FastVC: Fast Voice Conversion with non-parallel data

Oriol Barbany[1,2] Milos Cernak[1]

INTERSPEECH 2020

[1]Logitech Europe S.A., 1015, Lausanne, Switzerland
[2]École Polytechnique Fédérale de Lausanne (EPFL), 1015, Lausanne, Switzerland

Introduction

The motivation of this work:

- **Simple** approach (Occam's razor).
- **Fast** inference.
- **Competent** system in terms of quality.

Baseline

AutoVC [Qian et al., 2019]: Conditional AutoEncoder for zero-shot VC.

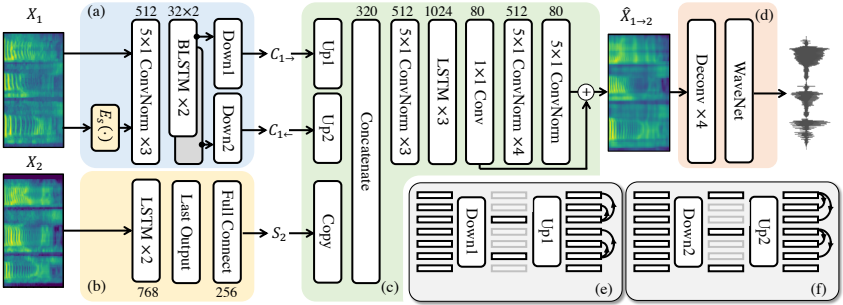


Figure 1: AutoVC model architecture in conversion mode. Figure from [Qian et al., 2019].

FastVC

```
$ diff autovc.pt fastvc.pt
```

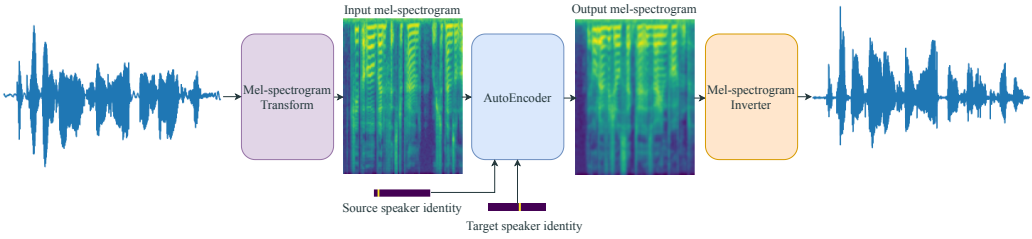


Figure 2: FastVC model architecture in conversion mode.

Raw data as input:

- No pre-processing: No Butterworth high pass filter [Butterworth, 1930] nor noise addition.
- Mel filter-bank using all frequency bins.

CNN-based mel-spectrogram: Allow learning the transformation.

Use speaker one-hot encoded identities instead of computing speaker embeddings.

Different bottleneck hyper-parameters:

- Double temporal downsampling factor, which matches design choice in [Qian et al., 2020].
- Half latent dimension.

MelGAN [Kumar et al., 2019] instead of WaveNet [van den Oord et al., 2016].

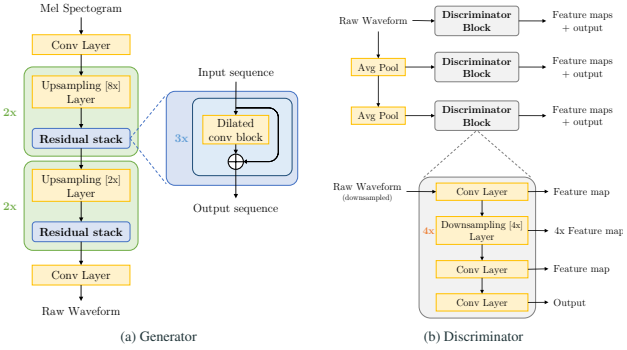


Figure 3: MelGAN model architecture. Figure from [Kumar et al., 2019].

Results - Audio samples

Example: English Female to German Male conversion.

- Source

Results - Audio samples

Example: English Female to German Male conversion.

- Source
- Target

Results - Audio samples

Example: English Female to German Male conversion.

- Source
- Target
- ASR+TTS [Hayashi et al., 2019, Inaguma et al., 2020, Watanabe et al., 2018]

Results - Audio samples

Example: English Female to German Male conversion.

- Source
- Target
- ASR+TTS [Hayashi et al., 2019, Inaguma et al., 2020, Watanabe et al., 2018]
- CycleVAE [Tobing et al., 2019]

Results - Audio samples

Example: English Female to German Male conversion.

- Source
- Target
- ASR+TTS [Hayashi et al., 2019, Inaguma et al., 2020, Watanabe et al., 2018]
- CycleVAE [Tobing et al., 2019]
- AutoVC [Qian et al., 2019]

Results - Audio samples

Example: English Female to German Male conversion.

- Source
- Target
- ASR+TTS [Hayashi et al., 2019, Inaguma et al., 2020, Watanabe et al., 2018]
- CycleVAE [Tobing et al., 2019]
- AutoVC [Qian et al., 2019]
- FastVC

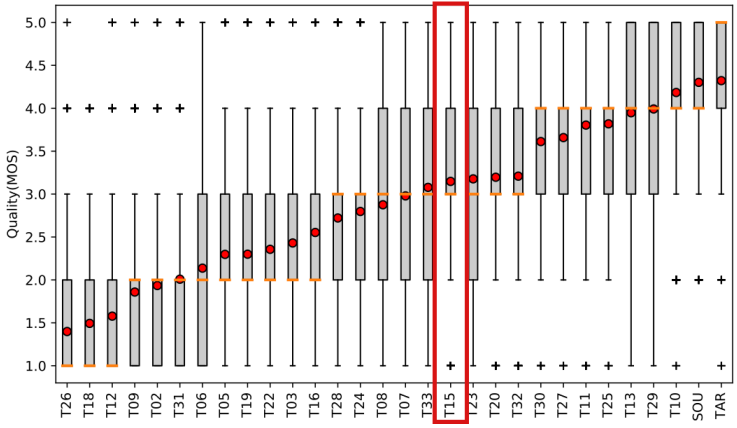
Results - Objective evaluation

Assess self-reconstruction (valid VC instance).
Related to quality for VC systems trained on self-reconstruction with **speaker-independent latents**.

Experiment	PESQ
AutoVC – baseline [Qian et al., 2019]	$\mathbf{2.56 \pm 0.23}$
FastVC with information bottleneck proposed in [Qian et al., 2019]	2.57 ± 0.25
FastVC (VCC20 submission)	$\mathbf{2.68 \pm 0.22}$
FastVC with end-to-end training	1.56 ± 0.29

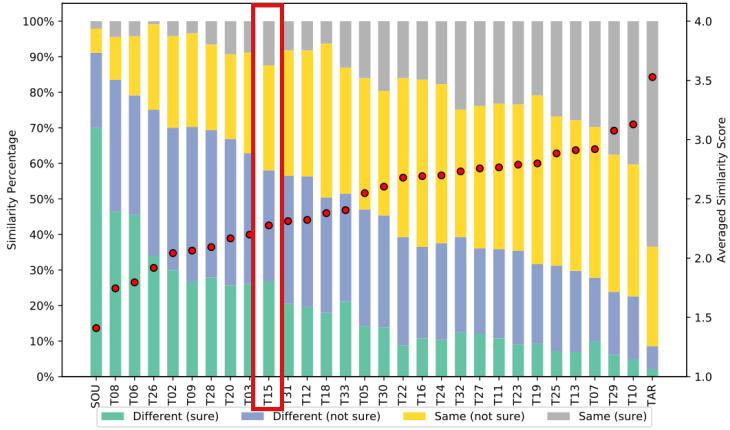
Results - Subjective evaluation

VC Challenge 2020 - Cross-lingual VC task.



Results - Subjective evaluation

VC Challenge 2020 - Cross-lingual VC task.



Conclusions

FastVC:

- **Simple**: Only need speech waveform and speaker ID.
- **Fast**[1]: 500x faster than AutoVC [Qian et al., 2019] and 4x than real time.
- **Competent**: Beats VC Challenge 2020 baselines in terms of quality.

[1] Results on Intel(R) Core(TM) i7-8700K @ 3.70GHz CPU.

References

Butterworth, S. (1930).
On the Theory of Filter Amplifiers.
In *Experimental Wireless and the Wireless Engineer*, pages 536–541.

Hayashi, T., Yamamoto, R., Inoue, K., Yoshimura, T., Watanabe, S., Toda, T., Takeda, K., Zhang, Y., and Tan, X. (2019).
ESPnet-TTS: Unified, Reproducible, and Integratable Open Source End-to-End Text-to-Speech Toolkit.

Inaguma, H., Kiyono, S., Duh, K., Karita, S., Soplin, N. E. Y., Hayashi, T., and Watanabe, S. (2020).
ESPnet-ST: All-in-One Speech Translation Toolkit.
arXiv preprint arXiv:2004.10234.

Kumar, K., Kumar, R., de Boissiere, T., Gestin, L., Teoh, W. Z., Sotelo, J., de Brébisson, A., Bengio, Y., and Courville, A. C. (2019).
MelGAN: Generative Adversarial Networks for Conditional Waveform Synthesis.
In *Advances in Neural Information Processing Systems 32*, pages 14910–14921. Curran Associates, Inc.

Qian, K., Jin, Z., Hasegawa-Johnson, M., and Mysore, G. J. (2020).
F0-Consistent Many-To-Many Non-Parallel Voice Conversion Via Conditional Autoencoder.
ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).

Qian, K., Zhang, Y., Chang, S., Yang, X., and Hasegawa-Johnson, M. (2019).
AUTOVC: Zero-Shot Voice Style Transfer with Only Autoencoder Loss.

Tobing, P. L., Wu, Y.-C., Hayashi, T., Kobayashi, K., and Toda, T. (2019).
Non-Parallel Voice Conversion with Cyclic Variational Autoencoder.

van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A. W., and Kavukcuoglu, K. (2016).
WaveNet: A Generative Model for Raw Audio.
CoRR, abs/1609.03499.

Watanabe, S., Hori, T., Karita, S., Hayashi, T., Nishitoba, J., Unno, Y., Enrique Yalta Soplin, N., Heymann, J., Wiesner, M., Chen, N., Renduchintala, A., and Ochiai, T. (2018).
ESPnet: End-to-End Speech Processing Toolkit.
In *Interspeech*, pages 2207–2211.

Thank you!