ATTS2S-VC: SEQUENCE-TO-SEQUENCE VOICE CONVERSION WITH ATTENTION AND CONTEXT PRESERVATION MECHANISMS

Kou Tanaka, Hirokazu Kameoka, Takuhiro Kaneko, Nobukatsu Hojo

NTT Communication Science Laboratories, NTT Corporation, Japan

{tanaka.ko, kameoka.hirokazu, kaneko.takuhiro, hojo.nobukatsu}@lab.ntt.co.jp

ABSTRACT

This paper describes a method based on a sequence-to-sequence learning (Seq2Seq) with attention and context preservation mechanism for voice conversion (VC) tasks. Seq2Seq has been outstanding at numerous tasks involving sequence modeling such as speech synthesis and recognition, machine translation, and image captioning. In contrast to current VC techniques, our method 1) stabilizes and accelerates the training procedure by considering guided attention and proposed context preservation losses, 2) allows not only spectral envelopes but also fundamental frequency contours and durations of speech to be converted, 3) requires no context information such as phoneme labels, and 4) requires no time-aligned source and target speech data in advance. In our experiment, the proposed VC framework can be trained in only one day, using only one GPU of an NVIDIA Tesla K80, while the quality of the synthesized speech is higher than that of speech converted by Gaussian mixture model-based VC and is comparable to that of speech generated by recurrent neural network-based text-to-speech synthesis, which can be regarded as an upper limit on VC performance.

Index Terms— Voice conversion, deep learning, sequence-to-sequence, attention mechanism, context preservation mechanism

1. INTRODUCTION

Voice conversion (VC) systems aim to convert para/non-linguistic information included in a given speech waveform while preserving its linguistic information. VC has been applied to various tasks, such as speaker conversion [1–3] for impersonating or hiding a speaker's identity, as a speaking aid [4, 5] for overcoming speech impairments, as a style conversion [6, 7] for controlling speaking styles including emotion, and for pronunciation/accent conversion [8, 9] in language learning.

A popular form of VC is a statistical one based on a Gaussian mixture model (GMM) [10]; it requires time-aligned parallel data of the source and target speech for training the conversion models. For frameworks requiring time-aligned parallel data, other researchers have proposed exemplar-based VCs using non-negative matrix factorization (NMF) [11, 12] and neural network (NN)-based VCs using restricted Boltzmann machines [13, 14], feed-forward NNs [15, 16], recurrent NNs [17, 18], variational autoencoders [19, 20], and generative adversarial nets [9]. On the other hand, frameworks requiring no parallel data, called parallel-data-free VCs, have been proposed [1, 3] to avoid the time-consuming job of recording speech for parallel data collection. Notably, the drawbacks of these VCs are the prerequisite of a large number of transcripts and/or difficulty converting the durations of the source speech.

Recently, sequence-to-sequence (Seq2Seq) learning [21, 22] has proved to be outstanding at various research tasks such as text-to-speech synthesis (TTS) [23–25] and automatic speech recognition (ASR) [26, 27]. The early Seq2Seq model [21] has encoder and decoder architectures for mapping an input sequence to an encoded representation used by the decoder network to generate an output sequence. To select critical information from the encoded representation in accordance with the output sequence representation, later Seq2Seq models [22, 28] introduce an attention mechanism. The key advantages of the Seq2Seq learning approach are the ability to train a single end-to-end model directly on the source and target sequences and the capacity to handle input and output sequences of different lengths. In particular, we expect that the Seq2Seq model makes it possible to convert not only acoustic features but also the durations of the source speech to those of the target speech. Moreover, Seq2Seq learning is extensible to semi-supervised learning [29], where it can avoid the time-consuming task of collecting parallel data. In a supervised learning task, Seq2Seq learning requires parallel data of the source and target sequences rather than time-aligned parallel data. Considering dual learning [30, 31], Seq2Seq learning can be trained with a small amount of parallel data and a large amount of non-parallel data.

In this paper, we propose a Seq2Seq-based VC with attention and context preservation mechanisms[1]. Our contributions are as follows:

- Our VC method makes it possible to stabilize and accelerate the training procedure by considering guided attention and context preservation losses.
- It makes it possible to convert not only spectral envelopes but also fundamental frequency contours and durations of the speech.
- It requires no context information such as phoneme labels, unlike [32, 33] which introduced the Seq2Seq model and used context information.

[1] Audio samples can be accessed on our web page: http://www.kecl.ntt.co.jp/people/tanaka.ko/projects/atts2svc/attention_based_seq2seq_voice_conversion.html

- It requires no time-aligned source or target speech data in advance.

We conducted an our experiment demonstrating that the quality of the synthesized speech generated by our VC framework is higher than that of speech generated by the conventional GMM-based VC, and it is comparable to that of speech generated by recurrent-NN based TTS in terms of both naturalness and speaker similarity. Note that the proposed model was trained in only one day, using only one GPU of an NVIDIA Tesla K80.

2. CONVENTIONAL VC

2.1. Frame/Sequence- based VC

There are two types of frame/sequence- based VC: VCs requiring parallel data [10, 34, 35] and parallel-data-free VC [1, 3]. The first framework has different procedures of training and conversion, as shown in Fig. 1a. The conversion procedure does not have a time warping function, despite that the training procedure includes a time-alignment step to handle source and target sequences having different lengths. The second framework is a parallel-data-free VC that does not require parallel source and target speech data. To realize parallel-data-free VC, the second framework uses context information [36, 37], adaptation techniques [38, 39], a pre-constructed speaker space [40, 41], and cycle consistency [1, 3]. Although these VC techniques have various training procedures, the conversion procedure does not involve the time warping function. Consequently, the frame/sequence- based VC frameworks do not allow us to convert the durations and the acoustic features of the source speech at the same time. In contrast, our model allows both the acoustic features and the durations to be converted at the same time.

2.2. Seq2Seq-based VC

In contrast to the frame/sequence- based VC frameworks, Seq2Seq-based VC frameworks make it possible to convert not only the acoustic features but also the durations of the source speech. Most Seq2Seq-based VCs consist of ASR and TTS modules which are trainable with pairs of speech and its transcript rather than the source and target speech. The ASR module converts the acoustic feature sequence of the source speech into a sequence of context information such as phoneme labels and context posterior probabilities [32, 37], and the TTS module generates the acoustic feature sequence of the desired speech from the sequence of context information. One approach to changing the duration of the source speech uses a re-generation method that generates the duration information from the text symbols after converting the acoustic features of the source speech into text symbols once. Namely, the duration information is erased once and re-generated. Another approach [32, 33] involves Seq2Seq learning, as shown in Fig. 1b. In this approach, the context posterior probability sequence of the source speech including the duration information is directly converted into a context posterior probability sequence of the desired speech including the duration information. Both approaches work well if ASR performs robustly and accurately enough, but they require a large number of transcripts to train each module. In contrast, our model does not use any transcript.

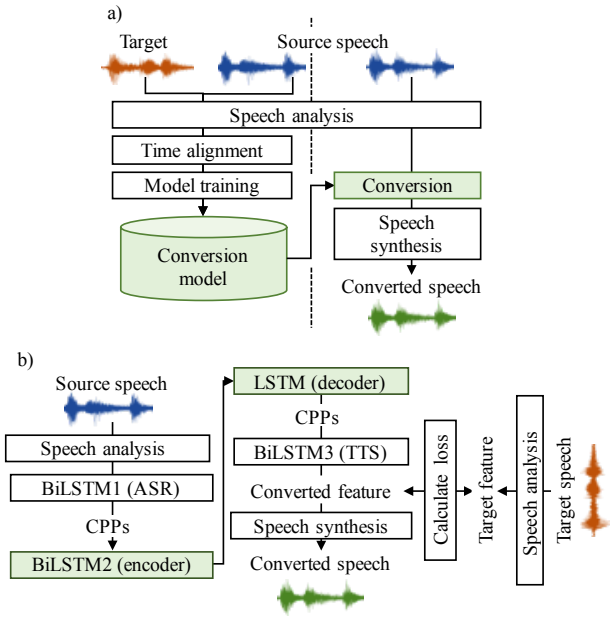

Fig. 1. System overviews of conventional VC, a) frame/sequence- based VC using parallel data (see Sec. 2.1) and b) Seq2Seq-based VC (see Sec. 2.2). "CPPs" and "BiLSTM" denote context posterior probabilities and bidirectional LSTM, respectively.

3. ATTS2S-VC

Our method consists of 1) four basic components of the Seq2Seq model and 2) two additional components as a context preservation mechanism. The four basic components are a source encoder, target encoder, target autoregressive (AR) decoder, and attention mechanism. The two additional components are a source decoder and another target decoder to keep linguistic information of the source speech. Figure 2 is an overview of the system.

3.1. Seq2Seq Model with Attention Mechanism

Let us use $\boldsymbol{X} = [\boldsymbol{x}_1, \cdots, \boldsymbol{x}_I]$ and $\boldsymbol{Y} = [\boldsymbol{y}_1, \cdots, \boldsymbol{y}_J]$ to denote sequences of acoustic features of the source and target speech, respectively. The source encoder network f_{SrcEnc} and target encoder network f_{TarEnc} encode the input sequences $\boldsymbol{X}$ and $\boldsymbol{Y}$ to the embeddings $\boldsymbol{K} = [\boldsymbol{k}_1, \cdots, \boldsymbol{k}_I]$ and $\boldsymbol{Q} = [\boldsymbol{q}_1, \cdots, \boldsymbol{q}_J]$, as follows:

$$\boldsymbol{K} = f_{\mathrm{SrcEnc}}(\boldsymbol{X}), \quad (1)$$

$$\boldsymbol{Q} = f_{\mathrm{TarEnc}}(\boldsymbol{Y}). \quad (2)$$

In order to accurately predict the output sequence, [22,28] introduced an attention mechanism. At each time frame of the embeddings $\boldsymbol{Q}$, the attention mechanism gives a probability distribution that describes the relationship between the given

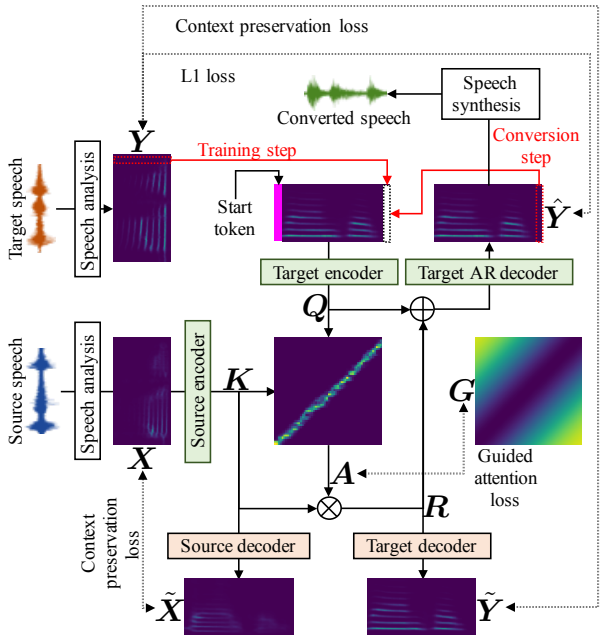

Fig. 2. System overviews of proposed VC. The solid lines indicate the training and conversion procedures. The dashed lines indicate calculations of the differences during training. Green boxes and red boxes respectively denote the original components and the proposed additional components.

time frame feature $\boldsymbol{q}_j$ and the embeddings $\boldsymbol{K}$. Consequently, the attention matrix $\boldsymbol{A}$ can be written as

$$e_{i,j} = f_{\mathrm{FFNN}}(\boldsymbol{k}_i, \boldsymbol{q}_j), \tag{3}$$

$$a_{i,j} = \frac{\exp(e_{i,j})}{\sum_i \exp(e_{i,j})}. \tag{4}$$

where f_{FFNN} indicates a function described by feed-forward NNs and $a_{i,j}$ is an element (i, j) of the attention matrix $\boldsymbol{A}$.

A seed $\boldsymbol{R} = [\boldsymbol{r}_1, \cdots, \boldsymbol{r}_J]$ of the target AR decoder is obtained by considering the long-range temporal dependencies between the source and target sequences as follows:

$$\boldsymbol{r}_j = \boldsymbol{K}\boldsymbol{a}_j. \tag{5}$$

As the name implies, the target AR decoder involves all previous outputs of itself. Hence, the input of the target AR decoder is $\boldsymbol{R}'$ combined with the seed $\boldsymbol{R}$ and the embeddings $\boldsymbol{Q}$. The output $\hat{\boldsymbol{Y}}$ of the Seq2Seq model is obtained through the target AR decoder f_{TarDecAR},

$$\hat{\boldsymbol{Y}} = f_{\mathrm{TarDecAR}}(\boldsymbol{R}'). \tag{6}$$

Finally, we minimize the objective function $\mathcal{L}_{\mathrm{Seq2Seq}}$ of Seq2Seq learning:

$$\mathcal{L}_{\mathrm{Seq2Seq}} = ||\hat{\boldsymbol{Y}} - \boldsymbol{Y}||_1. \tag{7}$$

3.2. Stabilizing and Accelerating Training Procedure

3.2.1. Guided Attention Loss

To accelerate the training of an attention module, [25] introduced a guided attention loss. Generally speaking, most speech signal processing applications, such as ASR, TTS, and VC, are time incremental algorithms. It is natural to assume that the time frame i of the source speech waveform progresses nearly linearly with respect to the time frame j of the target speech waveform, i.e., $i \sim \alpha j$, where $\alpha \sim \frac{I}{J}$. Therefore, the attention matrix $\boldsymbol{A}$ should be a nearly diagonal. A penalty matrix $\boldsymbol{G}$ is designed, as follows:

$$g_{i,j} = 1 - \exp\left\{\frac{-(\frac{i}{I} - \frac{j}{J})^2}{2\sigma_g^2}\right\}, \tag{8}$$

where σ_g controls how close $\boldsymbol{A}$ is to a diagonal matrix. The guided attention loss $\mathcal{L}_{\mathrm{ga}}$ is defined as

$$\mathcal{L}_{\mathrm{ga}} = ||\boldsymbol{G} \odot \boldsymbol{A}||_1, \tag{9}$$

where $\odot$ indicates an element-wise product.

3.2.2. Context Preservation Loss

To stabilize the training procedure, we propose a context preservation loss. In preliminary experiments, we found that the training procedure sometimes failed even if it took into account the guided attention loss (see speech samples on our web page 1). In particular, the converted speech seemed like randomly generated speech or speech repeating several phonemes. One possible reason is that minimizing the objective function $\mathcal{L}_{\mathrm{Seq2Seq}}$ sometimes makes the target AR decoder a network just reconstructing the input of the target encoder. It is because we use $\boldsymbol{Y}$ rather than $\hat{\boldsymbol{Y}}$ as the input of the target encoder in the training. As a result, the source encoder is not required to control the output of the target AR decoder and preserve the context information of the source speech.

To make the source encoder meaningful, we introduce two additional networks to the original Seq2Seq model as a context preservation mechanism. One is a source decoder f_{SrcDec} for reconstructing the source speech $\tilde{\boldsymbol{X}}$ from the embeddings $\boldsymbol{K}$. The other is a target decoder f_{TarDec} for predicting the target speech $\tilde{\boldsymbol{Y}}$ from the seed $\boldsymbol{R}$.

$$\tilde{\boldsymbol{X}} = f_{\mathrm{SrcDec}}(\boldsymbol{K}), \tag{10}$$

$$\tilde{\boldsymbol{Y}} = f_{\mathrm{TarDec}}(\boldsymbol{R}). \tag{11}$$

From another point of view, the source decoder f_{SrcDec} helps the source encoder to preserve the linguistic information of the source speech $\boldsymbol{X}$, while the target decoder f_{TarDec} helps the source encoder to encode the source speech $\boldsymbol{X}$ to the shared space of the source and target speech. Note that in the preliminary experiments, the target decoder was more important than the source decoder. The full objective function of our model is formulated as

$$\begin{aligned}\mathcal{L}_{\mathrm{proposed}} = \mathcal{L}_{\mathrm{Seq2Seq}} + \lambda_{\mathrm{ga}}\mathcal{L}_{\mathrm{ga}} \\ + \lambda_{\mathrm{cp}}(||\tilde{\boldsymbol{X}} - \boldsymbol{X}||_1 + ||\tilde{\boldsymbol{Y}} - \boldsymbol{Y}||_1),\end{aligned} \tag{12}$$

where λ_{cp} controls the context preservation loss.

4. EXPERIMENTS

4.1. Experimental Conditions

Datasets: We used the CMU Arctic database [42] consisting of utterances by two male speakers (**rms** and **bdl**) and two female speakers (**clb** and **slt**). To train the models, we used about 1,000 sentences (speech section of 50 min) of each speaker. To evaluate the performance, we used 132 sentences of each speaker. The sampling rate of the speech signals was 16 kHz. We treated **rms** and **clb** as source speakers and **bdl** and **slt** as target speakers. For the evaluations, we conducted intra-gender pairs, **rms**-**bdl** and **clb**-**slt**, and cross-gender pairs, **rms**-**slt** and **clb**-**bdl**. Note that we trained the conversion models for every speaker pair, independently.

Baseline system 1 (GMM-VC-wGV): We used a GMM-based VC method [10] as the baseline for frame/sequence-based VC described in Sec. 2.1. To train the conversion models, we used an open source VC toolkit sprocket [43] and its default settings, except for F_0 ranges and power thresholds. Note that a global variance (GV) [10] was also considered.

Baseline system 2 (LSTM-TTS): By assuming the ASR module and the encoder part of the encoder-decoder module in [32] work perfectly, we can focus on the TTS module. Therefore, we used an LSTM-based TTS method as the baseline of Seq2Seq-based VC described in Sec. 2.2. The contextual features used as input were 416-dimensional linguistic features obtained using the default question set of the open source TTS toolkit Merlin [44]. From the speech data, 60 Mel-cepstral coefficients, logarithmic F_0, and coded aperiodicities were extracted every 5 ms with the WORLD analysis system [45]. As the duration model, we stacked three LSTMs with 256 cells followed by a linear projection. As the acoustic model, we stacked three bidirectional LSTMs with 256 cells followed by a linear projection.

Proposed system (Proposed): Inspired by Tacotron [23], we used the architecture described in open Tacotron [46]. Note that we replaced all ReLU activations [47] with a gating mechanism of gated linear units [48]. Although the proposed method worked well for not only acoustic features of the WORLD vocoder but also raw spectral features, we chose to use the acoustic features of WORLD vocoder to balance the experimental conditions of **LSTM-TTS**. Note that the target AR decoder also generated the stop tokens. As the additional source decoder and target decoder networks, we used the same architectures as in the source encoder. The hyperparameters σ_g, λ_{ga}, λ_{cp} were 0.4, 10,000, and 10, respectively. The batch size, number of epochs, and reduction factor [49] were 32, 1,000 and 5. We used the Adam optimizer [50] and varied the learning rate over the course of training [51].

4.2. Experimental Results

As shown in Fig. 3, we conducted two subjective evaluations, preference tests on naturalness and speaker similarity. The number of listeners was 15, and each listener evaluated 80 shots consisting of randomly selected 10 speech samples $\times$ 4 pairs of intra/cross- gender $\times$ 2 comparisons, v.s. **GMM-VC-wGV** and v.s. **LSTM-TTS**.

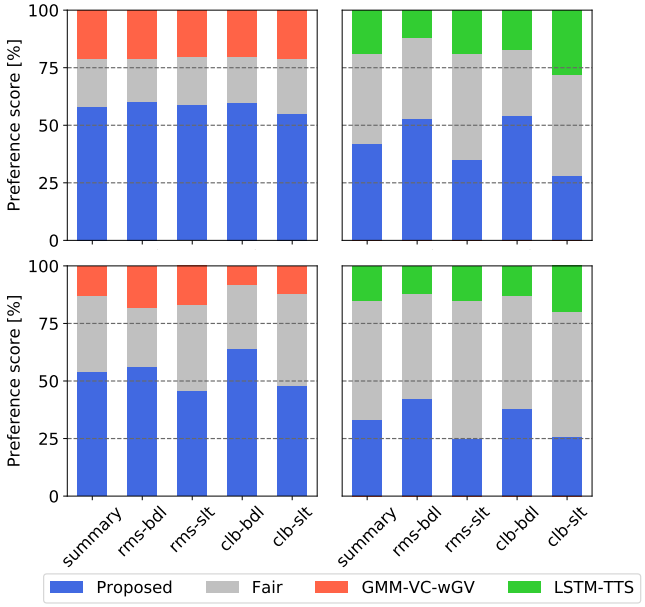

Fig. 3. Results of preference tests of naturalness (upper) and speaker similarity (lower).

The evaluations indicated that **Proposed** outperformed **GMM-VC-wGV** in terms of both naturalness and speaker similarity. This is because our method makes it possible to convert not only the acoustic features but also the durations of speech. In contrast, baseline system 1 forces the conversion while preserving the durations of the source speech. Consequently, durations not used in the target speech make the conversion errors larger.

Moreover, **Proposed** was comparable to **LSTM-TTS**. This result demonstrates that our method makes it possible to learn the key components for changing the individuality of the speaker while preserving the linguistic information. Notably, our model was trained without any transcript while [32, 33] used a large number of transcripts.

5. CONCLUSIONS

We proposed a method based on Seq2Seq learning with attention and context preservation mechanisms for VC tasks. Experimental results demonstrated that the proposed method outperformed the conventional GMM-based VC and was comparable to LSTM-based TTS. Extending the proposed method so that it can be used in semi-supervised learning tasks is ongoing work. Note that since we also progressed in a convolutional version of the proposed method [52] simultaneously, we will conduct further evaluations and report the results.

Acknowledgements: This work was supported by a grant from the Japan Society for the Promotion of Science (JSPS KAKENHI 17H01763).

6. REFERENCES

[1] Takuhiro Kaneko and Hirokazu Kameoka, "Parallel-data-free voice conversion using cycle-consistent adversarial networks," *arXiv preprint arXiv:1711.11293*, 2017. 1, 2

[2] Yang Gao, Rita Singh, and Bhiksha Raj, "Voice impersonation using generative adversarial networks," *arXiv preprint arXiv:1802.06840*, 2018. 1

[3] Hirokazu Kameoka, Takuhiro Kaneko, Kou Tanaka, and Nobukatsu Hojo, "StarGAN-VC: Non-parallel many-to-many voice conversion with star generative adversarial networks," *arXiv preprint arXiv:1806.02169*, 2018. 1, 2

[4] Alexander B Kain, John-Paul Hosom, Xiaochuan Niu, Jan PH van Santen, Melanie Fried-Oken, and Janice Staehely, "Improving the intelligibility of dysarthric speech," *Speech communication*, vol. 49, no. 9, pp. 743–759, 2007. 1

[5] Keigo Nakamura, Tomoki Toda, Hiroshi Saruwatari, and Kiyohiro Shikano, "Speaking-aid systems using GMM-based voice conversion for electrolaryngeal speech," *Speech Communication*, vol. 54, no. 1, pp. 134–146, 2012. 1

[6] Zeynep Inanoglu and Steve Young, "Data-driven emotion conversion in spoken english," *Speech Communication*, vol. 51, no. 3, pp. 268–283, 2009. 1

[7] Tomoki Toda, Mikihiro Nakagiri, and Kiyohiro Shikano, "Statistical voice conversion techniques for body-conducted unvoiced speech enhancement," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 20, no. 9, pp. 2505–2517, 2012. 1

[8] Daniel Felps, Heather Bortfeld, and Ricardo Gutierrez-Osuna, "Foreign accent conversion in computer assisted pronunciation training," *Speech communication*, vol. 51, no. 10, pp. 920–932, 2009. 1

[9] Takuhiro Kaneko, Hirokazu Kameoka, Kaoru Hiramatsu, and Kunio Kashino, "Sequence-to-sequence voice conversion with similarity metric learned using generative adversarial networks," *Proc. Interspeech 2017*, pp. 1283–1287, 2017. 1

[10] Tomoki Toda, Alan W Black, and Keiichi Tokuda, "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 8, pp. 2222–2235, 2007. 1, 2, 4

[11] Ryoichi Takashima, Tetsuya Takiguchi, and Yasuo Ariki, "Exemplar-based voice conversion using sparse representation in noisy environments," *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, vol. 96, no. 10, pp. 1946–1953, 2013. 1

[12] Zhizheng Wu, Tuomas Virtanen, Eng Siong Chng, and Haizhou Li, "Exemplar-based sparse representation with residual compensation for voice conversion," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 10, pp. 1506–1521, 2014. 1

[13] Ling-Hui Chen, Zhen-Hua Ling, Li-Juan Liu, and Li-Rong Dai, "Voice conversion using deep neural networks with layer-wise generative training," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 22, no. 12, pp. 1859–1872, 2014. 1

[14] Toru Nakashika, Tetsuya Takiguchi, and Yasuo Ariki, "Voice conversion based on speaker-dependent restricted boltzmann machines," *IEICE TRANSACTIONS on Information and Systems*, vol. 97, no. 6, pp. 1403–1410, 2014. 1

[15] Srinivas Desai, Alan W Black, B Yegnanarayana, and Kishore Prahallad, "Spectral mapping using artificial neural networks for voice conversion," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 5, pp. 954–964, 2010. 1

[16] Yuki Saito, Shinnosuke Takamichi, and Hiroshi Saruwatari, "Voice conversion using input-to-output highway networks," *IEICE Transactions on Information and Systems*, vol. 100, no. 8, pp. 1925–1928, 2017. 1

[17] Lifa Sun, Shiyin Kang, Kun Li, and Helen Meng, "Voice conversion using deep bidirectional long short-term memory based recurrent neural networks," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 4869–4873. 1

[18] Toru Nakashika, Tetsuya Takiguchi, and Yasuo Ariki, "High-order sequence modeling using speaker-dependent recurrent temporal restricted boltzmann machines for voice conversion," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014. 1

[19] Chin-Cheng Hsu, Hsin-Te Hwang, Yi-Chiao Wu, Yu Tsao, and Hsin-Min Wang, "Voice conversion from non-parallel corpora using variational auto-encoder," in *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2016 Asia-Pacific*. IEEE, 2016, pp. 1–6. 1

[20] Yuki Saito, Yusuke Ijima, Kyosuke Nishida, and Shinnosuke Takamichi, "Non-parallel voice conversion using variational autoencoders conditioned by phonetic posteriorgrams and d-vectors," in *Proc. ICASSP*, 2018, pp. 5274–5278. 1

[21] Ilya Sutskever, Oriol Vinyals, and Quoc V Le, "Sequence to sequence learning with neural networks," in *Advances in neural information processing systems*, 2014, pp. 3104–3112. 1

[22] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014. 1, 2

[23] Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, et al., "Tacotron: Towards end-to-end speech synthesis," *arXiv preprint arXiv:1703.10135*, 2017. 1, 4

[24] Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, et al., "Natural TTS synthesis by conditioning wavenet on mel spectrogram predictions," in *ICASSP*. IEEE, 2018, pp. 4779–4783. 1

[25] Hideyuki Tachibana, Katsuya Uenoyama, and Shunsuke Aihara, "Efficiently trainable text-to-speech system based on deep convolutional networks with guided attention," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4784–4788. 1, 3

[26] William Chan, Navdeep Jaitly, Quoc Le, and Oriol Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *ICASSP*. IEEE, 2016, pp. 4960–4964. 1

[27] Chung-Cheng Chiu, Tara N Sainath, Yonghui Wu, Rohit Prabhavalkar, Patrick Nguyen, Zhifeng Chen, Anjuli Kannan, Ron J Weiss, Kanishka Rao, Ekaterina Gonina, et al., "State-of-the-art speech recognition with sequence-to-sequence models," in *ICASSP*. IEEE, 2018, pp. 4774–4778. 1

[28] Colin Raffel, Minh-Thang Luong, Peter J Liu, Ron J Weiss, and Douglas Eck, "Online and linear-time attention by enforcing monotonic alignments," *arXiv preprint arXiv:1704.00784*, 2017. 1, 2

[29] Andros Tjandra, Sakriani Sakti, and Satoshi Nakamura, "Listening while speaking: Speech chain by deep learning," *CoRR*, vol. abs/1707.04879, 2017. 1

[30] Di He, Yingce Xia, Tao Qin, Liwei Wang, Nenghai Yu, Tieyan Liu, and Wei-Ying Ma, "Dual learning for machine translation," in *Advances in Neural Information Processing Systems*, 2016, pp. 820–828. 1

[31] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," *arXiv preprint arXiv:1703.10593*, 2017. 1

[32] Hiroyuki Miyoshi, Yuki Saito, Shinnosuke Takamichi, and Hiroshi Saruwatari, "Voice conversion using sequence-to-sequence learning of context posterior probabilities," *arXiv preprint arXiv:1704.02360*, 2017. 1, 2, 4

[33] Jing-Xuan Zhang, Zhen-Hua Ling, Li-Rong Dai, Li-Juan Liu, and Yuan Jiang, "Sequence-to-sequence acoustic modeling for voice conversion," *arXiv preprint arXiv:1810.06865*, 2018. 1, 2, 4

[34] Yannis Stylianou, Olivier Cappé, and Eric Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Transactions on speech and audio processing*, vol. 6, no. 2, pp. 131–142, 1998. 2

[35] Elina Helander, Tuomas Virtanen, Jani Nurminen, and Moncef Gabbouj, "Voice conversion using partial least squares regression," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 5, pp. 912–921, 2010. 2

[36] Sun Lifa, Li Kun, Wang Hao, Kang Shiyin, and Meng Helen, "Phonetic posteriorgrams for many-to-one voice conversion without parallel data training," in *ICME*, 2016. 2

[37] Li-Juan Liu, Zhen-Hua Ling, Yuan Jiang, Ming Zhou, and Li-Rong Dai, "Wavenet vocoder with limited training data for voice conversion," *Proc. Interspeech 2018*, pp. 1983–1987, 2018. 2

[38] Athanasios Mouchtaris, Jan Van der Spiegel, and Paul Mueller, "Nonparallel training for voice conversion based on a parameter adaptation approach," 2006. 2

[39] Chung-Han Lee and Chung-Hsien Wu, "Map-based adaptation for speech conversion using adaptation data selection and non-parallel training," in *Ninth International Conference on Spoken Language Processing*, 2006. 2

[40] Toda Tomoki, Ohtani Yamato, and Shikano Kiyohiro, "Eigenvoice conversion based on gaussian mixture model," in *INTERSPEECH*, 2006, pp. 2446–2449. 2

[41] Daisuke Saito, Keisuke Yamamoto, Nobuaki Minematsu, and Keikichi Hirose, "One-to-many voice conversion based on tensor representation of speaker space," in *Twelfth Annual Conference of the International Speech Communication Association*, 2011. 2

[42] John Kominek and Alan W Black, "The CMU Arctic speech databases," in *Fifth ISCA workshop on speech synthesis*, 2004. 4

[43] Kazuhiro Kobayashi and Tomoki Toda, "sprocket: Open-source voice conversion software," in *Proc. Odyssey 2018 The Speaker and Language Recognition Workshop*, 2018, pp. 203–210. 4

[44] Zhizheng Wu, Oliver Watts, and Simon King, "Merlin: An open source neural network speech synthesis system," *Proc. SSW, Sunnyvale, USA*, 2016. 4

[45] Masanori Morise, Fumiya Yokomori, and Kenji Ozawa, "WORLD: a vocoder-based high-quality speech synthesis system for real-time applications," *IEICE TRANSACTIONS on Information and Systems*, vol. 99, no. 7, pp. 1877–1884, 2016. 4

[46] Ito Keith, "Tacotron speech synthesis implemented in tensorflow," in *https://github.com/keithito/tacotron*. 4

[47] Vinod Nair and Geoffrey E Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th international conference on machine learning (ICML-10)*, 2010, pp. 807–814. 4

[48] Yann N Dauphin, Angela Fan, Michael Auli, and David Grangier, "Language modeling with gated convolutional networks," *arXiv preprint arXiv:1612.08083*, 2016. 4

[49] Heiga Zen, Yannis Agiomyrgiannakis, Niels Egberts, Fergus Henderson, and Przemysław Szczepaniak, "Fast, compact, and high quality LSTM-RNN based statistical parametric speech synthesizers for mobile devices," *arXiv preprint arXiv:1606.06061*, 2016. 4

[50] Diederik Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014. 4

[51] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," *CoRR*, vol. abs/1706.03762, 2017. 4

[52] Hirokazu Kameoka, Kou Tanaka, Takuhiro Kaneko, and Nobukatsu Hojo, "Convs2s-vc: Fully convolutional sequence-to-sequence voice conversion," *arXiv preprint arXiv:1811.01609*, 2018. 4