INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

VC-T: Streaming Voice Conversion Based on Neural Transducer

Hiroki Kanagawa , Takafumi Moriya*, Yusuke Ijima*

NTT Corporation, Japan
hiroki.kanagawa@ntt.com

Abstract

A conventional sequence-to-sequence voice conversion (seq2seq VC), i.e., attentional encoder-decoder, can be trained without the speech sequence pre-aligning normally used to counter the different lengths of the source and target speakers. However, if alignments rendered by attention are not monotonic, speech drops and repeats will happen, and the linguistic contents will not be kept. To address this issue, we propose VC-T, a novel streaming VC framework based on a neural transducer (RNNT); RNNT is effective in the automatic speech recognition field as it offers robust alignment against collapse. We also introduce an alignment design scheme for VC-T training. Experiments show that our offline and streaming VC-T variants outperform two modern seq2seq parallel VCs while offering a lower character error rate as a result of the proposal robust alignment. Our VC-T also achieves better naturalness the drastic degradation suffered by the conventional alternatives, especially for streaming VC.

Index Terms: streaming voice conversion, neural transducer

1. Introduction

Voice conversion (VC) is a technique that converts the source speaker's characteristics into those of the target speaker while preserving the linguistic content of the input speech. With the introduction of the statistical model-based approach, VC has been aggressively studied [1, 2, 3]. These studies can be roughly categorized into non-parallel and parallel VC approaches.

The non-parallel VC approach, which has been was actively studied in recent years, allows the training data to consist of different speech content from the source and target speakers. This advantage is well utilized by variational auto-encoder (VAE)-based approaches [4, 5, 6] and generative adversarial network (GAN)-based ones [7, 8, 9], as it offers large amounts of these data. However, comprehensive coverage in the form of speaker and utterance variations are required even for non-parallel data. If the quantity of these variations are insufficient, the non-parallel VC has difficulty in reproducing adequate speaker characteristics while retaining the intended linguistic content.

The parallel VC approach requires the same speech content between the source and target speakers in training, but can realize high-quality VC while requiring less training data than the non-parallel VC. The traditional offsets [3, 10, 11] aligned the source and target speaker's duration length differences with dynamic time warping (DTW) that is performed frame-wise. VC performance heavily depends on DTW accuracy, so if acoustic features are mapped between mismatched phonemes, the quality is degraded. The pre-alignment also drops the target speaker's duration information, thus speaking rate conversion cannot be performed. On the other hand, modern sequence-to-sequence (seq2seq) based approaches model not only the spectra differences but also duration differences across source and target speakers by the attention mechanism within an attentional encoder-decoder. This means that these approaches are DTW-free, so VC with speaking rate alteration is possible. However, if monotonic attention is not obtained, the results are corrupted by speech dropouts and content repetition. To alleviate this problem, [12] introduced a loss term that diagonalized the attention [13]. However, since this loss term sets a diagonal constraint over the whole utterance, the attention, which maps the source and target speaker's phonemes, training becomes problematic.

To achieve robust VC, we propose the novel VC framework, called VC-T. It is an advanced VC model based on the neural transducer framework (RNNT) [14]. RNNT is promising for developing accurate automatic speech recognition (ASR) [15] schemes. RNNT learns a mapping between the input acoustic feature and output token sequences even if they have different lengths. The difference from seq2seq-based ASR is that RNNT performs time synchronous decoding, not token synchronous decoding. Thus the alignments generated by RNNT are definitely monotonic and diagonal. In [16], they applied RNNT to realize a text-to-speech (TTS) model, i.e., Speech-T. The Speech-T naturally avoids the attention collapse problem and transduces the phoneme sequence into the acoustic feature sequence. Motivated by these studies, we apply the RNNT framework to VC for the first time. To this end, we set two goals; 1) fitting VC into the RNNT learning framework and 2) realizing correspondences between phonemes that would be impossible with the simple monotonic-attention constraint [13] employed by ConvS2S-VC. Our proposal VC-T, achieves both goals by; 1) utilizing the Speech-T's lazy forward algorithm, and 2) designing explicit phoneme-by-phoneme alignments between the source and target speaker. Objective and subjective evaluations show the effectiveness of VC-T is due to its ability to produce stable alignments. [1]

2. Related work

2.1. Modern seq2seq parallel VC

2.1.1. ConvS2S-VC [12]

The traditional seq2seq model adopts RNNs for both encoder and decoder model structures, and the encoder's final hidden states are fed to the decoder. Applying this to VC can directly convert speech without DTW between source and tar-

* Equal contribution.

[1] Sample audios are available here: https://ntt-hilab-gensp.github.io/is2023vct/

10.21437/Interspeech.2023-2383

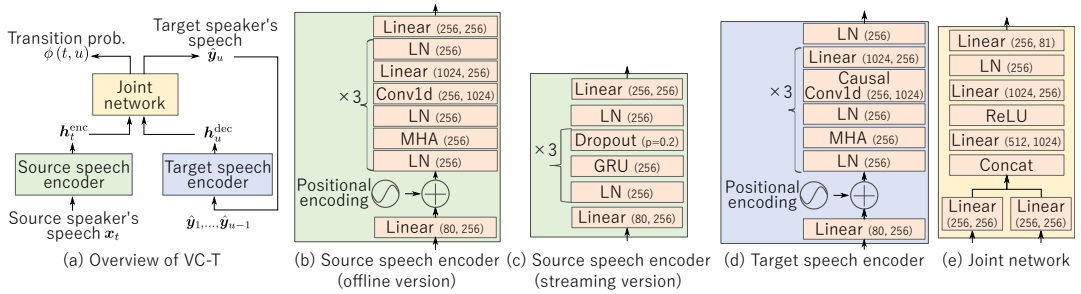

Figure 1: *(a) overviews the overall proposed VC-T. (b) and (c) are two types of source speech encoders used for offline and streaming, (d) and (e) illustrate the target speech encoder and joint network, respectively. LN and MHA stand for the layer normalization and multi-head self-attention modules, respectively. Each weighted module is tagged with input and unit size.*

get speakers [17]. Its drawbacks are that RNNs cannot compute the hidden state for each time step in parallel, and as sequence length increases, the mismatch between RNNs during training and inference increases, resulting in attention collapse. ConvS2S replaces RNNs with CNNs to overcome these issues [18]. ConvS2S-VC [12] applies this to VC and has demonstrated better performance than frame-wise VC [19] and RNN-based seq2seq one [17]. Replacing non-causal CNNs with causal ones yields streaming operation. Note, the loss term introduced by [13], constrains the alignment of the whole utterance to be monotonic. This makes it difficult to guarantee phoneme-by-phoneme correspondence because the source and target speakers do not speak each phoneme at the same speed.

2.1.2. Phonetic posteriorgram (PPG)-based approach

Speaker characteristics of the source speaker may leak into the VC's decoder and degrade the conversion performance. This problem is especially noticeable when there are multiple source speakers (e.g., many-to-many, any-to-many VC). A phonetic posteriorgram (PPG)-based approach is promising for removing these characteristics and extracting only linguistic information [20, 21, 22, 23]. This is done by pre-training an ASR model targeting phonemes and shared state IDs; the resulting model is used as the VC encoder. The final or bottleneck output from this ASR model is given to the decoder along with speaker information. BNE-S2SMoL-VC follows this framework, where the encoder is a hybrid CTC-attention based ASR model [24]; its bottleneck features are mapped to the target speaker's acoustic features using a mixed logistic attention decoder [25].

2.2. Speech-T: Neural transducer (RNNT) for TTS [16]

Tacotron2 [26] and TransformerTTS [27] are encoder-decoder models in TTS, and they can suffer alignment collapse. While FastSpeech2 [28] solves this problem by incorporating an explicit duration predictor, it cannot work in streaming mode. We note that Speech-T development was focused on the ability to obtain robust alignment and streaming operation of RNNT [14]; it has been utilized for TTS [16]. Regarding transition probabilities for RNNT, ASR can model them as a single categorical distribution along with a blank symbol and token labels. The blank label is used for the transition probability in the Speech-T model. However, TTS has a difficult trade-off between these transition probabilities and the generation probability of spectral, which is continuous variables. They proposed the forward algorithm for generative RNNT that separates transition probability computation and spectral prediction. Speech-T with this algorithm can synthesize natural speech without the alignment collapse observed in TransformerTTS.

3. Proposed RNNT-based VC model (VC-T)

3.1. Model architecture and forward propagation

Encouraged by RNNT's success in TTS, we propose an RNNT-based VC (VC-T) that is resists alignment collapse. Figure 1 (a) overviews VC-T; it consists of three modules, a source speech encoder, a target speech encoder, and a joint network. Figure 1 (b) and (c) depict offline and streaming source speech encoder networks[2], respectively. Figure 1 (d) and (e) illustrate the target speech encoder and the joint networks, respectively. The source speech encoder embeds the source speaker's speech into an intermediate representation. The target speech encoder also receives the spectral part of the past joint network's outputs and emits another intermediate representation. Feeding these outputs to the joint network, yields prediction of the spectra and its transition probability at the next time step. Here, we use the lazy forward algorithm as in [16], and the alignment can be obtained by using the following recurrence relation:

$$\begin{aligned}\alpha(t,u) = \alpha(t-1,u)\,\phi(t-1,u) \\ + \alpha(t,u-1)\,\{1-\phi(t,u-1)\}\,,\end{aligned} \tag{1}$$

where T and U are the number of frames in the source and target speaker's spectrum, respectively. t and u are their indices. Also, $\alpha(t,u)$ and $\phi(t,u)$ are the forward variables and transition probability on the trellis, respectively. $\phi(t,u)$ is obtained from a value preprocessed by a sigmoid function of the VC-T output vector. Each lattice of the trellis is computed according to Eq. (1), and the objective function is given by:

$$\mathcal{L} = \sum_{t=1}^{T}\sum_{u=1}^{U} \mathbb{I}\{(t,u)\in\tau\}\alpha(t,u)\{1-\phi(t,u)\}\,|\boldsymbol{y}_{u+1} - \boldsymbol{f}(t,u)|\,, \tag{2}$$

where $\boldsymbol{y}_{u+1}$ denotes the $u+1$'th frame target speaker's spectrum. $\boldsymbol{f}(t,u)$ is the t'th, and u'th predicted spectrum on the lattice generated by VC-T. Note that, as we can see $\phi(t,u)$ and $\boldsymbol{f}(t,u)$, the VC-T outputs three dimensional tensor $T\times U\times D$ in the training step. D is the number of VC-T output dimensions that concatenate the predicted transition probability and spectrum. Since filling all lattices is too computationally expensive, we omit the redundant predictive path computations except for those neighboring τ band frames that are close to the target alignment following [16]. Here $\mathbb{I}\{(t,u)\in\tau\}$ is an indicator function that tells whether or not index (t,u) is within the constraint τ band of the target alignment detailed in 3.2.

3.2. Target alignment design for VC-T

As mentioned in Section 2.1.1, ConvS2S-VC does not guarantee phoneme-by-phoneme correspondence between source and

[2]The streaming version can be built with causal convolution as is done for the target speech encoder, but we used a simple GRU in this paper.

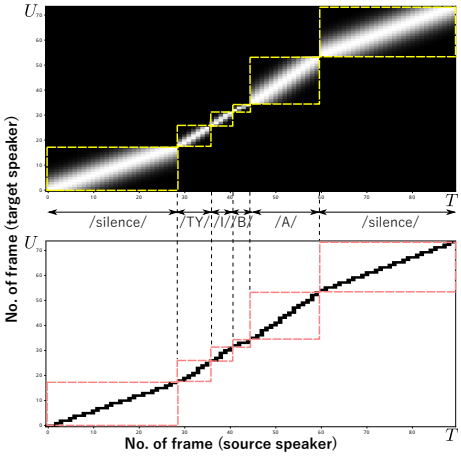

Figure 2: *Each dashed block in the upper and lower heatmaps represents actual $\boldsymbol{W}_n$ and $\boldsymbol{A}_n$, respectively. The horizontal and vertical axes correspond to the spectrum frames of the source and target speakers, respectively. Each duration in middle legends means actual phoneme segmentation, and quoted tokens by "/" mean phonemes.*

Table 1: *Our used VC models and the number of its parameters.*

Method	No. of parameters [million]
BNE-S2SMoL-VC [25]	20.8
ConvS2S-VC (offline) [12]	59.9
ConvS2S-VC (stream)	59.9
Proposed VC-T (offline)	19.7
Proposed VC-T (stream)	12.3

target speakers. Also, as stated in the previous section, the RNNT training framework requires target alignment. To address both of these issues and to implement VC-T, we design the phoneme-by-phoneme frame correspondences to be represented by a rectangular alignment matrix derived from the penalty matrix $\boldsymbol{W}_n \in \mathbb{R}^{T_n \times U_n}$ as in [13]. T_n, U_n are the frame lengths of the source and target speakers at the n'th phoneme, and each element of $\boldsymbol{W}_n$ is given by $W_n\left(t_n, u_n\right) = 1 - \exp\left\{-\left(t_n/T_n - u_n/U_n\right)^2 / 2g^2\right\}$, where t_n and u_n are the frame indices of the source and target speaker at the n'th phoneme, respectively. g is a hyperparameter set to 0.2. In this way, the n'th phoneme's alignment matrix $\boldsymbol{A}_n \in \mathbb{R}^{T_n \times U_n}$ is given by:

$$A_n\left(t_n, u_n\right) = \begin{cases} 1 & \left(t_n = \bar{t}_n\right) \\ 0 & \left(t_n \neq \bar{t}_n\right) \end{cases}, \tag{3}$$

$$\bar{t}_n = \underset{t_n}{\operatorname{argmin}} \left\{\boldsymbol{w}_n\left(u_n\right)\right\}, \tag{4}$$

where $A_n\left(\cdot\right)$ and $\boldsymbol{w}_n\left(u_n\right) \in \mathbb{R}^{T_n}$ denote $\boldsymbol{A}_n$'s element and the u_n'th row vector of $\boldsymbol{W}_n$, respectively. The upper and lower heatmaps in Fig. 2 show the actual $\boldsymbol{W}_n$ and $\boldsymbol{A}_n$ values obtained from the same training data, respectively. Incorporating such target alignments for RNNT improves phonemes mapping, which is mentioned in Section 2.1.1.

4. Experiments

4.1. Setup

For training VC models, we used multi-speaker speech data from three professional Japanese narrators, one male and two females. The sampling rate was 22.05 kHz. Each speaker had 1000 parallel utterances, forty were used as the evaluation set (about 3.3 minutes) and the rest as the training data (about 2.9 hours). We used an 80-dimensional logarithmic mel-spectrograms for the acoustic feature. The analysis frame shift was 12.5 ms.

Table 2: *Averaged mel-cepstrum distortion (MCD) and character error rate (CER). "F2F", "M2F", "F2M" denote female−to−female, male−to−female, and female−to−male scenario, respectively. The scores written in bold signify the column-wise best. Note that the evaluation utterances are different in each gender scenario.*

	MCD [dB]				CER [%]			
Method	F2F	M2F	F2M	Avg.	F2F	M2F	F2M	Avg.
GT	-	-	-	-	12.9	14.1	14.5	14.0
Resyn	-	-	-	-	13.2	14.8	15.0	14.5
BNE-S2SMoL-VC	5.7	5.5	5.9	5.7	25.6	23.8	26.8	25.4
ConvS2S-VC (offline)	**4.9**	**5.0**	**5.2**	**5.1**	16.9	21.3	24.4	21.4
ConvS2S-VC (stream)	5.5	5.5	5.8	5.6	16.9	27.4	28.5	25.2
VC-T (offline)	5.1	**5.0**	5.2	5.2	**14.3**	**16.7**	**17.9**	**16.6**
VC-T (stream)	5.3	5.1	5.3	5.3	14.4	18.5	19.4	17.9

Three types of one-to-one VC models were trained for each method: homo-gender conversion (Female2-to-Female1), hetero-gender ones (Male1-to-Female1 and Female1-to-Male1). Our proposed method was implemented as two variants; an offline model with non-causal transformer encoders and a streaming model with simple GRU encoders as shown in Fig. 1 (b) and (c), respectively. For designing the alignment described in Section 3.2, we used manually annotated phoneme labels, and set τ to 1. As the conventional methods, ConvS2S-VC [12][3] and BNE-S2SMoL-VC [25][4], mentioned in Section 2.1, were employed. ConvS2S-VC was built not only in its offline version, but also in a streaming version with zero look-ahead, i.e., the causal encoder[5]. These models were optimized by Adam [29] in 200k steps with a batch size of 16 by following [30]'s learning rate schedule. Note that BNE-S2SMoL-VC has an encoder that predicts 56 kinds of phonemes. We trained it in 3000k steps with the same learning rate schedule as the VC in advance, using data from our 1,050 internal Japanese speakers (about 312.9 hours), including VC's training data. Afterwards, only its VC decoder was trained under the same conditions as the other VCs, with the encoder weights frozen. For waveform generation from spectrograms, we used the speaker-independent HiFi-GAN vocoder (v1) [31] [6], which was trained on same data as BNE-S2SMoL-VC. Table 1 shows the number of the above model parameters. Note that we did not investigate streaming BNE-S2SMoL-VC because it adopted a naive encoder-decoder that cannot run in a streaming manner.

4.2. Objective evaluations

We objectively evaluated each VC's spectral reconstruction error and linguistic information correctness by the metrics of mel-cepstrum distortion (MCD) and character error rate (CER), respectively. MCD was calculated by converting the mel-spectrogram from VC into a 40-dimensional mel-cepstrum and then aligning the sequence length with that of the target speaker's natural speech by DTW. To evaluate CER, we used the wav2vec2.0 ASR model [32][7] [8]. The ASR model was fed

[3] https://github.com/kamepong/ConvS2S-VC.git

[4] https://github.com/liusongxiang/ppg-vc.git

[5] Although by forcing the attention matrix to be diagonal, the output frame length of the target speaker can be equal to that of the source speaker, we adopted the predicted raw attention as well as the other methods.

[6] https://github.com/jik876/hifi-gan.git

[7] https://huggingface.co/ttop324/wav2vec2-live-japanese

[8] The model has 100 vocabulary entries including blank symbols, Japanese kana, long vowel symbols, and alphabetic characters.

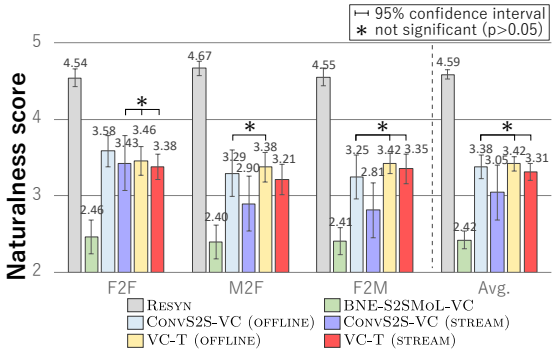

Figure 3: *Subjective evaluation results of naturalness.*

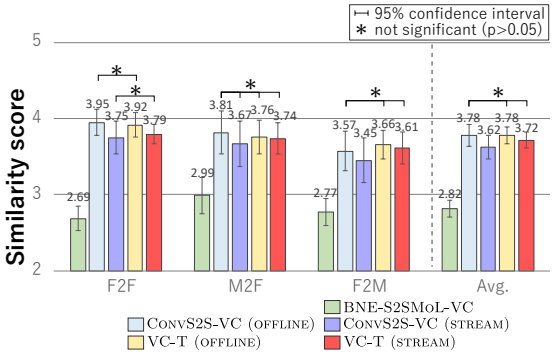

Figure 4: *Subjective evaluation results of speaker similarity.*

with natural or vocoded speech, and a greedy search was applied to the obtained logits to yield the recognition results.

Table 2 lists the objective evaluation results. As oracles, CERs for the ground truth speech (GT) and the analytic resynthesized one by HiFi-GAN (RESYN) were also calculated. First, RESYN did not trigger any remarkable decreases in CER, confirming that the vocoder posed no critical impediment to a linguistic content. Next, BNE-S2SMOL-VC has similar MCD and CER scores regardless of homo- and hetero-gender conversion, because the encoder removes the speaker characteristic of the source speaker. However, since no constraint is applied to make the alignment monotonic, speech dropouts due to alignment skips were often observed, and the MCD and CER scores were the worst among all methods. CONVS2S-VC (OFFLINE) scored better than this in both MCD and CER. Its streaming version, CONVS2S-VC (STREAM), is still better than BNE-S2SMOL-VC, while its performance is degraded compared to CONVS2S-VC (OFFLINE). However, although less frequent than in BNE-S2SMOL-VC, speech dropouts were still observed, which compromised the CER, especially in the hetero-gender cases. On the other hand, VC-T (OFFLINE) attained significantly better CER, even albeit the average MCD of the three models was slightly worse than that of CONVS2S-VC (OFFLINE). The proposed method achieved robust alignment, as there were no speech skips in the evaluation data. The MCD and CER of VC-T (STREAM) were worse than those of VC-T (OFFLINE), but the degree of deterioration was smaller and indeed superior to those of the same streaming model, CONVS2S-VC (STREAM). We performed the MAPSSWE significance test [33], and the differences of the CERs between CONVS2S-VC and VC-T in offline and streaming modes were statistically significant, $p < 0.001$. Thanks to their robust alignments generated from our VC-T, it could better convert source speaker's speech to target speaker's one than that of the CONVS2S-VC while preserving linguistic information. Moreover, our VC-T achieved the above results although the model size was much smaller than CONVS2S-VC (see Table 1).

4.3. Subjective evaluations

We subjectively evaluated the naturalness of converted speech, including RESYN. Seventeen listeners participated in the test, and the evaluation used a mean opinion score (MOS) on a five-point scale ranging from 5: very natural to 1: very unnatural. Eight sentences were randomly selected for each VC's gender setting, with a total of 108 utterances across all methods.

Figure 3 shows the naturalness evaluation results. While RESYN had a very high score, VC methods received lower scores because the spectra were degraded from those of the original speech. In particular, BNE-S2SMOL-VC scored the worst among all methods. As indicated by the CER evaluation in the previous section, this was due to collapsed alignments leading to unclear speech. CONVS2S-VC (OFFLINE) outperformed it, but scored slightly lower than homo-gender conversions, possibly owing to increased difficulty in hetero-gender conversion. Its streaming version, CONVS2S-VC (STREAM), was found to be inferior to CONVS2S-VC (OFFLINE), especially in hetero-gender conversions. We suspect this was because streaming operation suffers if future context is missing, since CONVS2S-VC only depends on the source speaker's speech unlike our VC-T. Thus, CONVS2S-VC (STREAM) led to produce unclear spectra due to unstable alignments, got a lager confidence interval than that of the others. Contrary to CONVS2S-VC (STREAM), VC-T (STREAM) exhibited no remarkable degradation compared to VC-T (OFFLINE). These results reveal that VC-T also works robustly in streaming mode thanks to its robust alignment, which again reflects RNNT's strength.

The similarity of converted speech to the target speaker was compared to that of the reference GT by using degradation mean opinion score (DMOS) using a five-point scale ranging from 5: very similar to 1: very dissimilar. Participants and evaluation utterances were same as the naturalness evaluation. Figure 4 presents the subjective evaluation results of speaker similarity. The overall tendency was similar to that found in the naturalness evaluation, with BNE-S2SMOL-VC exhibiting the worst speaker similarity in all gender conditions. CONVS2S-VC (STREAM) exhibited significant degradation. On the other hands, VC-T (OFFLINE) roughly matched CONVS2S-VC (OFFLINE). Unlike CONVS2S-VC (STREAM), VC-T (STREAM) was close to VC-T (OFFLINE), with almost no speaker similarity degradation. These overall results demonstrate that our proposal, VC-T, can attain robust alignments and improved naturalness and speaker similarity, especially under severe conditions such as streaming inferencing and heterosexual conversions.

5. Conclusions

This work proposed a novel RNNT-based parallel VC, i.e., VC-T, to obtain robust alignment. We also presented an alignment design method that allows RNNT training to be used in VC. We showed that the proposed VC achieved better CER than the conventional seq2seq VC as well as contributing to the preservation of speech content. Subjective evaluations also showed that the proposed method achieved better naturalness in streaming mode while achieving comparable speaker similarity to the conventional streaming ConvS2S-VC. Our future works include 1) improving VC-T performance with a pretraining approach [34], 2) extending the proposed method to many-to-many VC using target-speaker embedding [35, 36] and 3) evaluating speed enhancement by utilizing a cache [37] or downsampling [38].

6. References

[1] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, "Voice conversion through vector quantization," *Journal of the Acoustical Society of Japan (E)*, vol. 11, no. 2, pp. 71–76, 1990.

[2] A. Kain and M. W. Macon, "Spectral voice conversion for text-to-speech synthesis," *Proc. ICASSP*, pp. 285–288, 1998.

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

[4] Y.-C. W. Y. T. Chin-Cheng Hsu, Hsin-Te Hwang and H. Wang, "Voice conversion from non-parallel corpora using variational auto-encoder," *Proc. APSIPA*, pp. 1–6, 2016.

[5] Y. Li, K. A. Lee, Y. Yuan, H. Li, and Z. Yang, "Many-to-many voice conversion based on bottleneck features with variational autoencoder for non-parallel training data," *Proc. APSIPA*, pp. 829–833, 2018.

[6] K. Qian, Y. Zhang, S. Chang, X. Yang, and M. Hasegawa-Johnson, "AutoVC: Zero-shot voice style transfer with only autoencoder loss," *Proc. ICML*, pp. 5210–5219, 2019.

[7] T. Kaneko and H. Kameoka, "CycleGAN-VC: Non-parallel voice conversion using cycle-consistent adversarial networks," *Proc. EUSIPCO*, pp. 2100–2104, 2018.

[8] H. Kameoka, T. Kaneko, K. Tanaka, and N. Hojo, "StarGAN-VC: Non-parallel many-to-many voice conversion using star generative adversarial networks," *Proc. SLT*, pp. 266–273, 2018.

[9] B. Nguyen and F. Cardinaux, "NVC-Net: End-to-end adversarial voice conversion," *Proc. ICASSP*, pp. 7012–7016, 2022.

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

[11] K. L. H. M. Lifa Sun, Shiyin Kang, "Voice conversion using deep bidirectional long short-term memory based recurrent neural networks," *Proc. ICASSP*, pp. 4869–4873, 2015.

[12] H. Kameoka, K. Tanaka, D. Kwaśny, T. Kaneko, and N. Hojo, "ConvS2S-VC: Fully convolutional sequence-to-sequence voice conversion," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 28, pp. 1849–1863, 2020.

[13] H. Tachibana, K. Uenoyama, and S. Aihara, "Efficiently trainable text-to-speech system based on deep convolutional networks with guided attention," *Proc. ICASSP*, 2018.

[14] A. Graves, "Sequence transduction with recurrent neural networks," *Proc. ICML Representation Learning Workshop*, 2012.

[15] A. Gulati, J. Qin, C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, "Conformer: Convolution-augmented transformer for speech recognition," *Proc. Interspeech*, pp. 5036–5040, 2020.

[16] J. Chen, X. Tan, Y. Leng, J. Xu, G. Wen, T. Qin, and T.-Y. Liu, "Speech-T: Transducer for text to speech and beyond," *Proc. NeurIPS*, vol. 34, pp. 6621–6633.

[17] K. Tanaka, H. Kameoka, T. Kaneko, and N. Hojo, "AttS2S-VC: Sequence-to-sequence voice conversion with attention and context preservation mechanisms," *Proc. ICASSP*, pp. 6805–6809, 2019.

[18] J. Gehring, M. Auli, D. Grangier, D. Yarats, and Y. N. Dauphin, "Convolutional sequence to sequence learning," *Proc. ICML*, pp. 1243–1252, 2017.

[19] K. Kobayashi and T. Toda, "sprocket: Open-source voice conversion software." *Proc. Odyssey*, pp. 203–210, 2018.

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

[21] H. Miyoshi, Y. Saito, S. Takamichi, and H. Saruwatari, "Voice conversion using sequence-to-sequence learning of context posterior probabilities," *Proc. Interspeech*, pp. 1268–1272, 2017.

[22] Y. Saito, Y. Ijima, K. Nishida, and S. Takamichi, "Non-parallel voice conversion using variational autoencoders conditioned by phonetic posteriorgrams and d-vectors," *Proc. ICASSP*, pp. 5274–5278, 2018.

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

[24] S. Watanabe, T. Hori, S. Kim, J. R. Hershey, and T. Hayashi, "Hybrid CTC/attention architecture for end-to-end speech recognition," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1240–1253, 2017.

[25] S. Liu, Y. Cao, D. Wang, X. Wu, X. Liu, and H. Meng, "Any-to-many voice conversion with location-relative sequence-to-sequence modeling," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 29, pp. 1717–1728, 2021.

[26] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerry-Ryan, R. A. Saurous, Y. Agiomyrgiannakis, and Y. Wu, "Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions," *Proc. ICASSP*, pp. 4779–4783, 2018.

[27] N. Li, S. Liu, Y. Liu, S. Zhao, and M. Liu, "Neural speech synthesis with transformer network," *Proc. AAAI*, vol. 33, no. 01, pp. 6706–6713, 2019.

[28] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "Fastspeech 2: Fast and high-quality end-to-end text to speech," *Proc. ICLR*, 2021.

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

[30] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *Proc. NIPS*, vol. 33, 2017.

[31] J. Kong, J. Kim, and J. Bae, "HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis," *Proc. NeurIPS*, vol. 33, pp. 17 022–17 033, 2020.

[32] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Proc. NeurIPS*, vol. 33, pp. 12 449–12 460, 2020.

[33] L. Gillick and S. Cox, "Some statistical issues in the comparison of speech recognition algorithms," *Proc. ICASSP*, pp. 532–535, 1989.

[34] T. Moriya, T. Ashihara, T. Tanaka, T. Ochiai, H. Sato, A. Ando, Y. Ijima, R. Masumura, and Y. Shinohara, "Simpleflat: A simple whole-network pre-training approach for rnn transducer-based end-to-end speech recognition," *Proc. ICASSP*, pp. 5664–5668, 2021.

[35] T. Moriya, H. Sato, T. Ochiai, M. Delcroix, and T. Shinozaki, "Streaming target-speaker ASR with neural transducer," *Proc. Interspeech*, pp. 2673–2677, 2022.

[36] ——, "Streaming end-to-end target-speaker automatic speech recognition and activity detection," *IEEE Access*, vol. 11, pp. 13 906–13 917, 2023.

[37] X. Chen, Y. Wu, Z. Wang, S. Liu, and J. Li, "Developing real-time streaming transformer transducer for speech recognition on large-scale dataset," *Proc. ICASSP*, pp. 5904–5908, 2021.

[38] G. Pundak and T. N. Sainath, "Lower frame rate neural network acoustic models," *Proc. Interspeech*, pp. 22–26, 2016.