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 information, thus speaking rate conversion cannot be performed. On the other hand, modern sequence-to-sequence (seq2seq) based approaches model not only the spectrogram 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 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.

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 target speakers’ duration information, thus speaking rate conversion. This approach is advantageous for streaming VC as it offers robust alignment against collapse.

To address this issue, 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.

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

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 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, 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:

\[ \alpha(t, u) = \alpha(t-1, u) \phi(t-1, u) + \alpha(t, u-1) \{1 - \phi(t, u-1)\}, \]

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:

\[ L = \sum_{t=1}^{T} \sum_{u=1}^{U} I\{ (t, u) \in \tau \} \alpha(t, u) \{1 - \phi(t, u)\} (y_{u+1} - f(t, u)), \]

where \(y_{u+1}\) denotes the \(u+1\)’th frame target speaker’s spectrum. \(f(t, u)\) 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 \(f(t, u)\), 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 \(\tau\) band frames that are close to the target alignment following [16]. Here \(I\{ (t, u) \in \tau \}\) is an indicator function that tells whether or not index \((t, u)\) is within the constraint \(\tau\) 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 target 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.

2.2. 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.
The model has 100 vocabulary entries including blank symbols, Japanese kana, long vowel symbols, and alphabetic characters.

Table 2: Averaged mel-cepstrum distortion (MCD) and character error rate (CER). "F2F", "F2M", "M2F" 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.

<table>
<thead>
<tr>
<th>Method</th>
<th>MCD [dB]</th>
<th>CER [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>F2F</td>
<td>M2F</td>
</tr>
<tr>
<td>GT</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>RESYN</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>BNE-S2SMoL-VC</td>
<td>5.7</td>
<td>5.5</td>
</tr>
<tr>
<td>ConvS2S-VC (offline)</td>
<td>4.9</td>
<td>5.0</td>
</tr>
<tr>
<td>ConvS2S-VC (stream)</td>
<td>5.5</td>
<td>5.5</td>
</tr>
<tr>
<td>VC-T (offline)</td>
<td>5.1</td>
<td>5.0</td>
</tr>
<tr>
<td>VC-T (stream)</td>
<td>5.3</td>
<td>5.1</td>
</tr>
</tbody>
</table>

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] and BNE-S2SMoL-VC [25], 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. 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], 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 wav2vec2.0.0 ASR model [32]. The ASR model was fed
The table lists the objective evaluation results. As oracles, CERs for the ground truth speech (GT) and the analytic re-synthesized one by HiFi-GAN (R E S Y N ) were also calculated. First, R E S Y N 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. C O N V S 2 S - V C ( o f f l i n e ) scored better than this in both MCD and CER. Its streaming version, C O N V S 2 S - V C ( s t r e a m ) , is still better than BNE-S2SMoL-VC, while its performance is degraded compared to C O N V S 2 S - V C ( o f f l i n e ) . 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, V C - T ( o f f l i n e ) attained significantly better CER, even albeit the average MCD of the three models was slightly worse than that of C O N V S 2 S - V C ( o f f l i n e ) . The proposed method achieved robust alignment, as there were no speech skips in the evaluation data. The MCD and CER of V C - T ( s t r e a m ) were worse than those of V C - T ( o f f l i n e ) , but the degree of deterioration was smaller and indeed superior to those of the same streaming model, C O N V S 2 S - V C ( s t r e a m ) . We performed the MAPSSWE significance test [33], and the differences of the CERs between C O N V S 2 S - V C and V C - T in offline and streaming modes were statistically significant, $p < 0.001$. Thanks to their robust alignments generated from our V C - T , it could better convert source speaker’s speech to target speaker’s one than that of the C O N V S 2 S - V C while preserving linguistic information. Moreover, our V C - T achieved the above results although the model size was much smaller than C O N V S 2 S - V C (see Table 1).

4.3. Subjective evaluations

We subjectively evaluated the naturalness of converted speech, including R E S Y N . 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 R E S Y N 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. C O N V S 2 S - V C ( o f f l i n e ) outperformed it, but scored slightly lower than homo-gender conversions, possibly owing to increased difficulty in hetero-gender conversion. Its streaming version, C O N V S 2 S - V C ( s t r e a m ) , was found to be inferior to C O N V S 2 S - V C ( o f f l i n e ) , especially in hetero-gender conversions. We suspect this was because streaming operation suffers if future context is missing, since C O N V S 2 S - V C only depends on the source speaker’s speech unlike our V C - T . Thus, C O N V S 2 S - V C ( s t r e a m ) led to produce unclear spectra due to unstable alignments, got a larger confidence interval than that of the others. Contrary to C O N V S 2 S - V C ( s t r e a m ) , V C - T ( s t r e a m ) exhibited no remarkable degradation compared to V C - T ( o f f l i n e ) . These results reveal that V C - 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 worse speaker similarity in all gender conditions. C O N V S 2 S - V C ( s t r e a m ) exhibited significant degradation. On the other hands, V C - T ( o f f l i n e ) roughly matched C O N V S 2 S - V C ( o f f l i n e ) . Unlike C O N V S 2 S - V C ( s t r e a m ) , V C - T ( s t r e a m ) was close to V C - T ( o f f l i n e ) , with almost no speaker similarity degradation. These overall results demonstrate that our proposal, V C - 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., V C - 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


