The Andrew and Erna Viterbi Faculty of
ELECTRICAL & COMPUTER
ENGINEERING

SIPL Signal and Image Processing Lab.

TECHNION
Israel Institute
of Technology

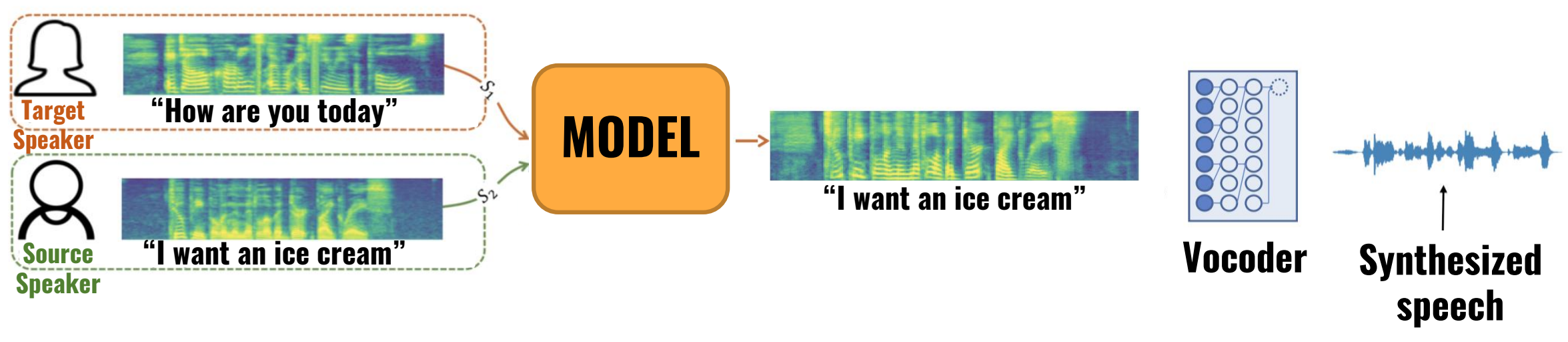
Work in Progress

Voice DeepFake

Idan Roth and Zahi Cohen, Supervised by Yair Moshe

Introduction

- Voice Conversion (VC) is a method of converting one's voice into the sound of another without changing its linguistic content
- May be of use in many real-world applications
- VC research has witnessed an important technological milestone with the appearance of deep learning techniques
- Many VC models suffer from the limitation of only converting voices they have been trained on
- One-shot VC approach solve this limitation
  - VC is performed with only one utterance received from the source and target speakers



Goals

- Convert one's voice to sound like that of another without changing the linguistic content
  - Reproduce results of a state-of-the-art VC one shot model
  - Suggest new ideas for improvement

Challenges

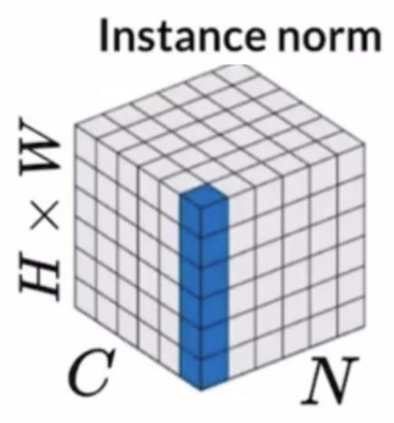
- Establish the mapping between the unparallel training data
  - Source and target speaker speech with different linguistic content
- Perform VC with only receiving one utterance from the source and target speakers
  - Speakers are not seen during training

Speech Information Disentanglement

- Speech is composed of the following:
  - Speaker voice identity information - time invariant characteristics
  - Linguistic content information - time variant characteristics
- Instance Normalization (IN) and Adaptive Instance Normalization (AdaIN) layers perform style transfer
  - Transfers the style of the source speaker to the style of target speaker while maintaining the linguistic content of source speaker.

$$\mathrm{IN}(\boldsymbol{Z}) = \frac{\boldsymbol{Z} - \mu(\boldsymbol{Z})}{\sigma(\boldsymbol{Z})}$$

Instance norm

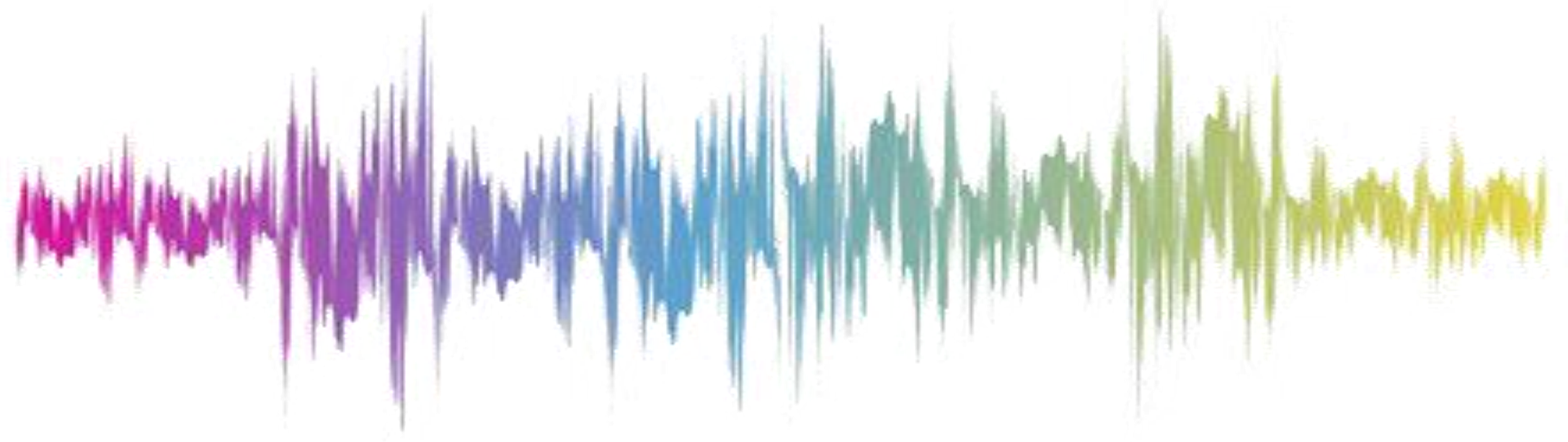


$$\mathrm{AdaIN}(\boldsymbol{H}, \mu(\boldsymbol{Z}), \sigma(\boldsymbol{Z})) = \sigma(\underbrace{\boldsymbol{Z}}_{\text{Target}})\mathrm{IN}(\underbrace{\boldsymbol{H}}_{\text{Source}}) + \mu(\boldsymbol{Z})$$

Latent Space

The Baseline Model – AGAIN-VC

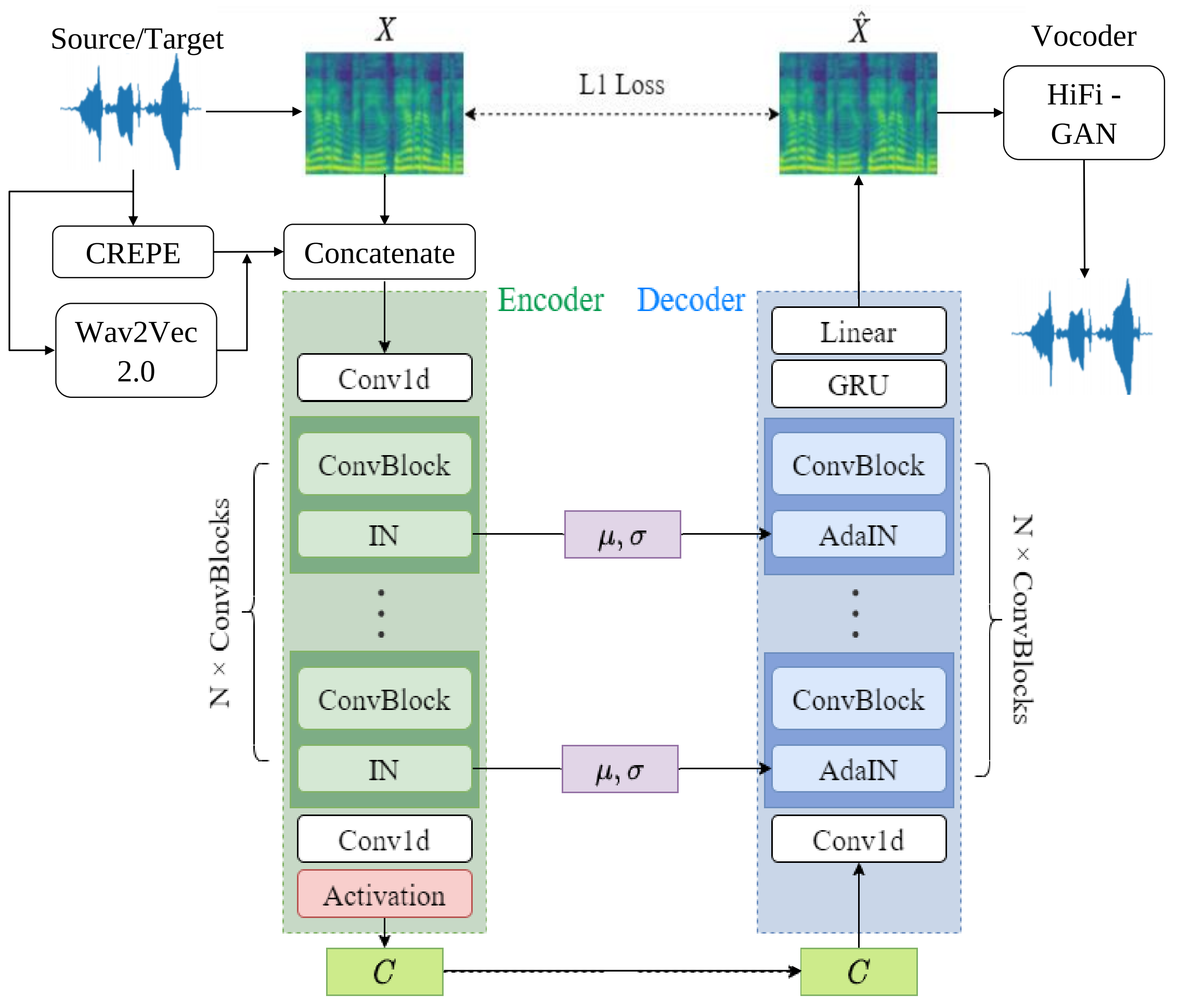
- One-shot VC model using activation guidance (AG) and adaptive instance normalization.
  - AG is an activation function that boosts the VC performance
  - Used as an information bottleneck to prevent the content embedding from leaking speaker information
- Auto-Encoder based with a Single encoder to extract the speaker and the content information
- Uses a pretrained MelGAN model as a vocoder
  - Synthesis of the converted speech back from mel-spectrogram to the waveform



Chen, Yen-Hao, et al. "Again-Vc: A One-Shot Voice Conversion Using Activation Guidance and Adaptive Instance Normalization." ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2021.

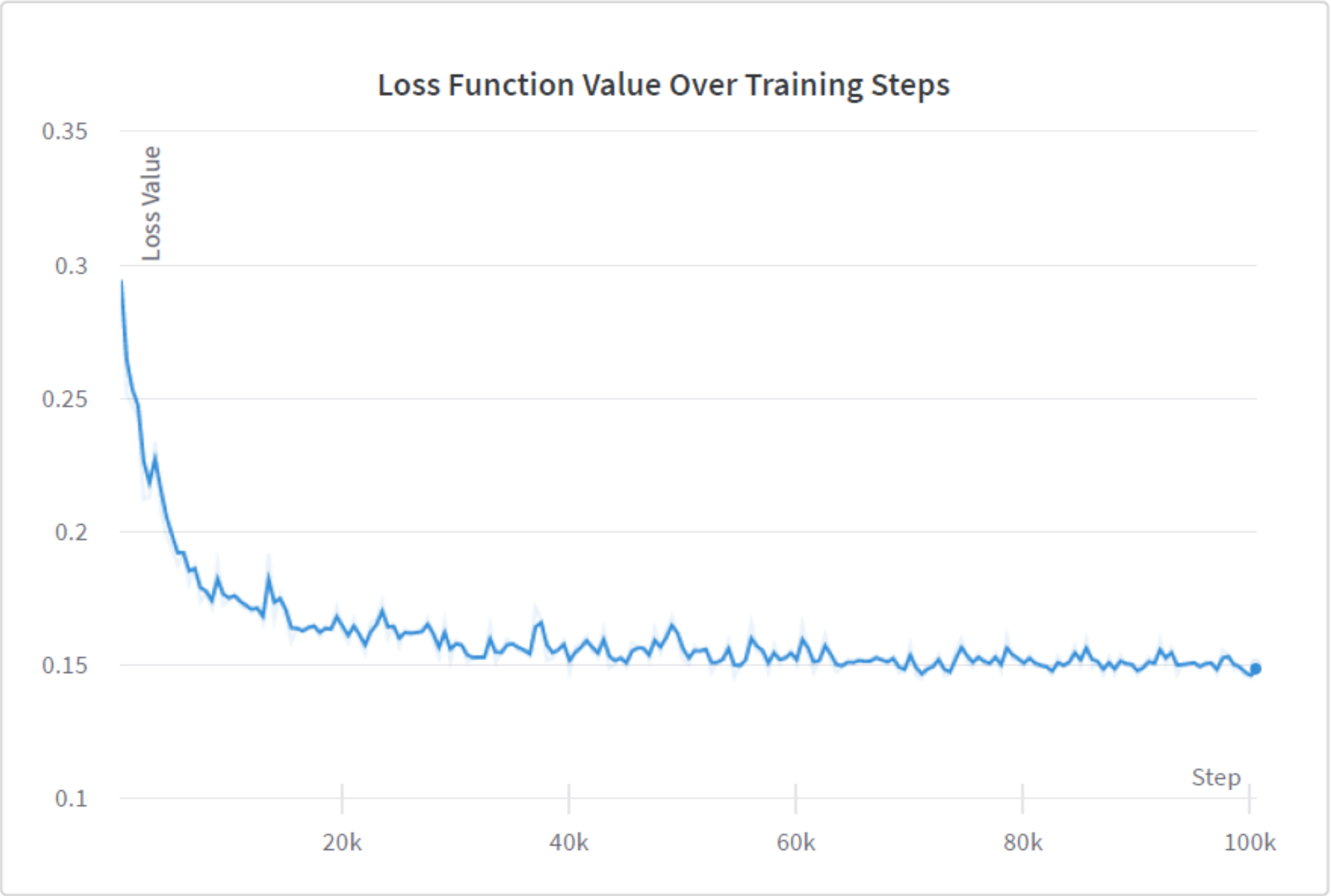
Our Proposed Model

- AGAIN-VC as a baseline model
- The Mel-spectrogram may lack important speaker and content information.
- Adding extracted information from the speech waveform for generating a richer information input:
  - Pitch estimation using CREPE pretrained model, a novel deep learning method for mono-phonic pitch tracking
  - Waveform feature extraction using Wav2Vec 2.0 pretrained model
- Switching the vocoder to HiFi-GAN, a state-of-the-art pretrained model
  - Speech audio can be synthesized efficiently
  - Outperforms the MelGAN results in terms of speech synthesis quality

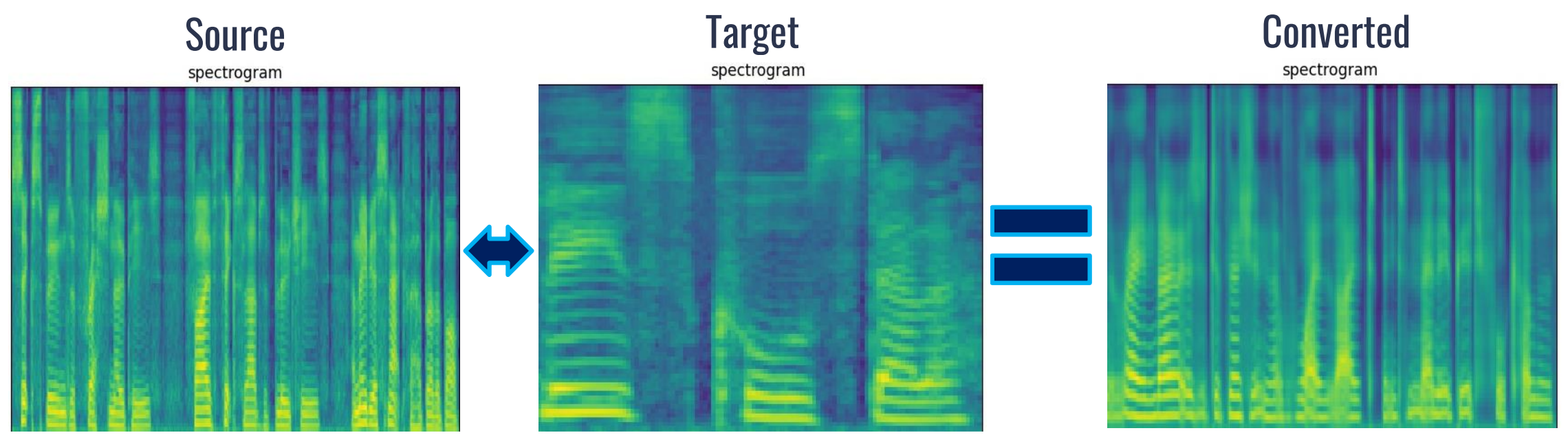


Results

- Model training results:



- AGAIN-VC inference Mel-spectrogram:



- AGAIN-VC + HiFi-GAN inference – listening to the results:

Conclusions

- We proposed a new VC model, based on AGAIN-VC, which includes additional extracted speech waveform information to improve VC performance.
- Pitch estimation and waveform feature extraction were added to the Mel-spectrogram to generate richer information
- The experimental results showed that One-shot VC performance haven't reached yet an applicable level, additional research should be conducted