VOICE CONVERSION WITH TRANSFORMER NETWORK

Ruolan Liu, Xiao Chen, Xue Wen

Samsung Research China – Beijing (SRC-B)
{ruolan.liu, xiao.chen, xue.wen}@samsung.com

ABSTRACT

This paper describes an end-to-end voice conversion system, which involves three main ideas: transformer, context preservation mechanisms, and model adaptation. Self-attention in the transformer architecture directly connects all positions, making it easier to learn long range dependencies and improve training efficiency. Context preservation mechanisms accelerate and stabilize training. Adaptation techniques are conductive to the training of the conversion mapping with limited training data. The results show that the proposed method obtains a higher MOS and the training speed is 2.72 times faster than LSTM based baseline system.

Index Terms— voice conversion, transformer, self-attention, guided attention, context preservation

1. INTRODUCTION

Voice conversion (VC) modifies a speech signal uttered by a source speaker to sound like a target speaker, while keeping the linguistic contents unchanged [1]. VC can be applied to various tasks, such as impersonating or hiding a speaker's identity [2], speaking aid for people with vocal impairments [3], and personalized text-to-speech (TTS) system using limited training data from the desired speaker [4].

Depending on whether parallel sentences are available for training, VC can be parallel or non-parallel [1]. Parallel VC makes it easy to align acoustic units between source and target speech, e.g. by dynamic time warping (DTW). Once alignment is established, learning a spectral mapping is relatively straightforward, e.g. via joint density Gaussian mixture model (JD-GMM) [5], exemplar-based model [6], or a deep neural network (DNN) [7]. In nonparallel VC unit alignment can still be estimated using an automatic speech recognition (ASR) module (e.g. [8]) but at lower accuracy. [9] takes a TTS-like approach that converts phonetic posteriorgram to mel-cepstrum sequence of target speaker. Alternatively, [10]-[13] back-off to matching target speaker marginals using a generative adversarial network (GAN), but enforce content transfer by cycle consistency and other regularizers. Despite advancements in nonparallel VC, parallel data should be used whenever available for maximum acoustic modelling power.

While most conventional VC operate on frame-by-frame basis, sequence-to-sequence (seq2seq) models [14] match sequences of variable lengths. First explored for neural machine translation (NMT), seq2seq has recently been extended to speech tasks like TTS [15], ASR [16] and VC [17]-[19]. A seq2seq model contains an encoder and a decoder, often employing convolutional or recurrent neural networks (CNN/RNNs) for sequence modelling. The more recent transformer network [20] employs self-attention that improves dataflow between both nearby and faraway neurons, and has achieved impressive results compared to CNN/RNN predecessors in NMT and other natural language processing (NLP) tasks.

In this paper, we focus on one-to-one voice conversion using parallel recordings without transcripts. This scenario is particularly useful for quickly adopting a pre-trained single-speaker TTS to generate speech in new voice. Our system combines the transformer architecture, context preservation and model adaptation in an attentional seq2seq VC. Results show that the proposed method achieves better naturalness and speaker similarity than our long short-term memory (LSTM) RNN baseline. Besides, using transformer instead of RNN speeds up the training process about 2.72 times.

2. BACKGROUND

2.1. Attentional models for voice conversion

Attentional seq2seq VC models are capable of converting both acoustic and prosodic features of the source speech. An early example of attentional VC is found in [21]. However, a long autoregressive path makes the model vulnerable to the so-called exposure bias problem [22], observed as failure to preserve language content. Recently text supervision [17] and context preservation [18] methods are proposed to stabilize training, and bottleneck features [19] from an ASR model are also used to condition the conversion process to generate correct content.

2.2. Transformer

Transformer [20] is a seq2seq network based solely on attention mechanisms. It contains an encoder and a decoder, both composed of a stack of *N* identical layers. Each encoder layer has a multi-headed self-attention (MHA) sub-layer and

978-1-5090-6631-5/20/$31.00 ©2020 IEEE

a fully connected feed-forward sub-layer; residual connection and layer normalization are applied to both. Each decoder layer has one more masked MHA sub-layer than encoder. Transformer has been shown to outperform many other models in NMT tasks. Recently, it is also successfully used for prosody modelling in TTS [23]. In this work we use transformer for seq2seq conversion of spectral features.

3. PROPOSED METHOD

The overall transformer-based VC network is shown in Fig. 1. It contains pre-nets (A, B), "usual" encoder and decoder with cross-attention (C), source and target decoders (E, F) for context preservation, and guided attention module (G). Blue blocks (EFG) are only used in training to accelerate and stabilize the training process. Red blocks (CEFG) contain modifications to the original transformer to fit our VC task. We train the conversion model using adaptation techniques.

3.1. Pre-net

Source and target mel-spectrograms first pass through a pre-net (Fig. 1 A and B, 256 hidden units, fully connected). This fits the input feature to the input dimension of the transformer (d_{model}). Pre-net outputs are summed with triangular position encodings (PE). This is written as

$$\mathbf{X}' = f_{EncPrenet}(\mathbf{X}) + PE \quad (1)$$

$$\mathbf{Y}' = f_{DecPrenet}(\mathbf{Y}) + PE \quad (2)$$

where **X** and **Y** are log-scale mel-spectrograms of source and target speech, **X'** and **Y'** are the sum of the pre-net outputs and PEs.

3.2. Encoder and decoder

Encoder and decoder in our system are almost the same as the original transformer, except that we use single-headed cross-attention (Fig.1 C) instead of multi-headed.

Self-attention in the encoder and decoder directly connects every pair of frames in the sequence by dot product:

$$\text{Attention}(\mathbf{Q},\mathbf{K},\mathbf{V}) = softmax\left(\frac{\mathbf{QK^T}}{\sqrt{d_k}}\right)\mathbf{V} \quad (3)$$

Doing so makes it easier to build long-range dependencies.

Multi-headed attention allows attending different aspects, such as linguistic information, voice color and prosody. With guided cross-attention (next section), however, we found that MHA takes longer to compute than we'd like. We therefore opt for single-headed cross-attention between encoder and decoder.

The predicted mel-spectrogram $\hat{\mathbf{Y}}$ is computed using the final linear layer. The seq2seq loss is defined as in Eq. 4.

$$\mathcal{L}_{seq2seq} = \left\|\hat{\mathbf{Y}} - \mathbf{Y}\right\|_1 \quad (4)$$

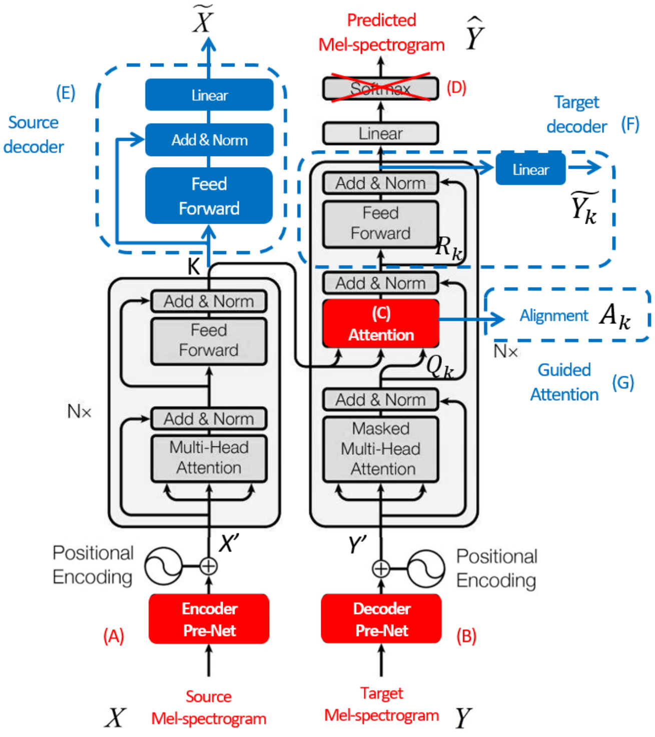

Figure 1. The overall Transformer based VC network

3.3. Guided attention and context preservation

In preliminary experiments we trained our VC network with the seq2seq objective but observed no sign of convergence. Inspired by [18], we apply guided attention and context preservation mechanisms to our transformer VC network to accelerate and stabilize the training process.

3.3.1. Guided attention loss (GA loss)

Guided attention loss [18] defines a guide matrix **G** to make the encoder-decoder alignment matrix **A** 'nearly diagonal':

$$g_{ij} = 1 - exp\left\{\frac{-\left(\frac{i}{I}-\frac{j}{J}\right)^2}{2\sigma_g^2}\right\} \quad (5)$$

where *I* and *J* are the lengths of source and target sequences, *i* and *j* are indices to encoder and decoder states, and σ_g is a hyper-parameter to control the "beam width". The guided attention loss is then computed as

$$\mathcal{L}_{ga} = \sum_k^N ||\mathbf{G}\odot\mathbf{A_k}||_1 \quad (6)$$

where the sum is over all decoder layers, with $\mathbf{A_k}$ being the cross-attention matrix of the kth layer.

3.3.2. Context preservation loss (CP loss)

The context preservation mechanisms [18] employ a source decoder and a target decoder. The source decoder predicts source input from encoder states (Fig.1 E):

$$\mathbf{K} = f_{Encoder}(\mathbf{X}') \tag{7}$$

$$\widetilde{\mathbf{X}} = f_{SrcDec}(\mathbf{K}) \tag{8}$$

where $\mathbf{K}$ is the encoder output. The target decoder predicts target input from cross-attention contexts only (Fig.1 F). As each decoder layer has one set of attentional contexts, we can compute N predictions:

$$\mathbf{Q_k} = f_{\mathrm{DecMHA}}(\mathbf{Y'_k}) \tag{9}$$

$$\mathbf{R_k} = f_{\mathrm{EncDecAtt}}(\mathbf{K}, \mathbf{Q_k}) \tag{10}$$

$$\widetilde{\mathbf{Y_k}} = f_{TrtDec}(\mathbf{R_k}) \tag{11}$$

where $\mathbf{Q_k}$ is the multi-headed attention query at k^{th} decoder layer. Notice that $\widetilde{\mathbf{Y_k}}$ is independent of $\mathbf{Q_k}$ (hence of $\mathbf{Y'_k}$) conditioned on $\mathbf{R_k}$.

The full loss function of our model is given in Eq. (12). GA and CP losses are weighted by hyper-parameters λ_{ga} and λ_{cp}:

$$\mathcal{L} = \mathcal{L}_{seq2seq} + \lambda_{ga}\mathcal{L}_{ga} + \lambda_{cp}\left(\left\|\widetilde{\mathbf{X}} - \mathbf{X}\right\|_1 + \sum_k^N \left\|\widetilde{\mathbf{Y_k}} - \mathbf{Y}\right\|_1\right) \tag{12}$$

3.4. Model Adaptation

Training high-quality VC from only source and target speaker data generally requires large amount of data from both. In this paper we leverage available multi-speaker dataset by first training four prototype VC models based on source and target genders (M2M, M2F, F2M, F2F) using all training set speakers. Given source and target speakers, one-to-one VC is initialized from the pre-trained model corresponding to source and target speaker gender and fine-tuned on source and target speaker data.

4. EXPERIMENT

4.1. Experimental setup

4.1.1. Dataset

Our experiments are conducted on an internal Mandarin Chinese dataset. It contains 105 speakers, 50 male and 55 female, and 105,000 utterances. Speakers are divided into 5 groups. Among the 1000 sentences of each speaker, 200 sentences are parallel for all speakers; the remaining 800 are parallel in each group.

4.1.2. Baseline system (LSTM-VC)

A LSTM based VC method is used as the baseline. Its design is based on Tacotron2 [15], to which we fit guided attention as in 3.3.1. Hyper-parameters σ_g and λ_{ga} are set to 0.4 and 100. Each one-to-one conversion model is trained using 800 pairs of parallel sentences from one pair of source and target speakers. Waveform is reconstructed using WaveNet conditioned on mel-spectrogram.

4.1.3. Proposed system (Transformer-VC)

The recordings are sampled at 24 kHz. 80-dimension mel-scale spectrograms are extracted every 10ms, followed by a logarithmic dynamic range compression. Hyper-parameters σ_g, λ_{ga} and λ_{cp} are set to 0.4, 10 and 0.14 in both pre-training and adaptation stages. We use the transformer architecture with d_{model} = 256, stacked layers N = 6, multi-head h = 8, and max sequence length maxlen = 500. 97 speakers are used to pre-train four initial models. Each model is trained by about 500k parallel speech data. The training data of M2M and F2F model does not contain the speech pairs of the same speaker (e.g. F1 to F1). Pre-training on one P40 GPU with batch size 28 (sentences) takes about 200k steps to converge. The remaining 8 speakers (4 male and 4 female) are used to train one-to-one VC models with 800 pairs of examples, each taking about 20k steps to converge.

4.1.4. WaveNet based vocoder

We train a WaveNet vocoder conditioned on mel-spectrogram with the same Mandarin Chinese dataset. Given limited training data of target speakers, we try adaptation [25] and WaveNet with speaker embedding [26]. Results show that both of them have the ability to model speaker characteristics from limited data. We choose WaveNet with speaker embedding for our experiments in this paper.

4.2. Results

We conduct several experiments to show our improvements with the two VC systems. Because the length of the converted speech and ground truth are different, the root mean square error (RMSE) of mel-spectrogram is not adopted in our experiments.

4.2.1. Training time

Without any recurrent connection, our model can be trained in parallel. With the same batch size (28) and the same GPU settings (1 NVIDIA P40 GPU), the time taken by each training step of the baseline and proposed methods is shown in Table 1. The proposed model trains 2.72x faster with slightly fewer parameters.

Since the proposed model is trained with adaptation, 5.5 hours with 20k steps is enough to converge for adaptation, comparing to 28.5 hours with 35k steps training from scratch of the baseline system.

Table 1 Comparison of time consuming per training step

Method	Time consuming(s)
Baseline	2.94
Proposed	1.08

Table 2 Comparison of MOS with 95% confidence intervals on naturalness (N) and similarity (S)

Conversion Pairs		Baseline	Proposed
M2F	N	3.81± 0.32	**4.29**± 0.29
	S	3.87± 0.20	**4.31**± 0.26
F2M	N	3.5± 0.28	**3.84**± 0.37
	S	3.67± 0.50	**4.14**± 0.35
M2M	N	3.75± 0.35	**3.93**± 0.35
	S	3.89± 0.46	**4.28**± 0.25
F2F	N	3.64± 0.42	**4.15**± 0.29
	S	3.81± 0.27	**4.11**± 0.32

4.2.2. MOS evaluations

We investigate mean option score (MOS) of both naturalness and similarity against Tacotron2 baseline. In this evaluation, 20 converted utterances in each intra/cross-gender category are selected randomly. 15 native speakers are asked to listen to original and converted utterances and rate naturalness and speaker similarity separately on a scale between 1 and 5.

Results are given in Table 2. We see that the proposed method achieves higher scores on both naturalness and similarity. We believe that utilizing the pre-trained model from large amount of data can lead to better pronunciation. Meanwhile, multi-headed structure in Transformer allows attending different aspects, and self-attention connects every pair of frames directly, both of which have contributed to smoother and more natural prosody. We also found that similarity is often scored slightly higher than naturalness. Since our dataset speakers are non-professionals, we conjecture that listeners would find the overall quality less natural and comfortable than professional broadcastings, but the tone of the converted voices are sounded very similar to the original. This is confirmed by informal interviews with raters.

4.2.3. Ablation test

Two ablation tests are conducted related to regularizers and model pre-training.

Table 3 lists the number of sentences with critical errors (loss or insertion of content) of 100 test sentences, correlated to four models trained with and without guided attention and context preservation losses. The reason of causing those errors is probably from attention mechanisms. Training without CP loss has failed both with and without GA. Training with CP loss, on the other hand, is shown to benefit from GA by producing fewer critical errors, often in the form of content skipping and repetition.

Table 3 Comparison of error sentences between different methods on the 100 sentences

Method	Error sentences
Without GA and CP loss	Failed to convert
Only GA loss	Failed to convert
Only CP loss	25
With GA and CP loss	4

Table 4 Comparison of MOS on Naturalness of different layer and head numbers

Hyper-parameters	3-layer	6-layer
4-head	-	4.06± 0.34
8-head	4.18± 0.29	4.29± 0.29

Table 5 Comparison of time consuming (in second) per training step of different numbers of head and layer

Hyper-parameters	3-layer	6-layer
4-head	-	0.75
8-head	0.54	1.08

In order to verify the benefits of pre-training, we try to train one-to-one M2F model with data of one pair of source and target speakers only. With the help of the regularizers, the converted utterances can be understood. However, the pronunciation is not full and natural. Compared to the M2F model initialized with the pre-trained model, it obtains a high MOS of 4.29 as shown in Table 2. This confirms that using adaptation with large amount of data can be beneficial to pronunciation.

4.2.4. Transformer with different modelling capacity

The original transformer uses 6 stacked layers and 8 attention heads. We test the performance and training time with different numbers of heads and layers with the M2F model. The results are given in Tables 4 and 5. Reducing the numbers of heads or layers can both increase the training speed, but are shown to be harmful to VC quality. This shows that most layers and attention heads are being exploited by the model, at least loosely.

5. CONCLUSION AND FUTURE WORK

We propose an end-to-end VC model based on the transformer architecture, regularized by guided attention and context preservation mechanisms. The proposed system can be trained in parallel and generates high quality speech with high similarity to target speaker.

While self-attention and the regularizers can effectively stabilize and accelerate the training process, the proposed system still suffers slow inference (as in other AR generators) and sometimes produces critical errors. We shall consider non-autoregressive sequence models in our future work to address both issues.

6. REFERENCES

[1] S. H. Mohammadi, A. Kain, “An overview of voice conversion systems,” *Speech Communication,* vol.88, pp.65-82, 2017

[2] Y. Gao, R. Singh, B. Raj, “Voice impersonation using generative adversarial networks,” *Proc. ICASSP*, pp.2506-2510, 2018

[3] A. B. Kain, J. Hosom, X. Niu, J. P. H. V. Santen, M. Fried-Oken, J. Staehely, “Improving the intelligibility of dysarthric speech,” *Speech Communication*, vol.49, no.9, pp.743-759, 2007.

[4] A. Kain, M. W. Macon “Spectral voice conversion for text-to-speech synthesis,” *Proc. ICASSP*, pp.285-288, 1998

[5] T. Toda, A.W. Black, K. Tokuda, “Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory,” *IEEE/ACM Transactions on Audio Speech and Language Processing*, vol.15, no.8, pp. 2222-2235, 2007.

[6] Z. Wu, T. Virtanen, E. S. Chng, H. 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.

[7] S. Desai, A. W. Black, B. Yegnanarayana, K. Prahallad, “Spectral mapping using artificial neural networks for voice conversion,” *IEEE/ACM Transactions on Audio Speech and Language Processing*, vol. 18, no. 5, pp. 954-964, 2010.

[8] F. Xie, F. K. Soong, and H. Li, “A KL divergence and DNN-based approach to voice conversion without parallel training sentences,” *Proc. INTERSPEECH*, pp. 287–291, 2016.

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

[10] 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

[11] T. Kaneko, H. Kameoka, K. Tanaka, N. Hojo. “StarGAN-VC2: rethinking conditional methods for starGAN-based voice conversion,” *Proc. INTERSPEECH*, pp. 679-683, 2019

[12] T. Kaneko, H. Kameoka, “Parallel-data-free voice conversion using cycle-consistent adversarial networks,” *arXiv preprint arXiv:1711.11293*, 2017.

[13] T. Kaneko, H. Kameoka, K. Tanaka, N. Hojo, “Cyclegan-VC2: Improved cyclegan-based non-parallel voice conversion”, *Proc. ICASSP,* pp. 6820-6824, 2019

[14] D. Bahdanau, K. Cho and Y. Bengio, “Neural machine translation by jointly learning to align and translate,” *Proc. ICLR*, 2015

[15] J. Shen, R. Pang, R. Weiss, M. Schuster, N. Jaitly and et al, “Natural TTS synthesis by conditioning wavenet on mel spectrogram predictions,” *Proc. ICASSP*, pp. 4779-4783, 2018.

[16] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, “Listen, attend and spell: a neural network for large vocabulary conversational speech recognition,” *Proc. ICASSP*, pp. 4960-4964, 2016.

[17] J. Zhang, Z. Ling, Y. Jiang, L. Liu, C. Liang and L. Dai, “Improving sequence-to-sequence voice conversion by adding text-supervision,” *Proc. ICASSP*, pp.6785-6789, 2019.

[18] 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.

[19] 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.

[20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, “Attention is all you need,” *Proc. NIPS*, pp. 5998–6008, 2017.

[21] M. V. Ramos, “Voice conversion with deep learning,” *Master’s Thesis*, Instituto Superior T´ecnico, 2016.

[22] S. Bengio, O. Vinyals, N. Jaitly, and N. Shazeer, “Scheduled sampling for sequence prediction with recurrent neural networks,” in *Advances in Neural Information Processing Systems*, pp.1171–1179, 2015.

[23] N. Li, S. Liu, Y. Liu, S. Zhao, M. Liu and M. Zhou, Ming, “Neural speech synthesis with transformer network,” *Proc. AAAI*, pp.6706-6713, 2019.

[24] H. Tachibana, K. Uenoyama and S. Aihara, “Efficiently trainable text-to-speech system based on deep convolutional Networks with guided attention,” *Proc. ICASSP,* pp.4784-4788, 2017.

[25] L. Liu, Z. Ling and L. Dai, “WaveNet vocoder with limited training data for voice conversion,” *Proc. INTERSPEECH*, pp. 1983–1987, 2018.

[26] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior and K. Kavukcuoglu, “WaveNet: a generative model for raw audio,” *arXiv preprint arXiv:*1609.03499, 2016.