and motivates the study of perceptual evaluations without the need of reference speech. Those metrics that don't require reference speech are called non-intrusive evaluation metrics. For example, Fu *et al.* [279] propose Quality-Net [279] that is an end-to-end model to predict PESQ ratings, that are the proxy for human ratings. Yoshimura *et al.* [280], Patton *et al.* [281] propose a CNN-based naturalness predictor to predict human MOS ratings, among other non-intrusive assessment metrics [282]–[284].

Lo *et al.* [285] propose MOSNet, another non-intrusive assessment technique based on deep neural networks, that learns to predict human MOS ratings. MOSNet scores are highly correlated with human MOS ratings at system level, and fairly correlated at utterance level. While it is a non-intrusive evaluation metric for naturalness, MOSNet can also be modified and re-purposed to predict the similarity scores between target speech and converted speech. It provides similarity scores with fair correlation values to human ratings on VCC 2018 dataset. MOSNet marks a recent advancement towards automatic perceptual quality evaluation [286], which is free and open-source.

Last but not least, both Frechet DeepSpeech Distance (FDSD, cFDSD) and Kernel DeepSpeech Distance (KDSD, cKDSD) have been found to be well correlated with MOS for speech generation [287]. We note that Frechet DeepSpeech Distance is motivated by Frechet Inception Distance (FID) [288], whereas Kernel DeepSpeech Distance is motivated by Kernel Inception Distance (KID) [289]. In both of these frameworks, the Inception image recognition network has been replaced with the DeepSpeech audio recognition network for evaluation of speech generation.

VII. Voice Conversion Challenges

In this section, we would like to give an overview of the series of voice conversion challenges, that provide shared tasks with common data sets and evaluation metrics for fair comparison of algorithms. The voice conversion challenge (VCC) is a biannual event since 2016. In a challenge, a common database is provided by the organizers. The participants build voice conversion systems using their own technology, and the organizers evaluate the performance of the converted speech. The main evaluation methodology is a listening test in which crowd-sourced evaluators rank the naturalness and speaker similarity.

The 2016 challenge offers a standard voice conversion task using a parallel training database was adopted [67]. The 2018 challenge features a more advanced conversion scenario using a non-parallel database [290]. The 2020 challenge puts forward a cross-lingual voice conversion research problem. A summary of VCC 2016, VCC 2018 and VCC 2020 is also provided in Table I.

TABLE I
SUMMARY OF VCC 2016, VCC 2018 AND VCC 2020

Challenge	Language	Task	Training Data	# Speakers	Testing Data
VCC 2016	monolingual	parallel	162 paired utterances	4 source, 4 target	54 utterances
VCC 2018	monolingual	parallel	81 paired utterances	4 source, 4 target	35 utterances
	monolingual	nonparallel	81 unpaired utterances	4 source, 4 target	35 utterances
VCC 2020	monolingual	parallel + nonparallel	20 paired, 50 unpaired utterances	4 source, 4 target	25 utterances
	crosslingual	nonparallel	70 unpaired utterances	4 source, 6 target	25 utterances

A. Why Is the Challenge Needed?

As described earlier, many of the voice conversion approaches are data-driven, hence speech data are required to train models and for conversion evaluation. To compare such data-driven methods each other precisely, a common database that specifies training and evaluation data explicitly is needed. However, such common database did not exist until 2016. Without common databases, researchers have to re-implement others' system with their own databases before trying any new ideas. In such situation, it is not guaranteed that the re-implemented system achieves the expected performance in the original work.

To address the same problem, the TTS community gave birth to the first Blizzard challenge in 2005. Since then, the challenge has defined various standard databases for TTS and has made comparisons of TTS much fairer and easier. The motivations of VCC are exactly the same as those of the Blizzard challenges. VCC introduced a few standard databases for voice conversion and also defined the common training and evaluation protocols. All the converted speech submitted by the participants for the challenges have been released publicly. In this way, researchers can compare the performance of their voice conversion system with that of other state-of-the-art systems without the need of re-implementation.

Another need on voice conversion standard databases arose from biometric speaker recognition community. As the voice conversion technology could be misused for attacking speaker verification systems, anti-spoofing countermeasures are required [291]. This is also called presentation attack detection. Anti-spoofing techniques aim at discriminating between fake artificial inputs presented to biometric authentication systems and genuine inputs. If sufficient knowledge and data regarding the spoofed data is available, a binary classifier can be constructed to reject artificial inputs. Therefore, the common VCC databases are also important for anti-spoofing research. With many converted speech data from advanced voice conversion systems, researchers in the biometric community can develop anti-spoofing models to strengthen the defence of speaker recognition systems, and to evaluate their vulnerabilities.

B. Overview of the 2016 Voice Conversion Challenge

We first overview the 2016 voice conversion challenge [67] and its datasets.[1] As the first shared task in voice conversion, a parallel voice conversion task and its evaluation protocol are defined for VCC 2016. The parallel dataset consists of 162 common sentences uttered by both source and target speakers. Target and source speakers are four native speakers of American English (two females and two males), respectively. In the challenge, the participants develop the conversion systems and produce converted speech for all possible source-target pair combinations. In total, eight speakers (plus two unused speakers) are included in the VCC 2016 database. The number of test sentences for evaluation is 54.

The main evaluation methodology adopted for the ranking is subjective evaluation on perceived naturalness and speaker similarity of the converted samples to target speakers. The naturalness is evaluated using the standard five-point scale mean-opinion score (MOS) test ranging from 1 (completely unnatural) to 5 (completely natural). The speaker similarity was evaluated using the Same/Different paradigm [292]. Subjects are asked to listen to two audio samples and to judge if they are speech signals produced by the same speaker in a four point scale: "Same, absolutely sure", "Same, not sure", "Different, not sure" and "Different, absolutely sure." As the perceived speaker similarity to a target speaker, and the perceived voice quality are not necessarily correlated, it is important to use a scatter-plot to observe the trade-off between the two aspects.

In the 2016 challenge, 17 participants submitted their conversion results. Two hundreds native listeners of English joined the listening tests. It is reported that the best system using GMM and waveform filtering obtained an average of 3.0 in the five-point scale evaluation for the naturalness judgement, and about 70% of its converted speech samples are judged to be the same as target speakers by listeners. However, it is also confirmed that there is still a huge gap between target natural speech and the converted speech. We observe that it remains a unsolved challenge to achieve good quality and speaker similarity at that time. More details of VCC 2016 can be found at [292]. Details of best performing systems are reported in [13].

C. Overview of the 2018 Voice Conversion Challenge

Next we give an overview of the 2018 voice conversion challenge [290] and its datasets.[2] VCC 2018 offers two tasks, parallel and non-parallel voice conversion tasks. A dataset and its evaluation protocol are defined for each task. The dataset for the parallel conversion task is similar to that of the 2016 challenge, except that it has a smaller number of common utterances uttered by source and target speakers. Target and source speakers are four native speakers of American English (two females and two males), respectively, but, they are different speakers from those used for the 2016 challenge. Like the 2016 challenge,

[1]The VCC2016 dataset is available at https://doi.org/10.7488/ds/1575.

[2]The VCC2018 dataset is available at https://doi.org/10.7488/ds/2337.

the participants were asked to develop conversion systems and to produce converted data for all possible source-target pair combinations.

VCC 2018 introduced a non-parallel voice conversion task for the first time. The same target speakers' data in the parallel task are used as the target. However, the source speakers are four native speakers of American English (2 females and 2 males) different from those of the parallel conversion task and their utterances are also all different from those of the target speakers. Like the parallel voice conversion task, converted data for all possible source-target pair combinations needed to be produced by the participants. In total twelve speakers are included in the VCC 2018 database. Each of the source and target speakers has a set of 81 sentences as training data, which is half of that for VCC 2016. The number of test sentences for evaluation is 35.

In the 2018 challenge, 23 participants submitted their conversion results to the parallel conversion task, with 11 of them additionally participating in the non-parallel conversion task. The same evaluation methodology as the 2016 challenge was adopted for the 2018 challenge and 260 crowd-sourced native listeners of English have joined the listening tests. It was reported that in both tasks, the best system using phone encoder and neural vocoder obtained an average of 4.1 in the five-point scale evaluation for the naturalness judgement and about 80% of its converted speech samples were judged to be the same as target speakers by listeners. It was also reported that the best system has similar performance in both the parallel and non-parallel tasks in contrast to results reported in literature.

In VCC 2018, the spoofing countermeasure was introduced as an supplement to subjective evaluation of voice quality, that brought together the voice conversion and speaker verification research community. More details of the 2018 challenge can be found at [290]. Details of best performing systems are reported in [293], [294].

From this challenge, we observed that new speech waveform generation paradigms such as WaveNet and phone encoding have brought significant progress to the voice conversion field. Further improvements have been achieved in the follow up papers [295], [296] and new VC systems that exceed the challenge's best performance have already been reported.

D. Overview of the 2020 Voice Conversion Challenge

The 2020 voice conversion challenge [297] consists of two tasks: 1) non-parallel training in the same language (English); and 2) non-parallel training over different languages (English-Finnish, English-German, and English-Mandarin).

In the first task, each participant trains voice conversion models for all source and target speaker pairs using up to 70 utterances, including 20 parallel utterances and 50 non-parallel utterances in English, for each speaker as the training data. Overall, 16 voice conversion models (i.e., 4 sources by 4 targets) are to be developed. In the second task, each participant develops voice conversion models for all source and target speaker pairs using up to 70 utterances for each speaker (i.e., in English for the source speakers, and in Finnish, German, or Mandarin for the target speakers) as the training data. Overall, 24 conversion systems (i.e., 4 sources by 6 targets) are to be developed.

In the 2020 challenge, 31 participants submitted their results to the first task, and 28 participants submitted their results to the second task. The participants were allowed to mix and combine different source speaker's data to train speaker-independent models. Moreover, the participants can also use orthographic transcriptions of the released training data to develop their voice conversion systems. Last but not least, the participants were free to perform manual annotations of the released training data, which can effectively improves the quality of the converted speech. The 2020 challenge organizers also built several baseline systems including the top system of the previous challenge on the new database. The codes of CycleVAE-based baseline[3] and Cascade ASR + TTS based VC[4] are released so that participants can build the basic systems easily and focus on their own innovation. The 2020 challenge features a multifaceted evaluation. In addition to the traditional evaluation metrics, the challenge also reports the speech recognition, speaker recognition, and anti-spoofing evaluation results on the converted speech.

According to the final report, it was encouraging to see that the speaker similarity scores of several systems are very close to that of natural speech of target speakers in the first task. However, none of the systems achieved human-level naturalness. The second task is a more challenging one. While we observed lower overall naturalness and similarity scores than those of the first task, the MOS scores of the best systems were higher than 4.0.

E. Relevant Challenges – ASVspoof Challenge

The spoofing capability against automatic speaker verification is a related topic to voice conversion, that has also been organized as technology challenges. The ASVspoof series of challenges are such biannual events, which started in 2013. Like in the voice conversion challenges, the organizers release a common database including many pairs of spoofed audio (converted, generated audio or replay audio) and genuine audio to the participants, who build anti-spoofing models using their own technology. The organizers rank the detection accuracy of the anti-spoofing results submitted by the participants.

In 2015, the first anti-spoofing database including various types of spoofed audio using voice conversion and TTS systems was constructed. This database became a reference standard in the automatic speaker verification (ASV) community [298], [299]. The main focus of the 2017 challenge was a replay task, where a large quantity of real-world replay speech data were collected [300]. In 2019, an even larger database including converted, generated, and replay speech data was constructed [301]. The best performing systems in the 2016 and 2018 voice conversion challenges were also used for generating advanced spoofed audio [302]. The challenges revealed that some anti-spoofing systems outperform human listeners in detecting spoofed audio.

[3][Online]. Available: https://github.com/bigpon/vcc20_baseline_cyclevae

[4][Online]. Available: https://github.com/espnet/espnet/tree/master/egs/vcc20

VIII. Resources

In addition to the voice conversion challenge databases described above, the CMU-Arctic database [303] and the VCTK databases [304] are also popular for voice conversion research. The current version of the CMU-Arctic database[5] has 18 English speakers and each of them reads out the same set of around 1,150 utterances, which are carefully selected from out-of-copyright texts from Project Gutenberg. This is suitable for parallel voice conversion since sentences are common to all the speakers. The current version (ver. 0.92) of the CSTR VCTK corpus[6] has speech data uttered by 110 English speakers with various dialects. Each speaker reads out about 400 sentences, which are selected from newspapers, the rainbow passage and an elicitation paragraph used for the speech accent archive. Since the rainbow passage and an elicitation paragraph are common to all the speakers, this database can be used for both parallel and non-parallel voice conversion.

Since neural networks are data hungry and generalization to unseen speakers is a key for successful conversion, large-scale, but, low-quality databases such as LibriTTS and VoxCeleb are also used for training some components required (e.g. speaker encoder) for voice conversion. The LibriTTS corpus [305] has 585 hours of transcribed speech data uttered by total of 2,456 speakers. The recording condition and audio quality are less than ideal, but, this corpus is suitable for training speaker encoder networks or generalizing any-to-any speaker mapping network. The VoxCeleb database [306] is further a larger scale speech database consisting of about 2,800 hours of untranscribed speech from over 6,000 speakers. This is an appropriate database for training noise-robust speaker encoder networks.

There are many open-source codes for training VC models. For instance, spocket [307] supports GMM-based conversions and ESPnet [308] supports cascaded ASR and TTS system. In addition, there are many open-source codes for neural-network based voice conversion written by the community at github.[7]

IX. Conclusion

This article provides a comprehensive overview of the voice conversion technology, covering the fundamentals and practice till August 2020. We reveal the underlying technologies and their relationship from the statistical approaches to deep learning, and discuss their promise and limitations. We also study the evaluation techniques for voice conversion. Moreover, we report the series of voice conversion challenges and resources that are useful information for researchers and engineers to start voice conversion research.

References

[1] J. Q. Stewart, "An electrical analogue of the vocal organs," *Nature*, vol. 110, pp. 311–312.

[2] A. Kain and M. W. Macon, "Spectral voice conversion for text-to-speech synthesis," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process., (ICASSP)*, 1998, pp. 285–288.

[3] M. Zhang, X. Wang, F. Fang, H. Li, and J. Yamagishi, "Joint training framework for text-to-speech and voice conversion using multi-source tacotron and wavenet," 2019, *arXiv:1903.12389*.

[4] C. Veaux, J. Yamagishi, and S. King, "Towards personalised synthesised voices for individuals with vocal disabilities: Voice banking and reconstruction," Aug. 2013.

[5] B. Srivastava, N. Vauquier, M. Sahidullah, A. Bellet, M. Tommasi, and E. Vincent, "Evaluating voice conversion-based privacy protection against informed attackers," Nov. 2019.

[6] Z. Wu and H. Li, "Voice conversion versus speaker verification: An overview," *APSIPA Trans. Signal Inf. Process.*, vol. 3, p. e17, 2014.

[7] C. Y. Huang, Y. Y. Lin, H. Y. Lee, and L. S. Lee, "Defending your voice: Adversarial attack on voice conversion," vol. abs/2005.08781, 2020.

[8] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, "Voice conversion through vector quantization," *J. Acoust. Soc. Jpn. (E)*, vol. 11, no. 2, pp. 71–76, 1990.

[9] K. Shikano, S. Nakamura, and M. Abe, "Speaker adaptation and voice conversion by codebook mapping," in *IEEE Int. Symp. Circuits Syst.*, 1991, pp. 594–597.

[10] E. Helander, J. Schwarz, J. Nurminen, H. Silen, and M. Gabbouj, "On the impact of alignment on voice conversion performance," in *Proc. 9th Annu. Conf. Int. Speech Commun. Assoc.*, 2008.

[11] T. Toda, A. W. Black, and K. Tokuda, "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 15, no. 8, pp. 2222–2235, Nov. 2007.

[12] H. Zen, Y. Nankaku, and K. Tokuda, "Probabilistic feature mapping based on trajectory HMMs," in *Proc. 9th Annu. Conf. Int. Speech Commun. Assoc.*, 2008.

[13] K. Kobayashi, S. Takamichi, S. Nakamura, and T. Toda, "The NU-naist voice conversion system for the voice conversion challenge 2016," in *Proc. Interspeech*, pp. 1667–1671.

[14] E. Helander, T. Virtanen, J. Nurminen, and M. Gabbouj, "Voice conversion using partial least squares regression," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 5, pp. 912–921, Jul. 2010.

[15] E. Helander, H. Silén, T. Virtanen, and M. Gabbouj, "Voice conversion using dynamic kernel partial least squares regression," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 3, pp. 806–817, Mar. 2012.

[16] Y. Luan, D. Saito, Y. Kashiwagi, N. Minematsu, and K. Hirose, "Semi-supervised noise dictionary adaptation for exemplar-based noise robust speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2014, pp. 1745–1748.

[17] R. Takashima, T. Takiguchi, and Y. Ariki, "Exemplar-based voice conversion in noisy environment," in *Proc. IEEE Spoken Lang. Technol. Workshop (SLT)*, 2012, pp. 313–317.

[18] Z. Wu, T. Virtanen, E. S. Chng, and H. Li, "Exemplar-based sparse representation with residual compensation for voice conversion," *IEEE/ACM Trans. Audio, Speech Lang. Process.*, vol. 22, no. 10, pp. 1506–1521, Oct. 2014.

[19] R. Aihara, K. Masaka, T. Takiguchi, and Y. Ariki, "Parallel dictionary learning for multimodal voice conversion using matrix factorization," in *Proc. INTERSPEECH*, 2016, pp. 27–40.

[20] Z. Jin, A. Finkelstein, S. DiVerdi, J. Lu, and G. J. Mysore, "Cute: A concatenative method for voice conversion using exemplar-based unit selection," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2016, pp. 5660–5664.

[21] R. Aihara, T. Nakashika, T. Takiguchi, and Y. Ariki, "Voice conversion based on non-negative matrix factorization using phoneme-categorized dictionary," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2014, pp. 7894–7898.

[22] B. Sisman, H. Li, and K. C. Tan, "Sparse representation of phonetic features for voice conversion with and without parallel data," in *Proc. IEEE Autom. Speech Recognit. Understanding Workshop*, 2017, pp. 677–684.

[23] M. Mashimo, T. Toda, H. Kawanami, K. Shikano, and N. Campbell, "Cross-language voice conversion evaluation using bilingual databases," in *Proc. 7th Int. Conf. Spoken Lang. Process.*, 2002.

[24] D. Sundermann, H. Hoge, A. Bonafonte, H. Ney, A. Black, and S. Narayanan, "Text-independent voice conversion based on unit selection," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2006, vol. 1.

[25] H. Wang, F. Soong, and H. Meng, "A spectral space warping approach to cross-lingual voice transformation in HMM-based TTS," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2015, pp. 4874–4878.

[5] [Online]. Available: http://www.festvox.org/cmu_arctic/

[6] [Online]. Available: https://doi.org/10.7488/ds/2645

[7] [Online]. Available: https://paperswithcode.com/task/voice-conversion

[26] D. Sundermann, H. Ney, and H. Hoge, "VTLN-based crosslanguage voice conversion," in *Proc. IEEE Autom. Speech Recognit. Understanding Workshop*, 2003.

[27] D. Erro, A. Moreno, and A. Bonafonte, "INCA algorithm for training voice conversion systems from nonparallel corpora," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 5, pp. 944–953, Jul. 2010.

[28] D. Erro and A. Moreno, "Frame alignment method for cross-lingual voice conversion," in *Proc. Interspeech*, 1972.

[29] J. Tao, M. Zhang, J. Nurminen, J. Tian, and X. Wang, "Supervisory data alignment for text-independent voice conversion," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 5, pp. 932–943, Jul. 2010.

[30] H. Silen, J. Nurminen, E. Helander, and M. Gabbouj, "Voice conversion for non-parallel datasets using dynamic kernel partial least squares regression," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 3, pp. 806–817, 2012.

[31] P. Song, Y. Jin, W. Zheng, and L. Zhao, "Text-independent voice conversion using speaker model alignment method from non-parallel speech," in *Proc. Annu. Conf. Int. Speech Commun. Assoc., Interspeech*, 2014, pp. 2308–2312.

[32] L. Sun, K. Li, H. Wang, S. Kang, and H. Meng, "Phonetic posteriorgrams for many-to-one voice conversion without parallel data training," in *Proc. IEEE Int. Conf. Multimedia Expo.*, 2016, pp. 1–6.

[33] T. J. Hazen, W. Shen, and C. White, "Query-by-example spoken term detection using phonetic posteriorgram templates," in *IEEE Autom. Speech Recognit. Understanding Workshop*, 2009, pp. 421–426.

[34] K. Kintzley, A. Jansen, and H. Hermansky, "Event selection from phone posteriorgrams using matched filters," in *Proc. Interspeech*, 2011, pp. 1905–1908.

[35] S. H. Mohammadi and A. Kain, "An overview of voice conversion systems," *Speech Commun.*, vol. 88, pp. 65–82, 2017.

[36] M. Narendranath, H. A. Murthy, S. Rajendran, and B. Yegnanarayana, "Transformation of formants for voice conversion using artificial neural networks," *Speech Commun.*, vol. 16, no. 2, pp. 207–216, 1995.

[37] K. Hornik, M. Stinchcombe, and H. White, "Multilayer feedforward networks are universal approximators," *Neural Netw.*, vol. 2, no. 5, pp. 359–366, 1989.

[38] R. H. Laskar, D. Chakrabarty, F. A. Talukdar, K. S. Rao, and K. Banerjee, "Comparing ANN and GMM in a voice conversion framework," *Appl. Soft Comput.*, vol. 12, no. 11, pp. 3332–3342, 2012.

[39] H. Q. Nguyen, S. W. Lee, X. Tian, M. Dong, and E. S. Chng, "High quality voice conversion using prosodic and high-resolution spectral features," *Multimedia Tools Appl.*, vol. 75, no. 9, pp. 5265–5285, 2016.

[40] L. Sun, S. Kang, K. Li, and H. Meng, "Voice conversion using deep bidirectional long short-term memory based recurrent neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2015, pp. 4869–4873.

[41] J. Wu, Z. Wu, and L. Xie, "On the use of I-vectors and average voice model for voice conversion without parallel data," in *Proc. Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA)*, 2016, pp. 1–6.

[42] F. L. Xie, F. K. Soong, and H. Li, "A KL divergence and DNN-based approach to voice conversion without parallel training sentences," in *Proc. Annu. Conf. Int. Speech Commun. Assoc., Interspeech*, pp. 287–291, 2016.

[43] C.-C. Hsu, H.-T. Hwang, Y.-C. Wu, Y. Tsao, and H.-M. Wang, "Voice conversion from non-parallel corpora using variational auto-encoder," in *Proc. IEEE, Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf.*, 2016, pp. 1–6.

[44] X. Tian, J. Wang, X. Haihua, E. S. Chng, and H. Li, "Average modeling approach to voice conversion with non-parallel data," in *Proc. Odyssey Speaker Lang. Recognit. Workshop*, 2018, pp. 1–10.

[45] L. Sun, H. Wang, S. Kang, K. Li, and H. Meng, "Personalized, cross-lingual TTS using phonetic posteriorgrams," in *Proc. Interspeech*, 2016, pp. 322–326.

[46] J. Serrà, S. Pascual, and C. S. Perales, "Blow: A single-scale hyperconditioned flow for non-parallel raw-audio voice conversion," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 6793–6803.

[47] C.-C. Hsu, H.-T. Hwang, Y.-C. Wu, Y. Tsao, and H.-M. Wang, "Voice conversion from unaligned corpora using variational autoencoding Wasserstein generative adversarial networks," 2017, *arXiv:1704.00849*.

[48] T. Kaneko and H. Kameoka, "Parallel-data-free voice conversion using cycle-consistent adversarial networks," 2017, *arXiv:1711.11293*.

[49] F. Fang, J. Yamagishi, I. Echizen, and J. Lorenzo-Trueba, "High-quality nonparallel voice conversion based on cycle-consistent adversarial network," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2018, pp. 5279–5283.

[50] J. Lorenzo-Trueba, F. Fang, X. Wang, I. Echizen, J. Yamagishi, and T. Kinnunen, "Can we steal your vocal identity from the Internet?: Initial investigation of cloning Obama's voice using GAN, WaveNet and low-quality found data," 2018, *arXiv:1803.00860*.

[51] H. Kameoka, T. Kaneko, K. Tanaka, and N. Hojo, "StarGAN-VC: Non-parallel many-to-many voice conversion with star generative adversarial networks," 2018, *arXiv:1806.02169*.

[52] M. Airaksinen, L. Juvela, B. Bollepalli, J. Yamagishi, and P. Alku, "A comparison between straight, glottal, and sinusoidal vocoding in statistical parametric speech synthesis," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 9, pp. 1658–1670, Sep. 2018.

[53] X. Wang, J. Lorenzo-Trueba, S. Takaki, L. Juvela, and J. Yamagishi, "A comparison of recent waveform generation and acoustic modeling methods for neural-network-based speech synthesis," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Calgary, Canada, Apr. 2018, pp. 4804–4808.

[54] T. Hayashi, A. Tamamori, K. Kobayashi, K. Takeda, and T. Toda, "An investigation of multi-speaker training for wavenet vocoder," in *Proc. IEEE Autom. Speech Recognit. Understanding Workshop*, 2017, pp. 712–718.

[55] A. Tamamori, T. Hayashi, K. Kobayashi, K. Takeda, and T. Toda, "Speaker-dependent wavenet vocoder," in *Proc. Interspeech*, vol. 2017, 2017, pp. 1118–1122.

[56] Y.-C. Wu, T. Hayashi, P. L. Tobing, K. Kobayashi, and T. Toda, "Quasi-periodic wavenet vocoder: A pitch dependent dilated convolution model for parametric speech generation," 2019, *arXiv:1907.00797*.

[57] Y.-C. Wu, P. L. Tobing, T. Hayashi, K. Kobayashi, and T. Toda, "Statistical voice conversion with quasi-periodic wavenet vocoder," 2019, *arXiv:1907.08940*.

[58] B. Sisman, M. Zhang, S. Sakti, H. Li, and S. Nakamura, "Adaptive wavenet vocoder for residual compensation in GAN-based voice conversion," in *Proc. IEEE Spoken Lang. Technol. Workshop*, 2018, pp. 282–289.

[59] H. Du, X. Tian, L. Xie, and H. Li, "Wavenet factorization with singular value decomposition for voice conversion," in *Proc. IEEE Autom. Speech Recognit. Understanding Workshop*, 2019, pp. 152–159.

[60] W.-C. Huang *et al.*, "Refined wavenet vocoder for variational autoencoder based voice conversion," in *Proc. IEEE 27th Eur. Signal Process. Conf.*, 2019, pp. 1–5.

[61] B. Sisman, M. Zhang, and H. Li, "A voice conversion framework with tandem feature sparse representation and speaker-adapted wavenet vocoder," in *Proc. Interspeech*, 2018, pp. 1978–1982.

[62] B. Sisman, M. Zhang, and H. L, "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.

[63] N. Kalchbrenner *et al.*, "Efficient neural audio synthesis," 2018, *arXiv:1802.08435*.

[64] R. Prenger, R. Valle, and B. Catanzaro, "WaveGlow: A flow-based generative network for speech synthesis," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Brighton, U.K., May 2019, pp. 3617–3621.

[65] S. Kim, S.-G. Lee, J. Song, J. Kim, and S. Yoon, "Flowavenet: A generative flow for raw audio," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 3370–3378.

[66] A. van den Oord *et al.*, "Wavenet: A generative model for raw audio," 2016, *arXiv:1609.03499*.

[67] T. Toda *et al.*, "The voice conversion challenge," in *Proc. Interspeech*, 2016, pp. 1632–1636.

[68] M. Wester, Z. Wu, and J. Yamagishi, "Multidimensional scaling of systems in the voice conversion challenge," in *SSW*, 2016, pp. 38–43.

[69] M. Wester, Z. Wu, and J. Yamagishi, "Analysis of the voice conversion challenge 2016 evaluation results," in *Proc. Interspeech*, 2016, pp. 1637–1641.

[70] J. Lorenzo-Trueba *et al.*, "The voice conversion challenge: Promoting development of parallel and nonparallel methods," 2018, *arXiv:1804.04262*.

[71] J. Lorenzo-Trueba *et al.*, "The voice conversion challenge 2018: Database and results," 2018.

[72] P. L. Tobing, Y.-C. Wu, T. Hayashi, K. Kobayashi, and T. Toda, "NU voice conversion system for the voice conversion challenge 2018," *Odyssey*, 2018, pp. 219–226.

[73] D. Griffin and J. Lim, "Signal estimation from modified short-time Fourier transform," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 32, no. 2, pp. 236–243, Apr. 1984.

[74] E. Moulines and F. Charpentier, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," *Speech Commun.*, vol. 9, no. 5–6, pp. 453–467, 1990.

[75] H. Valbret, E. Moulines, and J.-P. Tubach, "Voice transformation using psola technique," *Speech Commun.*, vol. 11, no. 2–3, pp. 175–187, 1992.

[76] L. M. Arslan, "Speaker transformation algorithm using segmental codebooks (STASC)," *Speech Commun.*, vol. 28, no. 3, pp. 211–226, 1999.

[77] Y. Stylianou, "Applying the harmonic plus noise model in concatenative speech synthesis," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 1, pp. 21–29, Jan. 2001.

[78] Y. Stylianou and O. Cappe, "A system for voice conversion based on probabilistic classification and a harmonic plus noise model," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process., (ICASSP)*, 1998, pp. 281–284.

[79] D. Erro and A. Moreno, "Weighted frequency warping for voice conversion," in *Proc. 8th Annu. Conf. Int. Speech Commun. Assoc.*, 2007.

[80] "Mel log spectrum approximation (MLSA) filter for speech synthesis," *Electron. Commun. Jpn. (Part I: Commun.)*, vol. 66, no. 2, pp. 10–18, 1983.

[81] M. Airaksinen, L. Juvela, B. Bollepalli, J. Yamagishi, and P. Alku, "A comparison between straight, glottal, and sinusoidal vocoding in statistical parametric speech synthesis," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 9, pp. 1658–1670, Sep. 2018.

[82] 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.

[83] S. Desai, E. V. Raghavendra, B. Yegnanarayana, A. W. Black, and K. Prahallad, "Voice conversion using artificial neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2009, pp. 3893–3896.

[84] B. Sisman and H. Li, "Wavelet analysis of speaker dependent and independent prosody for voice conversion," in *Proc. Interspeech*, 2018, pp. 52–56.

[85] W.-N. Hsu, Y. Zhang, and J. Glass, "Unsupervised learning of disentangled and interpretable representations from sequential data," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 1878–1889.

[86] W.-N. Hsu, Y. Zhang, and J. Glass, "Learning latent representations for speech generation and transformation," 2017, *arXiv:1704.04222*.

[87] S. Furui, "Digital speech processing, synthesis, and recognition(revised and expanded)," *Digit. Speech Process., Synth., Recognit.*, 2000.

[88] J. Shen *et al.*, "Natural TTS synthesis by conditioning wavenet on MEL spectrogram predictions," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 4779–4783.

[89] R. Liu, B. Sisman, J. Li, F. Bao, G. Gao, and H. Li, "Teacher-student training for robust tacotron-based TTS," in *Proc. 2020 IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2019, pp. 6274–6278.

[90] Z. Hanzlíček, J. Vít and D. Tihelka, "Wavenet-based speech synthesis applied to czech," in *Proc. Int. Conf. Text, Speech, Dialogue*. Berlin, Germany: Springer, 2018, pp. 445–452.

[91] S. Ö. Arik *et al.*, "Deep voice: Real-time neural text-to-speech," in *Proc. 34th Int. Conf. Mach.*, 2017, pp. 195–204.

[92] B. Sisman, "Machine learning for limited data voice conversion," Ph.D. dissertation, 2019.

[93] K. Chen, B. Chen, J. Lai, and K. Yu, "High-quality voice conversion using spectrogram-based wavenet vocoder," in *Proc. Interspeech*, 2018, pp. 1993–1997.

[94] N. Adiga, V. Tsiaras, and Y. Stylianou, "On the use of wavenet as a statistical vocoder," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 5674–5678.

[95] Y. Zhao, S. Takaki, H.-T. Luong, J. Yamagishi, D. Saito, and N. Minematsu, "Wasserstein GAN and waveform loss-based acoustic model training for multi-speaker text-to-speech synthesis systems using a wavenet vocoder," *IEEE Access*, vol. 6, pp. 60478–60488, 2018.

[96] X. Tian, E. S. Chng, and H. Li, "A speaker-dependent wavenet for voice conversion with non-parallel data," in *Proc. Interspeech, Graz, Austria*, 2019, pp. 15–19.

[97] H. Lu, Z. Wu, R. Li, S. Kang, J. Jia, and H. Meng, "A compact framework for voice conversion using wavenet conditioned on phonetic posteriorgrams," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 6810–6814.

[98] H. Du, X. Tian, L. Xie, and H. Li, "Wavenet factorization with singular value decomposition for voice conversion," in *Proc. IEEE Autom. Speech Recognit. Understanding Workshop*, 2019, pp. 152–159.

[99] S. Liu, Y. Cao, X. Wu, L. Sun, X. Liu, and H. Meng, "Jointly trained conversion model and wavenet vocoder for non-parallel voice conversion using mel-spectrograms and phonetic posteriorgrams," in *Proc. Interspeech*, 2019, pp. 714–718.

[100] J. Chorowski, R. Weiss, S. Bengio, and A. Oord, "Unsupervised speech representation learning using wavenet autoencoders," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 12, pp. 2041–2053, Dec. 2019.

[101] J. Lorenzo-Trueba *et al.*, "Towards achieving robust universal neural vocoding," in *Proc. Interspeech*, 2019, pp. 181–185.

[102] P. Govalkar, J. Fischer, F. Zalkow, and C. Dittmar, "A comparison of recent neural vocoders for speech signal reconstruction," in *Proc. 10th ISCA Speech Synth. Workshop*, 2019, pp. 7–12.

[103] Y.-H. Yi, Y. Ai, Z.-H. Ling, and L.-R. Dai, "Singing voice synthesis using deep autoregressive neural networks for acoustic modeling," 2019, *arXiv:1906.08977*.

[104] T. Okamoto, T. Toda, Y. Shiga, and H. Kawai, "Real-time neural text-to-speech with sequence-to-sequence acoustic model and waveglow or single Gaussian wavernn vocoders," in *Proc. Interspeech*, 2019, pp. 1308–1312.

[105] S. Maiti and M. I. Mandel, "Parametric resynthesis with neural vocoders," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, 2019, pp. 303–307.

[106] X. Wang, S. Takaki, and J. Yamagishi, "Neural source-filter-based waveform model for statistical parametric speech synthesis," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 5916–5920.

[107] X. Wang and J. Yamagishi, "Neural harmonic-plus-noise waveform model with trainable maximum voice frequency for text-to-speech synthesis," 2019, *arXiv:1908.10256*.

[108] X. Wang, S. Takaki, and J. Yamagishi, "Neural source-filter waveform models for statistical parametric speech synthesis," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 402–415, 2020.

[109] R. Yamamoto, E. Song, and J.-M. Kim, "Parallel wavegan: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 6199–6203.

[110] K. Kumar *et al.*, "Melgan: Generative adversarial networks for conditional waveform synthesis," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 14 910–14 921.

[111] X. Tian, S. W. Lee, Z. Wu, E. S. Chng, S. Member, and H. Li, "An exemplar-based approach to frequency warping for voice conversion," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 10, pp. 1–10, 2016.

[112] H. Kuwabara and Y. Sagisak, "Acoustic characteristics of speaker individuality: Control and conversion," *Speech Commun.*, vol. 16, no. 2, pp. 165–173, 1995.

[113] Y. Stylianou, O. Cappé, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Trans. Speech Audio Process.*, vol. 6, no. 2, pp. 131–142, Mar. 1998.

[114] H. Matsumoto and Y. Yamashita, "Unsupervised speaker adaptation from short utterances based on a minimized fuzzy objective function," *J. Acoust. Soc. Jpn. (E)*, vol. 14, no. 5, pp. 353–361, 1993.

[115] T. Toda, H. Saruwatari, and K. Shikano, "Voice conversion algorithm based on Gaussian mixture model with dynamic frequency warping of straight spectrum," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. Proc.*, 2001, pp. 841–844.

[116] T. Toda, J. Lu, S. Nakamura, and K. Shikano, "Voice conversion algorithm based on Gaussian mixture model applied to straight," 2000.

[117] T. Toda, A. W. Black, and K. Tokuda, "Spectral conversion based on maximum likelihood estimation considering global variance of converted parameter," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2005, pp. I–9.

[118] T. K. Moon, "The expectation-maximization algorithm," *IEEE Signal Process. Mag.*, vol. 13, no. 6, pp. 47–60, Nov. 1996.

[119] C. B. Do and S. Batzoglou, "What is the expectation maximization algorithm?," *Nat. Biotechnol.*, vol. 26, no. 8, pp. 897–899, 2008.

[120] G. Xuan, W. Zhang, and P. Chai, "Em algorithms of Gaussian mixture model and hidden Markov model," in *Proc. Int. Conf. Image Process.*, 2001, pp. 145–148.

[121] M. R. Gupta *et al.*, "Theory and use of the em algorithm," *Found. Trends Signal Process.*, vol. 4, no. 3, pp. 223–296, 2011.

[122] S. Takamichi, T. Toda, A. W. Black, and S. Nakamura, "Modulation spectrum-based post-filter for GMM-based voice conversion," in *Proc. IEEE Signal Inf. Process. Assoc. Annu. Summit Conf., Asia-Pacific*, 2014, pp. 1–4.

[123] Y. Ohtani, T. Toda, H. Saruwatari, and K. Shikano, "Maximum likelihood voice conversion based on GMM with straight mixed excitation," 2006.

[124] H. Kawanami, Y. Iwami, T. Toda, H. Saruwatari, and K. Shikano, "GMM-based voice conversion applied to emotional speech synthesis," in *Proc. 8th Eur. Conf. Speech Commun. Technol.*, 2003.

[125] R. Aihara, R. Takashima, T. Takiguchi, and Y. Ariki, "GMM-based emotional voice conversion using spectrum and prosody features," *Amer. J. Signal Process.*, vol. 2, no. 5, pp. 134–138, 2012.

[126] H.-T. Hwang, Y. Tsao, H.-M. Wang, Y.-R. Wang, and S.-H. Chen, "Incorporating global variance in the training phase of GMM-based voice conversion," in *Proc. IEEE Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf.*, 2013, pp. 1–6.

[127] T.-C. Zorilă, D. Erro, and I. Hernáez, "Improving the quality of standard GMM-based voice conversion systems by considering physically motivated linear transformations," in *Proc. Adv. Speech Lang. Technol. Iberian Lang*. Berlin, Germany: Springer, 2012, pp. 30–39.

[128] M. Ghorbandoost, A. Sayadiyan, M. Ahangar, H. Sheikhzadeh, A. S. Shahrebabaki, and J. Amini, "Voice conversion based on feature combination with limited training data," *Speech Commun.*, vol. 67, pp. 113–128, 2015.

[129] M. S. Ribeiro, J. Yamagishi, and R. A. Clark, "A perceptual investigation of wavelet-based decomposition of f0 for text-to-speech synthesis," in *Proc. 16th Annu. Conf. Int. Speech Commun. Assoc.*, 2015.

[130] M. S. Ribeiro, O. Watts, J. Yamagishi, and R. A. Clark, "Wavelet-based decomposition of f0 as a secondary task for DNN-based speech synthesis with multi-task learning," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2016, pp. 5525–5529.

[131] C.-C. Wang, Z.-H. Ling, B.-F. Zhang, and L.-R. Dai, "Multi-layer f0 modeling for HMM-based speech synthesis," in *Proc. IEEE 6th Int. Symp. Chin. Spoken Lang. Process.*, 2008, pp. 1–4.

[132] G. Sanchez, H. Silen, J. Nurminen, and M. Gabbouj, "Hierarchical modeling of F0 contours for voice conversion," in *Proc. Annu. Conf. Int. Speech Commun. Assoc., INTERSPEECH*, 2014, pp. 2318–2321.

[133] D. Erro, A. Moreno, and A. Bonafonte, "Voice conversion based on weighted frequency warping," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 5, pp. 922–931, Jul. 2009.

[134] D. Sundermann and H. Ney, "VTLN-based voice conversion," in *Proc. 3rd IEEE Int. Symp. Signal Process. Inf. Technol.*, 2003, pp. 556–559.

[135] M. Eichner, M. Wolff, and R. Hoffmann, "Voice characteristics conversion for TTS using reverse VTLN," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2004, pp. I–17.

[136] A. Přibilová and J. Přibil, "Non-linear frequency scale mapping for voice conversion in text-to-speech system with cepstral description," *Speech Commun.*, vol. 48, no. 12, pp. 1691–1703, 2006.

[137] R. Vích and M. Vondra, "Pitch synchronous transform warping in voice conversion," in *Cognitive Behaviour System*, Berlin, Germany: Springer-Verlag, 2012, pp. 280–289.

[138] E. Godoy, O. Rosec, and T. Chonavel, "Voice conversion using dynamic frequency warping with amplitude scaling, for parallel or nonparallel corpora," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 4, pp. 1313–1323, May 2012.

[139] D. Lee and H. Seung, "Algorithms for non-negative matrix factorization," *Adv. Neural Inf. Process. Syst.*, no. 1, pp. 556–562, 2001.

[140] S.-S. Wang *et al.*, "Wavelet speech enhancement based on nonnegative matrix factorization," *IEEE Signal Process. Lett.*, vol. 23, no. 8, pp. 1101–1105, Aug. 2016.

[141] N. Mohammadiha, P. Smaragdis, and A. Leijon, "Supervised and unsupervised speech enhancement using nonnegative matrix factorization," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 21, no. 10, pp. 2140–2151, Oct. 2013.

[142] K. A. Akarsh, "Speech enhancement using non negative matrix factorization and enhanced NMF," in *Proc. Int. Conf. Circuit, Power Comput. Technol.*, 2015.

[143] K. W. Wilson, B. Raj, P. Smaragdis, and A. Divakaran, "Speech denoising using nonnegative matrix factorization with priors," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2008, pp. 4029–4032.

[144] M. Sun, Y. Li, J. F. Gemmeke, and X. Zhang, "Speech enhancement under low SNR conditions via noise estimation using sparse and low-rank NMF with Kullback-Leibler divergence," *IEEE/ACM Trans. Audio, Speech Lang. Process.*, vol. 23, no. 7, pp. 1233–1242, Jul. 2015.

[145] Z. Wu, T. Virtanen, T. Kinnunen, E. S. Chng, and H. Li, "Exemplar-based voice conversion using non-negative spectrogram deconvolution," in *Proc. 8th ISCA Speech Synth. Workshop*, 2013.

[146] Y. C. Wu, H. T. Hwang, C. C. Hsu, Y. Tsao, and H. M. Wang, "Locally linear embedding for exemplar-based spectral conversion," in *Proc. Annu. Conf. Int. Speech Commun. Assoc., INTERSPEECH*, 2016, pp. 1652–1656.

[147] H. Ming, D. Huang, L. Xie, S. Zhang, M. Dong, and H. Li, "Exemplar-based sparse representation of timbre and prosody for voice conversion," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2016, pp. 5175–5179.

[148] B. Şişman, H. Li, and K. C. Tan, "Transformation of prosody in voice conversion," in *Proc. Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf.*, 2017, pp. 1537–1546.

[149] C.-C. Hsu, H.-T. Hwang, Y.-C. Wu, Y. Tsao, and H.-M. Wang, "Dictionary update for NMF-based voice conversion using an encoder-decoder network," in *Proc. 10th Int. Symp. Chin. Spoken Lang. Process.*, 2016, pp. 293–297.

[150] H. Ney, D. Suendermann, A. Bonafonte, and H. Höge, "A first step towards text-independent voice conversion," in *Proc. 8th Int. Conf. Spoken Lang. Process.*, 2004.

[151] H. Ye and S. J. Young, "Voice conversion for unknown speakers," in *Proc. Interspeech 8th Int. Conf. Spoken Lang. Process. (ICSLP)*, Jeju Island, Korea, Oct. 4–8, 2004. [Online]. Available: http://www.isca-speech.org/archive/interspeech_2004/i04_1161.html

[152] H. Ye and S. Young, "Quality-enhanced voice morphing using maximum likelihood transformations," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 4, pp. 1301–1312, Jul. 2006.

[153] A. W. Black and N. Campbell, "Optimising selection of units from speech databases for concatenative synthesis," 1995.

[154] K. Fujii, J. Okawa, and K. Suigetsu, "High individuality voice conversion based on concatenative speech synthesis," *Int. J. Elect., Comput., Energetic, Electron. Commun. Eng.*, vol. 1, no. 11, pp. 1617–1622, 2007.

[155] Y. Sagisaka, N. Kaiki, N. Iwahashi, and K. Mimura, "Atr μ-talk speech synthesis system," in *Proc. 2nd Int. Conf. Spoken Lang. Process.*, 1992.

[156] D. Erro, F. Diego, and A. Bonafonte, "Voice conversion of non-aligned data using unit selection," 2006.

[157] A. Mouchtaris, J. Van der Spiegel, and P. Mueller, "Nonparallel training for voice conversion based on a parameter adaptation approach," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 3, pp. 952–963, May 2006.

[158] T. Toda, Y. Ohtani, and K. Shikano, "Eigenvoice conversion based on Gaussian mixture model," in *Proc. Interspeech*, 2006.

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

[160] Z. Wu, T. Kinnunen, E. S. Chng, and H. Li, "Mixture of factor analyzers using priors from non-parallel speech for voice conversion," *IEEE Signal Process. Lett.*, vol. 19, no. 12, pp. 914–917, Dec. 2012.

[161] Y. Stylianou, O. Cappé, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Trans. Speech Audio Process.*, vol. 6, no. 2, pp. 131–142, Mar. 1998.

[162] Y. Zhou, X. Tian, H. Xu, R. K. Das, and H. Li, "Cross-lingual voice conversion with bilingual phonetic posteriorgrams and average modeling," in *Proc. Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2019, pp. 6790–6794.

[163] T. Kinnunen, L. Juvela, P. Alku, and J. Yamagishi, "Non-parallel voice conversion using i-vector PLDA: Towards unifying speaker verification and transformation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2017, pp. 5535–5539.

[164] H. Miyoshi, Y. Saito, S. Takamichi, and H. Saruwatari, "Voice conversion using sequence-to-sequence learning of context posterior probabilities," 2017, *arXiv:1704.02360*.

[165] S.-W. Park, D.-Y. Kim, and M.-C. Joe, "Cotatron: Transcription-guided speech encoder for any-to-many voice conversion without parallel data," 2020, *arXiv:abs/2005.03295*.

[166] F.-L. Xie, Y. Qian, F. K. Soong, and H. Li, "Pitch transformation in neural network based voice conversion," in *Proc. IEEE 9th Int. Symp. Chin. Spoken Lang. Process.*, 2014, pp. 197–200.

[167] T. Nakashika, R. Takashima, T. Takiguchi, and Y. Ariki, "Voice conversion in high-order eigen space using deep belief NETS," in *Proc. Interspeech*, 2013, pp. 369–372.

[168] S. H. Mohammadi and A. Kain, "Voice conversion using deep neural networks with speaker-independent pre-training," in *Proc. IEEE Spoken Lang. Technol. Workshop*, 2014, pp. 19–23.

[169] F.-L. Xie, Y. Qian, Y. Fan, F. K. Soong, and H. Li, "Sequence error (SE) minimization training of neural network for voice conversion," in *Proc. 15th Annu. Conf. Int. Speech Commun. Assoc.*, 2014.

[170] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2000, pp. 1315–1318.

[171] L.-H. Chen, Z.-H. Ling, L.-J. Liu, and L.-R. Dai, "Voice conversion using deep neural networks with layer-wise generative training," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 22, no. 12, pp. 1859–1872, Dec. 2014.

[172] T. Nakashika, T. Takiguchi, and Y. Ariki, "High-order sequence modeling using speaker-dependent recurrent temporal restricted Boltzmann machines for voice conversion," in *Proc. Annu. Conf. Int. Speech Commun. Assoc., INTERSPEECH*, 2014, pp. 2278–2282.

[173] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[174] F. A. Gers, J. Schmidhuber, and F. Cummins, "Learning to forget: Continual prediction with LSTM," 1999.

[175] K. Greff, R. K. Srivastava, J. Koutník, B. R. Steunebrink, and J. Schmidhuber, "LSTM: A search space odyssey," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 10, pp. 2222–2232, Oct. 2017.

[176] H. Ming, D. Huang, L. Xie, J. Wu, M. Dong, and H. Li, "Deep bidirectional LSTM modeling of timbre and prosody for emotional voice conversion," in *Proc. Annu. Conf. Int. Speech Commun. Assoc., INTERSPEECH*, Sep. 2016, pp. 2453–2457.

[177] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 2014, *arXiv:1409.0473*.

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

[179] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2016, pp. 4960–4964.

[180] Y. Wang *et al.*, "Tacotron: Towards end-to-end speech synthesis," in *Proc. Interspeech*, 2017, pp. 4006–4010.

[181] W. Ping, "Deep voice 3: 2000-speaker neural text-to-speech," in *Proc. ICLR*, 2018, pp. 214–217.

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

[183] J.-X. Zhang, Z.-H. Ling, L.-J. Liu, Y. Jiang, and L.-R. Dai, "Sequence-to-sequence acoustic modeling for voice conversion," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 3, pp. 631–644, Mar. 2019.

[184] K. Tanaka, H. Kameoka, T. Kaneko, and N. Hojo, "AttS2S-VC: Sequence-to-sequence voice conversion with attention and context preservation mechanisms," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 6805–6809.

[185] K. Cho *et al.*, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," 2014, *arXiv:1406.1078*.

[186] M.-T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," 2015, *arXiv:1508.04025*.

[187] J. Gehring, M. Auli, D. Grangier, D. Yarats, and Y. Dauphin, "Convolutional sequence to sequence learning," 2017, *arXiv:abs/1705.03122*.

[188] H. Kameoka, K. Tanaka, T. Kaneko, and N. Hojo, "ConvS2S-VC: Fully convolutional sequence-to-sequence voice conversion," 2018, *arXiv:abs/1811.01609*.

[189] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1125–1134.

[190] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2223–2232.

[191] Y. Choi, M. Choi *et al.*, "Stargan: Unified generative adversarial networks for multi-domain image-to-image translation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 8789–8797.

[192] M.-Y. Liu *et al.*, "Few-shot unsupervised image-to-image translation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 10 551–10 560.

[193] K. E. Ak, A. A. Kassim, J. Hwee Lim, and J. Yew Tham, "Learning attribute representations with localization for flexible fashion search," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7708–7717.

[194] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2223–2232.

[195] K. E. Ak, J. H. Lim, J. Y. Tham, and A. A. Kassim, "Attribute manipulation generative adversarial networks for fashion images," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 10 541–10 550.

[196] K. E. Ak, "Deep learning approaches for attribute manipulation and text-to-image synthesis," Ph.D. dissertation, 2019.

[197] K. E. Ak, J. H. Lim, J. Y. Tham, and A. A. Kassim, "Efficient multi-attribute similarity learning towards attribute-based fashion search," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2018, pp. 1671–1679.

[198] Y. Choi, Y. Uh, J. Yoo, and J.-W. Ha, "StarGAN v2: Diverse image synthesis for multiple domains," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 8188–8197.

[199] I. Goodfellow *et al.*, "Generative adversarial NETS," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

[200] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz, "Multimodal unsupervised image-to-image translation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 172–189.

[201] J.-Y. Zhu *et al.*, "Toward multimodal image-to-image translation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 465–476.

[202] K. E. Ak, J. H. Lim, J. Y. Tham, and A. A. Kassim, "Semantically consistent text to fashion image synthesis with an enhanced attentional generative adversarial network," *Pattern Recognit. Lett.*, 2020.

[203] K. Emir Ak, J. Hwee Lim, J. Yew Tham, and A. Kassim, "Semantically consistent hierarchical text to fashion image synthesis with an enhanced-attentional generative adversarial network," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, 2019.

[204] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional gans," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 8798–8807.

[205] Z. Meng, J. Li, Y. Gong, and B.-H. F. Juang, "Cycle-consistent speech enhancement," *Proc. Interspeech*, 2018.

[206] M. Mimura, S. Sakai, and T. Kawahara, "Cross-domain speech recognition using nonparallel corpora with cycle-consistent adversarial networks," *IEEE Autom. Speech Recognit. Understanding Workshop*, 2017, pp. 134–140.

[207] D. Yook, I.-C. Yoo, and S. Yoo, "Voice conversion using conditional cyclegan," in *Proc. IEEE Int. Conf. Comput. Sci. Comput. Intell.*, 2018, pp. 1460–1461.

[208] N. Jia, C. Zheng, and W. Sun, "Speech synthesis of children's reading based on cycleGAN model," in *J. Phys.: Conf. Ser.*, vol. 1607, no. 1, 2020, Art. no. 012046.

[209] S. Huang, Q. Li, C. Anil, X. Bao, S. Oore, and R. B. Grosse, "TimbreTron: A wavenet (cycleGAN (CQT (audio))) pipeline for musical timbre transfer," in *Proc. Int. Conf. Learn. Representations*, 2018.

[210] T. Kaneko and H. Kameoka, "CycleGAN-VC: Non-parallel voice conversion using cycle-consistent adversarial networks," in *Proc. 26th IEEE Eur. Signal Process. Conf.*, 2018, pp. 2100–2104.

[211] T. Kaneko, H. Kameoka, K. Tanaka, and N. Hojo, "Cycle GAN-VC2: Improved cycle GAN-based non-parallel voice conversion," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 6820–6824.

[212] Y. Taigman, A. Polyak, and L. Wolf, "Unsupervised cross-domain image generation," 2016, *arXiv:abs/1611.02200*.

[213] P. L. Tobing, Y.-C. Wu, T. Hayashi, K. Kobayashi, and T. Toda, "Voice conversion with cyclic recurrent neural network and fine-tuned wavenet vocoder," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 6815–6819.

[214] B. Sisman, M. Zhang, M. Dong, and H. Li, "On the study of generative adversarial networks for cross-lingual voice conversion," in *Proc. IEEE Autom. Speech Recognit. Understanding Workshop*, 2019, pp. 144–151.

[215] K. Zhou, B. Sisman, and H. Li, "Transforming spectrum and prosody for emotional voice conversion with non-parallel training data," 2020, *arXiv:2002.00198*.

[216] K. Zhou, B. Sisman, M. Zhang, and H. Li, "Converting anyone's emotion: Towards speaker-independent emotional voice conversion," 2020, *arXiv:2005.07025*.

[217] C.-C. Yeh, P.-C. Hsu, J.-C. Chou, H.-Y. Lee, and L.-S. Lee, "Rhythm-flexible voice conversion without parallel data using cycle-GAN over phoneme posteriorgram sequences," in *Proc. IEEE Spoken Lang. Technol. Workshop.*, 2018, pp. 274–281.

[218] R. Liu, B. Sisman, F. Bao, G. Gao, and H. Li, "WaveTTS: Tacotron-based TTS with joint time-frequency domain loss," 2020, *arXiv:2002.00417*.

[219] R. Liu, B. Sisman, F. Bao, G. Gao, and H. Li, "Modeling prosodic phrasing with multi-task learning in tacotron-based TTS," *IEEE Signal Process. Lett.*, vol. 27, pp. 1470–1474, 2020.

[220] R. Liu, B. Sisman, G. Gao, and H. Li, "Expressive TTS training with frame and style reconstruction loss," 2020, *arXiv:2008.01490*.

[221] M. Zhang, Y. Zhou, L. Zhao, and H. Li, "Transfer learning from speech synthesis to voice conversion with non-parallel training data," 2020.

[222] W.-C. Huang, T. Hayashi, Y.-C. Wu, H. Kameoka, and T. Toda, "Voice transformer network: Sequence-to-sequence voice conversion using transformer with text-to-speech pretraining," 2019, *arXiv:1912.06813*.

[223] J.-X. Zhang, Z.-H. Ling, Y. Jiang, L.-J. Liu, C. Liang, and L.-R. Dai, "Improving sequence-to-sequence voice conversion by adding text-supervision," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 6785–6789.

[224] H.-T. Luong and J. Yamagishi, "Bootstrapping non-parallel voice conversion from speaker-adaptive text-to-speech," in *Proc. IEEE Autom. Speech Recognit. Understanding Workshop.*, 2019, pp. 200–207.

[225] H.-T. Luong and J. Yamagishi, "Nautilus: A versatile voice cloning system," 2020, *arXiv:2005.11004*.

[226] F. Biadsy, R. J. Weiss, P. J. Moreno, D. Kanvesky, and Y. Jia, "Parrotron: An end-to-end speech-to-speech conversion model and its applications to hearing-impaired speech and speech separation," in *Proc. Interspeech*, 2019, pp. 4115–4119.

[227] S. Liu, Y. Cao, and H. Meng, "Multi-target emotional voice conversion with neural vocoders," 2020, *arXiv:2004.03782*.

[228] M. Zhang, B. Sisman, S. S. Rallabandi, H. Li, and L. Zhao, "Error reduction network for DBLSTM-based voice conversion," in *Proc. IEEE Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf.*, 2018, pp. 823–828.

[229] A. B. L. Larsen, S. K. Sonderby, H. Larochelle, and O. Winther, "Autoencoding beyond pixels using a learned similarity metric," in *Proc. 33rd Int. Conf. Mach. Learn., PMLR*, 2016.

[230] J.-C. Chou, C. chieh Yeh, and H. yi Lee, "One-shot voice conversion by separating speaker and content representations with instance normalization," 2019, *arXiv:abs/1904.05742*.

[231] D.-Y. Wu and H.-Y. Lee, "One-shot voice conversion by vector quantization," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 7734–7738.

[232] D.-Y. Wu, Y.-H. Chen, and H.-Y. Lee, "Vqvc+: One-shot voice conversion by vector quantization and u-net architecture," 2020, *arXiv:2006.04154*.

[233] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *Stat*, vol. 1050, p. 10, 2014.

[234] A. Van den oord *et al.*, "Neural discrete representation learning," in *Adv. Neural Inf. Process. Syst.*, 2017, pp. 6306–6315.

[235] A. Razavi, A. van den Oord, and O. Vinyals, "Generating diverse high-fidelity images with vq-vae-2," *Adv. Neural Inf. Process. Syst.*, 2019, pp. 14 866–14 876.

[236] S. Ding and R. Gutierrez-Osuna, "Group latent embedding for vector quantized variational autoencoder in non-parallel voice conversion," in *Proc. Interspeech*, 2019, pp. 724–728.

[237] W.-C. Huang, H.-T. Hwang, Y.-H. Peng, Y. Tsao, and H.-M. Wang, "Voice conversion based on cross-domain features using variational auto encoders," in *Proc. IEEE 11th Int. Symp. Chin. Spoken Lang. Process.*, 2018, pp. 51–55.

[238] Y. Li, K. A. Lee, Y. Yuan, H. Li, and Z. Yang, "Many-to-many voice conversion based on bottleneck features with variational autoencoder for non-parallel training data," in *Proc. IEEE Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf.*, 2018, pp. 829–833.

[239] Y. Saito, Y. Ijima, K. Nishida, and S. Takamichi, "Non-parallel voice conversion using variational autoencoders conditioned by phonetic posteriorgrams and D-vectors," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 5274–5278.

[240] W.-C. Huang *et al.*, "Unsupervised representation disentanglement using cross domain features and adversarial learning in variational autoencoder based voice conversion," *IEEE Trans. Emerg. Top. Comput. Intell.*, vol. 4, no. 4, pp. 468–479, Aug. 2020. [Online]. Available: http://dx.doi.org/10.1109/TETCI.2020.2977678

[241] T. Kaneko, H. Kameoka, K. Hiramatsu, and K. Kashino, "Sequence-to-sequence voice conversion with similarity metric learned using generative adversarial networks," in *Proc. Interspeech*, 2017, pp. 1283–1287.

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

[243] M.-Y. Liu and O. Tuzel, "Coupled generative adversarial networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 469–477.

[244] H. Zou, K. E. Ak, and A. A. Kassim, "Edge-GAN: Edge conditioned multi-view face image generation," in *Proc. IEEE Int. Conf. Image Process.*, 2020, pp. 2401–2405.

[245] S. Liu *et al.*, "Transferring source style in non-parallel voice conversion," 2020, *arXiv:2005.09178*.

[246] R. Kubichek, "Mel-cepstral distance measure for objective speech quality assessment," *Commun., Comput. Signal Process.*, vol. 1, pp. 125–128, 1993.

[247] J. Benesty, J. Chen, Y. Huang, and I. Cohen, "Pearson correlation coefficient," in *Noise Reduction Speech Processing*, Berlin, Germany: Springer-Verlag, 2009, pp. 1–4.

[248] T. Chai and R. R. Draxler, "Root mean square error (RMSE) or mean absolute error (MAE)?–Arguments against avoiding RMSE in the literature," *Geosci. Model Develop.*, vol. 7, no. 3, pp. 1247–1250, 2014.

[249] C. J. Willmott and K. Matsuura, "Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance," *Climate Res.*, vol. 30, no. 1, pp. 79–82, 2005.

[250] V. Grancharov and W. B. Kleijn, "Speech quality assessment," in *Springer Handbook Speech Process*, Berlin, Germany: Springer-Verlag, 2008, pp. 83–100.

[251] R. C. Streijl, S. Winkler, and D. S. Hands, "Mean opinion score (MOS) revisited: Methods and applications, limitations and alternatives," *Multimedia Syst.*, vol. 22, no. 2, pp. 213–227, 2016.

[252] M. Chu, H. Peng, and Y. Zhao, "Optimization of an objective measure for estimating mean opinion score of synthesized speech," US Patent 7,386,451, Jun. 10, 2008.

[253] M. Viswanathan and M. Viswanathan, "Measuring speech quality for text-to-speech systems: Development and assessment of a modified mean opinion score (MOS) scale," *Comput. Speech Lang.*, vol. 19, no. 1, pp. 55–83, 2005.

[254] A. Kain and M. W. Macon, "Design and evaluation of a voice conversion algorithm based on spectral envelope mapping and residual prediction," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2001, vol. 2, pp. 813–816.

[255] T. N. Flynn and A. A. J. Marley, "Best worst scaling: Theory and methods," in *Handbook Choice Modelling*, Northampton, Massachusetts, MA, USA: Edward Elgar Publishing, 2014, ch. 8, pp. 178–201.

[256] M. Zhang, B. Sisman, L. Zhao, and H. Li, "Deepconversion: Voice conversion with limited parallel training data," *Speech Commun.*, 2020.

[257] J. Lai, B. Chen, T. Tan, S. Tong, and K. Yu, "Phone-aware LSTM-RNN for voice conversion," in *Proc. IEEE 13th Int. Conf. Signal Process.*, 2016, pp. 177–182.

[258] A. W. Black *et al.*, "Articulatory features for expressive speech synthesis," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2012, pp. 4005–4008.

[259] B. Logan *et al.*, "Mel frequency cepstral coefficients for music modeling," in *Proc. Int. Symp. Music Inf. Retrieval*, 2000, pp. 1–11.

[260] 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. 577–586.

[261] W. Chu and A. Alwan, "Reducing F0 frame error of F0 tracking algorithms under noisy conditions with an unvoiced/voiced classification frontend," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2009, pp. 3969–3972.

[262] T. Nakatani, S. Amano, T. Irino, K. Ishizuka, and T. Kondo, "A method for fundamental frequency estimation and voicing decision: Application to infant utterances recorded in real acoustical environments," *Speech Commun.*, vol. 50, no. 3, pp. 203–214, 2008.

[263] R. Skerry-Ryan *et al.*, "Towards end-to-end prosody transfer for expressive speech synthesis with tacotron," 2018, *arXiv:1803.09047*.

[264] B. Sisman, G. Lee, H. Li, and K. C. Tan, "On the analysis and evaluation of prosody conversion techniques," in *Proc. IEEE Int. Conf. Asian Lang. Process.*, 2017, pp. 44–47.

[265] T. Watanabe, T. Murakami, M. Namba, T. Hoya, and Y. Ishida, "Transformation of spectral envelope for voice conversion based on radial basis function networks," in *Proc. 7th Int. Conf. Spoken Lang. Process.*, 2002.

[266] B. Ramani, M. A. Jeeva, P. Vijayalakshmi, and T. Nagarajan, "Cross-lingual voice conversion-based polyglot speech synthesizer for indian languages," in *Proc. 15th Annu. Conf. Int. Speech Commun. Assoc.*, 2014.

[267] O. Turk and L. M. Arslan, "Robust processing techniques for voice conversion," *Comput. Speech Lang.*, vol. 20, no. 4, pp. 441–467, 2006.

[268] S. Desai, A. W. Black, B. Yegnanarayana, and K. Prahallad, "Spectral mapping using artificial neural networks for voice conversion," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 5, pp. 954–964, Jul. 2010.

[269] M. Tamura, T. Masuko, K. Tokuda, and T. Kobayashi, "Speaker adaptation for HMM-based speech synthesis system using MLLR," in *Proc. 3rd ESCA/COCOSDA Workshop (ETRW) Speech Synth.*, 1998.

[270] V. Grancharov, D. Y. Zhao, J. Lindblom, and W. B. Kleijn, "Low-complexity, nonintrusive speech quality assessment," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 6, pp. 1948–1956, Nov. 2006.
[271] M. Wester, C. Valentini-Botinhao, and G. E. Henter, "Are we using enough listeners? No!–an empirically-supported critique of interspeech 2014 TTS evaluations," in *Proc. 16th Annu. Conf. Int. Speech Commun. Assoc.*, 2015.
[272] S. Zielinski, P. Hardisty, C. Hummersone, and F. Rumsey, "Potential biases in MUSHRA listening tests," in *Proc. 123rd Audio Eng. Soc. Convers.*, Audio Engineering Society, 2007.
[273] H. Benisty and D. Malah, "Voice conversion using GMM with enhanced global variance," in *Proc. 12th Annu. Conf. Int. Speech Commun. Assoc.*, 2011.
[274] J. Vít, Z. Hanzlíček, and J. Matoušek, "On the analysis of training data for wavenet-based speech synthesis," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 5684–5688.
[275] M. Zhang, J. Tao, J. Tian, and X. Wang, "Text-independent voice conversion based on state mapped codebook," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2008, pp. 4605–4608.
[276] H. Mizuno and M. Abe, "Voice conversion algorithm based on piecewise linear conversion rules of formant frequency and spectrum tilt," *Speech Commun.*, vol. 16, no. 2, pp. 153–164, 1995.
[277] I. Recommendation, "1534-1, Method for the subjective assessment of intermediate sound quality (MUSHRA)," *Int. Telecommun. Union*, Geneva, Switzerland, 2001.
[278] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2001, pp. 749–752.
[279] S.-W. Fu, Y. Tsao, H.-T. Hwang, and H.-M. Wang, "Quality-net: An end-to-end non-intrusive speech quality assessment model based on BLSTM," 2018, *arXiv:1808.05344*.
[280] T. Yoshimura *et al.*, "A hierarchical predictor of synthetic speech naturalness using neural networks," in *Proc. Interspeech*, 2016, pp. 342–346.
[281] B. Patton *et al.*, "AutoMOS: Learning a non-intrusive assessor of naturalness-of-speech," 2016, *arXiv:1611.09207*.
[282] M. Cernak and M. Rusko, "An evaluation of synthetic speech using the PESQ measure," in *Proc. Eur. Congr. Acoust.*, 2005, pp. 2725–2728.
[283] D.-Y. Huang, "Prediction of perceived sound quality of synthetic speech," in *Proc. Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf.*, 2011.
[284] U. Remes, R. Karhila, and M. Kurimo, "Objective evaluation measures for speaker-adaptive HMM-TTS systems," in *Proc. 8th ISCA Workshop Speech Synth.*, 2013.
[285] C.-C. Lo *et al.*, "MOSNET: Deep learning based objective assessment for voice conversion," 2019, *arXiv:1904.08352*.
[286] J. Williams, J. Rownicka, P. Oplustil, and S. King, "Comparison of speech representations for automatic quality estimation in multi-speaker text-to-speech synthesis," 2020, *arXiv:2002.12645*.
[287] M. Bińkowski *et al.*, "High fidelity speech synthesis with adversarial networks," 2019, *arXiv:1909.11646*.
[288] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local Nash equilibrium," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 6626–6637.
[289] M. Bińkowski, D. J. Sutherland, M. Arbel, and A. Gretton, "Demystifying MMD GANs," 2018, *arXiv:1801.01401*.
[290] J. Lorenzo-Trueba *et al.*, "The voice conversion challenge 2018: Promoting development of parallel and nonparallel methods," in *Proc. Odyssey, Speaker Lang. Recognit. Workshop*, 2018, pp. 195–202. [Online]. Available: http://dx.doi.org/10.21437/Odyssey.2018-28
[291] Z. Wu *et al.*, "Spoofing and countermeasures for speaker verification: A survey," *Speech Commun.*, vol. 66, pp. 130–153, 2015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167639314000788
[292] M. Wester, Z. Wu, and J. Yamagishi, "Analysis of the voice conversion challenge 2016 evaluation results," in *Proc. Interspeech*, 2016, pp. 1637–1641.
[293] Y. Wu, P. L. Tobing, T. Hayashi, K. Kobayashi, and T. Toda, "The NU non-parallel voice conversion system for the voice conversion challenge 2018," in *Proc. Odyssey, Speaker Lang. Recognit. Workshop*, 2018, pp. 211–218. [Online]. Available: http://dx.doi.org/10.21437/Odyssey.2018-30
[294] L.-J. Liu, Z.-H. Ling, Y. Jiang, M. Zhou, and L.-R. Dai, "Wavenet vocoder with limited training data for voice conversion," in *Proc. Interspeech*, 2018, pp. 1983–1987. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2018-1190
[295] J. Zhang, Z. Ling, L. Liu, Y. Jiang, and L. Dai, "Sequence-to-sequence acoustic modeling for voice conversion," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 3, pp. 631–644, Mar. 2019.
[296] J. Zhang, Z. Ling, and L. Dai, "Non-parallel sequence-to-sequence voice conversion with disentangled linguistic and speaker representations," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 540–552, 2020.
[297] Y. Zhao *et al.*, "Voice conversion challenge 2020: Intra-lingual semi-parallel and cross-lingual voice conversion," 2020, *arXiv:2008.12527*.
[298] Z. Wu *et al.*, "ASVspoof 2015: The first automatic speaker verification spoofing and countermeasures challenge," in *Proc. Interspeech*, 2015, pp. 2037–2041.
[299] Z. Wu *et al.*, "ASVspoof: The automatic speaker verification spoofing and countermeasures challenge," *IEEE J. Sel. Top. Signal Process.*, vol. 11, no. 4, pp. 588–604, Jun. 2017.
[300] T. Kinnunen *et al.*, "The ASVspoof 2017 challenge: Assessing the limits of replay spoofing attack detection," in *Proc. Interspeech*, 2017, pp. 2–6.
[301] M. Todisco *et al.*, "ASVspoof 2019: Future horizons in spoofed and fake audio detection," in *Proc. Interspeech*, 2019, pp. 1008–1012. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2019-2249
[302] X. Wang *et al.*, "ASVspoof 2019: A large-scale public database of synthetic, converted and replayed speech," 2019.
[303] J. Kominek and A. W. Black, "The CMU arctic speech databases," in *Proc. 5th ISCA Workshop Speech Synth.*, 2004.
[304] C. Veaux *et al.*, "CSTR VCTK corpus: English multi-speaker corpus for cstr voice cloning toolkit," 2016.
[305] H. Zen *et al.*, "LibriTTS: A corpus derived from LibriSpeech for text-to-speech," in *Proc. Interspeech*, 2019, pp. 1526–1530. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2019-2441
[306] A. Nagrani, J. S. Chung, W. Xie, and A. Zisserman, "VoxCeleb: Large-scale speaker verification in the wild," *Comput. Speech Lang.*, vol. 60, 2020, Art. no. 101027. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0885230819302712
[307] K. Kobayashi and T. Toda, "sprocket: Open-source voice conversion software," in *Proc. Odyssey, Speaker Lang. Recognit. Workshop*, 2018, pp. 203–210. [Online]. Available: http://dx.doi.org/10.21437/Odyssey.2018-29
[308] S. Watanabe *et al.*, "ESPnet: End-to-end speech processing toolkit," in *Proc. Interspeech*, 2018, pp. 2207–2211. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2018-1456

Berrak Sisman (Member, IEEE) received the Ph.D. degree from the National University of Singapore in 2020, fully funded by A*STAR Graduate Academy. She is currently an Assistant Professor with the Singapore University of Technology and Design (SUTD). She is also an Affiliated Researcher and Team Leader with the National University of Singapore (NUS). She was an exchange Ph.D. Student with the University of Edinburgh and a Visiting Scholar with The Centre for Speech Technology Research, University of Edinburgh in 2019. She was attached to RIKEN Advanced Intelligence Project, Japan in 2018. She has authored coauthored in leading journals and conferences, including IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, ASRU, INTERSPEECH and ICASSP. Her research interests include speech information processing, machine learning, speech synthesis and voice conversion.

Junichi Yamagishi (Senior Member, IEEE) received the Ph.D. degree from the Tokyo Institute of Technology (Tokyo Tech), Tokyo, Japan, in 2006. He is currently a Professor with the National Institute of Informatics, Tokyo, Japan, and also a Senior Research Fellow with The Centre for Speech Technology Research, The University of Edinburgh, Edinburgh, U.K. Since 2006, he has authored or coauthored more than 250 refereed papers in international journals and conferences. Prof. Yamagishi was the recipient of the Tejima Prize as the best Ph.D. thesis of Tokyo Tech in 2007. He was the recipient of the Itakura Prize from the Acoustic Society of Japan in 2010, the Kiyasu Special Industrial Achievement Award from the Information Processing Society of Japan in 2013, the Young Scientists' Prize from the Minister of Education, Science and Technology in 2014, the JSPS Prize from the Japan Society for the Promotion of Science in 2016, and the 17th DOCOMO Mobile Science Award from the Mobile Communication Fund, Japan in 2018. He was one of the organizers for special sessions on Spoofing and Countermeasures for the Automatic Speaker Verification at INTERSPEECH 2013, the 1st/2nd/3rd ASVspoof Evaluation, the Voice Conversion Challenge 2016/2018/2020, and the VoicePrivacy Challenge 2020. He was an Associate Editor for the IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, a Lead Guest Editor of the IEEE JOURNAL OF SELECTED TOPICS IN SIGNAL PROCESSING Special Issue on Spoofing and Countermeasures for Automatic Speaker Verification, and a member of the Technical Committee of the IEEE Signal Processing Society Speech and Language. He is currently the Chairperson of ISCA Special Interest Group: Speech Synthesis (SynSig), a member of the Technical Committee for the Asia-Pacific Signal and Information Processing Association Multimedia Security and Forensics, an IEEE Senior Area Editor of the IEEE/ACM TRANSACTION ON AUDIO, SPEECH, AND LANGUAGE PROCESSING.

Simon King (Fellow, IEEE) received the M.A. (Cantab) and M.Phil. degrees from the University of Cambridge, Cambridge, U.K., and the Ph.D. degree from University of Edinburgh, Edinburgh, U.K. He has been with the Centre for Speech Technology Research, University of Edinburgh, since 1993, where he is currently Professor of Speech Processing and the Director of the Centre. His research interests include speech synthesis, recognition and signal processing and he has around 230 publications across these areas. Prof. King has served on the ISCA SynSIG Board and currently co-organises the Blizzard Challenge. He has previously served on the IEEE SLTC and as an Associate Editor for the IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, and is currently an Associate Editor for *Computer Speech and Language*.

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 Professor with the Department of Electrical and Computer Engineering, National University of Singapore (NUS). Prior to joining NUS, he taught in the University of Hong Kong (1988–1990) and the South China University of Technology (1990–1994). He was a Visiting Professor with CRIN in France (1994–1995), Research Manager with the Apple-ISS Research Centre (1996–1998), Research Director in Lernout & Hauspie Asia Pacific (1999–2001), Vice President in InfoTalk Corp. Ltd. (2001–2003), and the Principal Scientist and Department Head of Human Language Technology with the Institute for Infocomm Research, Singapore (2003–2016). His research interests include automatic speech recognition, speaker and language recognition, and natural language processing. Dr Li was the Editor-in-Chief of IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING (2015–2018), a member of the Editorial Board of *Computer Speech and Language* (2012–2018), an elected member of IEEE Speech and Language Processing Technical Committee (2013–2015), the President of the International Speech Communication Association (2015–2017), the President of Asia Pacific Signal and Information Processing Association (2015–2016), and the President of Asian Federation of Natural Language Processing (2017–2018). He was the General Chair of ACL 2012, INTERSPEECH 2014, and ASRU 2019. He is a Fellow of the ISCA. He was the recipient of the National Infocomm Award 2002 and the President's Technology Award 2013 in Singapore. He was named one of the two Nokia Visiting Professors in 2009 by the Nokia Foundation, and U Bremen Excellence Chair Professor in 2019.