INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Reverberation-Controllable Voice Conversion Using Reverberation Time Estimator

Yeonjong Choi, Chao Xie, Tomoki Toda

Nagoya University, Japan

{yeonjong.choi,xie.chao}@g.sp.m.is.nagoya-u.ac.jp, tomoki@icts.nagoya-u.ac.jp

Abstract

Recent trends have emerged to implement voice conversion (VC) in real-world scenarios where background sounds and reverberation are inevitable. However, most VC studies mainly focus on clean speech conversion, where high-quality speech data are required for training and testing. Moreover, the background sounds and reverberation are treated as interferences to be discarded, despite being informative to be retained in some scenarios, such as movie dubbing and singing VC. In this paper, we propose a reverberation-robust VC framework consisting of a reverberation time (T60) estimation module and a VC module. The T60 estimator is introduced to provide the VC module with the reverberation information to model the reverberant speech. Experimental results show that 1) our framework can disentangle and control the speaker identity and reverberation from the speech, and 2) we can get acceptable VC performances dealing with reverberation, even when clean training data are not available.

Index Terms: voice conversion, reverberant speech, reverberation modeling

1. Introduction

Voice conversion (VC) aims to convert a speaker (Source speaker) of a speech to sound like another speaker (Target speaker) while retaining its linguistic content [1]. Many researchers have dedicated to the studies of VC even before the advent of deep learning. These early proposed methods, such as vector quantization (VQ) [2] and Gaussian mixture modeling [3], lay a foundation for the deep-learning-based ones. Despite deep learning significantly improves the performance of VC in terms of speech naturalness and speaker similarity [4] and makes it possible to use nonparallel data in VC model training, clean speech data from source and target speakers are essentially required. Moreover, new challenges have emerged when implementing VC into real-world applications, such as communication aids for the speech-impaired [5], and voice dubbing for movies [6], where high-quality speech data are not always available and data in real-world often entangled with noise and reverberation. On the other hand, these background sounds are also informative and should be retained in the converted samples in specific applications, such as singing VC [7][8][9] and data augmentation for speech recognition [10].

Compared to tremendous studies in conventional VC, only a few works focus on VC in noisy and reverberant conditions. However, most methods still require clean speech data from source/target speakers for training. Moreover, background sounds are considered interferences to be discarded. For example, Takashima *et al.* [11] proposed a sparse representation-based VC using nonnegative matrix factorization to filter out the noise. Miao *et al.* [12] proposed a noise-robust VC that introduces two filtering methods at the pre- and post-processing stages, respectively, to suppress the noise. In our previous work [13], we have developed a three-stage VC framework towards noisy-reverberant scenarios, consisting of a pre-trained denoising model, a pre-trained dereverberation model, and a nonparallel VC model. Although clean speech data for the VC model training are unnecessary, the background noise and reverberation are filtered as interferences.

Recently, some researchers have started concentrating on noise-controllable VC, where the VC system is noise-robust, and the background noise can be either discarded or retained in the converted samples based on specific scenarios. Furthermore, clean speech data from source/target speakers are not necessary. Xie *et al.* [14] and Yao *et al.* [15] have proposed two noisy-to-noisy (N2N) VC frameworks to model the noisy speech directly by introducing separated noise as conditions. However, despite two methods being capable of controlling noise, reverberation is not considered, and how to address it is unclear.

In this paper, we proposed a reverberation-controllable VC framework to disentangle and control the speaker identity and reverberation factors in a reverberant speech. A pre-trained reverberation time (T60) estimator is introduced to extract T60-related information. Inspired by those noisy-to-noisy VC works [14][15] where using separated noise to model noisy speech data can achieve better VC performance, our VC module also takes the estimated T60 information as the condition to model the reverberant speech data. Furthermore, by introducing a pre-trained TasNet for dereverberation task [16][17], our proposed framework is explored in a more extreme condition where clean training data are unavailable and instead the dereverbered speech is used as input. Objective and subjective evaluations are conducted, and the experimental results show that our method achieves acceptable VC performance with reverberation control, even when clean source/target data is not provided during training. Our contributions to this work are as follows:

- We propose a reverberation-controllable VC framework consisting of a T60 estimator and a nonparallel VC model based on variational auto encoder (VAE). A pre-trained dereverberation model is also introduced for the scenario where clean speech data are unavailable.
- We evaluate our framework with both of dereverbered- and reverberant-speech inputs to confirm whether the framework can disentangle the reverberation of the speech inputs, and model the reverberation totally based on the T60-related information conditioning.
- We investigate how choosing the training data pairs (whether input speech and/or output ground truth are clean-

10.21437/Interspeech.2023-1356

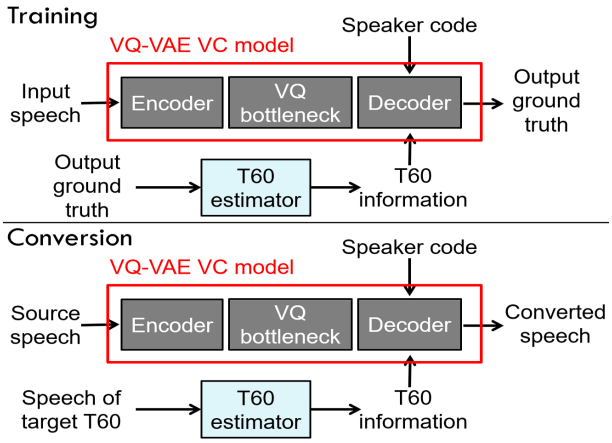

Figure 1: *Overview of the proposed VC framework for controlling reverberation. T60 estimator is a pre-trained module, and is fixed during the VC training stage.*

, reverberant-, dereverbered-speech) in the denoising VAE training framework will affect the VC performances.

2. VC framework for controlling reverberation

2.1. Framework overview

Figure 1 illustrates the proposed framework consisting of a VC model and a T60 estimator. The T60 estimator is pre-trained on the publicly available mega dataset [18] and fixed during the VC training. A dereverberation model is prepared and can be optionally used to process the reverberant data.

In the training stage, the VC model receives a speech with a clean, dereverberant, or reverberant attribute as input. In contrast, the same speech with a different reverberation property is used as the ground truth to calculate the reconstruction loss, and also the input of the T60 estimator to extract the reverberation information, which is used as a condition in the decoder of the VC model. We will discuss the different combinations of the training data pairs (whether input speech and/or output ground truth are clean-, reverberant-, dereverbered-speech) in Section 3.3.

During conversion stage, source-speaker's speech is given as the VC model input, where the model generates a converted speech, conditioned on a target-speaker code and a T60-related information. Here, the T60 information is extracted from the T60 estimator, where a speech of target T60 is given as its input.

2.2. Voice conversion model

We implemented the VC model as VQ-VAE [19], which consists of an encoder, a vector quantizer, and a decoder. The encoder consists of five convolution blocks composed of a one-dimensional convolution (Conv1D) layer followed by a batch normalization layer and a ReLU activation function. Given log mel-spectrogram sequence $\{\mathbf{x}_t|t = 1, ..., T\}$ as input, the encoder transforms it to a sequence of latent representation denoted as $\{\mathbf{z}_j|j = 1, ..., N\}$. Then the latent representation is quantized by the vector quantizer managing a trainable codebook $\{\mathbf{e}_i|i = 1, ..., 512\}$, where $\mathbf{e}_i$ is a 64-dimensional vector. Each vector of the latent representation $\mathbf{z}_j$ is mapped into the vector $\mathbf{e}_k$ in the codebook by the nearest distance:

$$\hat{\mathbf{z}}_j = \mathbf{e}_k \tag{1}$$

$$k = \arg\min_i ||\mathbf{z}_j - \mathbf{e}_i||^2 \tag{2}$$

where $\{\hat{\mathbf{z}}_j|j = 1, ..., N\}$ is a replaced discrete quantized representation. The decoder consists of a lightweight recurrent network to reconstruct the waveform based on $\hat{\mathbf{z}}_j$ and the embedded speaker identity (speaker code) in an auto-regressive way.

During the training stage, the VQ-VAE is trained to minimize the sum of the negative log-likelihood of the reconstruction loss and the commitment loss :

$$\mathcal{L} = -\frac{1}{T}\sum_{t=1}^{T}\log p(\mathbf{x}_t|\hat{\mathbf{z}}, \mathbf{c}, \mathbf{r}) + \beta\frac{1}{N}\sum_{j=1}^{N}||\mathbf{z}_j - \mathrm{sg}(\hat{\mathbf{z}}_j)||^2 \tag{3}$$

where $\mathbf{c}$ is the speaker code, $\mathbf{r}$ is the T60-related conditioning, β is the commitment weight, and $\mathrm{sg}(\cdot)$ indicates the stop-gradient operation. Since this function is not differentiable, the gradient of the loss function through the codebook is approximated by the straight-through estimator [20]. The codebook is updated by exponential moving averages [21].

2.3. T60 estimator

The T60 estimator is implemented by the speaker encoder of the zero-shot VC method AdaIN-VC [22], because reverberation information shares the same characteristic as speaker identity that is time-invariant. The T60 estimator receives an arbitrary-length speech and extracts the reverberation-related information into a fixed-dimensional embedding vector. In our experiments, the estimator is pre-trained with the mean square error loss, to outputs a single T60 value linearly transformed from the embedding vector. During the training and conversion stage of VC, the output linear layer of the estimator is removed, so that the embedding vector extracted from the last hidden layer can be used as the T60-related conditioning for the decoder of the VC model. Since we do not use additional loss terms to guarantee the disentanglement between speaker identity and reverberation, the different utterances of each speaker in the training set possess randomly sampled T60 values so that the disentanglement can be learned in a self-supervised way.

2.4. TasNet dereverberation model

TasNet [16][17] is a neural network that directly models the time-domain waveform signal using a convolutional encoder-decoder architecture. TasNet consists of an encoder that has a non-negativity constraint on its output, a separator with deep LSTM architecture, and a linear decoder. The separator performs the enhancement step by estimating a proper weighting function for the encoder output at each time step. Then the decoder converts the enhanced latent representation into the waveform signal. TasNet's performance and its effectiveness have been demonstrated by being compared with a number of systems that work on time-frequency domain representations [16][17]. We use the SD-SDR loss [23] as an objective function for training.

3. Experimental Evaluations

3.1. Dataset

All the experiments were conducted on single-channel, 8kHz sampled speech data.

For the training of the dereverberation model, we used WHAMR! dataset [18], which was originally designed for

speech separation under noisy-reverberant conditions. It contained 58.03 hours of speech data for training and 14.65 hours for the validation set. We used `s1_reverb` set as the input signals, which is a set of reverberated single speaker sources, and used `s1_anechoic` set as the ground truth signals, which is a clean set. Since the clips consist of 2 channels originally, we extracted the left channel of those to make it single channel sources.

The same dataset was used for the T60 estimator training. We used `s1_reverb` and `s1_anechoic` sets as the input signals, and used the corresponding T60 labels as output targets.

We used VCC2018 dataset [24] for VQ-VAE VC models training and evaluation. There were 8 source speakers and 4 target speakers in the data, where each speaker uttered 81 training clips and 35 evaluation clips. For our experimental purpose, we generated reverberant versions of VCC2018 by convolving with room impulse responses (RIR). We used RIR settings of WHAMR! [18], which the T60 range was $0.1 \sim 1.0$ sec. Here, we followed the settings of evaluation set, so those RIRs were not seen by dereverberation model and T60 estimator during its training. We divided the RIR settings into two groups, one to be convolved with training sets of VCC2018, and the other with evaluation sets. Note that different RIRs were used for generating each utterance of reverberant speech. For evaluation data, four speakers (VCC2SM3, VCC2SM4, VCC2SF3, VCC2SF4) were selected as source speakers, and two speakers (VCC2TF2, VCC2TM2) as target speakers.

3.2. Model training

For VQ-VAE VC model, we used a model that was implemented by [19], where we followed the same hyperparameter settings with [13]. For TasNet dereverberation model, we used Asteroid toolkit [25] with the same settings as [13]. For T60 estimator, we used the codes implemented by [22]. For the estimator training, the same hyperparameter settings with the VC model were used. For our purpose, a linear layer with a single output was added during its training, and removed again during the VC training and conversion stage.

3.3. Methods to be evaluated

For the experiments, we prepared the following VQ-VAE VC frameworks, which were similar to the denoising training:

- Model-A (trained on C-C, C-R, R-C, and R-R pairs)
- Model-B (trained on C-C, C-R, D-C, and D-R pairs)
- Model-C (trained on D-D, D-R, R-D, and R-R pairs)

Here, "C" indicates clean speech data, "R" indicates reverberant speech data, and "D" means dereverbered speech data, processed by the TasNet dereverberation model. For example, "C-R" indicates that the pair is composed of clean input speech and reverberant output ground truth speech. Model-A and C were evaluated on the dereverbered input case and the reverberant input case separately, in order to find out whether the VC models can disentangle the reverberation from input speech. Model-B was introduced to find out whether the VC performance will be improved than Model-A if we optimize the model to the clean- and dereverbered inputs. Model-C was introduced for more extreme assumption, where clean data are unavailable.

In addition, we also prepared our previous VC framework [13] denoted as "Dereverb-TasNet $\rightarrow$ VC", where the conversion model was trained on dereverbered speech only, without any T60-related information conditioning.

We used three kinds of reverberation conditioning for our experiments, which were extracted from a single simulated reverberant utterance for each. The utterances were not uttered by source- or target-speakers. The actual and the estimated T60, estimated from the utterances, were as follows:

- Rev-1 - Actual : 0.00 sec, Estimated : -0.000227 sec
- Rev-2 - Actual : 0.502 sec, Estimated : 0.442 sec
- Rev-3 - Actual : 0.956 sec, Estimated : 0.831 sec

Note that the above reverberations were never seen by the dereverberation model, the T60 estimator, and the VC model, during its training.

For the T60 estimator performance, when we input the target-speakers' evaluation utterances that were simulated with the above three kinds of RIRs, mean errors for Rev-1, 2, and 3 were $1.09 \cdot 10^{-3}$, $3.66 \cdot 10^{-2}$, and $1.08 \cdot 10^{-1}$, respectively.

Table 1: *The objective speech intelligibility (STOI) of the converted speech (Compared with clean source-speaker speech).*

Model		Target reverberation			
		Rev-1	Rev-2	Rev-3	Average
Source-speaker speech		1.00	0.95	0.86	0.93
Model-A	Dereverb input	0.73	0.70	0.66	0.69
	Reverb input	0.72	0.68	0.64	0.68
Model-B	Dereverb input	0.72	0.69	0.65	0.69
Model-C	Dereverb input	0.72	0.68	0.65	0.68
	Reverb input	0.71	0.67	0.64	0.67
Dereverb-TasNet $\rightarrow$ VC		0.71	-	-	-

Table 2: *The objective speaker similarities of the converted speech (Compared with clean target-speaker speech).*

Model		Target reverberation			
		Rev-1	Rev-2	Rev-3	Average
Target-speaker speech		1.00	0.84	0.72	0.85
Model-A	Dereverb input	0.83	0.70	0.62	0.72
	Reverb input	0.83	0.70	0.62	0.72
Model-B	Dereverb input	0.84	0.70	0.63	0.72
Model-C	Dereverb input	0.80	0.66	0.61	0.69
	Reverb input	0.79	0.68	0.61	0.69
Dereverb-TasNet $\rightarrow$ VC		0.80	-	-	-

3.4. Experimental results

3.4.1. Results of objective evaluation

Tables 1 and 2 show the speech intelligibility and speaker similarities of the converted speech, respectively. Here, STOI [26] compared with clean source-speaker reference is used for Table 1. End-to-end speaker verification system [27] is used for the speaker similarities, compared with clean target-speaker reference, for Table 2. The first row of each table indicates the results of reference speech, which show the maximum values that each column (target reverberation) can have. Next, from the results of Model-A, using dereverbered- or reverberant-speech as VC model input doesn't seem to show significant differences, which indicates that our framework can disentangle the reverberation from input speech, so that it doesn't affect the VC performance much. Next, we investigate whether optimizing the VC model to clean- and dereverbered-speech would increase the VC performance (Model-B). Comparing it with Model-A (dereverbered input case), the speaker similarities of Model-B

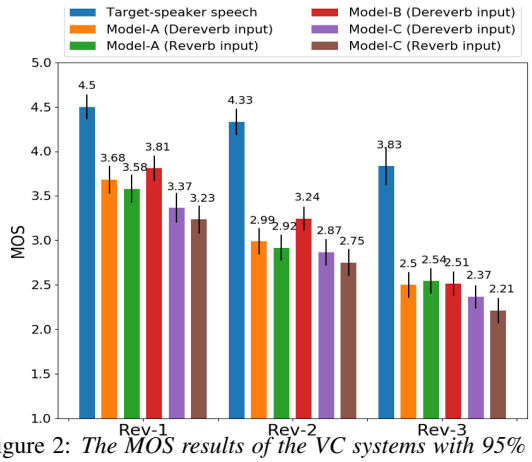

Figure 2: *The MOS results of the VC systems with 95% confidence interval.*

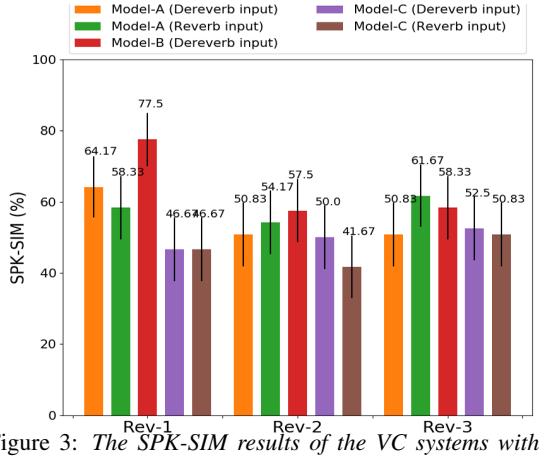

Figure 3: *The SPK-SIM results of the VC systems with 95% confidence interval.*

seem to slightly increase with the score 0.84, 0.70, and 0.63, for Rev-1, 2, and 3, respectively, while for the speech intelligibility, Model-A shows slightly better with the score 0.73, 0.70, and 0.66. But basically the differences seem to be negligibly small. Also, for the scenario where no clean data are assumed to be available, Model-C is investigated. From both tables, Model-C shows slightly worse results than A, which means using no clean data for VC training is still a challenging task. However, if we compare our proposed method with the previous work (Dereverb-TasNet → VC) [13], it is obvious that our proposed method is comparable in generating clean converted speech.

3.4.2. Results of subjective evaluation

For the naturalness test, we conducted a mean opinion score (MOS) test, where 15 listeners, recruited on Amazon Mechanical Turk, were asked to give a naturalness score from 1 to 5 (higher is better). We sampled three utterances from each of 4 conversion-pairs (two source- and two target-speakers), with 3 target reverberations, converted by five systems, 1) Model-A (Dereverb input), 2) Model-A (Reverb input), 3) Model-B (Dereverb input), 4) Model-C (Dereverb input), and 5) Model-C (Reverb input). We also sampled three reference speeches from each two target-speakers with three target reverberations, thus 198 samples were evaluated in total by every listener. For the speaker-similarity (SPK-SIM) and reverberation-similarity (REV-SIM) test, we conducted the SIM test [24]. In the two SIM tests, each listener listened to two kinds of samples: a converted speech, and a reference speech of the same target speaker and the target reverberation with a different sentence. Listeners

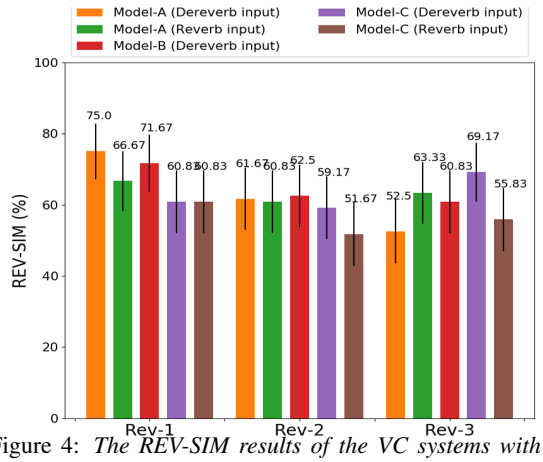

Figure 4: *The REV-SIM results of the VC systems with 95% confidence interval.*

were asked to determine whether those samples were uttered by the same speaker for SPK-SIM, and whether the samples have the same reverberation levels for REV-SIM: *Definitely the same, Maybe the same, Maybe different, Definitely different*. We sampled 2 utterances from each conversion-pairs and reverberations, thus 120 samples were evaluated in total by every listener.

The results are shown in Figure 2, 3, and 4. For the SIM tests, the percentages indicate the added percentages of *Definitely the same* and *Maybe the same*. As shown in Figure 2 and 3, Model-B shows the best performances among the dereverbered input cases, in terms of MOS and SPK-SIM. It is obvious that optimizing the VC model to clean- and dereverbered-speech is helpful for the VC performance improvement. Moreover from Figure 2, similarly from the objective results, our model can disentangle the reverberation from the input speech well, as the dereverbered input and reverbered input case for Model-A doesn't show significant differences, where we can get the same conclusion from Model-C as well. However, in Figure 3 and 4, there are some gaps between dereverbered input and reverbered input case, which means there might some reverberation information left from the input speech, affecting the VC performances. Finally reverberation-controllability can be checked from Figure 4. Though there are some gaps between the systems, it is observable that our model can control the reverberation well, especially that Model-B shows more than 60% of reverberation similarities for all the target reverberations. While Model-C has an advantage that it does not use clean training data, it still has many rooms to improve its performances, which we will focus on for our future works.

4. Conclusions

In this paper, we presented a voice conversion (VC) framework that is robust to reverberation, and can control it. Our framework consists of a pre-trained reverberation time estimator and a nonparallel VC model, where a pre-trained dereverberation model can be optionally used for VC training and conversion stage. The experimental results showed that our framework can disentangle the reverberation from the input speech, and model the reverberation based on the reverberation conditioning. Also, we can get acceptable performances dealing with reverberation, even when clean training data are unavailable. As future works, we plan to try our method for more recent VC systems and 16kHz sampled speech data.

Acknowledgements: This work was supported by JST SPRING Grant Number JPMJSP2125, and JST CREST Grant Number JPMJCR19A3.

5. References

[1] B. Sisman, J. Yamagishi, S. King, and H. Li, "An overview of voice conversion and its challenges: From statistical modeling to deep learning," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 132–157, 2021.

[2] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, "Voice conversion through vector quantization," in *ICASSP-88., International Conference on Acoustics, Speech, and Signal Processing*, 1988, pp. 655–658 vol.1.

[3] Y. Stylianou, O. Cappe, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 2, pp. 131–142, 1998.

[4] Z. Yi, W.-C. Huang, X. Tian, J. Yamagishi, R. K. Das, T. Kinnunen, Z.-H. Ling, and T. Toda, "Voice Conversion Challenge 2020 — Intra-lingual semi-parallel and cross-lingual voice conversion —," in *Proc. Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge 2020*, 2020, pp. 80–98. [Online]. Available: http://dx.doi.org/10.21437/VCC_BC.2020-14

[5] C. Veaux, J. Yamagishi, and S. King, "Towards personalised synthesised voices for individuals with vocal disabilities: Voice banking and reconstruction," in *Proceedings of the Fourth Workshop on Speech and Language Processing for Assistive Technologies*. Grenoble, France: Association for Computational Linguistics, Aug. 2013, pp. 107–111. [Online]. Available: https://aclanthology.org/W13-3917

[6] F. Mukhneri, I. Wijayanto, and S. Hadiyoso, "Voice conversion for dubbing using linear predictive coding and hidden markov model," *Journal of Southwest Jiaotong University*, vol. 55, 01 2020.

[7] C. Deng, C. Yu, H. Lu, C. Weng, and D. Yu, "Pitchnet: Unsupervised singing voice conversion with pitch adversarial network," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7749–7753.

[8] Y.-J. Luo, C.-C. Hsu, K. Agres, and D. Herremans, "Singing voice conversion with disentangled representations of singer and vocal technique using variational autoencoders," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 3277–3281.

[9] H. Guo, Z. Zhou, F. Meng, and K. Liu, "Improving adversarial waveform generation based singing voice conversion with harmonic signals," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 6657–6661.

[10] S. Shahnawazuddin, N. Adiga, K. Kumar, A. Poddar, and W. Ahmad, "Voice Conversion Based Data Augmentation to Improve Children's Speech Recognition in Limited Data Scenario," in *Proc. Interspeech 2020*, 2020, pp. 4382–4386.

[11] R. Takashima, R. Aihara, T. Takiguchi, and Y. Ariki, "Noise-robust voice conversion based on spectral mapping on sparse space," in *Proc. 8th ISCA Workshop on Speech Synthesis (SSW 8)*, 2013, pp. 71–75.

[12] X. Miao, M. Sun, X. Zhang, and Y. Wang, "Noise-robust voice conversion using high-quefrency boosting via sub-band cepstrum conversion and fusion," *Applied Sciences*, vol. 10, no. 1, 2020. [Online]. Available: https://www.mdpi.com/2076-3417/10/1/151

[13] Y. Choi, C. Xie, and T. Toda, "An Evaluation of Three-Stage Voice Conversion Framework for Noisy and Reverberant Conditions," in *Proc. Interspeech 2022*, 2022, pp. 4910–4914.

[14] C. Xie, Y.-C. Wu, P. L. Tobing, W.-C. Huang, and T. Toda, "Direct noisy speech modeling for noisy-to-noisy voice conversion," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 6787–6791.

[15] J. Yao, Y. Lei, Q. Wang, P. Guo, Z. Ning, L. Xie, H. Li, J. Liu, and D. Xie, "Preserving background sound in noise-robust voice conversion via multi-task learning," 2022. [Online]. Available: https://arxiv.org/abs/2211.03036

[16] Y. Luo and N. Mesgarani, "Tasnet: Time-domain audio separation network for real-time, single-channel speech separation," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 696–700.

[17] ——, "Real-time Single-channel Dereverberation and Separation with Time-domain Audio Separation Network," in *Proc. Interspeech 2018*, 2018, pp. 342–346.

[18] M. Maciejewski, G. Wichern, E. McQuinn, and J. L. Roux, "Whamr!: Noisy and reverberant single-channel speech separation," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 696–700.

[19] B. van Niekerk, L. Nortje, and H. Kamper, "Vector-Quantized Neural Networks for Acoustic Unit Discovery in the ZeroSpeech 2020 Challenge," in *Proc. Interspeech 2020*, 2020, pp. 4836–4840.

[20] Y. Bengio, N. Léonard, and A. Courville, "Estimating or propagating gradients through stochastic neurons for conditional computation," 2013. [Online]. Available: https://arxiv.org/abs/1308.3432

[21] A. van den Oord, O. Vinyals, and k. kavukcuoglu, "Neural discrete representation learning," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper/2017/file/7a98af17e63a0ac09ce2e96d03992fbc-Paper.pdf

[22] J. chieh Chou and H.-Y. Lee, "One-Shot Voice Conversion by Separating Speaker and Content Representations with Instance Normalization," in *Proc. Interspeech 2019*, 2019, pp. 664–668.

[23] J. L. Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "Sdr – half-baked or well done?" in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 626–630.

[24] J. Lorenzo-Trueba, J. Yamagishi, T. Toda, D. Saito, F. Villavicencio, T. Kinnunen, and Z. Ling, "The Voice Conversion Challenge 2018: Promoting Development of Parallel and Nonparallel Methods ," in *Proc. The Speaker and Language Recognition Workshop (Odyssey 2018)*, 2018, pp. 195–202.

[25] M. Pariente, S. Cornell, J. Cosentino, S. Sivasankaran, E. Tzinis, J. Heitkaemper, M. Olvera, F.-R. Stöter, M. Hu, J. M. Martín-Doñas, D. Ditter, A. Frank, A. Deleforge, and E. Vincent, "Asteroid: The PyTorch-Based Audio Source Separation Toolkit for Researchers," in *Proc. Interspeech 2020*, 2020, pp. 2637–2641.

[26] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2010, pp. 4214–4217.

[27] L. Wan, Q. Wang, A. Papir, and I. L. Moreno, "Generalized end-to-end loss for speaker verification," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 4879–4883.