Tsinghua University

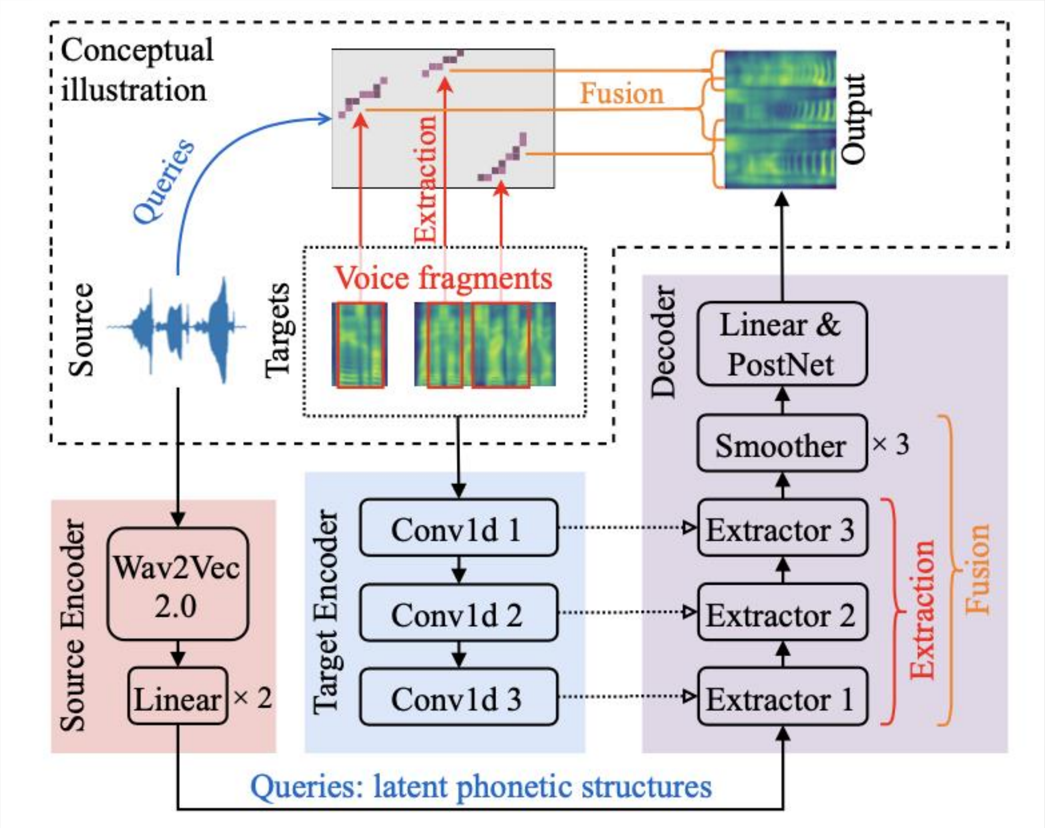
Weekly Reading

FRAGMENTVC: ANY-TO-ANY VOICE CONVERSION BY END-TO-END EXTRACTING AND FUSING FINE-GRAINED VOICE FRAGMENTS WITH ATTENTION

FragmentVC

- Any-to-any VC models aim to convert the voice from and to any speakers even unseen during training

- FragmentVC uses the latent phonetic structure of the utterance of the source speaker obtained with Wav2Vec 2.0 as the query to extract the fine-grained voice fragments in the utterances of the target speaker and fuse them into the desired utterance, all based on the attention mechanism of Transformer and achieved end-to-end.



Methodology

- Source Encoder

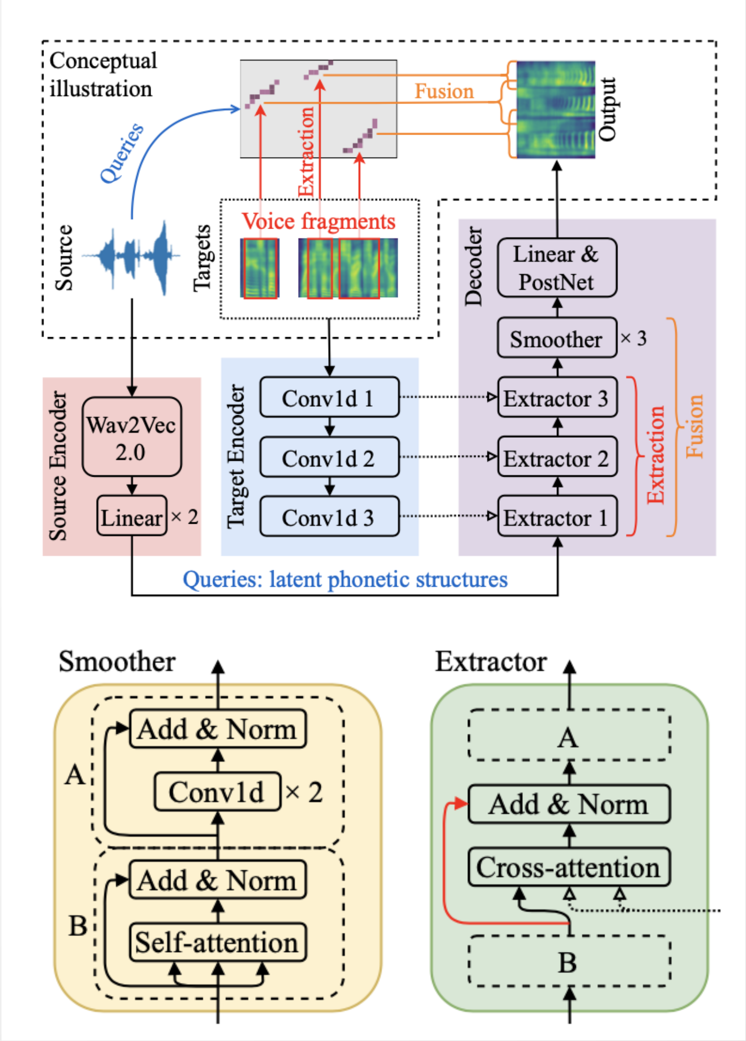
Wav2Vec is used as a pretrained feature extractor to extract 768-dimensional speech representations of the source utterance, then converted to 512-dimension by two linear layers with ReLU activation, to be used as the input to the decoder.

- Target Encoder

The log mel-spectrograms of utterances from the target speaker are concatenated and fed into the target encoder, which is composed of three ReLU-activated 1d-convolution layers, for extracting the voice fragments to be used below.

- Decoder

The decoder is composed of a stack of extractors and smoothers, followed by a linear projection and a PostNet to predict the spectrogram for the desired output voice. Both the extractors and smoothers are Transformer layers with two attention heads.



Methodology(Decoder)

- Extractor

The extractors are equipped with both self-attention and cross-attention that attend on the output of the target encoder, while the smoother contains self-attention only. The feed-forward layers in each Transformer layer are convolutional networks.

The extractors are based on the latent phonetic structure of the source speaker utterance by cross-attention to extract fine-grained voice fragments from target speaker utterances. The cross-attention is purposely designed to have a U-Net-like architecture.

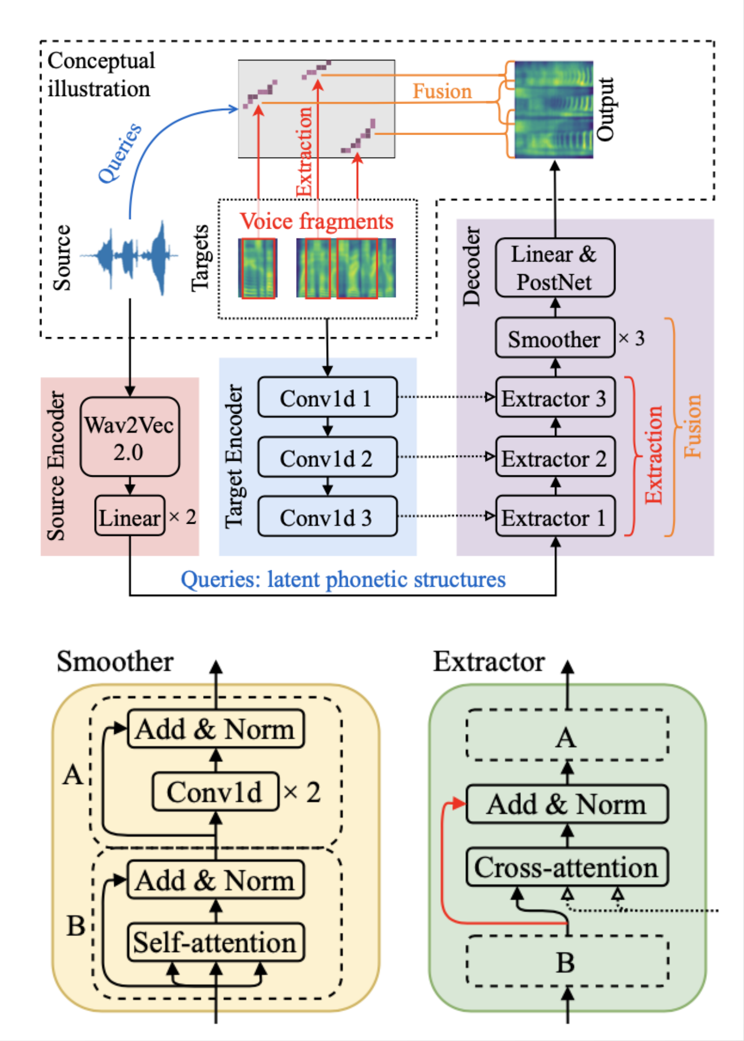
Since some residual speaker information is inevitably carried by the Wav2Vec features, we remove the residual connection over the cross-attention module in Extractor 1.

- Smoother

The smoothers finally take the output of the extractor stack to further smooth the output utterances.

- Loss Function

Only L1 loss between the predicted and the ground-truth mel-spectrograms is used to train the entire network.



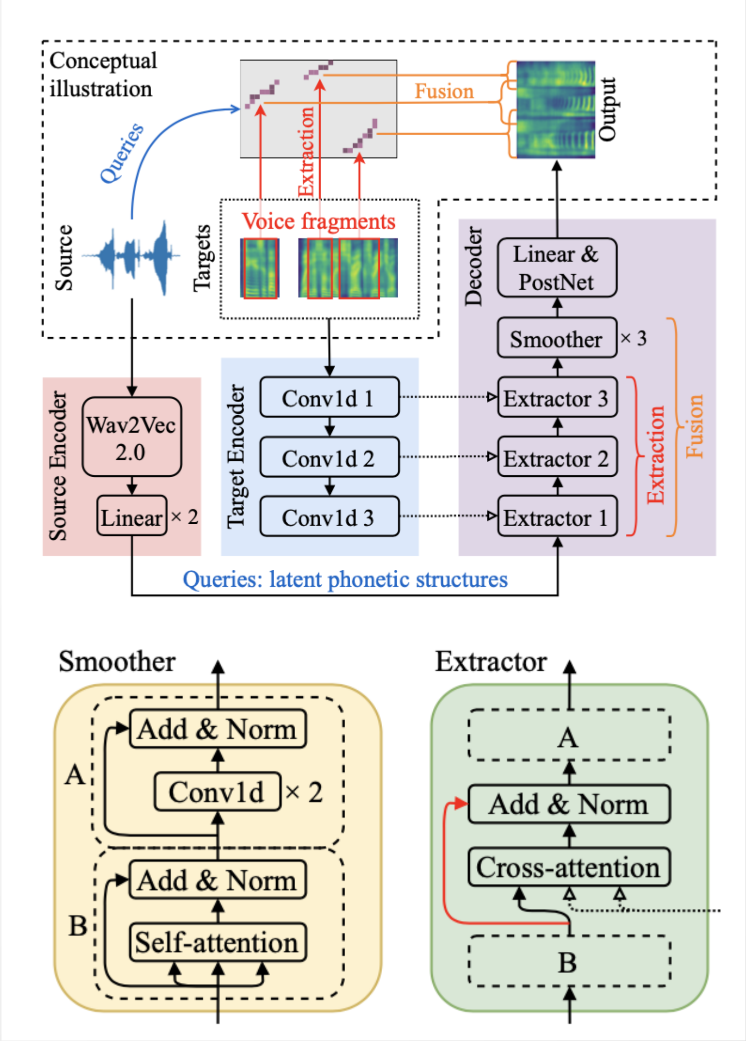
Methodology

- Two stage training

1st: The same single utterance from a training speaker is used as the input of both the source encoder and the target encoder, and the training goal is to reconstruct the log spectrogram of the utterance.

2nd: In the second stage, we concatenate the spectrograms of 10 utterances and feed them into the target encoder, while feeding a single utterance to the source encoder, all produced by the same speaker, with the training goal being reconstructing the spectrogram of the source utterance.

In the beginning, the source utterance is always included in the 10 target utterances, but the probability that it is included then linearly decays to zero as the training proceeds, so as to learn incrementally the scenario that the source and targets are getting more and more different.



Experiment

- Speaker Verification(SV)

The SV system took a converted utterance as input and generated a fix-dimensional embedding.
The conversion was considered successful if the cosine similarity between the embeddings of the target utterance and the converted utterance exceeded a predefined threshold.

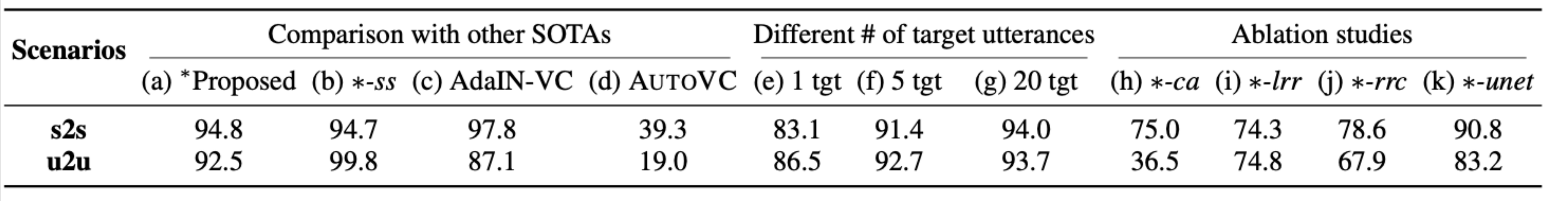
- Subjective evaluation Two Mean Opinion Score (MOS) tests.

In the first test, each subject was asked to listen to an authentic utterance from the target speaker and a converted result, and then to score from 1 to 5 regarding how confident they would consider these two utterances to be produced by the same speaker (5 being absolutely same and 1 absolutely different).

In the second test, the subjects were given a converted utterance or a vocoder-reconstructed authentic utterance and asked to score from 1 to 5 how natural the utterance sounded.

Results

The SV accuracy (%) for seen-to-seen (VCTK, with EER being 5.6%) and unseen-to-unseen (CMU, with EER being 2.6%) scen



Scenarios	Comparison with other SOTAs				Different # of target utterances			Ablation studies			
	(a) *Proposed	(b) ∗-*ss*	(c) AdaIN-VC	(d) AUTOVC	(e) 1 tgt	(f) 5 tgt	(g) 20 tgt	(h) ∗-*ca*	(i) ∗-*lrr*	(j) ∗-*rrc*	(k) ∗-*unet*
s2s	94.8	94.7	97.8	39.3	83.1	91.4	94.0	75.0	74.3	78.6	90.8
u2u	92.5	99.8	87.1	19.0	86.5	92.7	93.7	36.5	74.8	67.9	83.2

MOS tests

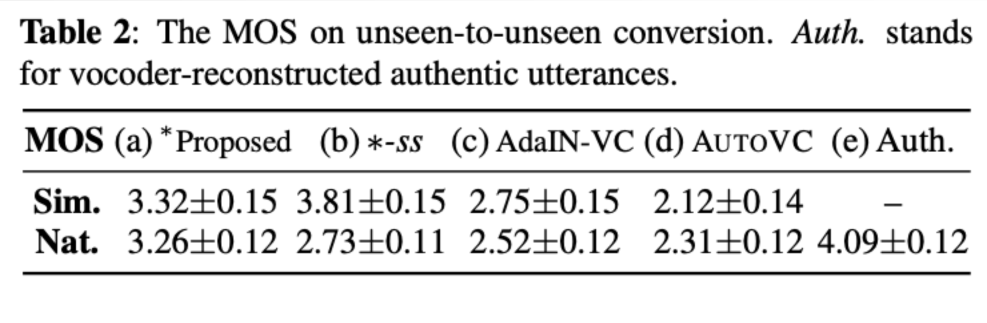


Table 2: The MOS on unseen-to-unseen conversion. *Auth.* stands for vocoder-reconstructed authentic utterances.

MOS	(a) *Proposed	(b) ∗-*ss*	(c) AdaIN-VC	(d) AUTOVC	(e) Auth.
Sim.	3.32±0.15	3.81±0.15	2.75±0.15	2.12±0.14	–
Nat.	3.26±0.12	2.73±0.11	2.52±0.12	2.31±0.12	4.09±0.12

Conclusion

The objective and subjective evaluations verified the proposed FragmentVC achieved comparable or even better performance than other SOTA approaches. How the Wav2Vec representations actually contributed to the model, if it can be jointly learned, or if it is possible to find some other pretrained representations for the purpose are yet to be investigated.

谢谢聆听

请大家批评指正！