LEARNING IN YOUR VOICE: NON-PARALLEL VOICE CONVERSION BASED ON SPEAKER CONSISTENCY LOSS

Yoohwan Kwon\textsuperscript{1,2}, Soo-Whan Chung\textsuperscript{1}, Hee-Soo Heo\textsuperscript{2}, Hong-Goo Kang\textsuperscript{1}

\textsuperscript{1}Department of Electrical & Electronic Engineering, Yonsei University, South Korea
\textsuperscript{2}Naver Corporation, South Korea

ABSTRACT

In this paper, we propose a novel voice conversion strategy to resolve the mismatch between the training and conversion scenarios when parallel speech corpus is unavailable for training. Based on auto-encoder and disentanglement frameworks, we design the proposed model to extract identity and content representations while reconstructing the input speech signal itself. Since we use other speaker’s identity information in the training process, the training philosophy is naturally matched with the objective of voice conversion process. In addition, we effectively design the disentanglement framework to reliably preserve linguistic information and to enhance the quality of converted speech signals. The superiority of the proposed method is shown in subjective listening tests as well as objective measures.

Index Terms— Voice conversion, speech disentanglement, speaker consistency loss

1. INTRODUCTION

Voice conversion (VC) is a technique for converting the speech of source speaker to the speech that sounds like being spoken by a target speaker while preserving the linguistic information of source speech. In early studies \cite{1, 2}, most voice conversion researches were performed using parallel speech corpus, i.e. speech signals with common scripts spoken by many different speakers. These methods effectively transfer ones’ speaking style to others while preserving speaking contents. However, they result in severely distorted sound if the test scripts to be converted are not included in the training stage. Although the problem can be relieved by using large-scale datasets, it requires a large amount of costs for manual annotations and parallel speech recording.

On the Internet, there are plethora of utterances spoken by a large number of people, and some large-scale datasets are publicly available. In recent, these non-parallel speech corpus is used for voice conversion researches that are implemented independent of speaking contents. It is a highly challenging task since there is no explicit target signal in supervision. Instead of relying on strong supervision, some works use additional data or extra systems helpful to the conversion process such as text scripts \cite{3} or speech recognition \cite{4, 5}. These approaches are somewhat effective when we can not use strong supervision, but it is still not a proper solution. To solve the problem, disentanglement \cite{6, 7} and generative model-based approaches \cite{8, 9} were studied. Especially, generative adversarial network (GAN)-based methods in \cite{10, 11, 12, 13} showed promising results in the conversion tasks. However, these methods are still suffered from low speech quality with low naturalness and inaccurate pronunciation on outputs.

AutoVC \cite{14} is the most prevalent work that are related to the proposed method, where it uses auto-encoder framework to generate speech signals. In the training phase, AutoVC decomposes input signals into identity and content features and reconstructs itself to build a speech generation system. In the conversion phase, it replaces identity feature with other’s representation to convert speech into target’s voice. AutoVC is classified into a zero-shot learning approach since it does not require any additional cues or systems and is applicable to unseen speakers. However, the training criterion of AutoVC is set to reconstruct input signal itself, not to convert speaking styles or maintain contextual information; thus, the converted speech has distortions or missing pronunciation.

In this paper, we present a novel training strategy, whose objective is well-matched with voice conversion. The trunk structure is the same as the one used in AutoVC, i.e. reconstructing input spectral features using disentangled embeddings. In addition to the self-reconstruction process, we replace the input identity embeddings with the ones taken from target speakers. Consequently, the self-reconstruction helps to learn speech generation, and the additional speaker consistency criterion provides a supervision for speaking style conversion. Additionally, we modify the disentanglement method for content embeddings to preserve linguistic expression on converted speech. To preserve reliable linguistic expression on converted speech, we also introduce an attention method to encode content embeddings instead of the conventional hard sampling strategy. Our contributions to the conversion task improve the voice conversion performance by addressing the conversion process on the training stage and introducing a powerful disentanglement method. The superiority of the proposed method would be examined in several objective and subjective measurements.
2. RELATED WORKS

2.1. AutoVC

AutoVC is a zero-shot non-parallel many-to-many voice conversion model based on a simple auto-encoder framework \[14\]. By decomposing speech inputs into content and identity representations with information-constraining bottleneck encoding layers, the model reconstructs the input speech in the decoding stage. The content encoder takes input speech and passes it through a bottleneck to produce a content embedding conditioned on source speaker embedding from a pretrained identity encoder. The content embeddings are downsampled to make it more difficult to present speaker characteristics of input signals. Then the decoder takes the content embedding and identity embedding to reconstruct spectral features. When training, source speaker embedding is fed to the decode for self-reconstruction supervision. The network is trained to minimize reconstructed features and input features, and the content embedding is extracted from the reconstructed features to make it resemble that of input speech by minimizing the distance between them. Although AutoVC shows powerful disentanglement and conversion performance, there are still rooms for improvement.

2.2. Style transfer with perceptual loss

On various generation tasks such as style transfer \[15\] \[16\], voice conversion \[19\], there is commonly a lack of ground truth data. Perceptual loss has been widely used to represent target information on the generation outputs, and it is advantageous on learning tasks using non-parallel data. Perceptual loss does not directly address the objective of tasks, e.g., minimizing differences between pixels or spectral features, but it regulates representations of outputs on the latent space. It minimizes the distance between high-level representations of outputs and that is extracted from other data presenting the desired feature. This learning strategy is powerful to provide supervision of perceptual and semantic expressions on outputs while not forcing outputs to match target data directly.

3. PROPOSED METHOD

In this section, we demonstrate the proposed learning strategy for voice conversion using a disentanglement framework. The common problem with using a non-parallel dataset on voice conversion is that there is no explicit ground-truth speech to generate. Our proposed method adopts the strategy of AutoVC and designs a consistency loss to resolve this problem.

3.1. Self-reconstruction loss

As described in \[14\], they set up an auto-encoder to generate input speech signal itself. Content embedding is extracted from input signal, and identity embedding is brought from another utterance spoken by the same speaker. Since both speech signals are spoken by the same person, the identity representation should be same. The decoder takes these content and identity embeddings to reconstruct input spectral features. Figure 1(a) shows our training strategy using the self-reconstruction loss. In the self-reconstruction, the objective function is given as follows,

\[ L_{\text{SRL}} = \| D(C_{i,A}, S_{j,A}) - X_{i,A} \|_2 + \| \hat{C}_i - C_i \|, \]

where \( D \) is the decoder network, \( C_{i,A} \) is content embedding extracted from \( i \)-th speech \( x_{i,A} \) spoken by speaker \( A \), and \( S_{j,A} \) is the identity embedding of the \( j \)-th utterance spoken by speaker \( A \), \( X_{i,A} \) is the spectral feature of \( x_{i,A} \), respectively.

3.2. Speaker consistency loss

In a machine learning-based approach, matching the test and training scenario always achieves better performance than an unmatched scenario. Since the learning criterion of AutoVC does not directly address the conversion task, we propose a new training strategy to learn spectral modeling in others’ voices for effective voice conversion, which is called a speaker consistency loss (SCL). Our method utilizes an implicit representation as a target for supervision instead of target spectral features, which does not exist in a non-parallel dataset. The content and speaker embeddings for the reconstruction should be kept on the converted speech.
Fig. 2: Structure of modified attention bottleneck block

signals. Thus, if we extract content and identity embeddings from converted signals, they should be represented in the same distributions of input embeddings. By using these characteristics of voice conversion, we exploit other’s speech signals for training. Figure 1(b) shows our training strategy of the speaker consistency loss. The learning criterion is to minimize distances between embeddings as below:

\[ L_{SCL} = |\hat{C}_i - C_i| + |\hat{S}_B - S_B|, \tag{2} \]

where \( S_B \) indicates speaker embedding of speaker \( B \). Here, the content encoder extracts linguistic representations, but its parameters are not updated with the consistency loss. Whole network is trained with below objective function:

\[ L_{total} = L_{SRL} + \lambda L_{SCL} \tag{3} \]

We use 0.5 for the value \( \lambda \), and the combination of multiple learning criteria is beneficial to its training by addressing both voice conversion and speech generation.

3.3. Attentive bottleneck

The other key factor of voice conversion is feature disentanglement. AutoVC disentangles content embeddings by an information bottleneck structure with temporal downsampling as described in [14]. However, due to its structural limitation of hard sampling, it frequently loses content information. Therefore, we modify the information bottleneck structure using attention mechanism. We exploit a self-attention structure to extract content embedding as in Figure 2. The enhanced content feature that contains linguistic information of near frames compensates the loss from temporal downsampling. Scaled dot-product [17] is used for attention mechanism. After that, we change the downsampling to a max-pooling procedure. Since max-pooling does not take a value from the fixed frames unlike downsampling, the attention mechanism works well to contain the linguistic information while maintaining its disentanglement effect of downsampling.

4. EXPERIMENTS

In this section, we conduct several objective and subjective experiments on a non-parallel many-to-many voice conversion task to confirm the performance of the proposed method.

Fig. 3: Speaker recognition accuracy. Left: the recognition performance of content bottleneck feature. Right: the recognition accuracy on voice conversion outputs

4.1. Experimental settings

We split this dataset into training and test sets; in the evaluation set, we reserve 10 speakers not included in the training set. Also, we randomly select 10% of utterances spoken by 99 speakers for the evaluation of seen speaker conversion. The input speech signal is transformed into an 80-dimensional Mel-spectrum in the logarithm scale, which is sliced at every 16ms with 64ms window size.

The network structure is exactly same as the one in AutoVC [14], and we use WaveNet vocoder [18] to generate speech signal from reconstructed Mel-spectrum. The speaker encoder consists of 2 LSTM layers, and it is pre-trained in the generalized end-to-end loss [19] but is not fine-tuned with a training process related to voice conversion. We conduct several baseline methods proposed in [14] and [20] in scratch.

4.2. Objective evaluation

We evaluate the performances of the disentanglement and voice conversion strategy for objective comparisons, respectively. In the evaluation of disentanglement, content bottleneck features are examined in speakers’ identities, where they should not specify the speaker if they are well-disentangled from the speaker’s identities. The speaker of converted speech should be recognized as the target speaker, not the speaker of the input speech.

For the speaker recognition on the bottleneck feature, we train an additional back-end classifier from the bottleneck feature, using speakers in the training set. Figure 3 shows the classification accuracy of seen speakers in the evaluation set. If there remains nothing related to identity on the content bottleneck, it would show 1% accuracy, whereas our proposed method shows 1.1%, and AutoVC results in 3.7%. The speaker recognition accuracy on the input speech shows 95.5%, while the results of conversion outputs are 72.5% for the proposed method and 74.8% for AutoVC baseline. These recognition results prove that the proposed method outperforms the baseline method in the conversion strategy as well as the improved disentanglement method.
4.3. Subjective evaluation

We conduct three subjective assessments for the quality of speech and its speaker similarity to the target speech: mean-opinion-score (MOS) on naturalness and similarity, and preference tests. We reserve two types of speaker sets, which are seen and unseen speakers to compare the effects on the conversion of not-enrolled speakers. Each dataset has another 4 gender pair source-target subsets, i.e., male-to-male (M-M), male-to-female (M-F), female-to-male (F-M), and female-to-female (F-F). We randomly generated 4 conversion samples for each subset, and 15 listeners were asked to rate the performance of a total of 32 samples per each method. In the MOS tests, listeners score the naturalness and the similarity of conversion outputs from 1 to 5. Besides, they compare the conversion methods by judging which sample is more similar to the target speaker for the preference test.

In Figure 4 and Figure 5, it reports the MOS results on the seen and unseen speaker evaluation sets. Baseline methods in [14] [20] show 2.81 and 3.30 MOS on naturalness respectively, while they score 2.62 and 3.09 on similarity. On the other hands, the proposed method shows significantly improved performance, scoring 3.52 on naturalness and 3.78 on similarity. In the unseen speaker set, the proposed method is also effective compared to the baselines, where the proposed method shows 3.20 on the naturalness and 3.64 on similarity while AutoVC scores 2.62 and 3.09, respectively.

On the preference test in Figure 6, the proposed method is more preferred than AutoVC for entire evaluation sets. Listeners responded that the outputs of the proposed method are better than those of AutoVC with 60.7% and 59.5% preference scores on seen and unseen speaker sets. However, they reported that if speakers are converted between the same genders, it is frequently difficult to tell the difference between the proposed method and the baseline, although the proposed one is still better than the baseline. The most impressive results on the preference test is that conversion between different genders with the proposed method is far preferred, while there is little preference on the baseline method. The conversion samples are available on the demo page.

5. CONCLUSION

In this paper, we proposed a novel non-parallel voice conversion model based on the disentanglement framework. By introducing the training method for matching conversion and training procedure, the proposed model can be trained on non-parallel data more efficiently. We designed a speaker consistency loss to directly address the conversion task on training phase. Also, we modified the content bottleneck extraction method, and it results in the improvement of the disentanglement with little missing linguistic information. Both objective and subjective evaluation results confirmed that the proposed method outperformed baseline methods.

[http://github.com/yoohwankwon/Learning_in_your_voice]
6. REFERENCES

[1] Seyed Hamidreza Mohammadi and Alexander Kain, “Voice conversion using deep neural networks with speaker-independent pre-training,” in IEEE Spoken Language Technology Workshop (SLT), 2014.

[2] 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, vol. 22, no. 12, pp. 1859–1872, 2014.

[3] Jing-Xuan Zhang, Zhen-Hua Ling, and Li-Rong Dai, “Non-parallel sequence-to-sequence voice conversion with disentangled linguistic and speaker representations,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 28, pp. 540–552, 2019.

[4] Feng-Long Xie, Frank K Soong, and Haifeng Li, “A kl divergence and dnn-based approach to voice conversion without parallel training sentences,” in Proceedings of INTERSPEECH, 2016.

[5] Yuki Saito, Yusuke Ijima, Kyosuke Nishida, and Shinnosuke Takamichi, “Non-parallel voice conversion using variational autoencoders conditioned by phonetic posteriorgrams and d-vectors,” in Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, 2018, pp. 5274–5278.

[6] Da-Yi Wu, Yen-Hao Chen, and Hung-Yi Lee, “Vqvc+: One-shot voice conversion by vector quantization and u-net architecture,” arXiv preprint arXiv:2006.04154, 2020.

[7] Da-Yi Wu and Hung-yi Lee, “One-shot voice conversion by vector quantization,” in Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, 2020.

[8] Chin-Cheng Hsu, Hsin-Te Hwang, Yi-Chiao Wu, Yu Tsao, and Hsin-Min Wang, “Voice conversion from unaligned corpora using variational autoencoding wasserstein generative adversarial networks,” arXiv preprint arXiv:1704.00849, 2017.

[9] Hirokazu Kameoka, Takuhiro Kaneko, Kou Tanaka, and Nobukatsu Hojo, “Acvae-vc: Non-parallel many-to-many voice conversion with auxiliary classifier variational autoencoder,” arXiv preprint arXiv:1808.05092, 2018.

[10] Takuhiro Kaneko and Hirokazu Kameoka, “CycleGAN-vc: Non-parallel voice conversion using cycle-consistent adversarial networks,” in Proceedings of European Signal Processing Conference (EUSIPCO), 2018.

[11] Takuhiro Kaneko, Hirokazu Kameoka, Kou Tanaka, and Nobukatsu Hojo, “CycleGAN-vc2: Improved cycleGAN-based non-parallel voice conversion,” in Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, 2019.

[12] Hirokazu Kameoka, Takuhiro Kaneko, Kou Tanaka, and Nobukatsu Hojo, “Stargan-vc: Non-parallel many-to-many voice conversion using star generative adversarial networks,” in IEEE Spoken Language Technology Workshop (SLT), 2018.

[13] Takuhiro Kaneko, Hirokazu Kameoka, Kou Tanaka, and Nobukatsu Hojo, “Stargan-vc2: Rethinking conditional methods for stargan-based voice conversion,” arXiv preprint arXiv:1907.12279, 2019.

[14] Kaizhi Qian, Yang Zhang, Shiyu Chang, Xuesong Yang, and Mark Hasegawa-Johnson, “Autovc: Zero-shot voice style transfer with only autoencoder loss,” in Proceedings of the International Conference on Machine Learning, 2019.

[15] Justin Johnson, Alexandre Alahi, and Li Fei-Fei, “Perceptual losses for real-time style transfer and super-resolution,” in Proceedings of the European Conference on Computer Vision, 2016.

[16] Dmytro Kotovenko, Arsiom Sanakoyeu, Sabine Lang, and Bjorn Ommer, “Content and style disentanglement for artistic style transfer,” in Proceedings of the International Conference on Computer Vision, 2019.

[17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin, “Attention is all you need,” in Advances in Neural Information Processing Systems, 2017.

[18] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu, “Wavenet: A generative model for raw audio,” arXiv preprint arXiv:1609.03499, 2016.

[19] Li Wan, Quan Wang, Alan Papir, and Ignacio Lopez Moreno, “Generalized end-to-end loss for speaker verification,” in Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, 2018.

[20] Ju-chieh Chou, Cheng-chieh Yeh, and Hung-yi Lee, “One-shot voice conversion by separating speaker and content representations with instance normalization,” arXiv preprint arXiv:1904.05742, 2019.