



Article

# Wav2wav: Wave-to-Wave Voice Conversion

Changhyeon Jeong <sup>1</sup>, Hyung-pil Chang <sup>2</sup>, In-Chul Yoo <sup>2</sup> and Dongsuk Yook <sup>2,\*</sup>

<sup>1</sup> Lotte Innovate Co., Ltd., Seoul 08500, Republic of Korea; changhyeon.jeong@lotte.net

<sup>2</sup> Artificial Intelligence Laboratory, Department of Computer Science and Engineering, Korea University, Seoul 02841, Republic of Korea; hpchang@korea.ac.kr (H.-p.C.); icyoo@ai.korea.ac.kr (I.-C.Y.)

\* Correspondence: yook@korea.ac.kr

**Abstract:** Voice conversion is the task of changing the speaker characteristics of input speech while preserving its linguistic content. It can be used in various areas, such as entertainment, medicine, and education. The quality of the converted speech is crucial for voice conversion algorithms to be useful in these various applications. Deep learning-based voice conversion algorithms, which have been showing promising results recently, generally consist of three modules: a feature extractor, feature converter, and vocoder. The feature extractor accepts the waveform as the input and extracts speech feature vectors for further processing. These speech feature vectors are later synthesized back into waveforms by the vocoder. The feature converter module performs the actual voice conversion; therefore, many previous studies separately focused on improving this module. These works combined the separately trained vocoder to synthesize the final waveform. Since the feature converter and the vocoder are trained independently, the output of the converter may not be compatible with the input of the vocoder, which causes performance degradation. Furthermore, most voice conversion algorithms utilize mel-spectrogram-based speech feature vectors without modification. These feature vectors have performed well in a variety of speech-processing areas but could be further optimized for voice conversion tasks. To address these problems, we propose a novel wave-to-wave (wav2wav) voice conversion method that integrates the feature extractor, the feature converter, and the vocoder into a single module and trains the system in an end-to-end manner. We evaluated the efficiency of the proposed method using the VCC2018 dataset.

**Keywords:** end-to-end; generative adversarial network; vocoder; voice conversion



**Citation:** Jeong, C.; Chang, H.-p.; Yoo, I.-C.; Yook, D. Wav2wav: Wave-to-Wave Voice Conversion. *Appl. Sci.* **2024**, *14*, 4251. <https://doi.org/10.3390/app14104251>

Academic Editor: Douglas O'Shaughnessy

Received: 20 April 2024

Revised: 13 May 2024

Accepted: 14 May 2024

Published: 17 May 2024

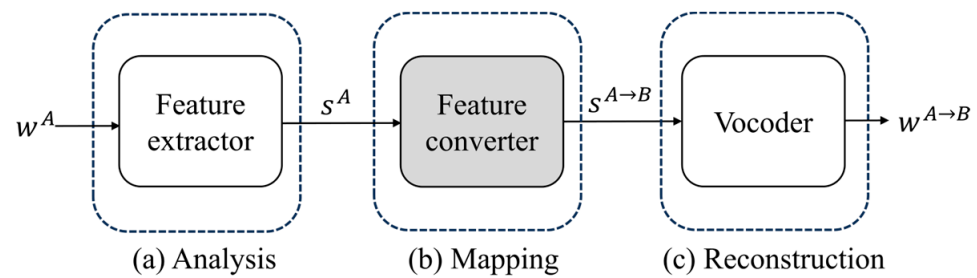


**Copyright:** © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

Voice conversion aims to convert the speaker-specific characteristics of input speech into the target speaker's characteristics while preserving the linguistic content of the input speech. The application areas of voice conversion include entertainment [1], education [2], and medicine [3] domains. With advances in deep learning, many studies on voice conversion using deep neural network (DNN) models have been attempted. In particular, generative models, such as variational autoencoders (VAEs) [4] and generative adversarial networks (GANs) [5], have shown promising results in many voice conversion tasks [6–11]. Recently proposed flow-based voice conversion [12] and diffusion model-based voice conversion [13] also showed good performance.

Most of these works on voice conversion generally consist of three steps: analysis, mapping, and reconstruction [14]. In the analysis step, feature vectors are extracted, which are easy to process while retaining relevant information from the input waveform. The mapping step is the actual voice conversion process that changes only the identity of the speaker while preserving the linguistic content of the input feature vectors. Finally, the reconstruction step uses a vocoder to synthesize a waveform using the converted feature vectors from the previous step. Figure 1 illustrates this process.



**Figure 1.** Analysis–mapping–reconstruction steps of a conventional voice conversion system.  $w^A$  and  $s^A$  represent the waveform and the feature vectors of speaker  $A$ , respectively. Similarly,  $w^{A \rightarrow B}$  and  $s^{A \rightarrow B}$  represent the waveform and the feature vectors of the converted speech from speaker  $A$  to speaker  $B$ .

The analysis steps of the conventional approaches use fixed feature extraction methods, such as Fourier transform and mel-scale filtering. Traditionally, these feature vectors have been widely adopted in various speech-processing areas, such as speech recognition, speaker identification, and text-to-speech. However, it could be further optimized [15,16] for voice conversion tasks, as the importance of each frequency can vary across conversion pairs. Additionally, most of the existing approaches focus on improving the feature converter model itself, as it performs the actual voice conversion task and provides its output to a pre-trained vocoder. This approach may lead to poor performance because the vocoder model is not explicitly trained to process the output feature vectors from the feature converter model. Therefore, some unnatural speech characteristics that arise from the feature converter model are not sufficiently compensated for by the vocoder model, resulting in poor speech quality. Recent studies [17–22] that proposed to jointly train feature converter models and vocoder models showed promising results. In this study, we extend the work in [22] and propose an architecture that integrates all three stages of analysis–mapping–reconstruction to optimize the quality of converted speech. Table 1 summarizes the existing voice conversion methods.

**Table 1.** Summary of the conventional voice conversion methods.

Authors	Year	Refs.	Dataset	Method	Model
P. L. Tobing et al.	2019	[6]	VCC2018	VAE with cycle loss	RNNT
D. Yook et al.	2020	[7]	VCC2018	Ref. [6] with multiple decoders	CNN
T. Kaneko et al.	2018	[8]	VCC2016	GAN with cycle loss	CNN
T. Kaneko et al.	2019	[9]	VCC2018	Improved [8] by modifying loss and models	CNN
T. Kaneko et al.	2020	[10]	VCC2018	Improved [9] by applying TFAN norm	CNN
T. Kaneko et al.	2021	[11]	VCC2018	Improved [9] by masking input feature	CNN
M. Proszewska et al.	2022	[12]	In-house	Flow-based model	LSTM
K. Kobayashi et al.	2016	[17]	VCC2016	Direct waveform modification	Diff VC
Y. Kurita et al.	2019	[18]	In-house	Applied [17] to a singing voice conversion task	Diff VC
K. Kobayashi et al.	2018	[19]	VCC2018	Open-source implementation of [17]	Diff VC
J.-W. Kim et al.	2020	[20]	TIDIGITS	Translation-based method using a transformer	Transformer
B. Nguyen et al.	2022	[21]	VCTK	Content and speaker disentanglement	CNN

The proposed method enables wave-to-wave (wav2wav) voice conversion by jointly training the entire analysis–mapping–reconstruction steps in an end-to-end manner. This allows the model to flexibly adjust model parameters to achieve high-quality speech output. However, training from scratch consumes too much training time and data, as it is difficult to find optimal feature representations. Furthermore, training the GAN-based models used in the mapping and reconstruction steps is known to be inherently difficult. Our proposed method overcomes these difficulties in two ways. The first is that the analysis step uses two-layer convolutional networks (CNNs) initialized with a discrete Fourier transform (DFT) matrix and mel-filterbank coefficients. This allows the analysis step to

start from the traditional mel-spectrogram extraction method and gradually optimize the parameters, thereby achieving efficient learning and flexible parameter adjustment. Second, the difficulties of training the GAN models in the proposed wav2wav method are addressed using pre-training and two-phase training techniques. This allowed us to train the proposed wav2wav model reliably on small datasets, such as VCC2018 [23]. To verify the performance of the proposed method, objective and subjective evaluations were performed using the mel-cepstral distance (MCD) metric and the mean opinion score (MOS) metric, respectively.

The contributions of this paper are as follows:

- We propose a novel wave-to-wave voice conversion architecture that jointly trains analysis–mapping–reconstruction modules for high-quality voice conversion.
- We provide an efficient training algorithm so that the proposed GAN-based integrated model can be reliably trained with very small amounts of training data, such as VCC2018.
- The supervised learning process of standalone vocoders were modified to accommodate unsupervised learning in the end-to-end learning of the integrated model.
- We demonstrate the usefulness of the proposed method using both objective and subjective measures.

The rest of the paper is organized as follows. In Section 2, relevant previous studies are reviewed. Section 3 describes the proposed wav2wav model. Section 4 analyzes the experimental results. Finally, Section 5 concludes the paper.

## 2. Related Works

In this section, we describe the CycleGAN-based voice conversion and a HiFi-GAN-based vocoder. CycleGAN-based algorithms are used as the mapping stage in many voice conversion systems due to the excellent quality of the converted speech. Recently, HiFi-GAN vocoders have been widely adopted because of the high quality of the synthesized speech and stability in training, while many other deep neural network-based vocoders exhibit high quality but are difficult to train.

### 2.1. CycleGAN-Based Voice Conversion

One way of training voice conversion systems requires the use of parallel training data consisting of identical transcription utterances from different speakers. However, collecting such data is expensive. The CycleGAN-based voice conversion algorithm uses two pairs of generator and discriminator for training without such parallel data. One pair converts the spectral feature from a source speaker to that of a target speaker and the other pair converts vice versa. Using both generators, it is possible to convert the spectral feature from the source speaker into that of the target speaker and then to convert it back to that of the original source speaker. Two constraints are given in the training process so that the speaker identity of the spectral feature is changed to the target speaker while the linguistic content of the utterance is preserved. The first constraint is that the spectral features before and after the cyclic conversion must have the same linguistic content, which is enforced by a cycle consistency loss [24]. The second constraint is that the speaker identity of the resulting spectral features from each generator must be indistinguishable from that of the target speaker, which is enforced by an adversarial loss from the discriminators [5].

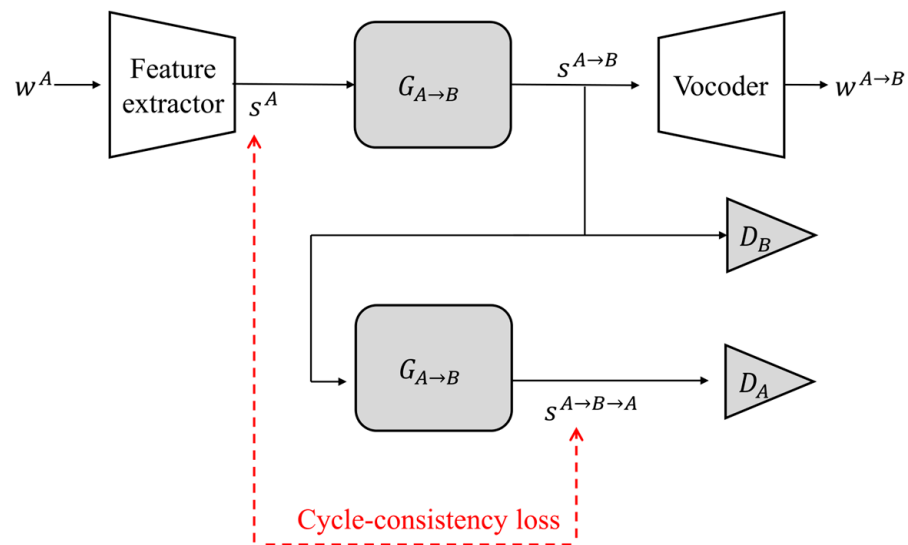
Figure 2 illustrates the training process of the CycleGAN-based mapping model. Source speaker  $A$ 's waveform  $w^A$  is converted to spectral feature  $s^A$  using spectral analysis, and the spectral feature is used as input into the generator model:

$$s^A = F(w^A), \quad (1)$$

where  $F$  represents the spectral analysis model. The resulting spectral feature  $s^A$  is converted into  $s^{A \rightarrow B}$  through the generator model where the identity of the speaker is changed from speaker  $A$  to speaker  $B$ :

$$s^{A \rightarrow B} = G_{A \rightarrow B}(s^A), \quad (2)$$

where  $G_{A \rightarrow B}$  represents the generator model that converts the speaker identity of the input spectral feature from speaker  $A$  to speaker  $B$ . The spectral feature  $s^A$  of the source speaker  $A$  is converted to  $s^{A \rightarrow B}$ ; then, it is converted back to  $s^{A \rightarrow B \rightarrow A}$ . The cycle consistency loss is measured between  $s^A$  and  $s^{A \rightarrow B \rightarrow A}$  to ensure the linguistic content preservation. Similarly, spectral feature  $s^B$  from the target speaker  $B$  is converted to  $s^{B \rightarrow A}$ ; then, it is converted back again to  $s^{B \rightarrow A \rightarrow B}$ , which provides another cycle consistency loss. The discriminators  $D_A$  and  $D_B$  provide the adversarial losses of  $s^{B \rightarrow A}$  and  $s^{A \rightarrow B}$ , respectively.



**Figure 2.** Overview of the CycleGAN-based voice conversion model. Greyed objects contain trainable parameters. The feature extractor and vocoder are not trained in this approach.

The recently proposed MaskCycleGAN-VC [11], which is a variant of CycleGAN-based voice conversion, achieved a state-of-the-art performance by using masked spectral features. The generator learns to restore the masked spectral parts, which improves the robustness of the generator. However, it should be noted that MaskCycleGAN-VC is a standalone mapping method that is not jointly optimized with the analysis stage nor with the reconstruction stage.

## 2.2. HiFi-GAN Vocoder

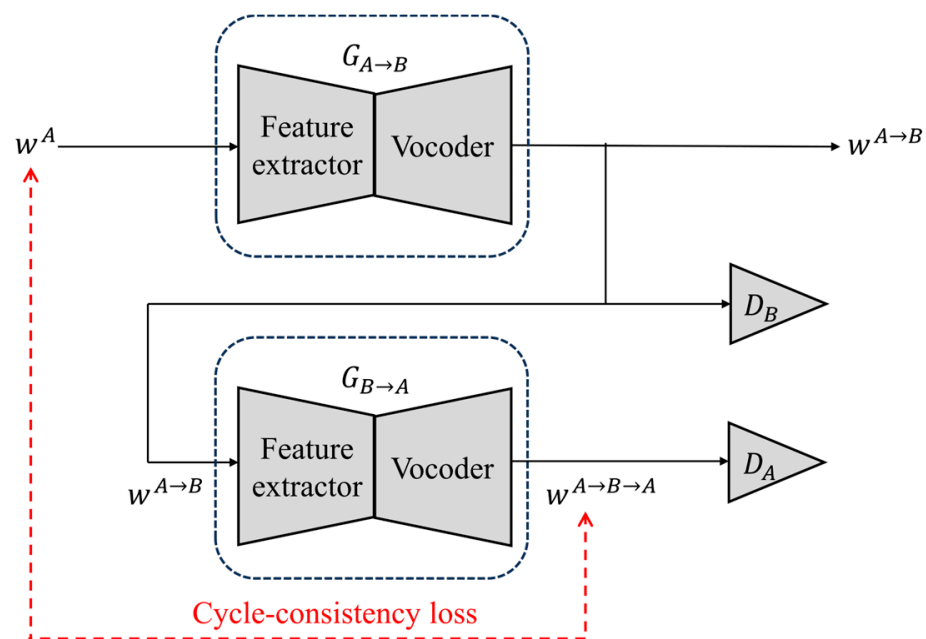
A vocoder reconstructs waveforms by synthesizing them from input spectral features. Rule-based vocoders, such as Griffin-Lim [25] and WORLD [26] vocoders, were widely used because of their simplicity, but the quality of the resulting speech was not satisfactory. Recently, many deep learning-based neural vocoders with excellent sound quality have been proposed. One such neural vocoder, WaveNet [27], produces high-quality speech signals, but is slow to train and inference due to its autoregressive structure. Many works attempted to minimize the computational overhead of WaveNet while maintaining high-quality outputs [28–32].

The recently proposed MelGAN [33] and HiFi-GAN [34] are GAN-based vocoders that can obtain high-quality speech at a much faster speed due to their non-autoregressive characteristics. HiFi-GAN uses two types of discriminators: multi-scale discriminator (MSD) [33] and multi-period discriminator (MPD) [34]. These two discriminators enable HiFi-GAN to produce state-of-the-art performance when generating high quality speech signals. However, it should be noted that the spectral feature conversion model and the

vocoder model are trained independently in the conventional voice conversion approaches, which means that the vocoder is not trained to handle any residual characteristics of the source speakers that may remain in the outputs of the spectral conversion model, leading to degradation of the output waveform for voice conversion.

### 3. Proposed Method

As discussed in the previous section, traditional voice conversion approaches train the spectral feature conversion models and vocoder models separately and use a fixed feature extraction scheme, leading to the degraded quality of converted speech. The proposed method, depicted in Figure 3, adopts a wave-to-wave approach, which uses waveforms directly as the input and output of the system and jointly trains the analysis, mapping, and reconstruction modules in an end-to-end manner, enabling more flexible feature analysis and natural speech generation.



**Figure 3.** Overview of the proposed voice conversion model. Greyed objects contain trainable parameters. The generators include both a feature extractor and vocoder, which can be jointly trained in an end-to-end manner.

However, training the feature extraction module from scratch in this end-to-end learning may require too much data and training time. Since the mel-spectral features are widely used in various speech-processing tasks, we decided to improve their performance further by learning detailed parameters suitable for voice conversion based on them. We first reproduced the mel-spectral feature extraction process with a two-layer CNN initialized with a DFT matrix and mel-filterbank coefficients; then, each weight was adjusted during training. Figure 4 shows the feature extraction network. The output of the first convolution layer, which initially consisted of real and imaginary components extracted using conventional feature extraction methods, was converted to magnitude vectors via square, summation, and square root operations. The resulting magnitude vectors were fed into the second convolutional layer, followed by rectified linear unit (ReLU) and logarithm operations. The CNN parameters were shared between the two generators and updated during training.

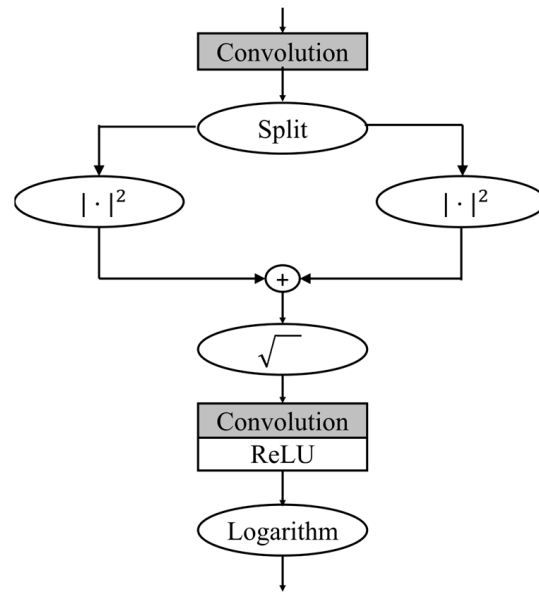


Figure 4. Architecture of the proposed feature extractor model.

By including the feature-extracting CNN and vocoder into a generator, the proposed voice conversion model enables waveform inputs and waveform outputs, resulting in wave-to-wave voice conversion, shown as follows:

$$w^{A \rightarrow B} = G_{A \rightarrow B}(w^A), \tag{3}$$

where  $G_{A \rightarrow B}$  is the generator model that converts the speaker identity of the input speech from speaker  $A$  to speaker  $B$ , and  $w^{A \rightarrow B}$  is the converted waveform from speaker  $A$ 's waveform  $w^A$  to match speaker  $B$ 's voice characteristics.

The loss function of the proposed method consists of adversarial loss [5], cycle consistency loss [35], identity-mapping loss [36], and feature-matching loss [34], similar to other CycleGAN-based methods, as shown below:

$$\mathcal{L}_{adv1}(G_{A \rightarrow B}, D_B) = \mathbb{E}_{w^B} [\log D_B(F(w^B))] + \mathbb{E}_{w^A} [\log(1 - D_B(F(G_{A \rightarrow B}(w^A))))], \tag{4}$$

$$\mathcal{L}_{adv2}(G_{A \rightarrow B}, D_B) = \mathbb{E}_{w^B} [\log D_B(F(w^B))] + \mathbb{E}_{w^{B \rightarrow A}} [\log(1 - D_B(F(G_{A \rightarrow B}(w^{B \rightarrow A}))))], \tag{5}$$

$$\mathcal{L}_{cyc}(G_{A \rightarrow B}) = \mathbb{E}_{w^B} [\|F(G_{A \rightarrow B}(G_{B \rightarrow A}(w^B))) - F(w^B)\|_1], \tag{6}$$

$$\mathcal{L}_{id}(G_{A \rightarrow B}) = \mathbb{E}_{w^B} [\|F(G_{A \rightarrow B}(w^B)) - F(w^B)\|_1], \tag{7}$$

$$\mathcal{L}_{fm}(G_{A \rightarrow B}) = \mathbb{E}_{w^B} \left[ \sum_i \|D_B^i(F(G_{A \rightarrow B}(G_{B \rightarrow A}(w^B)))) - D_B^i(F(w^B))\|_1 \right], \tag{8}$$

where  $\mathbb{E}$  stands for the expectation operation,  $G_{B \rightarrow A}$  is the generator model that converts the speaker identity of the input speech from speaker  $B$  to speaker  $A$ , and  $D_B^i$  represents the  $i$ -th layer of the discriminator. Since the output of the generator model is waveforms, they need to be converted to spectral features to calculate the losses. The losses for the other conversion direction, i.e.,  $\mathcal{L}_{adv1}(G_{B \rightarrow A}, D_A)$ ,  $\mathcal{L}_{adv2}(G_{B \rightarrow A}, D_A)$ ,  $\mathcal{L}_{cyc}(G_{B \rightarrow A})$ ,  $\mathcal{L}_{id}(G_{B \rightarrow A})$ , and  $\mathcal{L}_{fm}(G_{B \rightarrow A})$ , are similarly defined.

In the conventional analysis–mapping–reconstruction approaches, such as in Figure 1, the discriminator of the spectral feature conversion model and the vocoder model serve different purposes. The discriminator in the spectral feature conversion model aims to distinguish the speaker's identity from the target speaker's spectral feature and the converted

spectral feature. On the other hand, the discriminator in the vocoder model aims to distinguish the original and synthesized waveforms. Since the generator in the proposed method includes the function of a vocoder, the loss in the proposed method is calculated by using the discriminator of the spectral feature conversion model and the vocoder model together. Through this, the proposed voice conversion model can generate waveforms similar to those from the target speaker while maintaining the quality of the original waveforms. We used a modified version of HiFi-GAN as the generator of the proposed method. Therefore, the discriminators of the proposed method utilize the HiFi-GAN discriminator, as well as the conventional GAN discriminator.

One important limitation of GAN-based models is that they are difficult to train, especially for complex models like the proposed one, where the HiFi-GAN-style vocoder is used as the generator in the CycleGAN-style training. Since it is not easy to stabilize the training process and to avoid the collapsing problem, we applied two techniques to alleviate this problem. The first technique was to pre-train the generators using the speech data of the source and target speakers in the training data. Since the supervised training of the generator was relatively easy to converge, we first trained them to bootstrap the training in the CycleGAN-style training process. The second technique was to use a phase-wise model update method. The generator and discriminator models were trained in two alternating phases. The first phase used the following loss function:

$$\mathcal{L}_{\text{phase1}} = \mathcal{L}_{\text{adv1}}(G_{A \rightarrow B}, D_B) + \mathcal{L}_{\text{adv1}}(G_{B \rightarrow A}, D_A) + \lambda_{\text{id}} (\mathcal{L}_{\text{id}}(G_{A \rightarrow B}) + \mathcal{L}_{\text{id}}(G_{B \rightarrow A})), \quad (9)$$

where  $\lambda_{\text{id}}$  is the weight for the identity-mapping loss. The second phase utilized the following loss function:

$$\begin{aligned} \mathcal{L}_{\text{phase2}} = & \mathcal{L}_{\text{adv2}}(G_{A \rightarrow B}, D_B) + \mathcal{L}_{\text{adv2}}(G_{A \rightarrow B}, \text{MSD}_B) + \mathcal{L}_{\text{adv2}}(G_{A \rightarrow B}, \text{MPD}_B) \\ & + \mathcal{L}_{\text{adv2}}(G_{B \rightarrow A}, D_A) + \mathcal{L}_{\text{adv2}}(G_{B \rightarrow A}, \text{MSD}_A) + \mathcal{L}_{\text{adv2}}(G_{B \rightarrow A}, \text{MPD}_A) \\ & + \lambda_{\text{cyc}} (\mathcal{L}_{\text{cyc}}(G_{A \rightarrow B}) + \mathcal{L}_{\text{cyc}}(G_{B \rightarrow A})) \\ & + \lambda_{\text{fm}} (\mathcal{L}_{\text{fm}}(G_{A \rightarrow B}) + \mathcal{L}_{\text{fm}}(G_{B \rightarrow A})) \end{aligned} \quad (10)$$

where  $\lambda_{\text{cyc}}$  and  $\lambda_{\text{fm}}$  are the weights for the cycle consistency loss and feature-matching loss, respectively;  $\text{MSD}_A$ ,  $\text{MPD}_A$ ,  $\text{MSD}_B$ , and  $\text{MPD}_B$  are the MSD and MPD of speaker  $A$  and speaker  $B$ , respectively.

Algorithm 1 summarizes the alternating phase-wise training process of the proposed method. The parameters of the two generators are initialized with the pre-trained vocoder using the speech data from the source and target speakers. The parameters of the feature-extracting CNN are also initialized with the coefficients of the DFT matrix and mel-filterbanks. In the first training phase, speaker  $A$ 's waveform was fed into the generator  $G_{A \rightarrow B}$  to generate speaker  $B$ 's speech. Using this converted waveform  $w^{A \rightarrow B}$ , the generator  $G_{A \rightarrow B}$  and discriminator  $D_B$  were trained. Also, the generator  $G_{B \rightarrow A}$  and discriminator  $D_A$  were trained in a similar fashion by swapping the source and target speakers.

In the second phase, the converted waveform  $w^{B \rightarrow A}$  from the first phase was fed into the generator  $G_{A \rightarrow B}$  to generate waveform  $w^{B \rightarrow A \rightarrow B}$ . Then, the generator  $G_{A \rightarrow B}$ , discriminator  $D_B$ ,  $\text{MSD}_B$ , and  $\text{MPD}_B$  were trained. Similarly, the generator  $G_{B \rightarrow A}$ , discriminator  $D_A$ ,  $\text{MSD}_A$ , and  $\text{MPD}_A$  were trained by swapping the source and target speakers. These two phases were repeated until the model converged. It should be noted that the first phase was completely unsupervised learning using non-parallel data, that is,  $w^A$  and  $w^B$  were not the same transcription utterances. On the other hand, since the second phase required parallel data to compute the feature-matching loss, the pseudo-parallel data, that is,  $w^A$  and  $w^{A \rightarrow B}$  (as well as  $w^B$  and  $w^{B \rightarrow A}$ ), which were the same transcription utterances from different speakers, were utilized.

**Algorithm 1** wav2wav

---

```

1 Initialization: load pretrained parameters.
2 repeat
3   Select  $w^A$  and  $w^B$  from the training data randomly.
   /* Phase 1 */
4    $w^{A \rightarrow B} \leftarrow G_{A \rightarrow B}(w^A)$ 
5    $w^{B \rightarrow A} \leftarrow G_{B \rightarrow A}(w^B)$ 
6   Compute  $\nabla \mathcal{L}_{\text{phase1}}$  using  $w^A, w^{B \rightarrow A}, w^B,$  and  $w^{A \rightarrow B}$ .
7    $G_{A \rightarrow B} \leftarrow G_{A \rightarrow B} - \eta \nabla \mathcal{L}_{\text{phase1}}$ 
8    $D_B \leftarrow D_B + \eta \nabla \mathcal{L}_{\text{phase1}}$ 
9    $G_{B \rightarrow A} \leftarrow G_{B \rightarrow A} - \eta \nabla \mathcal{L}_{\text{phase1}}$ 
10   $D_A \leftarrow D_A + \eta \nabla \mathcal{L}_{\text{phase1}}$ 
   /* Phase 2 */
11   $w^{B \rightarrow A \rightarrow B} \leftarrow G_{A \rightarrow B}(w^{B \rightarrow A})$ 
12   $w^{A \rightarrow B \rightarrow A} \leftarrow G_{B \rightarrow A}(w^{A \rightarrow B})$ 
13  Compute  $\nabla \mathcal{L}_{\text{phase2}}$  using  $w^A, w^{A \rightarrow B}, w^{A \rightarrow B \rightarrow A}, w^B, w^{B \rightarrow A},$  and  $w^{B \rightarrow A \rightarrow B}$ .
14   $G_{A \rightarrow B} \leftarrow G_{A \rightarrow B} - \eta \nabla \mathcal{L}_{\text{phase2}}$ 
15   $D_B \leftarrow D_B + \eta \nabla \mathcal{L}_{\text{phase2}}$ 
16   $MSD_B \leftarrow MSD_B + \eta \nabla \mathcal{L}_{\text{phase2}}$ 
17   $MPD_B \leftarrow MPD_B + \eta \nabla \mathcal{L}_{\text{phase2}}$ 
18   $G_{B \rightarrow A} \leftarrow G_{B \rightarrow A} - \eta \nabla \mathcal{L}_{\text{phase2}}$ 
19   $D_A \leftarrow D_A + \eta \nabla \mathcal{L}_{\text{phase2}}$ 
20   $MSD_A \leftarrow MSD_A + \eta \nabla \mathcal{L}_{\text{phase2}}$ 
21   $MPD_A \leftarrow MPD_A + \eta \nabla \mathcal{L}_{\text{phase2}}$ 
22 until convergence

```

---

**4. Experiments**

In the experiments, we used a subset of the VCC2018 [35] dataset used in many existing voice conversion studies to show that the proposed method can be effective, even with a small amount of data. Two female speakers (SF3, TF1) and two male speakers (SM3, TM1) were used as both source speakers and target speakers. We denote these speakers as F1, F2, M1, and M2, respectively. We used 81 training sentences and 35 test sentences for each speaker. A preliminary experiment was conducted to decide the optimal values of the hyperparameters  $\lambda_{\text{id}}$ ,  $\lambda_{\text{cyc}}$ , and  $\lambda_{\text{fm}}$ , which were found to be 30, 45, and 0.5, respectively. We used MaskCycleGAN [11] as a baseline for comparison since it was the state-of-the-art among the CycleGAN-based voice conversion methods. For fair comparison, we trained a MelGAN to synthesize the waveforms from the output of MaskCycleGAN using the two speakers, i.e., the source and target speakers, for each conversion direction.

**4.1. Objective Evaluation**

We used MCD [37] as a measure for the objective evaluation. MCD measures the distance between a pair of spectral features. MCD is calculated as follows:

$$\text{MCD} = \frac{10}{\ln 10} \sqrt{2 \sum_{i=1}^N (m_{k,i}^t - m_{k,i}^c)^2}, \quad (11)$$

where  $N$ ,  $m^t$ ,  $m^c$ , and  $k$  denote the mel-cepstral coefficient (MCC) dimension, target speech MCC, converted speech MCC, and the frame index, respectively. We used a 16-dimensional MCC. Since the converted spectral features may have different lengths from the target, the dynamic time warping (DTW) algorithm was applied to compensate for the length difference. A lower MCD value between the converted speech and the target speech indicates a better voice conversion performance.

Table 2 summarizes the average MCDs and 95% confidence intervals of the converted speeches by MaskCycleGAN and the proposed wav2wav. The proposed method outperformed MaskCycleGAN in both intra-gender and inter-gender cases, with an average MCD



of 6.06 and a 95% confidence interval of  $\pm 0.08$ , demonstrating superior performance and more stable learning than the baseline method.

**Table 2.** Average MCDs and 95% confidence intervals of the speeches converted by MaskCycleGAN and wav2wav.

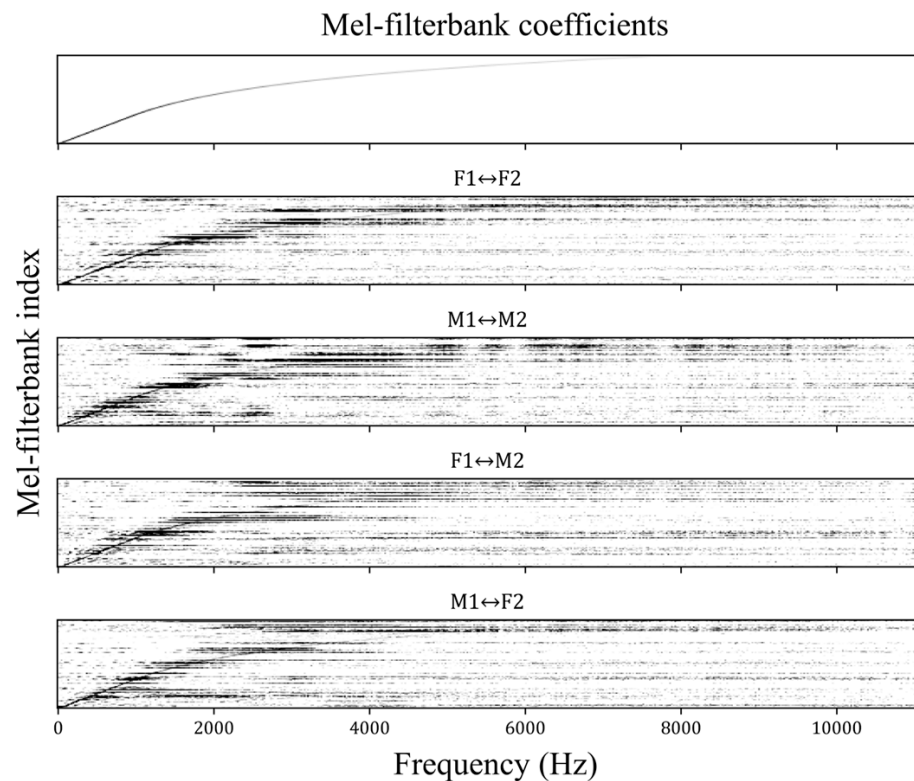
	Conversion Direction	MaskCycleGAN	wav2wav
Intra-gender	F1→F2	7.68 ± 0.29	6.08 ± 0.26
	F2→F1	7.44 ± 0.23	6.01 ± 0.26
	M1→M2	7.96 ± 0.27	6.04 ± 0.22
	M2→M1	7.04 ± 0.17	5.88 ± 0.18
	Average	7.53 ± 0.13	6.01 ± 0.11
Inter-gender	F1→M2	8.48 ± 0.21	5.86 ± 0.16
	M2→F1	8.58 ± 0.21	5.75 ± 0.16
	M1→F2	8.70 ± 0.28	6.47 ± 0.24
	F2→M1	8.36 ± 0.23	6.38 ± 0.22
	Average	8.53 ± 0.12	6.12 ± 0.11
Average		8.03 ± 0.10	6.06 ± 0.08

An ablation experiment was performed to analyze the proposed method further. First, we replaced the MelGAN with the HiFi-GAN for the baseline MaskCycleGAN, which we call “spec2spec” in Table 3. The same amount of training data was used to train both vocoders. Using the HiFi-GAN vocoder improved the MCD over the baseline MelGAN vocoder. Next, we replaced the analysis module of the “spec2spec” with the feature extracting CNN (Figure 4), which is called “wav2spec” in the table. That is, “wav2spec” took waveforms as input and outputs spectral features; then, it reconstructed waveforms using a separately trained HiFi-GAN vocoder. By integrating the analysis and mapping modules, the voice conversion performance was improved further. The third variant, called “spec2wav”, did not use the feature extracting CNN, but used the vocoder as the generator. That is, only the mapping and reconstruction modules were integrated, leaving the analysis module standalone. It can be seen in the table that “spec2wav” showed a better performance than “wav2spec”, indicating that merging the mapping and reconstruction modules was more effective than merging the analysis and mapping modules. The last row of Table 3 represents the proposed wav2wav that integrates all three modules: analysis, mapping, and reconstruction, which achieved the best performance among all the variants.

**Table 3.** Average MCDs and 95% confidence intervals for MaskCycleGAN, spec2spec, wav2spec, spec2wav, and wav2wav.

Method	Feature Extractor	Vocoder	MCD
MaskCycleGAN	DFT	MelGAN	8.03 ± 0.10
spec2spec	DFT	HiFi-GAN	7.88 ± 0.12
wav2spec	CNN	HiFi-GAN	7.52 ± 0.11
spec2wav	DFT	N/A	6.82 ± 0.10
wav2wav	CNN	N/A	6.06 ± 0.08

To analyze the spec2wav and wav2wav further, we looked into the learned weights of the feature-extracting CNN. The weights of the first CNN layer were not changed much from the initial weights. However, the second CNN layer, which initially corresponded to the mel-filterbanks, changed a lot. Figure 5 shows the learned weights of the second CNN layer after the training was completed. The learned weights were quite different from the original filterbank coefficients, where they covered wider frequency bands for each filter. Also, they were different from speaker to speaker.

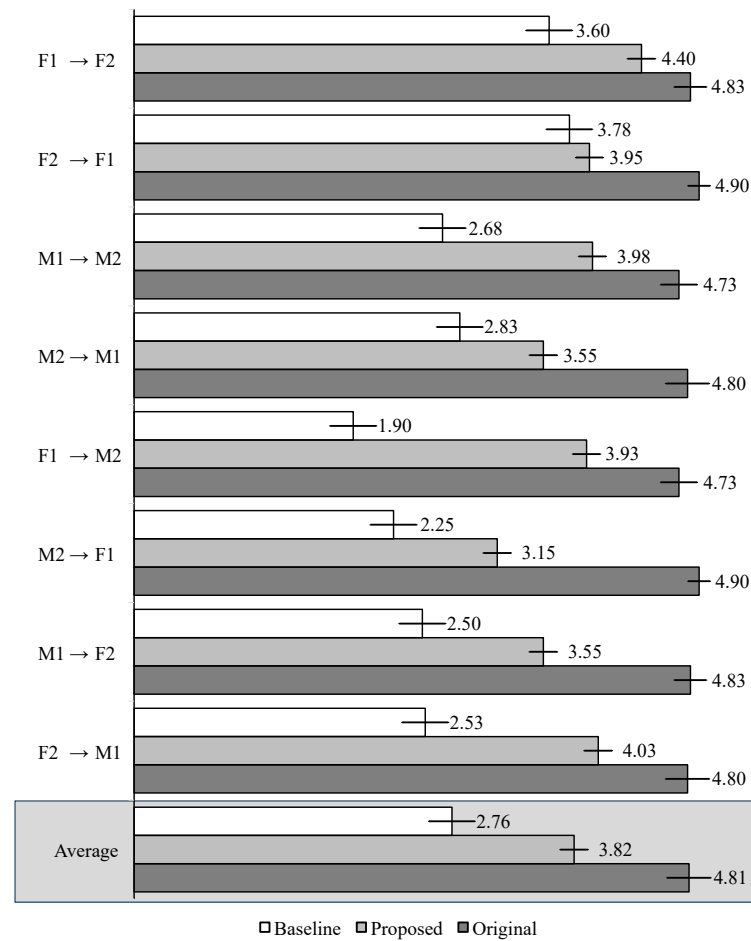


**Figure 5.** Mel-filterbank coefficients (top) and the learned weights of the feature-extracting CNN (the rest). The larger weight values are displayed in a darker color.

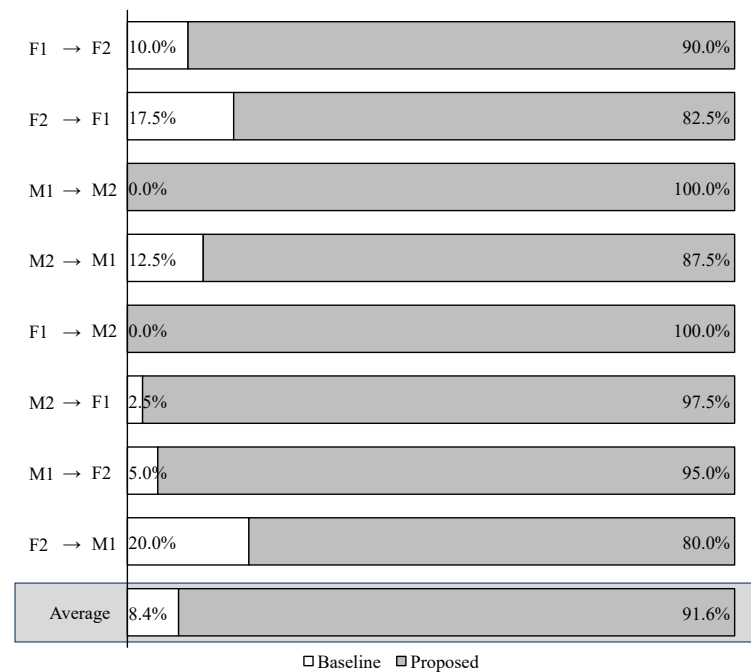
#### 4.2. Subjective Evaluation

Subjective evaluations of the sound quality and similarity for the proposed wav2wav and the baseline MaskCycleGAN were conducted using the MOS and XAB tests, respectively. A total of 80 utterances (4 utterances  $\times$  8 conversion directions  $\times$  2 methods + 16 original utterances) were used for the sound quality test, and 96 utterances (8 utterances  $\times$  4 conversion directions  $\times$  2 methods + 32 original utterances) were used for the similarity test. A total of 10 listeners participated in the experiment. For the sound quality test, the participants were asked to rate each utterance on a scale of 1 to 5, with 1 being “very bad” and 5 being “very good”. For the similarity test, the participants were asked to indicate which of the two converted speeches from the two methods sounded more similar to the target speaker’s utterance. Each time, a target speaker’s utterance was played first, then the two converted speeches were played in random order.

Figure 6 summarizes the sound quality test results. It shows that the proposed method significantly improved the sound quality of the converted voice compared with the baseline method, regardless of the conversion directions and conversion pairs. The proposed method achieved an average MOS of 3.82, which was significantly higher than the baseline score of 2.76. Figure 7 shows how similar the converted voice was to the target speaker’s voice. The numbers in the graph indicate the proportion chosen as more similar to the target speaker’s voice. On average, 92% of the utterances converted by the proposed method were selected as being more similar to the target speakers’ voices. Some sample speech waveforms produced by wav2wav are available online at <https://wav2wav.github.io/wav2wav> (accessed on 20 April 2024).



**Figure 6.** Sound quality test result. MOS and 95% confidence intervals of the original speech and the converted speeches by the baseline MaskCycleGAN and the proposed wav2wav.



**Figure 7.** The results of the similarity test (XAB) between the original and two types of converted voices using the baseline MaskCycleGAN and the proposed wav2wav.

## 5. Conclusions

In this paper, we introduce wav2wav, a wave-to-wave voice conversion approach enabling end-to-end training, which overcomes the limitations of traditional analysis–mapping–reconstruction methods. By incorporating a feature-extracting CNN into the existing vocoder and by modifying it to act as the generator of the CycleGAN-style voice conversion, the proposed method enables direct conversion from waveform to waveform. Since the proposed algorithm does not depend on any language-specific characteristics, it can be used for voice conversion for languages other than English. Moreover, the proposed method demonstrates the ability to learn effectively, even with a small-sized dataset of just 10 min. Both the objective and subjective evaluations confirmed that the proposed approach significantly improved the quality of converted speech and generated speech that closely resembled the target speech. Due to these characteristics, the proposed method can be used in a variety of applications, such as entertainment [1], education [2], and medicine [3] domains.

In future research, the proposed method may be extended to singing voice conversion. The method proposed in this study extracts global characteristics encompassing the entire voice of each speaker. Although this method is effective for converting general everyday conversational speech, it is expected that performance may deteriorate in cases where the pitch and rhythm change dynamically within a sentence, such as in a singing voice. Therefore, it will be possible to improve the performance of singing voice conversion by developing a new modeling method that takes into account the fact that speech characteristics can change even in a single sentence.

**Author Contributions:** Conceptualization and writing—review and editing, I.-C.Y. and D.Y.; methodology, D.Y.; experiment, C.J. (while he was a student at Korea University) and H.-p.C.; writing—original draft preparation, C.J. (while he was a student at Korea University); All authors have read and agreed to the published version of the manuscript.

**Funding:** This work was supported by the Basic Science Research Program through the National Research Foundation (NRF) of Korea funded by the Ministry of Science, ICT and Future Planning (NRF-2017R1E1A1A01078157). Also, it was supported by the NRF under project BK21 FOUR.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The raw data supporting the conclusions of this article will be made available by the authors on request.

**Conflicts of Interest:** The author Changhyeon Jeong was employed by the company Lotte Innovate Co., Ltd. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## References

1. Nose, T.; Igarashi, Y. Real-time talking avatar on the internet using kinect and voice conversion. *Int. J. Adv. Comput. Sci. Appl.* **2015**, *6*, 301–307. [\[CrossRef\]](#)
2. Felps, D.; Bortfeld, H.; Gutierrez-Osuna, R. Foreign accent conversion in computer assisted pronunciation training. *Speech Commun.* **2009**, *51*, 920–932. [\[CrossRef\]](#) [\[PubMed\]](#)
3. Zhao, Y.; Kuruvilla-Dugdale, M.; Song, M. Voice conversion for persons with amyotrophic lateral sclerosis. *IEEE J. Biomed. Health Inform.* **2019**, *24*, 2942–2949. [\[CrossRef\]](#) [\[PubMed\]](#)
4. Kingma, D.P.; Welling, M. Auto-encoding variational bayes. In Proceedings of the International Conference on Learning Representations, Banff, AB, Canada, 14–16 April 2014.
5. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. In Proceedings of the Neural Information Processing Systems, Montreal, QC, Canada, 8–13 December 2014; pp. 2672–2680.
6. Tobing, P.L.; Wu, Y.-C.; Hayashi, T.; Kobayashi, K.; Toda, T. Non-parallel voice conversion with cyclic variational autoencoder. In Proceedings of the Interspeech, Graz, Austria, 15–19 September 2019; pp. 659–663. [\[CrossRef\]](#)
7. Yook, D.; Leem, S.-G.; Lee, K.; Yoo, I.-C. Many-to-many voice conversion using cycle-consistent variational autoencoder with multiple decoders. The Speaker and Language Recognition Workshop (Odyssey 2020). In Proceedings of the Odyssey: The Speaker and Language Recognition Workshop, Tokyo, Japan, 1–5 November 2020; pp. 215–221. [\[CrossRef\]](#)

8. Kaneko, T.; Kameoka, H. CycleGAN-VC: Non-parallel voice conversion using cycle-consistent adversarial networks. In Proceedings of the European Signal Processing Conference, Rome, Italy, 3–7 September 2018; pp. 2100–2104. [[CrossRef](#)]
9. Kaneko, T.; Kameoka, H.; Tanaka, K.; Hojo, N. CycleGAN-VC2: Improved CycleGAN-based non-parallel voice conversion. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, Brighton, UK, 12–17 May 2019; pp. 6820–6824.
10. Kaneko, T.; Kameoka, H.; Tanaka, K.; Hojo, N. CycleGAN-VC3: Examining and improving CycleGAN-VCs for mel-spectrogram conversion. In Proceedings of the Interspeech, Shanghai, China, 25–29 October 2020; pp. 2017–2021. [[CrossRef](#)]
11. Kaneko, T.; Kameoka, H.; Tanaka, K.; Hojo, N. Maskcyclegan-VC: Learning non-parallel voice conversion with filling in frames. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, Toronto, ON, Canada, 6–11 June 2021; pp. 5919–5923. [[CrossRef](#)]
12. Proszewska, M.; Beringer, G.; Sa'ez-Trigueros, D.; Merritt, T.; Ezzerg, A.; Barra-Chicote, R. GlowVC: Mel-spectrogram space disentangling model for language-independent text-free voice conversion. In Proceedings of the Interspeech, Incheon, Republic of Korea, 18–22 September 2022; pp. 2973–2977.
13. Popov, V.; Vovk, I.; Gogoryan, V. Diffusion-based voice conversion with fast maximum likelihood sampling scheme. In Proceedings of the International Conference on Learning Representations, Virtual Event, 25–29 April 2022.
14. Sisman, B.; Yamagishi, J.; King, S.; Li, H. An overview of voice conversion and its challenges: From statistical modeling to deep learning. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2021**, *29*, 132–157. [[CrossRef](#)]
15. Sainath, T.N.; Weiss, R.J.; Senior, A.; Wilson, K.W.; Vinyals, O. Learning the speech front-end with raw waveform CLDNNs. In Proceedings of the Interspeech, Dresden, Germany, 6–10 September 2015; pp. 1–5. [[CrossRef](#)]
16. Sailor, H.B.; Patil, H.A. Filterbank learning using convolutional restricted boltzmann machine for speech recognition. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, Shanghai, China, 20–25 March 2016; pp. 5895–5899. [[CrossRef](#)]
17. Kobayashi, K.; Toda, T.; Nakamura, S. F0 transformation techniques for statistical voice conversion with direct waveform modification with spectral differential. In Proceedings of the IEEE Spoken Language Technology Workshop, San Diego, CA, USA, 13–16 December 2016; pp. 700–963. [[CrossRef](#)]
18. Kurita, Y.; Kobayashi, K.; Takeda, K.; Toda, T. Robustness of Statistical voice conversion based on direct waveform modification against background sounds. In Proceedings of the Interspeech, Graz, Austria, 15–19 September 2019; pp. 684–688. [[CrossRef](#)]
19. Kobayashi, K.; Toda, T. Sprocket: Open-source voice conversion software. In Proceedings of the Odyssey: The Speaker and Language Recognition Workshop, Les Sables d'Olonne, France, 26–29 June 2018; pp. 203–210. [[CrossRef](#)]
20. Kim, J.-W.; Jung, H.-Y.; Lee, M. Vocoder-free end-to-end voice conversion with transformer network. In Proceedings of the International Joint Conference on Neural Networks, Glasgow, UK, 19–24 July 2020; pp. 1–8. [[CrossRef](#)]
21. Nguyen, B.; Cardinaux, F. NVC-Net: End-to-end adversarial voice conversion. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, Singapore, 22–27 May 2022; pp. 7012–7016. [[CrossRef](#)]
22. Jeong, C. Voice Conversion Using Generative Adversarial Network Based Vocoder. Master's Thesis, Department of Computer Science and Engineering, Korea University, Seoul, Republic of Korea, 2023.
23. Lorenzo-Trueba, J.; Yamagishi, J.; Toda, T.; Saito, D.; Villavicencio, F.; Kinnunen, T.; Ling, Z. The voice conversion challenge 2018: Promoting development of parallel and nonparallel methods. In Proceedings of the Odyssey: The Speaker and Language Recognition Workshop, Les Sables d'Olonne, France, 26–29 June 2018; pp. 195–202. [[CrossRef](#)]
24. Zhu, J.-Y.; Park, T.; Isola, P.; Efros, A.A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In Proceedings of the International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2223–2232.
25. Griffin, D.W.; Lim, J.S. Multiband excitation vocoder. *IEEE Trans. Acoust. Speech Signal Process.* **1988**, *36*, 1223–1235. [[CrossRef](#)]
26. Morise, M.; Yokomori, F.; Ozawa, K. WORLD: A vocoder-based high-quality speech synthesis system for real-time applications. *IEICE Trans. Inf. Syst.* **2016**, *E99.D*, 1877–1884. [[CrossRef](#)]
27. Oord, A.; Dieleman, S.; Zen, H.; Simonyan, K.; Vinyals, O.; Graves, A.; Kalchbrenner, N.; Senior, A.; Kavukcuoglu, K. Wave-Net: A generative model for raw audio. In Proceedings of the ISCA Speech Synthesis Workshop, Sunnyvale, CA, USA, 13–15 September 2016; p. 125.
28. Kalchbrenner, N.; Elsen, E.; Simonyan, K.; Noury, S.; Casagrande, N.; Lockhart, E.; Stimberg, F.; Oord, A.; Dieleman, S.; Kavukcuoglu, K. Efficient neural audio synthesis. In Proceedings of the International Conference on Machine Learning, Stockholm, Sweden, 10–15 July 2018; pp. 2410–2419.
29. Jin, Z.; Finkelstein, A.; Mysore, G.J.; Lu, J. Fftnet: A real-time speaker-dependent neural vocoder. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, Calgary, AB, Canada, 15–20 April 2018; pp. 2251–2255. [[CrossRef](#)]
30. Valin, J.-M.; Skoglund, J. LPCNET: Improving neural speech synthesis through linear prediction. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, Brighton, UK, 12–17 May 2019; pp. 5891–5895. [[CrossRef](#)]
31. Oord, A.; Li, Y.; Babuschkin, I.; Simonyan, K.; Vinyals, O.; Kavukcuoglu, K.; Driessche, G.; Lockhart, E.; Cobo, L.C.; Stimberg, F.; et al. Parallel WaveNet: Fast high-fidelity speech synthesis. In Proceedings of the International Conference on Machine Learning, Stockholm, Sweden, 10–15 July 2018; pp. 3918–3926.
32. Prenger, R.; Valle, R.; Catanzaro, B. Waveglow: A flow-based generative network for speech synthesis. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, Brighton, UK, 12–17 May 2019; pp. 3617–3621. [[CrossRef](#)]

33. Kumar, K.; Kumar, R.; Boissiere, T.; Gestin, L.; Teoh, W.Z.; Sotelo, J.; Brebisson, A.; Bengio, Y.; Courville, A. MelGAN: Generative adversarial networks for conditional waveform synthesis. In Proceedings of the Neural Information Processing Systems, Vancouver, BC, Canada, 8–14 December 2019; pp. 14910–14921.
34. Kong, J.; Kim, J.; Bae, J. HiFi-GAN: Generative adversarial networks for efficient and high-fidelity speech synthesis. In Proceedings of the Neural Information Processing Systems, Online, 6–12 December 2020; pp. 17022–17033.
35. Zhou, T.; Krahenbuhl, P.; Aubry, M.; Huang, Q.; Efros, A.A. Learning dense correspondence via 3D-guided cycle consistency. In Proceedings of the IEEE/CVF Computer Vision and Pattern Recognition Conference, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 117–126.
36. Taigman, Y.; Polyak, A.; Wolf, L. Unsupervised cross-domain image generation. In Proceedings of the International Conference on Learning Representations, Toulon, France, 24–26 April 2017.
37. Kubichek, R. Mel-cepstral distance measure for objective speech quality assessment. In Proceedings of the IEEE Pacific Rim Conference on Communications, Computers and Signal Processing, Victoria, BC, Canada, 19–21 May 1993; Volume 1, pp. 125–128. [[CrossRef](#)]

**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.