

ASSEM-VC: REALISTIC VOICE CONVERSION BY ASSEMBLING MODERN SPEECH SYNTHESIS TECHNIQUES

Kang-wook Kim^{*†}

Seung-won Park[†]

Junhyeok Lee^{*}

Myun-chul Joe^{*}

^{*} MINDsLab Inc., Republic of Korea

[†] Seoul National University, Republic of Korea

ABSTRACT

Recent works on voice conversion (VC) focus on preserving the rhythm and the intonation as well as the linguistic content. To preserve these features from the source, we decompose current non-parallel VC systems into two encoders and one decoder. We analyze each module with several experiments and reassemble the best components to propose Assem-VC, a new state-of-the-art any-to-many non-parallel VC system. We also examine that PPG and Cotatron features are speaker-dependent, and attempt to remove speaker identity with adversarial training. Code and audio samples are available at <https://github.com/mindslab-ai/assem-vc>.

Index Terms— voice conversion, speech synthesis, speaker disentanglement, adversarial training

1. INTRODUCTION

Voice conversion (VC) is the process of converting the voice of the source speaker to the target speaker's voice without changing the linguistic information. Early attempts in VC [1, 2] used parallel training data, where the linguistic content of the source and the target speakers are the same. The rhythm and the intonation of converted speech needed not to be preserved. However, recent VC studies worked on also retaining the rhythm [3, 4, 5, 6] or the intonation [7, 8] of the source utterance without the need for a parallel dataset. In other words, they only modify timbre while maintaining linguistic content, rhythm, and intonation of the source speech. This imposes a stronger constraint than simply preserving linguistic content, and is more useful for entertainment purposes *e.g.* making an expressive audiobook with celebrities' voices or providing emotional VC service.

Some prior works that adopted autoencoder-based models [7, 9, 10] were able to utilize non-parallel datasets by disentangling speaker information from speech. However, these works did not focus on maintaining the rhythm and the intonation of the source speech. Other attempts used phonetic posteriors (PPG) [3, 4, 5], which are suitable for capturing the linguistic content and the rhythm of speech. Park *et al.* [11] proposed Cotatron, a linguistic feature extractor based on a multispeaker text-to-speech (TTS) model [12]. Valle *et al.*

[13] introduced Mellotron, that can generate emotional and singing voice without explicit training data. Cotatron and Mellotron have the drawback of requiring a transcript of the source speech, while Park *et al.* [11] reported using automatic speech recognition (ASR) transcription can solve this issue without degradation of quality.

In this paper, we decompose current VC systems, experiment with their components, and propose *Assem-VC*. Our contributions are as follows:

- We categorize and analyze the components of existing non-parallel any-to-many VC systems and select the best combination.
- We examine that PPG and Cotatron features are not speaker-independent since such features incorporate speaking rate.
- We experiment with applying adversarial training to linguistic features to remove speaker dependency.
- On naturality and the speaker similarity, Assem-VC achieves the mean opinion scores comparable to the scores of the natural speech in many-to-many settings.

2. APPROACH

We choose the three current state-of-the-art VC systems capable of any-to-many conversion while maintaining the rhythm and the intonation of the source: PPG-VC, Cotatron-VC [11], and Mellotron-VC [13]. We break these models down to a linguistic encoder, an intonation encoder, and a decoder and compare each component. We also suggest several modifications to enhance each component. Finally, we propose Assem-VC, based on the best combination of components.

2.1. Linguistic encoder

PPG Encoder. PPG are sequences that give the posterior probability distribution of phonemes for each time frame, and obtained from a pretrained speaker-independent ASR. Various works [3, 4, 5] employ PPG as speaker-independent linguistic features in VC.

Cotatron. Cotatron [11] is a linguistic encoder based on multispeaker Tacotron 2 [12]. Cotatron learns to predict the mel

spectrogram from the text as well as the alignment between the mel spectrogram and the text transcript:

$$\hat{M}_{1:i}, A_i = \underset{\text{tts}}{\text{Decoder}} \left(\underset{\text{text}}{\text{Encoder}}(T), M_{0:i-1}, z^{id} \right), \quad (1)$$

where T, M, A, z^{id} are text, log mel spectrogram, alignment, and speaker representation, respectively. The outputs of Cotatron are linguistic features L , calculated as following: $L = \text{matmul} \left(A, \underset{\text{text}}{\text{Encoder}}(T) \right)$.

Mellotron Encoder. Mellotron [13] is a multispeaker voice synthesis model based on Tacotron 2 [12] that can synthesize emotive and singing voice without explicit training data. We view the concatenation of context vector calculated from text encoding and estimated alignments over all decoder time steps as the linguistic encoding of Mellotron.

Adversarial Cotatron. We speculate that Cotatron features and PPG are not speaker-independent because the speaker identity can be leaked through rhythm. Nevertheless, the speaker information other than rhythm should not be embedded in the linguistic feature. Thus, we adopt domain adversarial training [14] to Cotatron features L . We train an additional speaker classifier with a gradient reversal layer using linguistic features.

2.2. Intonation encoder

Residual Encoder. Cotatron-VC introduces the residual encoder which encodes intonation into output residual features R [11]. Despite its careful design, whether the residual encoder outputs are speaker-independent is questionable. It is also uncertain whether the residual encoder can properly encode the speech of unseen speakers. The aforementioned risks may affect the quality of the output.

Normalized F0. Mellotron-VC uses fundamental frequency (F0) estimation algorithm, YIN [15] to encode intonation. To make encoding more successful, we made some modifications and applied them to PPG-VC and Assem-VC. We empirically observed that RAPT [16] is more precise than YIN for estimate pitch contours, use RAPT for intonation encoder of Assem-VC. After extracting the log F0 for each frame of all audios from a given speaker, we normalize them based on the mean and the standard deviation of the log F0 distribution of the speaker. Since the speaker identity cannot be fully removed from the pitch contour with different unvoiced values across different speakers, we fill unvoiced segments with a constant value of -10 after the normalization.

2.3. Decoder

Causal Decoder. Out of all the models we deal with, only Mellotron-VC used a causal decoder. During the decoding step, the output is generated in an autoregressive manner by using the linguistic encoding, the intonation encoding, and

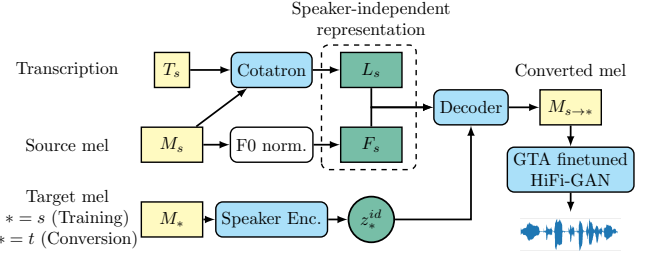


Fig. 1: The overall architecture of Assem-VC. Speaker representation z^{id} is extracted from the speaker encoder conditions the VC decoder via conditional batch normalization layer. s and t are short for the source and the target, respectively.

the previous mel frames. This architecture is capable of generating higher quality mel spectrograms than non-causal decoders. However, since a causal decoder is trained through teacher forcing, it may learn to cheat off of previous mel spectrogram frames, which are highly dependent on the source speaker. Thus, speaker disentanglement may not be achieved even if the speaker-independent features are used.

Non-causal Decoder. A fully convolutional non-causal decoder is used in PPG-VC, Cotatron-VC, and Assem-VC. The structure of the decoder is the same as Cotatron-VC [11] except the speaker conditioning method. Instead of using speaker embedding for conditioning, we use an additional speaker encoder to capture the variation of target speech in the entire corpus. Extracted speaker representation conditions the decoder via conditional batch normalization layer [17].

2.4. Assem-VC

We experiment with two different versions of Assem-VC; one uses Cotatron as the linguistic encoder and the other uses adversarial Cotatron. The linguistic features are concatenated with normalized F0 and fed to a non-causal VC decoder. The decoder uses an additional speaker encoder for speaker conditioning. For the vocoder, we use *ground truth alignment* (GTA) finetuned HiFi-GAN. Further details about GTA fine-tuning are described in 3.1. Fig. 1 shows the architecture of our proposed VC system, Assem-VC. Cotatron in the figure can be replaced with adversarial Cotatron.

3. EXPERIMENTS

3.1. Training

Dataset. VCTK [18] and LibriTTS [19] are utilized in our experiments. First, following the previous literature [11], we split VCTK into 80% train set, 10% validation set, and 10% test set, without any overlap of the transcript between them. Note that the speakers of VCTK are shared across the splits. Then, the systems are trained on the union of VCTK train set and LibriTTS *train-clean-100* subset. Finally, VCTK test

split is used to evaluate many-to-many conversion, while the *test-clean* subset of LibriTTS is used as the source audio for the unseen speakers to evaluate any-to-many conversion.

To enhance the stability of the training process, only the speakers with more than 5 minutes of utterances in LibriTTS *train-clean-100* split are utilized. We only used audios shorter than 10 seconds for efficient batching. All audios are down-sampled to 22.05 kHz and then normalized. The mel spectrogram configuration is the same as used in Park *et al.* [11].

PPG-VC. A pretrained wav2letter++ [20] is utilized to extract the speaker-independent PPG. The ASR model is trained using LibriSpeech [21] and Common Voice [22]. The CER and the WER for the LibriSpeech *dev-clean* split are 2.32% and 6.82% without language model. Since the frame rate of PPG is different from that of mel spectrogram, PPG are linearly interpolated to match the length of the mel spectrogram.

Cotatron-VC & Assem-VC. For Cotatron-VC and Assem-VC, we first train Cotatron and train the whole VC system with the pretrained Cotatron fixed. To stabilize the alignment learning of Cotatron, we initially train using only the LibriTTS *train-clean-100* subset and later add VCTK. For adversarial Cotatron, the structure and training strategy of the speaker classifier are identical to those proposed in Zhang *et al.* [23]. We follow the settings of Park *et al.* [11] for other training details.

Mellotron-VC. Unlike Cotatron-VC, Mellotron-VC [13] is trained in a single stage. The transfer learning strategy from Cotatron-VC is employed for stable alignment learning. Similar to Assem-VC, we normalize the pitch with respect to the speaker. Other settings and architecture are consistent with the original Mellotron.

HiFi-GAN Vocoder. All systems employ HiFi-GAN [24] as the vocoder. HiFi-GAN is pretrained with 123 hours of proprietary multi-speaker TTS dataset and then finetuned on the train split of VCTK.

GTA Finetuning. Inspired by Shen *et al.* [12], we utilize GTA mel spectrograms when finetuning HiFi-GAN after training the VC system. HiFi-GAN takes the reconstructed mel spectrogram from the VC system as the input and learns to predict the source speech. To the best of our knowledge, this is the first application of GTA finetuning for VC.

3.2. Conversion and evaluation

We evaluate the many-to-many scenario by testing the conversion between the speakers from VCTK. For the any-to-many scenario, we select a speech from *test-clean* subset of LibriTTS as the unseen source speakers. In both scenarios, only the speakers from VCTK are selected as target.

The systems are mainly evaluated with two subjective metrics with the same details as [11]: mean opinion score (MOS) and degradation mean opinion score (DMOS). MOS measures the naturalness of speech, and DMOS measures the

speaker similarity between the target speaker’s speech and the converted speech.

3.3. Quantifying the degree of disentanglement

To quantify the degree of speaker disentanglement of features, we measure the accuracy of the speaker classifier trained with such features. The speaker classification accuracy (SCA) is higher if the features are more entangled with the speaker identity. The classifier is trained and tested on VCTK, and SCA values are reported after the validation accuracy plateaus for each feature. The classifier consists of 4 convolutional layers and batch normalization layers, followed by a max-pooling layer and a dense layer with dropout. Unlike other features, the size of alignments is not fixed, due to varying lengths of mel spectrogram and text. To reduce the text index dimension, we compute weighted text indices for each frame using the alignments values to train a classifier.

4. RESULTS

4.1. Degree of disentanglement

We measure the SCA of the features to compare their degree of disentanglement, which is shown in Table 1. Since speaker information such as speaking rate still remains in linguistic features, SCA of PPG and Cotatron features L are higher than random guessing. Since the Cotatron features have SCA close to that of the PPG, we speculate that the PPG and the Cotatron features are equally capable in terms of speaker disentanglement. Lastly, the normalized pitch contour F is more speaker-independent than the residual encoding R from Cotatron-VC.

Table 1: Degree of speaker disentanglement. Rand. denotes the case of the classifier trained with the features filled with random numbers. *adv* indicates the features from adversarial Cotatron. L , F , R , A , M denote linguistic features, normalized F0, residual features, alignments, mel spectrogram.

Feature	Rand.	PPG	L	(L, F)	(L, R)
SCA	0.9%	31.6%	32.8%	41.7%	57.9%
Feature	A	A_{adv}	L_{adv}	M	
SCA	35.8%	24.0%	33.3%	99.5%	

To reveal the effect of adversarial training, we additionally measure the SCA of linguistic features and alignments of adversarial Cotatron. Adversarial training helped to reduce the speaker dependency of alignments A , but it does not affect the dependency of Cotatron features L . Also, adversarial training hinders Cotatron from learning the alignments between audio and its transcript, which in turn impairs the voice conversion model’s ability to learn the proper pronunciation. As shown

Table 2: Comparison of components and performances of the examined VC approaches. * *GTA* denotes whether the HiFi-GAN vocoder is finetuned with a ground-truth aligned mel spectrogram. Adv. Cotatron is short for adversarial Cotatron.

Approach	Components				Many-to-many		Any-to-many	
	Linguistic	Intonation	Decoder	GTA *	MOS	DMOS	MOS	DMOS
Source as target					4.14 ± 0.08	1.71 ± 0.14	–	–
Target as target					4.14 ± 0.08	3.94 ± 0.12	–	–
PPG-VC	PPG	F0 (RAPT)	Non-causal	✗	2.47 ± 0.09	2.99 ± 0.13	–	–
				✓	3.81 ± 0.09	3.78 ± 0.13	3.63 ± 0.11	3.66 ± 0.13
Cotatron-VC	Cotatron	Residual	Non-causal	✗	2.60 ± 0.12	3.44 ± 0.15	–	–
				✓	3.75 ± 0.09	3.79 ± 0.13	2.99 ± 0.15	3.31 ± 0.15
Mellotron-VC	Tacotron2	F0 (YIN)	Causal	✗	2.94 ± 0.11	2.87 ± 0.14	–	–
				✓	3.36 ± 0.10	3.15 ± 0.13	3.42 ± 0.11	2.84 ± 0.13
Assem-VC	Cotatron	F0 (RAPT)	Non-causal	✗	2.87 ± 0.11	3.27 ± 0.11	–	–
				✓	3.91 ± 0.10	3.86 ± 0.12	3.61 ± 0.12	3.86 ± 0.11
	Adv. Cotatron	F0 (RAPT)	Non-causal	✗	2.35 ± 0.13	3.24 ± 0.14	–	–
				✓	3.56 ± 0.11	3.69 ± 0.13	3.25 ± 0.13	3.48 ± 0.14

in Table 2, there is a degradation of quality when adopting adversarial training on Cotatron.

4.2. Many-to-many conversion

As presented in Table 2, our proposed combination Assem-VC outperforms all other methods in terms of both MOS with DMOS, and both of its scores close to those of natural speech. GTA finetuning plays a key role in removing noisy artifacts, improving MOS and DMOS of all models. We observe that the increase in MOS and DMOS after GTA finetuning is relatively low for Mellotron-VC. We suspect that causal decoders have an advantage in the amount of information available during generation of each mel frame. However, such setting also has a negative impact on speaker disentanglement, resulting in the lowest DMOS across the board. PPG-VC has the closest MOS to Assem-VC, but it still performs poorly on DMOS. Cotatron-VC shows worse performance than Assem-VC on both MOS and DMOS, with or without the GTA finetuning. These results suggest that a combination of Cotatron and F0 estimator is better than PPG encoder and the residual encoder as linguistic and intonation encoder from the source speech.

4.3. Any-to-many conversion

To evaluate the generalization to the unseen speakers, we also explore any-to-many situation, namely to convert the speech of an arbitrary speaker to the voices of multiple speakers. As shown in Table 2, the overall results of the any-to-many conversion are lower than the many-to-many, since any-to-many is inherently a more challenging task. Assem-VC using Cotatron outperforms other baseline models. PPG-VC yields a competitive MOS with Assem-VC, but a worse DMOS. The residual encoder cannot properly encode the intonation with unseen speakers, which causes MOS of Cotatron-VC to be

the lowest. Mellotron-VC’s causal decoder fails to preserve the speaker’s characteristics, resulting in the worst DMOS.

5. CONCLUSION

In this paper, we compared the existing methods on any-to-many non-parallel VC. We decompose these models into three components: a linguistic encoder, an intonation encoder, and a decoder. We plug different combinations of modules into these components and report the results. First, our results suggest that the Cotatron features are better than PPG at representing the linguistic features of a speech without sacrificing the performance on speaker similarity. Also, the traditional F0 estimator is better than the residual encoder at disentangling the intonation from a speech. We also show that the causal decoder has a negative effect on speaker disentanglement. For the first time, to the best of our knowledge, we apply GTA finetuning for VC, so that non-causal decoders generate a natural speech that captures the target speaker identity. We observed that adversarial training for speaker disentanglement interferes with Cotatron’s ability to learn alignments. We select the best features and re-assemble to present an any-to-many VC system, Assem-VC. Assem-VC achieves the level of natural speech recordings in both MOS and DMOS for any-to-many VC with 108 target speakers. Our results indicate that Assem-VC can be applied to real-world applications which require voice conversion indistinguishable from human speech.

6. ACKNOWLEDGEMENTS

The authors would like to thank Sang Hoon Woo, Seungyu Han, and Dongho Choi from MINDsLab Inc., Yoonhyung Lee from Seoul National University for providing beneficial feedback on the initial version of this paper.

7. REFERENCES

- [1] M.Zhang, X.Wang, F.Fang, H.Li, and J.Yamagishi, "Joint training framework for text-to-speech and voice conversion using multi-source tacotron and wavenet," *Proc. Interspeech 2019*, pp. 1298–1302, 2019.
- [2] J.Zhang, Z.Ling, Y.Jiang, L.Liu, C.Liang, and L.Dai, "Improving sequence-to-sequence voice conversion by adding text-supervision," in *ICASSP*. IEEE, 2019, pp. 6785–6789.
- [3] L.Sun, K.Li, H.Wang, S.Kang, and H.Meng, "Phonetic posteriorgrams for many-to-one voice conversion without parallel data training," in *ICME*. IEEE, 2016, pp. 1–6.
- [4] Y.Saito, Y.Ijima, K.Nishida, and S.Takamichi, "Non-parallel voice conversion using variational autoencoders conditioned by phonetic posteriorgrams and d-vectors," in *ICASSP*. IEEE, 2018, pp. 5274–5278.
- [5] H.Lu, Z.Wu, D.Dai, R.Li, S.Kang, J.Jia, and H.Meng, "One-shot voice conversion with global speaker embeddings," *Proc. Interspeech 2019*, pp. 669–673, 2019.
- [6] A.Polyak, L.Wolf, and Y.Taigman, "Tts skins: Speaker conversion via asr," *Proc. Interspeech 2020*, pp. 786–790, 2019.
- [7] K.Qian, Z.Jin, M.Hasegawa-Johnson, and G. J.Mysore, "F0-consistent many-to-many non-parallel voice conversion via conditional autoencoder," in *ICASSP*. IEEE, 2020, pp. 6284–6288.
- [8] S.Liu, Y.Cao, S.Kang, N.Hu, X.Liu, D.Su, D.Yu, and H.Meng, "Transferring Source Style in Non-Parallel Voice Conversion," *Proc. Interspeech 2020*, pp. 4721–4725, 2020.
- [9] K.Qian, Y.Zhang, S.Chang, X.Yang, and M.Hasegawa-Johnson, "AutoVC: Zero-shot voice style transfer with only autoencoder loss," in *ICML*, 2019, pp. 5210–5219.
- [10] K.Qian, Y.Zhang, S.Chang, M.Hasegawa-Johnson, and D.Cox, "Unsupervised speech decomposition via triple information bottleneck," in *ICML*, 2020, pp. 7836–7846.
- [11] S.-W.Park, D.-Y.Kim, and M.-C.Joe, "Cotatron: Transcription-guided speech encoder for any-to-many voice conversion without parallel data," *Proc. Interspeech 2020*, pp. 4696–4700, 2020.
- [12] J.Shen, R.Pang, R. J.Weiss, M.Schuster, N.Jaitly, Z.Yang, Z.Chen, Y.Zhang, Y.Wang, R.Skerry-Ryan, et al., "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," in *ICASSP*. IEEE, 2018, pp. 4779–4783.
- [13] R.Valle, J.Li, R. J.Prenger, and B.Catanzaro, "Melotron: Multispeaker expressive voice synthesis by conditioning on rhythm, pitch and global style tokens," in *ICASSP*. IEEE, 2020, pp. 6189–6193.
- [14] Y.Ganin, E.Ustinova, H.Ajkan, P.Germain, H.Larochelle, F.Lavolette, M.Marchand, and V.Lempitsky, "Domain-adversarial training of neural networks," *The journal of machine learning research*, vol. 17, no. 1, pp. 2096–2030, 2016.
- [15] A.De Cheveigné and H.Kawahara, "Yin, a fundamental frequency estimator for speech and music," *The Journal of the Acoustical Society of America*, vol. 111, no. 4, pp. 1917–1930, 2002.
- [16] D.Talkin, "A robust algorithm for pitch tracking (rapt)," *Speech coding and synthesis*, vol. 495, pp. 518, 1995.
- [17] V.Dumoulin, J.Shlens, and M.Kudlur, "A learned representation for artistic style," in *ICLR*, 2017.
- [18] J.Yamagishi, C.Veaux, K.MacDonald, et al., "Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit (version 0.80)," 2012.
- [19] H.Zen, R.Clark, R. J.Weiss, V.Dang, Y.Jia, Y.Wu, Y.Zhang, and Z.Chen, "Libritts: A corpus derived from librispeech for text-to-speech," *Proc. Interspeech 2019*, pp. 1526–1530, 2019.
- [20] V.Pratap, A.Hannun, Q.Xu, J.Cai, J.Kahn, G.Synnaeve, V.Liptchinsky, and R.Collobert, "Wav2letter++: A fast open-source speech recognition system," in *ICASSP*. IEEE, 2019, pp. 6460–6464.
- [21] V.Panayotov, G.Chen, D.Povey, and S.Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *ICASSP*. IEEE, 2015, pp. 5206–5210.
- [22] R.Ardila, M.Branson, K.Davis, M.Henretty, M.Kohler, J.Meyer, R.Morais, L.Saunders, F. M.Tyers, and G.Weber, "Common voice: A massively-multilingual speech corpus," *arXiv preprint arXiv:1912.06670*, 2019.
- [23] Y.Zhang, R. J.Weiss, H.Zen, Y.Wu, Z.Chen, R.Skerry-Ryan, Y.Jia, A.Rosenberg, and B.Ramabhadran, "Learning to Speak Fluently in a Foreign Language: Multilingual Speech Synthesis and Cross-Language Voice Cloning," *Proc. Interspeech 2019*, pp. 2080–2084, 2019.
- [24] J.Kong, J.Kim, and J.Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," in *NeurIPS*, 2020, pp. 17022–17033.