INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

CFVC: Conditional Filtering for Controllable Voice Conversion

Kou Tanaka, Takuhiro Kaneko, Hirokazu Kameoka, Shogo Seki

NTT Communication Science Laboratories, NTT Corporation, Japan

kou.tanaka.ef@hco.ntt.co.jp

Abstract

This paper describes a many-to-many voice conversion model that filters the speaker vector to control high-level attributes such as speaking rate while preserving voice timbre. In order to control only the speaking rate, it is essential to decompose the speaker vector into a speaking rate vector and others. The challenge is to train such disentangled representations with no/few annotation data. Motivated by this difficulty, we propose an approach combining the conditional filtering method with data augmentation. The experimental results showed that our method disentangled complex attributes without annotation and separately controlled speaking rate and voice timbre. Audio samples can be accessed on our web page[1].

Index Terms: Voice conversion, conditional filtering, disentanglement, speaker representation, sequence-to-sequence learning

1. Introduction

Voice conversion (VC) technology, which converts one speaking style to another without changing linguistic information, has been applied for various tasks; speaker conversion [1,2], singing conversion [3, 4], assistive systems [5, 6] to overcome speech and hearing impairments, and pronunciation and accent conversions [7] in language learning. Although the recent conversion quality has been improved thanks to sequence-to-sequence (S2S) learning approaches [2,4,6,8,9], the efficient way for optimal control of high-level attributes such as speaking rate and fundamental frequency (F_0) is still an active research area. In this work, we focus on speaking rate controllability.

A well-known pure signal processing (SP) method for modifying speaking rate is a waveform similarity-based synchronous overlap addition (WSOLA) method [10]. WSOLA performs temporal modification of speech waveforms by using a scale factor. For more flexible control, [11] proposed a phonetic posterior-gram (PPG) based VC, which was the best VC method in VC challenge 2020 [12]. The PPGs are extracted from the speech in advance using an external automatic speech recognition (ASR) system trained with a large amount of training data. Then, an inverter from the PPG to the Mel spectrogram is trained. In addition, to change the speaking rate, each context duration per speaker is calculated in advance. The context duration is manually modified along with the pre-calculated duration statistics during the inference (called *duration adjustment*). The converted speech sounds natural but requires an external model and duration statistics for each context.

To reduce tedious labeling tasks, a semi-supervised generative modeling method [13] for text-to-speech synthesis (TTS) has been proposed. They introduce a variational autoencoder (VAE) based speaker encoder to the S2S-based TTS system. By introducing the VAE, speech attributes such as speaking rate, which are essential but rarely labeled, can be discovered and controlled. However, their framework is in the scope of semi-supervised learning, requiring small amounts of labeled data as training data to make reasonable inferences, and their application is TTS.

To avoid the labeling task altogether, an unsupervised speech disentanglement method [14] has been proposed for the VC task. They perform aspect-specific voice conversion by disentangling speech into content, rhythm, pitch, and voice timbre using multiple autoencoders (AEs) in an unsupervised manner. Each AE constrains the flow of information on the speech components to be disentangled using efficient signal processing methods and resampling techniques. Unfortunately, data modification is necessary during the training and testing, and reference speech is also needed to estimate the rhythm corresponding to the speaking rate.

Inspired by [15], we propose a conditional filtered voice

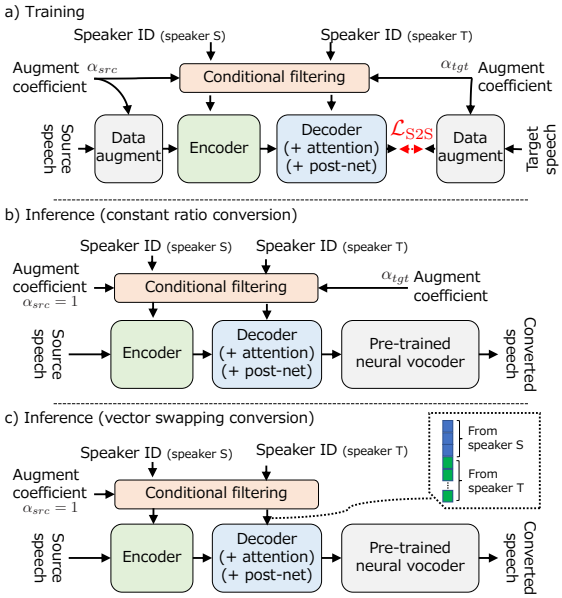

Figure 1: *System overview of the proposed method during the training (a) and the inference (b) and (c). The differences between (b) and (c) are described in Sec. 2.2. In (c), the converted speech has target speaker's voice timbre but source speaker's speaking rate, for example. (c) is important to confirm whether the proposed method makes is possible to decompose the speaker representation.*

[1] http://www.kecl.ntt.co.jp/people/tanaka.ko/projects/cfvc/

10.21437/Interspeech.2023-1193

conversion (CFVC) model for controlling high-level attributes such as speaking rate while preserving voice timbre. Unlike [15], we do not require annotation in advance to learn the attribute representation. The filtering coefficients are treated as continuous values rather than discontinuous values. Furthermore, the conversion framework is within sequence-to-sequence learning. Similar to [14], a data augmentation approach is adopted rather than data labeling. However, our approach makes it possible to represent and control the speaker's speaking rate as a vector by introducing a conditional filtering technique, unlike [14]. Our contributions are as follows:

- Our model training requires only speech data and no annotation for speaking rate.
- Combining the conditional filtering method with data augmentation makes it possible to learn the represention of each speaker's speaking rate as a multiple-dimensional vector, not a scalar.
- It is possible to convert the speech into a voice with the speaking rate of speaker S and the voice timbre of speaker T by manipulating the learned speaker vector, as shown in Fig. 1 (c).
- We briefly confirmed that our model also allows training and controlling other high-level attributes such as F_0 (see the bonus track in the supplemental material).

2. Method

The system overview is shown in Fig. 1. We employ an S2S model architecture where the source speech parameters are used as input, and the target speech parameters are generated as output. All we need for training is a parallel corpus of input and output paired speech.

As the speech parameters, we extract 80-dimensional log-Mel spectrogram features $\boldsymbol{X}_{src} = [\boldsymbol{x}_{1,src}, \cdots, \boldsymbol{x}_{i,src}]$ and $\boldsymbol{X}_{tgt} = [\boldsymbol{x}_{1,tgt}, \cdots, \boldsymbol{x}_{j,tgt}]$ over a range of 80-7600 Hz from the given source and target speech signals sampled at 16 kHz. The requirements for Short-Time Fourier Transform are the same as reported in [16]; a Hanning window, 64 ms frame length, eight ms frameshift, and 1024-point Fast Fourier Transform. Since shorter sequences make it easier to train each model, we shortened the length of the sequence by creating subframes containing two time frames.

2.1. Data Augmentation

This paper focuses on speaking rate control and augments the training data with an efficient signal processing method, WSOLA [10]. Time-domain speech waveforms $\boldsymbol{Y}_{src}$ and $\boldsymbol{Y}_{tgt}$ are manipulated before the log-Mel spectrogram extraction. Using the scale factors α_{src} and α_{tgt}, $\boldsymbol{Y}_{src}$ and $\boldsymbol{Y}_{tgt}$ are stretched to α_{src} and α_{tgt} times their original length. The scale factor of 1 indicates no data augmentation. Therefore, the extracted log-Mel spectrogram features $\boldsymbol{X}_{src}$ and $\boldsymbol{X}_{tgt}$ are already stretched. We sampled α_{src} and α_{tgt} from the uniform distribution over a range of 0.8-1.25 and performed the data augmentation with 50% probability during the training. In this paper, the uniform distribution range was determined from the fastest and slowest speakers in the training data. To avoid misunderstanding, WSOLA requires a scale factor and does not require precalculation of speaking rate.

2.2. Conditional Filtering

As shown in Fig. 2, input one-hot vectors $\boldsymbol{o}_{src}$ and $\boldsymbol{o}_{tgt}$ corresponding to speaker id are passed into a conditional filtering module f_{CF} consisting of a 1024-dimensional linear projection, followed by the hyperbolic tangent activation function and three linear projections to get 3-dimensional variable components $\boldsymbol{c}_{v,src}$ and $\boldsymbol{c}_{v,tgt}$ changed by the data augmentation, 3-dimensional constant components $\boldsymbol{c}_{c,src}$ and $\boldsymbol{c}_{c,tgt}$ not changed by the data augmentation, and 253-dimensional remaining components $\boldsymbol{c}_{r,src}$ and $\boldsymbol{c}_{r,tgt}$. 3-dimensional vectors $\boldsymbol{c}_{s,src}$ and $\boldsymbol{c}_{s,tgt}$ for representing speaking rate are generated as follows:

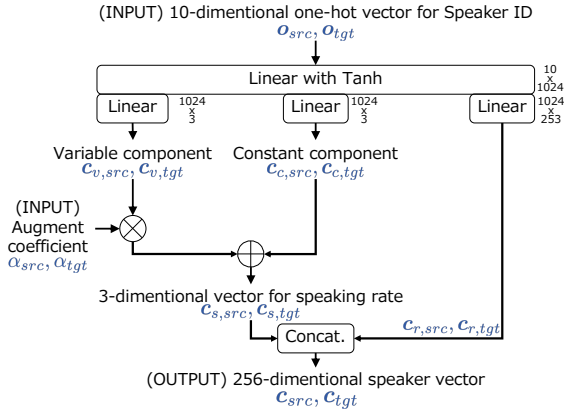

Figure 2: *Conditional filtering module for controlling the speaking rate. "Linear", "Tanh" and "Concat." indicate the linear transformation layer, the hyperbolic tangent function, and the vector concatenation, respectively. Augment coefficients α_{src} and α_{tgt} are defined at Sec. 2.1.*

$$\boldsymbol{c}_{s,src} = \boldsymbol{c}_{v,src} * \alpha_{src} + \boldsymbol{c}_{c,src}, \quad (1)$$

$$\boldsymbol{c}_{s,tgt} = \boldsymbol{c}_{v,tgt} * \alpha_{tgt} + \boldsymbol{c}_{c,tgt}. \quad (2)$$

Finally, a 256-dimensional speaker vectors $\boldsymbol{c}_{src}$ and $\boldsymbol{c}_{tgt}$ conditioning on the encoder and decoder are obtained by concatenating the speaking rate vectors $\boldsymbol{c}_{s,src}$ and $\boldsymbol{c}_{s,tgt}$ and the remainings $\boldsymbol{c}_{r,src}$ and $\boldsymbol{c}_{r,tgt}$.

At the test time, we can control the speaking rate of the converted speech $\hat{\boldsymbol{X}}_{crc}$ by changing α_{tgt}. Considering the conditional filtering module makes it possible to represent the speaking rate as $\boldsymbol{c}_{s,tgt}$, we can try to generate the speech $\hat{\boldsymbol{X}}_{vsc}$ with target speaker's voice timbre but source speaker's speaking rate by concatenating $\boldsymbol{c}_{s,src}$ and $\boldsymbol{c}_{r,tgt}$. We call these tasks 'constant ratio conversion' and 'vector swapping conversion' (see Fig. 1 (b) and (c)).

In the preliminary experiment, we confirmed that changing the speaking rate representation from 3-dimensional vector representation into a scalar representation resulted in performance degradation. Therefore, a multidimensional representation may help to model the high-level attributes. We also confirmed that the constant components were necessary because the vector swapping conversion failed by removing the constant components.

2.3. Encoder and Decoder

Inspired by [17], an input log-Mel spectrogram $\boldsymbol{X}_{src}$ and a source speaker vector $\boldsymbol{c}_{src}$ are passed into an encoder network f_{Enc} is composed of three 1D non-causal convolutional layers each containing 512 filters with the kernel size of 5, followed by a batch normalization [18] layer and a rectified linear unit (ReLU) activation. The output of the final convolutional layer is passed into a single bi-directional long short term memory (LSTM) layer containing 512 units to generate encoded features

$\boldsymbol{Z}$, as follows:

$$\boldsymbol{Z} = f_{\mathrm{Enc}}(\boldsymbol{X}_{src}, \boldsymbol{c}_{src}). \quad (3)$$

A decoder network f_{Dec} is the same as reported in [17], except for changing the LSTM unit size from 1024 to 512. [2] and taking the target speaker vector $\boldsymbol{c}_{tgt}$ as input. The decoder involving the attention mechanism and post-net predicts an attention matrix $\boldsymbol{A}$, the output log-Mel spectrogram $\hat{\boldsymbol{X}} = [\hat{\boldsymbol{x}}_1, \cdots, \hat{\boldsymbol{x}}_j]$, and its improved $\hat{\boldsymbol{X}}_{post}$ from $\boldsymbol{Z}$, $\tilde{\boldsymbol{X}} = [\boldsymbol{x}_0, \boldsymbol{x}_{1,tgt}, \cdots, \boldsymbol{x}_{j-1,tgt}]$, and $\boldsymbol{c}_{tgt}$, where $\boldsymbol{x}_0$ indicates a zero vector, known as *start token*, as follows:

$$\hat{\boldsymbol{X}}_{post}, \hat{\boldsymbol{X}}, \boldsymbol{A} = f_{\mathrm{Dec}}(\boldsymbol{Z}, \tilde{\boldsymbol{X}}, \boldsymbol{c}_{tgt}). \quad (4)$$

Finally, we minimize the objective function $\mathcal{L}_{\mathrm{S2S}}$ to train the encoder f_{Enc} and the decoder f_{Dec}:

$$\mathcal{L}_{\mathrm{S2S}} = ||\hat{\boldsymbol{X}} - \boldsymbol{X}_{tgt}||_2, \quad (5)$$

where $||\hat{\boldsymbol{X}}_{post} - \boldsymbol{X}_{tgt}||_2$ and *stop token loss* [17] are also used but omitted for brevity.

2.4. Neural Vocoder

In the proposed model, any neural vocoders by conditioning on Mel spectrograms as input can be used. In the experiment, HiFi-GAN [19] was used as the neural vocoder. We applied *V2 setting* described in [19] except for the eight ms frameshift condition.

3. Experiments

The most important thing to prove here is that the proposed method makes it possible to disentangle the speaker representation and allow the manipulation of each element of the speaker vector. Therefore, objective experimental evaluations demonstrate that 1) the length of the vector swapping converted speech $\hat{\boldsymbol{X}}_{vsc}$ using the speaking rate vector $\boldsymbol{c}_{s,src}$ from the source speaker and the remaining $\boldsymbol{c}_{r,tgt}$ from the target speaker is equivalent to the length of the source speech $\boldsymbol{X}_{src}$. Then, subjective experimental evaluations demonstrate that 2) the voice timbre of the converted speech $\hat{\boldsymbol{X}}_{vsc}$ is similar to that of the target speech $\boldsymbol{X}_{tgt}$.

3.1. Experimental Conditions

We conducted experimental evaluations using a phonetically balanced Japanese speech parallel dataset [20] consisting of utterances by six professional male speakers and four professional female speakers. The speech was recorded in a quiet room with minimal reverberation, and silent sections were removed using annotation by experts. Therefore, fluctuation of the length of the speech length caused by fluctuation of the length of the silence section need not be considered in this experiment. To train VC models, we used 450 sentences (speech section of around 0.5 hours) per speaker. Table 1 shows speech section length statistics. To evaluate the performance, we used 53 sentences per speaker. The encoder-decoder models were trained on *many-to-many* condition, which is 10-speaker input and 10-speaker output.

The number of training iterations is 100k. The learning rate and the exponential decay rate for the first moment for Adam [21] were set at 0.0002 and 0.9 after 4k step of warmup. The mini-batch size was 16. We evaluated the following converted speech.

[2] We confirmed no degradation in the preliminary objective experiments.

Table 1: *Averaged speech section length [sec] over the training dataset. Male speaker (upper) and female speaker (lower). The lower the value the faster the speaking rate.*

Speaker	*mho*	*mht*	*mmy*	*msh*	*mtk*	*myi*
Length	3.65	4.18	3.89	3.93	4.54	3.69

Speaker	*fkn*	*fks*	*ftk*	*fym*
Length	4.42	4.05	4.48	4.11

- Baseline: Converted speech by using a sequence-to-sequence VC model, which is equivalent to the proposed method without the data augmentation and the conditional filtering module.
- Proposed-swap: Converted speech $\hat{\boldsymbol{X}}_{vsc}$ by using the speaking rate vector $\boldsymbol{c}_{s,src}$ from the source speaker and the remaining $\boldsymbol{c}_{r,tgt}$ from the target speaker in the vector swapping conversion using the proposed method.
- Proposed-0.6∼1.4: Converted speech $\hat{\boldsymbol{X}}_{crc}$ when α_{tgt} is varied from 0.6 to 1.4 in the constant ratio conversion using the proposed method.

As the objective evaluation metrics, we used Mel-cepstral distortion (MCD) [dB], root mean square error of F_0 (F_0 RMSE) [Hz], and character error rate (CER) [%]. We used dynamic time warping to get the alignment between the converted sample and the reference sample. To calculate the MCD and F_0 RMSE, we extracted 1-24 order Mel-cepstrums and F_0s from the raw speech and the converted speech synthesized by the neural vocoder. Note that the range of F_0 RMSE was 22-24, and no difference was found, so it was omitted. The CER was calculated by the Transformer-based ASR model trained on the corpus of spontaneous Japanese (CSJ) [22], which was provided by ESPnet [23].

As the objective evaluation metrics to evaluate the speaking rate controllability, we define duration factor (DF) and duration ratio correlation coefficient (DRCC). DF is proposed to evaluate the controllability of speaking rate in constant ratio conversion. In constant ratio conversion, the length of the converted speech should vary in proportion to the given augment coefficient α_{tgt}. For example, the length of the converted speech with $\alpha_{tgt} = 0.8$ should be 0.8 times the length of the converted speech with $\alpha_{tgt} = 1.0$. Therefore, we define the DF as follows:

$$\mathrm{DF} = \frac{1}{N}\sum_{n=1}^{N} \frac{\mathrm{length(converted\ speech\ with}\ \alpha_{tgt})}{\mathrm{length(converted\ speech\ with}\ \alpha_{tgt} = 1.0)}, \quad (6)$$

where N is the number of the evaluated audio clips. In this experiment, N was 4770 (90 speaker pairs * 53 sentences). The closer the DF value is to α_{tgt}, the better the performance.

DRCC is also proposed to evaluate the controllability of speaking rate in vector swapping conversion. First, we define duration ratio 1 (DR1) and DR2 as follows:

$$\mathrm{DR1} = \frac{\mathrm{length(original\ speech}\ \boldsymbol{X}_{tgt})}{\mathrm{length(original\ speech}\ \boldsymbol{X}_{src})}, \quad (7)$$

$$\mathrm{DR2} = \frac{\mathrm{length(converted\ speech\ from}\ \boldsymbol{X}_{src})}{\mathrm{length(reconstrcuted\ speech\ from}\ \boldsymbol{X}_{src})}, \quad (8)$$

If the conversion model is trained well, DR2 should be the same value as DR1, which is the reference DR. As DRCC for "Baseline" and "Proposed-1.0", we calculated the correlation

Table 2: *Objective evaluation results. The lower the MCD and CER, the better the performance. The closer the DF and DRCC are to α and 1, respectively, the better the performance. $\star$ indicates the extrapolation because α is sampled from a uniform distribution over a range of 0.8-1.25 during the training.*

System	α	MCD	CER	DF	DRCC
Baseline	—	5.16	14.7	—	0.99
Proposed-swap	—	5.20	14.6	—	0.95
	0.6$\star$	4.44	16.1	0.70	—
(Faster)	0.7$\star$	4.81	15.5	0.75	—
↑	0.8	5.08	15.0	0.81	—
	0.9	5.18	14.8	0.90	—
Proposed-	1.0	5.16	14.6	—	0.98
	1.1	5.33	14.8	1.10	—
↓	1.2	5.44	15.0	1.20	—
(Slower)	1.3$\star$	5.49	15.3	1.31	—
	1.4$\star$	5.51	15.9	1.44	—

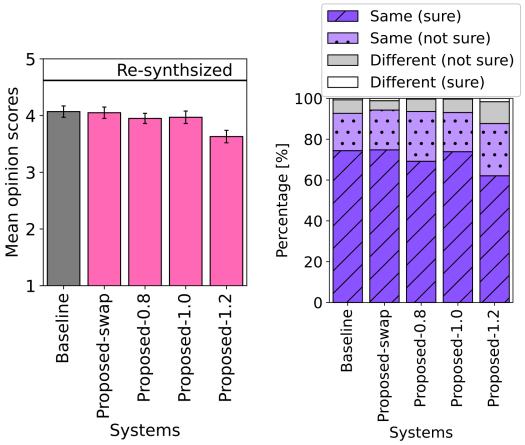

Figure 3: *Subjective evaluation results on sound quality (left) and voice timbre similarity (right). Re-synthesized indicates the speech synthesized from the ground-truth Mel-spectrogram.*

coefficient between DR1 and DR2 over 4770 audio clips. The closer the DRCC value is to 1, the better the performance. For "Proposed-swap", we used the following DR2 to calcurate the DRCC.

$$\mathrm{DR2} = \frac{\mathrm{length}(\text{converted speech } \hat{\boldsymbol{X}} \text{ from } \boldsymbol{X}_{src})}{\mathrm{length}(\text{converted speech } \hat{\boldsymbol{X}}_{vsc} \text{ from } \boldsymbol{X}_{src})}, \quad (9)$$

If the proposed method works as intended, the speech length of $\hat{\boldsymbol{X}}_{vsc}$ should be that of the reconstructed speech from $\boldsymbol{X}_{src}$ because of the converted speech $\hat{\boldsymbol{X}}_{vsc}$ using the source speaker's speaking rate vector $\boldsymbol{c}_{s,src}$. Of course, if the proposed method does not work well and the length of the converted speech is similar to that of the target speech, the DRCC of "Proposed-swap" will be close to 0.0.

As the subjective evaluation of sound quality, we conducted a 5-scaled mean opinion score (MOS): 5 for excellent, 4 for good, 3 for fair, 2 for poor, and 1 for bad. To confirm voice timbre similarity, we also conducted a 4-scaled preference test (PT): 4 for same (sure), 3 for same (not sure), 2 for different (not sure), and 1 for different (sure). 10 native Japanese speakers participated in each subjective evaluation. Each system was evaluated over 275 times.

3.2. No Negative Impact of Data Augmentation and Conditional Filtering Modules

The comparison of "Baseline" and "Proposed-1.0" showed that there is no difference in objective and subjective evaluation results, as shown in Table 2 and Fig. 3. Therefore, we found no problems with the addition of data augmentation and conditional filtering modules.

3.3. Disentangled Representation

There was no difference between "Baseline" and "Proposed-swap" for the objective evaluations, as shown in Table 2. Especially, the DRCC of "Proposed-swap" was still higher than 0.9 despite a part of the speaker vector being swapped. The results showed that the length of the converted speech "Proposed-swap" is approximately the same as the length of the source speech.

On the other hand, the sound quality and the voice timbre similarity of "Proposed-swap" were comparable to those of "Baseline," as shown in Fig. 3. "Proposed-swap" has the voice timbre of the target speech even though a part of the speaker vector was swapped. These results showed that the proposed method allowed us to decompose the speaker representation into a speaking rate component and other components.

3.4. Other Discussions

For "Proposed-0.6∼0.8", as shown in Table 2, the DFs behaved as if there was an upper limit to how fast we could speak. For "Proposed-0.8∼1.2", compared with "Baseline," the speaking rate was successfully controlled without degrading CERs. For "Proposed-1.2∼1.4", MCDs and CERs tended to be degraded, as in the case of "Proposed-0.6∼0.8".

On the other hand, as shown in Fig. 3, the subjective evaluations showed that the sound quality and the voice timbre similarity of the proposed method were comparable to those of the baseline method except for $\alpha = 1.2$. A possible reason is that slower speech may make the difference easier to understand, as in language learning. In particular, the converted speech further slowing down an initially slow speech may be out of the natural speech space.

4. Conclusions

We proposed a many-to-many voice conversion using data augmentation and conditional filtering to obtain a speaker representation decomposed into the speaking rate and other components without annotation, precomputation of statistics, or reference speech. Experimental results showed that the proposed method disentangled complex attributes of the speaker and controlled speaking rate and voice timbre separately. In the future, we plan to extend the proposed method to an any-to-many/any voice conversion method.

Acknowledgment
This work was supported by JST CREST Grant Number JPMJCR19A3, Japan.

5. References

[1] A. van den Oord, O. Vinyals, and K. Kavukcuoglu, "Neural discrete representation learning," in *NeurIPS*, 2017, pp. 6309–6318.

[2] J. Zhang, Z. Ling, L. Liu, Y. Jiang, and L. Dai, "Sequence-to-sequence acoustic modeling for voice conversion," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 3, pp. 631–644, 2019.

[3] K. Kobayashi, T. Toda, G. Neubig, S. Sakti, and S. Nakamura, "Statistical singing voice conversion with direct waveform modification based on the spectrum differential," in *INTERSPEECH*, 2014, pp. 2514–2518.

[4] R. Valle, J. Li, R. Prenger, and B. Catanzaro, "Mellotron: Multispeaker expressive voice synthesis by conditioning on rhythm, pitch and global style tokens," in *ICASSP*, 2020, pp. 6189–6193.

[5] K. Tanaka, T. Toda, G. Neubig, S. Sakti, and S. Nakamura, "A hybrid approach to electrolaryngeal speech enhancement based on noise reduction and statistical excitation generation," *IEICE Transactions on Information and Systems*, vol. 97, no. 6, pp. 1429–1437, 2014.

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

[7] D. Felps, H. Bortfeld, and R. Gutierrez-Osuna, "Foreign accent conversion in computer assisted pronunciation training," *Speech Communication*, vol. 51, no. 10, pp. 920–932, 2009.

[8] H. Kameoka, W.-C. Huang, K. Tanaka, T. Kaneko, N. Hojo, and T. Toda, "Many-to-many voice transformer network," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 656–670, 2020.

[9] K. Tanaka, H. Kameoka, T. Kaneko, and S. Seki, "Distilling sequence-to-sequence voice conversion models for streaming conversion applications," in *SLT*, 2022 (to appear).

[10] W. Verhelst and M. Roelands, "An overlap-add technique based on waveform similarity (WSOLA) for high quality time-scale modification of speech," in *ICASSP*, vol. 2, 1993, pp. 554–557.

[11] L.-J. Liu, Y.-N. Chen, J.-X. Zhang, Y. Jiang, Y.-J. Hu, Z.-H. Ling, and L.-R. Dai, "Non-parallel voice conversion with autoregressive conversion model and duration adjustment," in *Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge*, 2020, pp. 126–130.

[12] 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 *Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge*, 2020, pp. 80–98.

[13] R. Habib, S. Mariooryad, M. Shannon, E. Battenberg, R. Skerry-Ryan, D. Stanton, D. Kao, and T. Bagby, "Semi-supervised generative modeling for controllable speech synthesis," in *ICLR*, 2020.

[14] C. H. Chan, K. Qian, Y. Zhang, and M. Hasegawa-Johnson, "SpeechSplit2.0: Unsupervised speech disentanglement for voice conversion without tuning autoencoder bottlenecks," in *ICASSP*, 2022, pp. 6332–6336.

[15] T. Kaneko, K. Hiramatsu, and K. Kashino, "Generative attribute controller with conditional filtered generative adversarial networks," in *CVPR*, 2017, pp. 6089–6098.

[16] H. Kameoka, K. Tanaka, and T. Kaneko, "FastS2S-VC: Streaming non-autoregressive sequence-to-sequence voice conversion," *arXiv preprint arXiv:2104.06900*, 2021.

[17] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan *et al.*, "Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions," in *ICASSP*, 2018, pp. 4779–4783.

[18] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *ICML*, 2015, pp. 448–456.

[19] J. Kong, J. Kim, and J. Bae, "HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis," in *NeurIPS*, 2020, pp. 17 022–17 033.

[20] A. Kurematsu, K. Takeda, Y. Sagisaka, S. Katagiri, H. Kuwabara, and K. Shikano, "ATR Japanese speech database as a tool of speech recognition and synthesis," *Speech Communication*, vol. 9, no. 4, pp. 357–363, 1990.

[21] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR*, 2015.

[22] K. Maekawa, "Corpus of spontaneous Japanese: Its design and evaluation," in *Workshop on Spontaneous Speech Processing and Recognition*, 2003, pp. 7–12.

[23] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N. Enrique Yalta Soplin, J. Heymann, M. Wiesner, N. Chen, A. Renduchintala, and T. Ochiai, "ESPnet: End-to-end speech processing toolkit," in *INTERSPEECH*, 2018, pp. 2207–2211.