Voice Conversion Survey

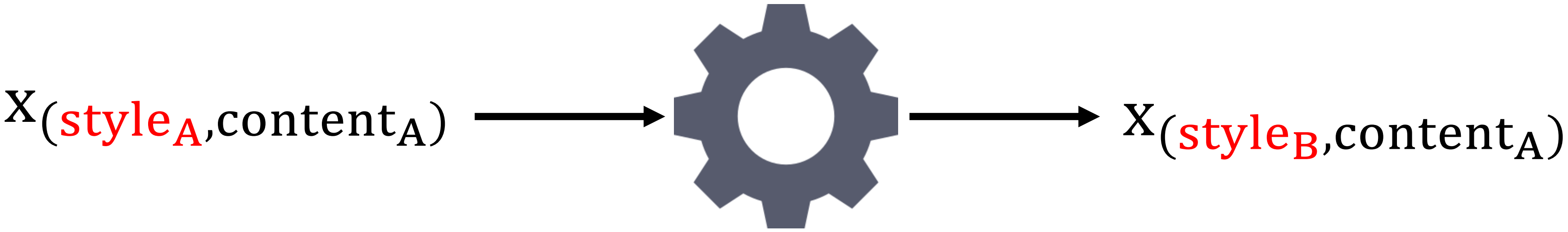
IELAB

정영석

Style Transfer

Style Transfer Task

- Definition of Style Transfer task
  - ✓ Style transfer is one of the most crucial task in Machine Learning. It is being studied regardless of field (Vision, NLP and audio).
  - ✓ Style transfer target to maintain the input data's contents and convert input data's style into other domains.



$$X_{(\text{style}_A,\text{content}_A)} \longrightarrow \quad \longrightarrow X_{(\text{style}_B,\text{content}_A)}$$

Voice Conversion

Voice Conversion

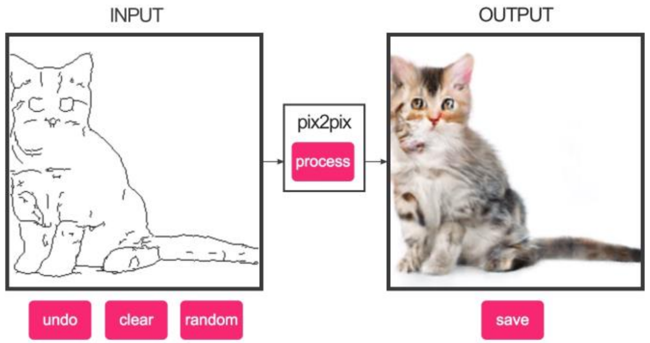
- Contents and Style information in Voice Conversion
  - ✓ In Voice Conversion task, people consider linguistic information as contents vector.
  - ✓ All information related to the speaker is defined as style information.
  - ✓ Convert A style Utterance to B style Utterance.

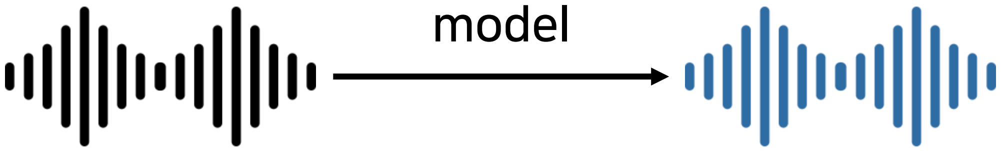
$$U_{(\mathrm{style}_{\mathrm{A}},\mathrm{content}_{\mathrm{A}})} \longrightarrow ⚙ \longrightarrow U_{(\mathrm{style}_{\mathrm{B}},\mathrm{content}_{\mathrm{A}})}$$

Voice Conversion

▪ The method for Voice Conversion

- The methods for parallel dataset.
  - ✓ Many traditional VC methods require parallel dataset.
  - ✓ This can be problematic since misalignment involved in parallel data can cause speech-quality degradation: thus, it require pre-screening method.
  - ✓ Moreover, collecting parallel data can be painstaking process in real application scenarios.



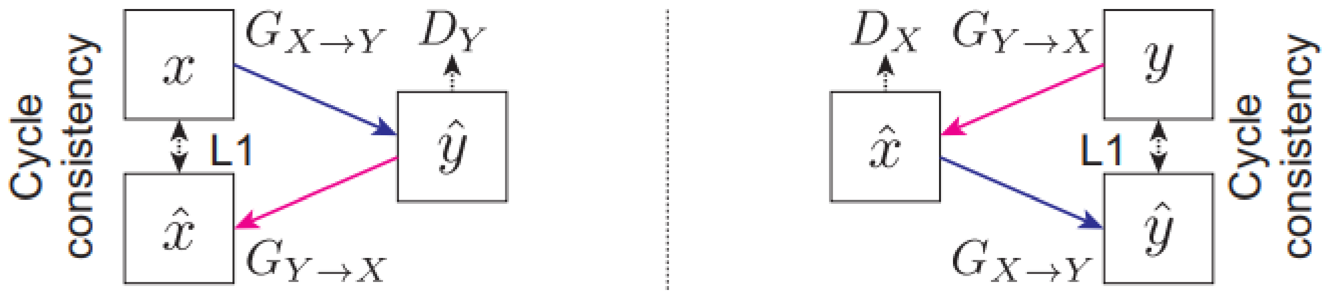


1. S. Desai, A. W. Black, B. Yegnanarayana, and K. Prahallad, "Spectral mapping using artificial neural networks for voice conversion," IEEE/ACM Trans. Audio Speech Lang. Process., vol. 18, no. 5, pp. 954– 964, 2010.
2. S. H. Mohammadi and A. Kain, "Voice conversion using deep neural networks with speaker-independent pre-training," in Proc. SLT, 2014, pp. 19–23.
3. K. Oyamada, H. Kameoka, T. Kaneko, H. Ando, K. Hiramatsu, and K. Kashino, "Non-native speech conversion with consistency-aware recursive network and generative adversarial network," in Proc. APSIPA ASC, 2017, pp. 182–188.
4. T. Nakashika, T. Takiguchi, and Y. Ariki, "High-order sequence modeling using speaker-dependent recurrent temporal restricted Boltzmann machines for voice conversion," in Proc. Interspeech, 2014, pp. 2278– 2282.
5. L. Sun, S. Kang, K. Li, and H. Meng, "Voice conversion using deep bidirectional long short-term memory based recurrent neural networks," in Proc. ICASSP, 2015, pp. 4869–4873.

Voice Conversion

▪ The method for Voice Conversion

- The methods for non-parallel voice.
  - ✓ The fore mentioned methods have some limitation.
  - ✓ Some studies try to apply GAN-based model (CycleGAN, StarGAN, …) already utilized in Vision domain.
  - ✓ The GAN-based models are not appropriate for zero-shot conversion.



(a) Forward-inverse mapping

(b) Inverse-forward mapping

Fig. 1. Training procedure of CycleGAN

5. T. Kaneko and H. Kameoka, "CycleGAN-VC: Non-parallel Voice Conversion Using Cycle-Consistent Adversarial Networks," *2018 26th European Signal Processing Conference (EUSIPCO)*, 2018, pp. 2100-2104
6. T. Kaneko, H. Kameoka, K. Tanaka and N. Hojo, "Cyclegan-VC2: Improved Cyclegan-based Non-parallel Voice Conversion," *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 6820-6824,
7. H. Kameoka, T. Kaneko, K. Tanaka and N. Hojo, "StarGAN-VC: non-parallel many-to-many Voice Conversion Using Star Generative Adversarial Networks," *2018 IEEE Spoken Language Technology Workshop (SLT)*, 2018
8. Kaneko, Takuhiro, et al. "StarGAN-VC2: Rethinking conditional methods for StarGAN-based voice conversion." *arXiv preprint arXiv:1907.12279* (2019).

Voice Conversion

▪ The method for Voice Conversion

- Disentangle features for zero-shot conversion

  ✓ Disentangle method

  1. Gradient reversal

  2. Sufficiently shallow bottleneck

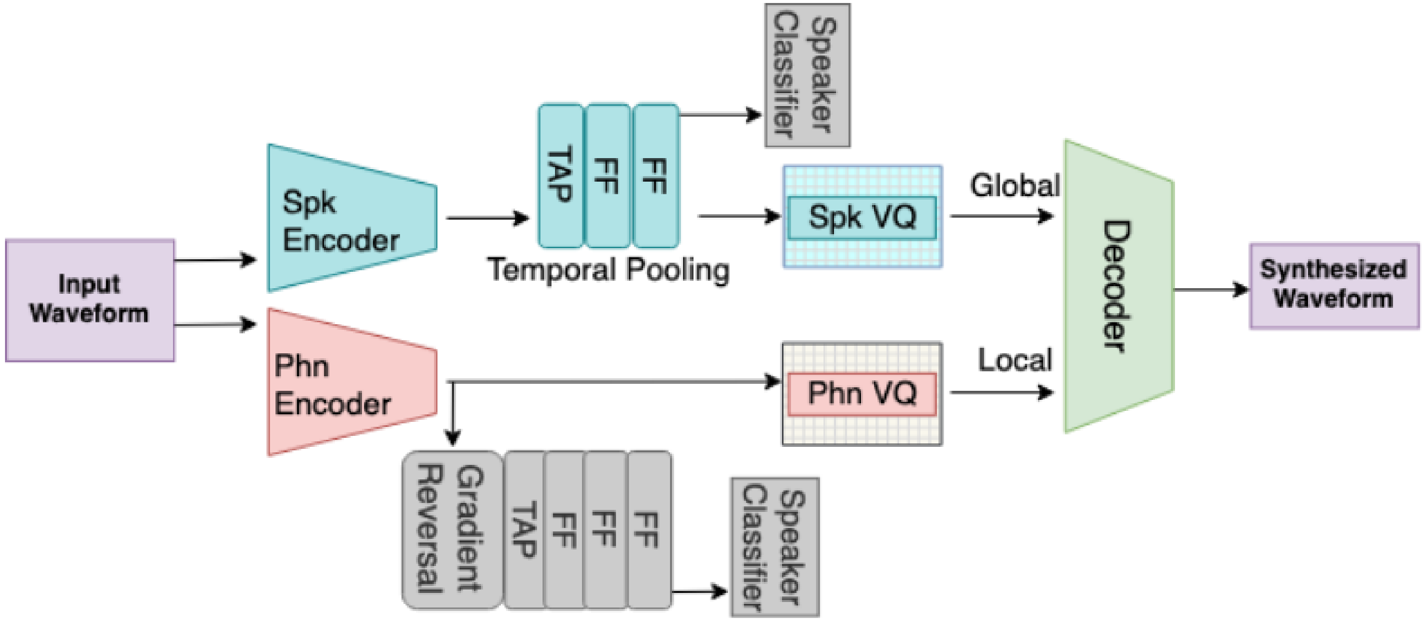
  3. Instance Normalization

  4. Minimize mutual information loss

Voice Conversion

▪ The method for Voice Conversion

- Disentangle features for zero-shot conversion (Gradient Reversal)

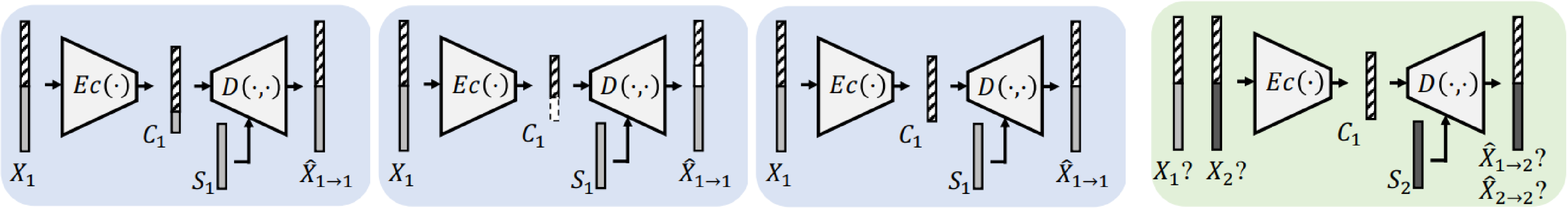


9. Williams, Jennifer, et al. "Learning Disentangled Phone and Speaker Representations in a Semi-Supervised VQ-VAE Paradigm." *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021.
10. Mor, Noam, et al. "A universal music translation network." *arXiv preprint arXiv:1805.07848* (2018).

Voice Conversion

The method for Voice Conversion

- Disentangle features for zero-shot conversion (shallow bottleneck, etc.,)



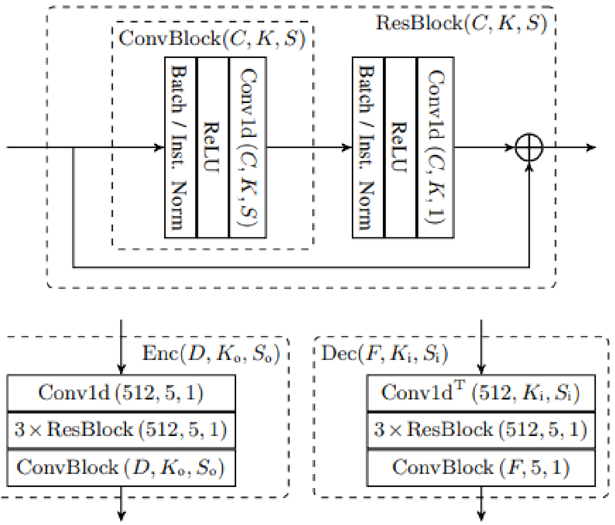
(a) Bottleneck too wide (b) Bottleneck too narrow (c) Bottleneck just right (d) Conversion

9. Williams, Jennifer, et al. "Learning Disentangled Phone and Speaker Representations in a Semi-Supervised VQ-VAE Paradigm." *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021.
10. Mor, Noam, et al. "A universal music translation network." *arXiv preprint arXiv:1805.07848* (2018).
11. Wang, Shijun, and Damian Borth. "NoiseVC: Towards High Quality Zero-Shot Voice Conversion." *arXiv preprint arXiv:2104.06074* (2021).
12. Oord, Aaron van den, Oriol Vinyals, and Koray Kavukcuoglu. "Neural discrete representation learning." *arXiv preprint arXiv:1711.00937* (2017).
13. Cífka, Ondřej, et al. "Self-Supervised VQ-VAE for One-Shot Music Style Transfer." *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021.

Voice Conversion

The method for Voice Conversion

- Disentangle features for zero-shot conversion (Instance Normalization)



$$\mu_{nc} = \frac{1}{HW}\sum_{j=1}^{H}\sum_{k=1}^{W} x_{ncjk}$$

$$\sigma^2_{nc} = \frac{1}{HW}\sum_{j=1}^{H}\sum_{k=1}^{W}(x_{ncjk} - \mu_{nc})^2$$

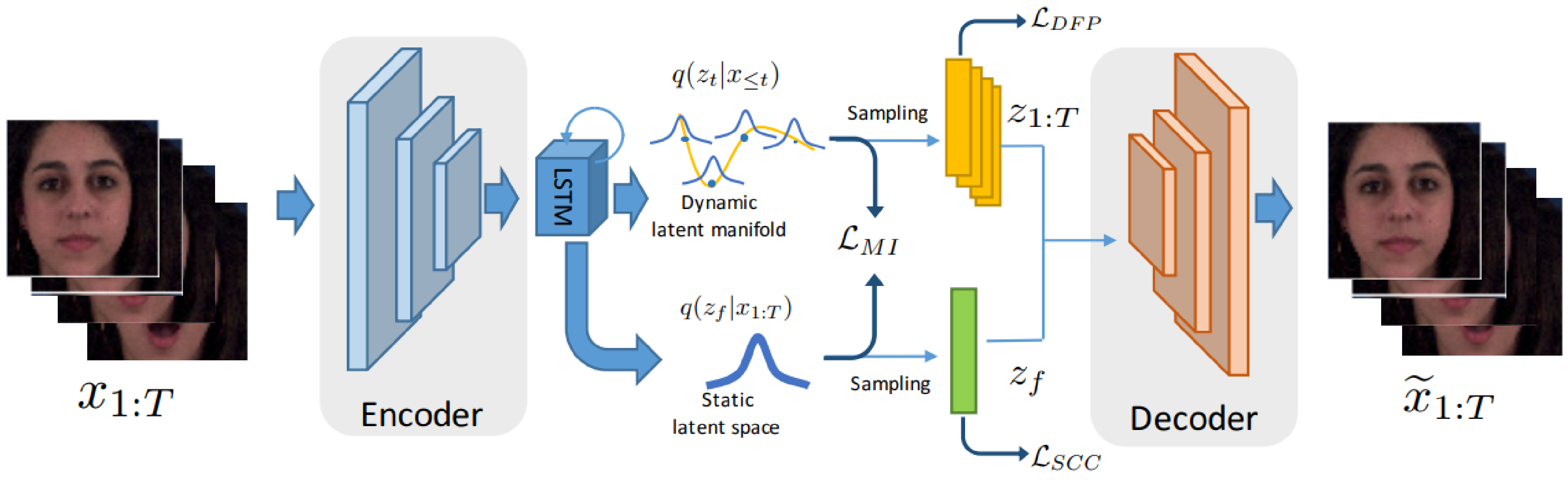
$$\hat{x} = \frac{x - \mu_{nc}}{\sqrt{\sigma^2_{nc} + \epsilon}}$$

14. Ebbers, Janek, et al. "Contrastive Predictive Coding Supported Factorized Variational Autoencoder for Unsupervised Learning of Disentangled Speech Representations." *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021.
15. Wang, Shijun, and Damian Borth. "NoiseVC: Towards High Quality Zero-Shot Voice Conversion." *arXiv preprint arXiv:2104.06074* (2021).

Voice Conversion

The method for Voice Conversion

- Disentangle features for zero-shot conversion (Minimize Mutual Information)

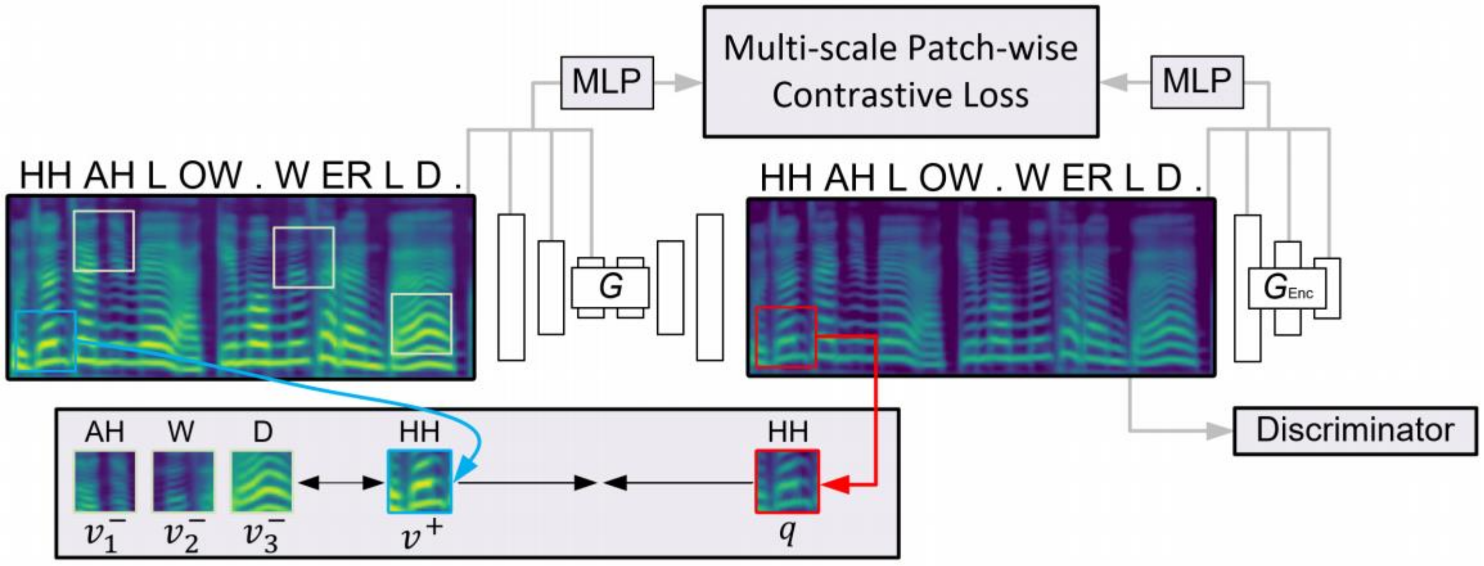


16. Zhu, Yizhe, et al. "S3VAE: Self-supervised sequential VAE for representation disentanglement and data generation." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020.
17. Liang, Shuang, et al. "Unsupervised Learning for Multi-Style Speech Synthesis with Limited Data." *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021.

Voice Conversion

▪ The method for Voice Conversion

- Other methods…



18. Li, Tingle, et al. "CVC: Contrastive Learning for Non-parallel Voice Conversion." *arXiv preprint arXiv:2011.00782* (2020).