INVERTIBLE VOICE CONVERSION WITH PARALLEL DATA

Zexin Cai[1], *Ming Li*[*1,2]

[1]Department of Electrical and Computer Engineering, Duke University, Durham, NC, 27708, USA
[2]Suzhou Municipal Key Laboratory of Multimodal Intelligent Systems, Duke Kunshan University, Kunshan, 215316, PR China

ABSTRACT

This paper introduces an innovative deep learning framework for parallel voice conversion to mitigate inherent risks associated with such systems. Our approach focuses on developing an invertible model capable of countering potential spoofing threats. Specifically, we present a conversion model that allows for the retrieval of source voices, thereby facilitating the identification of the source speaker. This framework is constructed using a series of invertible modules composed of affine coupling layers to ensure the reversibility of the conversion process. We conduct comprehensive training and evaluation of the proposed framework using parallel training data. Our experimental results reveal that this approach achieves comparable performance to non-invertible systems in voice conversion tasks. Notably, the converted outputs can be seamlessly reverted to the original source inputs using the same parameters employed during the forwarding process. This advancement holds considerable promise for elevating the security and reliability of voice conversion.

***Index Terms*—** Voice conversion, Audio security, Invertible neural networks, Anti-spoofing

1. INTRODUCTION

Voice conversion (VC) provides the capability to alter the vocal characteristics of a source audio signal to match a desired voice while keeping the linguistic content unchanged. Over the past decade, deep neural network-based VC techniques have outperformed traditional VC methods, yielding synthesis results with exceptional fidelity [1]. Particularly, when equipped with neural vocoders [2, 3], these advanced VC systems are capable of generating speech that is remarkably natural, rivalling human speech, under both parallel and non-parallel training conditions. Nevertheless, this high-fidelity synthesized speech poses challenges in distinguishing between authentic and synthesized speech segments. This challenge has been amplified by recent advancements in many-to-many VC, considerably broadening the range of convertible voices for voice cloning [4, 5]. Furthermore, investigations into zero-shot conversion are reducing the necessity for enrollment audio recordings, as voice cloning now only necessitates a brief audio sample from the target speaker [6, 7].

Many contemporary voice services depend on speaker verification — a biometric method to validate individuals' identities. Nevertheless, those aforementioned achievements in VC inevitably empower spoofing attacks on automatic speaker verification (ASV) systems [8]. Various studies have highlighted the vulnerabilities of speaker verification/recognition systems to spoofing attacks via VC [8, 9]. Accordingly, researchers have been investigating countermeasures to mitigate these vulnerabilities. Two predominant strategies have emerged: the development of more robust ASV systems and the incorporation of an independent spoofing detection mechanism. While both approaches aim to tackle spoofing threats, the latter has garnered greater attention within this field [10]. As a result, the biannual ASVspoof challenge was initiated in 2015, serving as a platform to stimulate and support researchers in their efforts to enhance spoofing detection performance [11, 12].

However, numerous countermeasures in the spoofing literature are aimed at discerning whether an audio signal is synthetic, yet they lack the capability to trace the origin of fraudulent activity. Particularly in instances where VC is exploited in criminal activities, identifying the true speaker behind the converted audio becomes pivotal for criminal investigations and legal proceedings. To address these potential threats emerging from VC systems, we put forth an alternative countermeasure. Our primary contribution lies in designing a novel and reliable conversion system equipped with the unique ability to reverse the conversion process. This facet enables the retrieval of the source speech from the converted result, a capability that remains absent in existing approaches. Specifically, our proposed system takes advantage of the affine coupling layer [13], integrated with a modified transformer encoder [14] to transform features within the coupling layers. Our proposed model is trained and evaluated on the parallel dataset CMU ARCTIC [15]. The experimental results demonstrate that the proposed architecture achieves successful and invertible VC. Moreover, our proposed system yields comparable performance with other VC approaches regarding naturalness and speaker similarity.

2. BACKGROUND: INVERTIBLE NETWORKS

In recent years, invertible neural networks (INNs) have gained considerable attention and made substantial advancements across diverse research fields, including image generation [13, 16] and speech synthesis [17, 18]. Originally conceived as generative models, INNs were designed to produce synthetic content from standard probability distributions. Typically, INNs consist of a series of bijective mapping [19]. Given a high-dimensional vector $\mathbf{x} \in \mathbb{R}^d$, the network functions as a bijective transformation denoted by f, involving K consecutive transformations that convert the input $\mathbf{x}$ to an output $\mathbf{y} \in \mathbb{R}^d$, as formulated in Equations 1 and 2.

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

$$f = f_1 \circ f_2 \circ \cdots \circ f_K \tag{2}$$

Importantly, each transformation is reversible. These transformations are often referred to as "forwards", while the corresponding inverse computations that map the output $\mathbf{y}$ back to $\mathbf{x}$ are termed "backwards". The corresponding backward process is mathematically defined in Equations 3 and 4.

$$\mathbf{x} = f^{-1}(\mathbf{y}) \tag{3}$$

$$f^{-1} = f_K^{-1} \circ \cdots \circ f_2^{-1} \circ f_1^{-1} \tag{4}$$

*Corresponding author: Ming Li, ming.li369@duke.edu

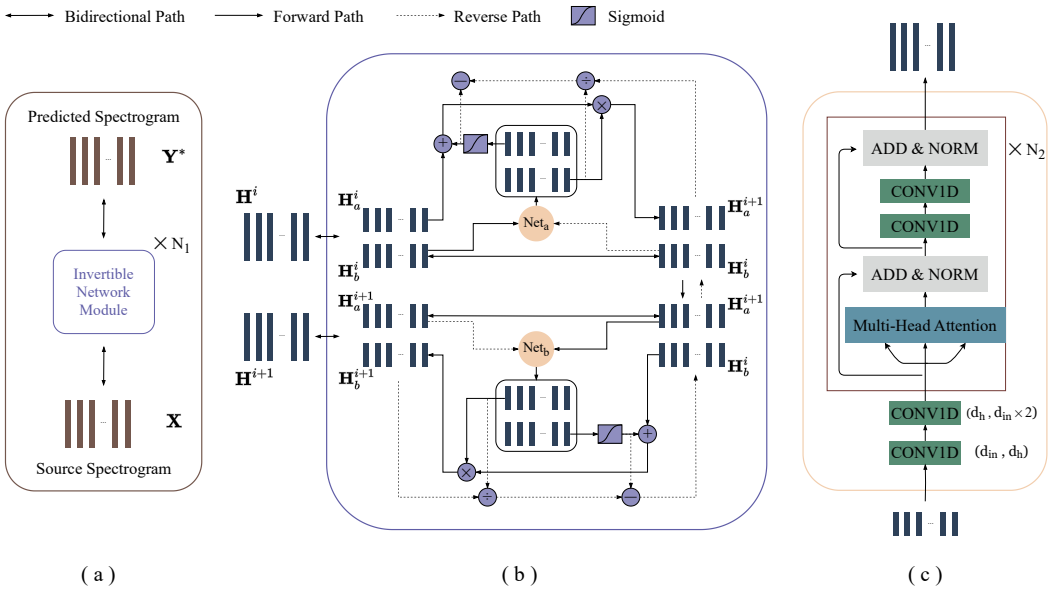

Fig. 1. Invertible Voice Conversion Architecture. (a) General overview of the proposed model. (b) Structure and data flow of the Invertible Network Module. (c) Structure of the nonlinear network component 'Net'

Bijective networks are powerful in tasks such as density estimation, variational inference, and generative modeling across diverse data modalities encompassing audio, images, and text. INNs with a tractable determinant of the Jacobian are commonly referred to as normalizing flows, which have gained widespread adoption in a multitude of fields. For instance, Gudovskiy et al. [20] and Yamaguchi et al. [21] have applied flows for anomaly detection. In speech signal processing, these flows have found utility in text-to-speech [18] and neural vocoder [17], demonstrating competitive results with other generative models like generative adversarial networks (GANs).

Recent studies have incorporated normalizing flows for VC [22]. With the addition of speech attributes like speaker representations as conditions, these flow-based models exhibit the capability for versatile any-to-any VC and even the generation of entirely new voices [23, 24]. Typically, during the training phase, flow-based VC models transform acoustic features into a latent representation denoted as $\mathbf{z}$, conditioned on attribute c. Here, $\mathbf{z}$ adheres to a normal distribution while c is extracted by speech encoders. In the inference phase, the converted voice is generated by drawing a random sample $\mathbf{z}$ from the latent distribution and a given condition c. Unlike those models, our proposed method leverages the invertibility of INNs, eliminating the necessity for tractable determinants as seen in traditional flows. Instead of transforming acoustic features into a latent space, we deploy INN modules to directly facilitate the reversible mapping of acoustic features between source and target speakers.

3. PROPOSED FRAMEWORK

The conversion framework of our proposed model is depicted in Figure 1 (a). Here, the transformation from the source spectrogram to the target counterpart is achieved by a series of invertible network modules. The input acoustic feature is the Mel-spectrogram extracted from audio signals. The number of invertible network modules is determined by the hyperparamater N_1. Each step of the invertible network module consists of a block containing two consecutive affine coupling layers, as visualized in Figure 1 (b). The nonlinear network component *Net* within the coupling layers is illustrated in Figure 1 (c).

3.1. Invertible Coupling Layer

Each step of the invertible module involves an alternating pattern executed by two consecutive affine coupling layers. This process is vividly illustrated in Figure 1 (b), wherein a module is distinctly divided into two parts: the upper portion, preserving half of the input hidden features $\mathbf{H}_b^i$ unchanged, and the symmetrical lower part, responsible for transforming $\mathbf{H}_b^i$, where $\mathbf{H}^i$ denotes the input of ith invertible network module. Note that the input of the first invertible network module is $\mathbf{X}$, the source acoustic feature, and the output of the last invertible network module is $\mathbf{Y}^*$, the predicted acoustic feature. The affine coupling layer serves as a robust bijective operation that plays a pivotal role in converting the source feature to the desired target feature [13]. Recognizing that maintaining invertibility necessitates certain values to remain unmodified in this process, we adopt such alternating manner to completely transform the input.

The forward and reverse operations of the upper coupling layer are formulated in Table 1, where $\textsc{split}()$ is the operation that chunks the input feature into two halves along the channel/feature dimension. Net_a is a nonlinear mapping structure that increases the channel dimension to obtain intermediate variables $\mathbf{U}$ and $\mathbf{B}$. If we assume that $\mathbf{H}^i \in \mathbb{R}^{T\times d}$ is a matrix with a channel size of d, then the size of $\mathbf{H}_a^i$ and $\mathbf{H}_b^i$ is $d/2$ after the $\textsc{split}()$ operation. The output vector of Net_a has a channel size of d, which is then split into two halves to obtain $\mathbf{U}$ and $\mathbf{B}$, each with a size of $d/2$. $\boldsymbol{S}$ is the element-wise scale vector, ϵ is a constant, and σ is the sigmoid function. Feature $\mathbf{H}_b^i$ remains unchanged in the first coupling layer. This allows us to obtain the same values $\boldsymbol{S}$ and $\mathbf{B}$ during the reverse phase by feeding $\mathbf{H}_b^i$ into Net_a. Then the original input $\mathbf{H}_a^i$ can be retrieved by inverse operations with respect to the way we obtain $\mathbf{H}_a^{i+1}$. The forward and reverse operations of the bottom coupling layer are similar, but they keeps $\mathbf{H}_a^{i+1}$ unchanged and convert $\mathbf{H}_b^i$ to $\mathbf{H}_b^{i+1}$.

Table 1. The forward and reverse process in the ith coupling layer

FORWARD	REVERSE
$\mathbf{H}_a^i, \mathbf{H}_b^i = \text{SPLIT}(\mathbf{H}^i)$	$\mathbf{H}_a^{i+1}, \mathbf{H}_b^i = \text{SPLIT}(\mathbf{H}')$
$\mathbf{U}, \mathbf{B} = \text{SPLIT}(\textit{Net}_a(\mathbf{H}_b^i))$	$\mathbf{U}, \mathbf{B} = \text{SPLIT}(\textit{Net}_a(\mathbf{H}_b^i))$
$\boldsymbol{S} = \sigma(\mathbf{U} + \epsilon)$	$\boldsymbol{S} = \sigma(\mathbf{U} + \epsilon)$
$\mathbf{H}_a^{i+1} = \boldsymbol{S} \odot \mathbf{H}_a^i + \mathbf{B}$	$\mathbf{H}_a^i = (\mathbf{H}_a^{i+1} - \mathbf{B}) \oslash \boldsymbol{S}$
$\mathbf{H}' = \text{CONCAT}(\mathbf{H}_a^{i+1}, \mathbf{H}_b^i)$	$\mathbf{H}^i = \text{CONCAT}(\mathbf{H}_a^i, \mathbf{H}_b^i)$

3.2. Conversion *Net*

Generally, the *Net* component can be any network structure that doubles the input feature's dimensionality to help obtain the scaling factor u and the affine component t for the feature conversion process. In our proposed model, we adopt a transformer-based structure to perform this task [14]. Depicted in Figure 1 (c), this architecture starts with two 1D convolutions designed to pre-encode the input, with the second convolution doubling the feature's dimensionality. Following this, we incorporate N_2 identical blocks, each composed of a multi-head attention module and a convolution module housing two 1D convolutional layers. After that, N_2 identical blocks are used, each containing a multi-head attention module and a convolution module with two 1D convolutional layers. Both modules integrate residual connections and culminate in layer normalization.

3.3. Loss Function

Training is optimized by minimizing the mean square error (MSE) between the predicted Mel-spectrogram and the ground-truth spectrogram. In addition, we incorporate mean absolute error (L1) losses between the means of the two Mel-spectrograms and L1 loss between their standard deviations as supplementary criteria. Consequently, our final training loss is formulated in Equation 5.

$$\begin{aligned} L_{\text{train}} &= \text{MSE}(\text{Predicted\_Mel}, \text{Target\_Mel}) \\ &+ L1(mean(\text{Predicted\_Mel}), mean(\text{Target\_Mel})) \\ &+ L1(std(\text{Predicted\_Mel}), std(\text{Target\_Mel})) \end{aligned} \tag{5}$$

4. EXPERIMENTS

4.1. Dataset

Our primary dataset for experimentation is drawn from the CMU ARCTIC English corpus[1] [15], a publicly accessible speech database featuring parallel recordings of various speakers reading textual content. From this database, we have chosen four speakers—two males and two females—for our experimental purposes. In our paper, we label them as 'bdl', 'slt', 'clb', and 'rms'. This phonetically balanced corpus encompasses 1132 parallel speech utterances for each individual speaker, all distributed in 16 kHz waveforms. In our experimental setup, 1000 of these utterances per speaker are allocated for training, while the remaining 132 are reserved for the test set.

4.2. Training Setup

In our experiment, we use the Mel-spectrogram as the acoustic feature. We extract 80-dimensional Mel-spectrograms with a window length of 25ms and a hop length of 12.5ms from the continuous speech. During the data preparation stage, we use Dynamic Time Warping (DTW) to align the acoustic features for parallel pairs. For every pair of selected speakers, we train and evaluate three corresponding voice conversion systems:

Invertible VC, our proposed model, is structured with 4 consecutive invertible modules ($N_1 = 4$). The number of the transformer-based encoder block N_2 is set to 4. Within the conversion *Net*, we use two pre-encoding convolutional layers with a kernel size of 3, while the channel size d_h is set to 512. We employ two heads for the attention mechanism. The internal convolutional layers within the transformer-based block have an intermediary channel size of 1024, with respective filter sizes of 9 and 1. Throughout the training process, the mini-batch size is set at 64. Optimization is performed using the Adam optimizer, initialized with a learning rate of 0.0001. We train a model for each pair of speakers until convergence, which typically takes more than 1000 epochs.

Transformer-VC follows a transformer-based voice conversion methodology akin to other models documented in the literature, including the Voice Transformer Network [25] and the non-autoregressive sequence-to-sequence VC [26]. Specifically, we adopt the basic structure of the transformer-based Fastspeech2[2] [27], but with the modification of omitting the variance adaptor and transitioning from phoneme embeddings to the Mel-spectrogram of the source speaker as input. The size of the PreNet in this model is set to 256. There are 4 encoder layers and 4 decoder layers, each incorporating 2 attention heads and a hidden layer size of 256. Both encoder dropout and decoder dropout are set to 0.2. The mini-batch size, optimizer and training settings remain consistent with those of the Invertible VC approach.

CycleGAN-VC3 is a GAN-based approach, which is another popular approach in VC. This model can be used for both parallel and non-parallel data. We implemented the CycleGAN-VC3 model as per the specifications outlined in the original paper [28]. We follow the hyperparameter settings from CycleGAN-VC3, except that all models are trained with parallel pairs for more than 1000 epochs until convergence. The batch size is set to 12 and the number of frames per training sample is set to 128.

All synthesized spectrograms from the VC systems mentioned above are subsequently converted into audio waveforms using a neural vocoder called HiFiGAN [3]. This vocoder undergoes training using the extracted Mel-spectrograms from the training dataset, which consists of 4000 utterances in total.

4.3. Results

For each VC approach mentioned above, we train twelve distinct pairwise conversion models constructed from the four chosen speakers. Our synthesized samples are accessible online for listening[3]. The confirmation of our model's invertibility can be achieved by both listening to the samples and referring to Table 3, which involves objective evaluation conducted using Mel-Spectrogram Distortion (MSD). This metric, resembling Mel Cepstral distortion (MCD) but applied to distinct features [29], measures the difference between the synthesized Mel-spectrograms and their natural counterparts. These MSD scores are computed across the test set encompassing all conversion speaker pairs. Note that our extracted Mel-spectrograms are normalized within the value range of $[-1, 1]$.

The MSD score between source speakers and the target speakers stands at 3.34. With our proposed Invertible VC model, the distortion rate to the target utterances drops to 2.52. A marginal performance enhancement is observed with the Transformer model, showcasing an MSD score of 2.41. However, the MSD score for

[1] http://festvox.org/cmu_arctic/

[2] https://github.com/ming024/FastSpeech2

[3] https://caizexin.github.io/parallel_invvc/index.html

Table 2. The mean opinion scores (MOS) with 95% confidence interval (CI) of three VC approaches, where **Invertible VC** is our proposed system. *p-value* is obtained by T-test comparing MOS samples between our proposed method and the target VC approach.

Speakers		**Naturalness**			**Similarity**		
source	target	Invertible VC	Transformer-VC	CycleGAN-VC3	Invertible VC	Transformer-VC	CycleGAN-VC3
bdl	clb	3.84±0.23	**4.01±0.19**	3.71±0.23	**4.13±0.18**	4.1±0.18	3.38±0.22
	rms	**4.21±0.18**	4.17±0.17	3.98±0.2	**4.12±0.19**	4.1±0.18	3.47±0.21
	slt	3.75±0.19	**4.02±0.19**	3.77±0.2	4.22±0.17	**4.24±0.17**	3.85±0.21
clb	bdl	3.35±0.22	3.2±0.24	**3.53±0.23**	3.83±0.22	**4.12±0.18**	3.48±0.22
	rms	3.81±0.21	**3.98±0.23**	3.39±0.23	4.03±0.18	**4.18±0.18**	2.47±0.2
	slt	3.31±0.24	3.93±0.22	**4.1±0.2**	3.83±0.2	**4.23±0.19**	4.22±0.19
rms	bdl	3.01±0.23	**3.11±0.25**	2.69±0.21	3.76±0.22	**3.82±0.19**	3.17±0.21
	clb	3.44±0.23	**3.47±0.24**	2.82±0.26	3.93±0.21	**3.95±0.19**	1.91±0.2
	slt	3.24±0.22	**3.47±0.2**	3.21±0.22	3.91±0.18	**4.03±0.2**	3.0±0.22
slt	bdl	3.21±0.23	**3.39±0.23**	3.36±0.23	3.97±0.19	**4.02±0.2**	3.72±0.22
	clb	4.02±0.21	4.08±0.2	**4.27±0.17**	4.35±0.18	**4.48±0.16**	4.27±0.18
	rms	4.01±0.2	**4.17±0.18**	3.58±0.21	4.05±0.19	**4.15±0.17**	2.75±0.2
All		3.59±0.07	**3.78±0.06**	3.52±0.07	4.01±0.06	**4.12±0.05**	3.31±0.07
p-values		-	6.2×10^{-5}	0.154	-	6.43×10^{-3}	$< 10^{-5}$

Table 3. Objective performance based on Mel-Spectrogram Distortion (dB), Src denotes source utterances, Tgt denotes target utterances, VC denotes the voice converted utterances and INV denotes the inverted utterances obtained from the converted results

Src - Tgt	VC - Tgt			Src - INV
	Invertible	Transformer	CycleGAN	Invertible
3.34	2.52	2.41	10.81	0.00

the CycleGAN model is unsatisfactory, even surpassing the MSD score for source-target pairs. Moreover, the credibility of our proposed model's invertibility is validated between the inverted results and the source Mel-spectrograms, resulting in an MSD score of 0. This reaffirms the efficacy of our proposed approach.

We contend that subjective evaluation of the inverted results against their natural counterparts is unnecessary since the network is entirely invertible. The inverted voices, as demonstrated by online listening and objective evaluation, perfectly mirror the input voices. Therefore, concerning subjective evaluation, we conduct an assessment to determine whether the invertible model can achieve comparable naturalness and similarity to non-invertible models. To this end, we randomly select five utterances from the test set and gather their corresponding conversion results for listening test. Thus, each pairwise conversion model contributes five utterances, resulting in a total of 60 utterances for each approach. The listening test, involving 15 participants, assesses naturalness and speaker similarity by evaluating 180 converted utterances and 180 pairs of converted and real voices. The participants rate the utterances on a MOS scale from 1 to 5 with 0.5 increments.

Table 2 presents the MOS results of the listening test conducted. Our proposed system scores around 3.59 on naturalness, which is slightly lower than the non-invertible VC approach Transformer-VC. Our proposed system's performance is similar to that of CycleGAN-V3, which scores around 3.52, as the *p-value* of 0.154 indicates low statistical significance. On the other hand, our proposed system achieves a speaker similarity score of 4.01, while Transformer-VC scores 4.12 and CycleGAN-VC3 scores 3.31. A score above 4 indicates impressive conversions of the target voices from our proposed system. CycleGAN-VC3 obtains a lower score due to the conversion systems between two speaker pairs, 'clb' and 'rms', 'slt' and 'rms'. This is also the underlying reason for CycleGAN-VC3's relatively poor MSD score in the objective evaluation. Although some of the pairs achieved fair performance in terms of naturalness, such as 'slt' to 'rms', the converted voices obtained by the models trained with these pairs are not satisfactory.

Overall, out of the three mentioned approaches, the Transformer-VC exhibits superior performance, followed closely by our proposed system. Despite this, our system is distinct in its ability to recover the source voice, rendering our invertible parallel VC approach proficient in restoring the converted voice to its original speaker. This is a distinctive feature absent in the other two methods, and it comes with only a marginal trade-off in performance compared to the transformer-based parallel voice conversion model.

5. DISCUSSION AND CONCLUSION

Similar to countermeasures against spoofing attacks found in existing literature, our proposed approach also exhibits certain limitations. While our system effectively traces the source speaker, its applicability is currently restricted to utterances synthesized by it, and the invertibility is presently available at the spectrogram level. One practical scenario for our system involves the storage of generated Mel-spectrograms. In cases where suspected synthesized speech is encountered, even if it exhibits some distortions, we can employ spectrogram-based matching algorithms such as audio fingerprinting to identify the corresponding stored Mel-spectrogram within the database. Subsequently, our proposed model can be employed to invert it, thus recovering the original speech from the source speaker.

It's important to note that, up to this point, achieving invertibility has necessitated the use of parallel data, which can be associated with significant costs. Our future endeavors will primarily revolve around the goal of achieving invertibility at the waveform level, using non-parallel data, with the aim of enhancing the robustness and cost-effectiveness of our approach.

In conclusion, we present a bijective architecture for voice conversion that possesses the unique ability to trace source speakers, potentially serving as a countermeasure against voice conversion spoofing attacks. Through the utilization of affine coupling layers, our model exhibits the capacity to revert the conversion output to its original input feature. While our proposed system demonstrates proficient conversion performance with parallel data, it's worth noting that our subjective evaluation results indicate a slight performance decrease when compared to a transformer-based conversion model.

6. REFERENCES

[1] Berrak Sisman, Junichi Yamagishi, Simon King, and Haizhou Li, "An Overview of Voice Conversion and Its Challenges: From Statistical Modeling to Deep Learning," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 132–157, 2020.

[2] Marcel de Korte, Jaebok Kim, and Esther Klabbers, "Efficient Neural Speech Synthesis for Low-Resource Languages Through Multilingual Modeling," in *Proc. Interspeech 2020*, pp. 2967–2971.

[3] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae, "HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis," *Proc. of NeurIPS 2020*, vol. 33, pp. 17022–17033.

[4] Ju chieh Chou and Hung-Yi Lee, "One-Shot Voice Conversion by Separating Speaker and Content Representations with Instance Normalization," in *Proc. Interspeech 2019*, pp. 664–668.

[5] Yist Y Lin, Chung-Ming Chien, Jheng-Hao Lin, Hung-yi Lee, and Lin-shan Lee, "Fragmentvc: Any-to-any Voice Conversion by End-to-end Extracting and Fusing Fine-grained Voice Fragments with Attention," in *2021 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 5939–5943.

[6] Kaizhi Qian, Yang Zhang, Shiyu Chang, Xuesong Yang, and Mark Hasegawa-Johnson, "Autovc: Zero-shot Voice Style Transfer with Only Autoencoder Loss," in *International Conference on Machine Learning*, 2019, pp. 5210–5219.

[7] Haozhe Zhang, Zexin Cai, Xiaoyi Qin, and Ming Li, "SIG-VC: A Speaker Information Guided Zero-Shot Voice Conversion System for Both Human Beings and Machines," in *2022 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 6567–65571.

[8] Tomi Kinnunen, Zhi-Zheng Wu, Kong Aik Lee, Filip Sedlak, Eng Siong Chng, and Haizhou Li, "Vulnerability of Speaker Verification Systems against Voice Conversion Spoofing Attacks: The Case of Telephone Speech," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 4401–4404.

[9] Zhizheng Wu, Nicholas Evans, Tomi Kinnunen, Junichi Yamagishi, Federico Alegre, and Haizhou Li, "Spoofing and Countermeasures for Speaker Verification: A Survey," *Speech Communication*, vol. 66, pp. 130–153, 2015.

[10] Zhizheng Wu, Junichi Yamagishi, Tomi Kinnunen, Cemal Hanilçi, Mohammed Sahidullah, Aleksandr Sizov, Nicholas Evans, Massimiliano Todisco, and Hector Delgado, "ASVspoof: The Automatic Speaker Verification Spoofing and Countermeasures Challenge," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 4, pp. 588–604, 2017.

[11] Zhizheng Wu, Tomi Kinnunen, Nicholas Evans, Junichi Yamagishi, Cemal Hanilçi, Md Sahidullah, and Aleksandr Sizov, "ASVspoof 2015: The First Automatic Speaker Verification Spoofing and Countermeasures Challenge," in *Sixteenth Annual Conference of The International Speech Communication Association*, 2015.

[12] Xingming Wang, Xiaoyi Qin, Tinglong Zhu, Chao Wang, Shilei Zhang, and Ming Li, "The DKU-CMRI System for the ASVspoof 2021 Challenge: Vocoder based Replay Channel Response Estimation," in *Proc. 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge*, pp. 16–21.

[13] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio, "Density Estimation Using Real NVP," in *5th International Conference on Learning Representations*, 2017.

[14] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention Is All You Need," in *Proc. of NeurIPS 2017*, pp. 5998–6008.

[15] John Kominek, Alan W Black, and Ver Ver, "CMU ARCTIC Databases for Speech Synthesis," 2003.

[16] Durk P Kingma and Prafulla Dhariwal, "Glow: Generative Flow with Invertible 1x1 Convolutions," in *Proc. of NeurIPS 2018*, vol. 31.

[17] Ryan Prenger, Rafael Valle, and Bryan Catanzaro, "Waveglow: A Flow-based Generative Network for Speech Synthesis," in *2019 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 3617–3621.

[18] Jaehyeon Kim, Sungwon Kim, Jungil Kong, and Sungroh Yoon, "Glow-TTS: A Generative Flow for Text-to-Speech via Monotonic Alignment Search," in *Proc. of NeurIPS 2020*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, Eds., vol. 33, pp. 8067–8077.

[19] Laurent Dinh, David Krueger, and Yoshua Bengio, "NICE: Non-linear Independent Components Estimation," in *3rd International Conference on Learning Representations*, 2015.

[20] Denis Gudovskiy, Shun Ishizaka, and Kazuki Kozuka, "CFLOW-AD: Real-Time Unsupervised Anomaly Detection with Localization via Conditional Normalizing Flows," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 98–107.

[21] Masataka Yamaguchi, Yuma Koizumi, and Noboru Harada, "AdaFlow: Domain-adaptive Density Estimator with Application to Anomaly Detection and Unpaired Cross-domain Translation," in *2019 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 3647–3651.

[22] Joan Serrà, Santiago Pascual, and Carlos Segura, "Blow: A Single-scale Hyperconditioned Flow for Non-parallel Raw-audio Voice Conversion," in *Proc. of NeurIPS 2019*, pp. 6790–6800.

[23] Piotr Bilinski, Thomas Merritt, Abdelhamid Ezzerg, Kamil Pokora, Sebastian Cygert, Kayoko Yanagisawa, Roberto Barra-Chicote, and Daniel Korzekwa, "Creating New Voices using Normalizing Flows," in *Proc. Interspeech 2022*, pp. 2958–2962.

[24] Jiahong Huang, Wen Xu, Yule Li, Junshi Liu, Dongpeng Ma, and Wei Xiang, "FlowCPCVC: A Contrastive Predictive Coding Supervised Flow Framework for Any-to-Any Voice Conversion," in *Proc. Interspeech 2022*, pp. 2558–2562.

[25] Wen-Chin Huang, Tomoki Hayashi, Yi-Chiao Wu, Hirokazu Kameoka, and Tomoki Toda, "Voice Transformer Network: Sequence-to-Sequence Voice Conversion Using Transformer with Text-to-Speech Pretraining," in *Proc. Interspeech 2020*, pp. 4676–4680.

[26] Tomoki Hayashi, Wen-Chin Huang, Kazuhiro Kobayashi, and Tomoki Toda, "Non-Autoregressive Sequence-To-Sequence Voice Conversion," in *2021 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2021, pp. 7068–7072.

[27] Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu, "FastSpeech 2: Fast and High-Quality End-to-End Text to Speech," in *International Conference on Learning Representations*, 2021.

[28] Takuhiro Kaneko, Hirokazu Kameoka, Kou Tanaka, and Nobukatsu Hojo, "CycleGAN-VC3: Examining and Improving CycleGAN-VCs for Mel-Spectrogram Conversion," in *Proc. Interspeech 2020*, pp. 2017–2021.

[29] Robert Kubichek, "Mel-cepstral Distance Measure for Objective Speech Quality Assessment," in *Proceedings of IEEE Pacific Rim Conference on Communications Computers and Signal Processing*. IEEE, 1993, vol. 1, pp. 125–128.