HiFi-VC: High Quality ASR-Based Voice Conversion

Anton Kashkin, Ivan Karpukhin, Svyatoslav Shishkin

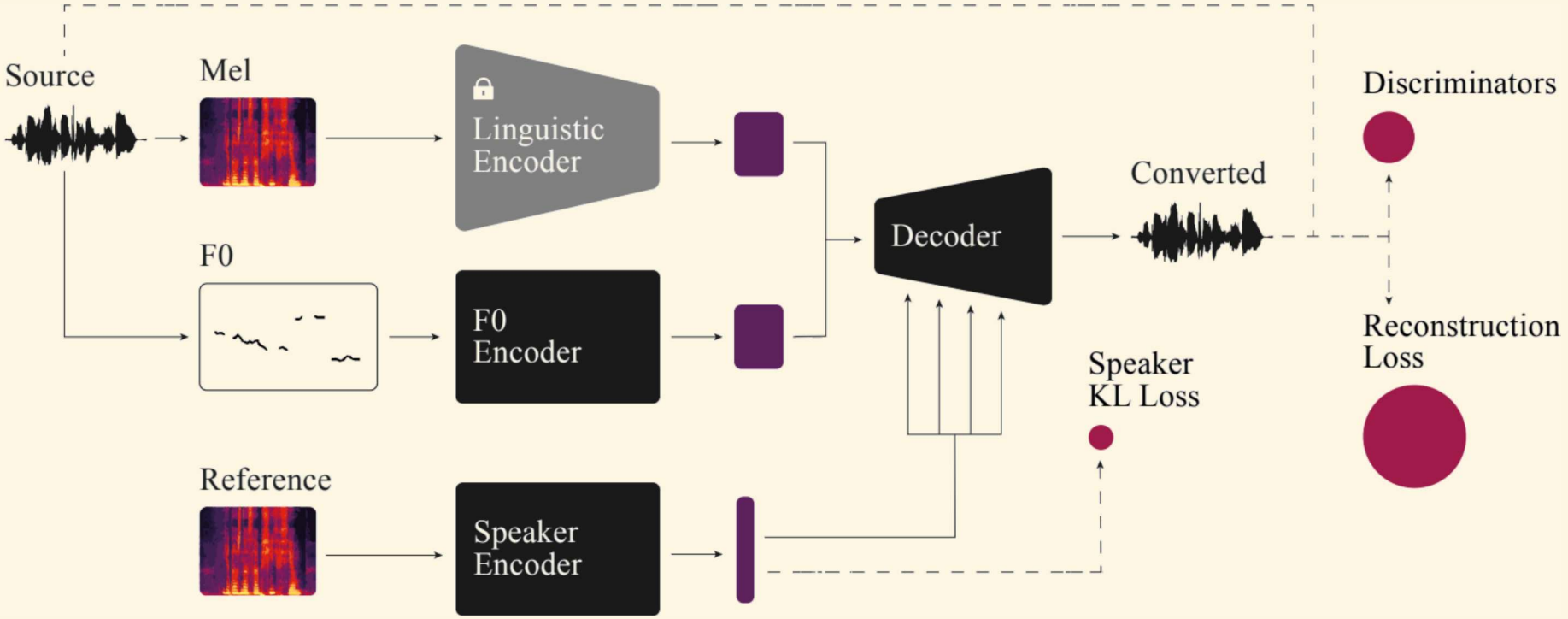
Introduction

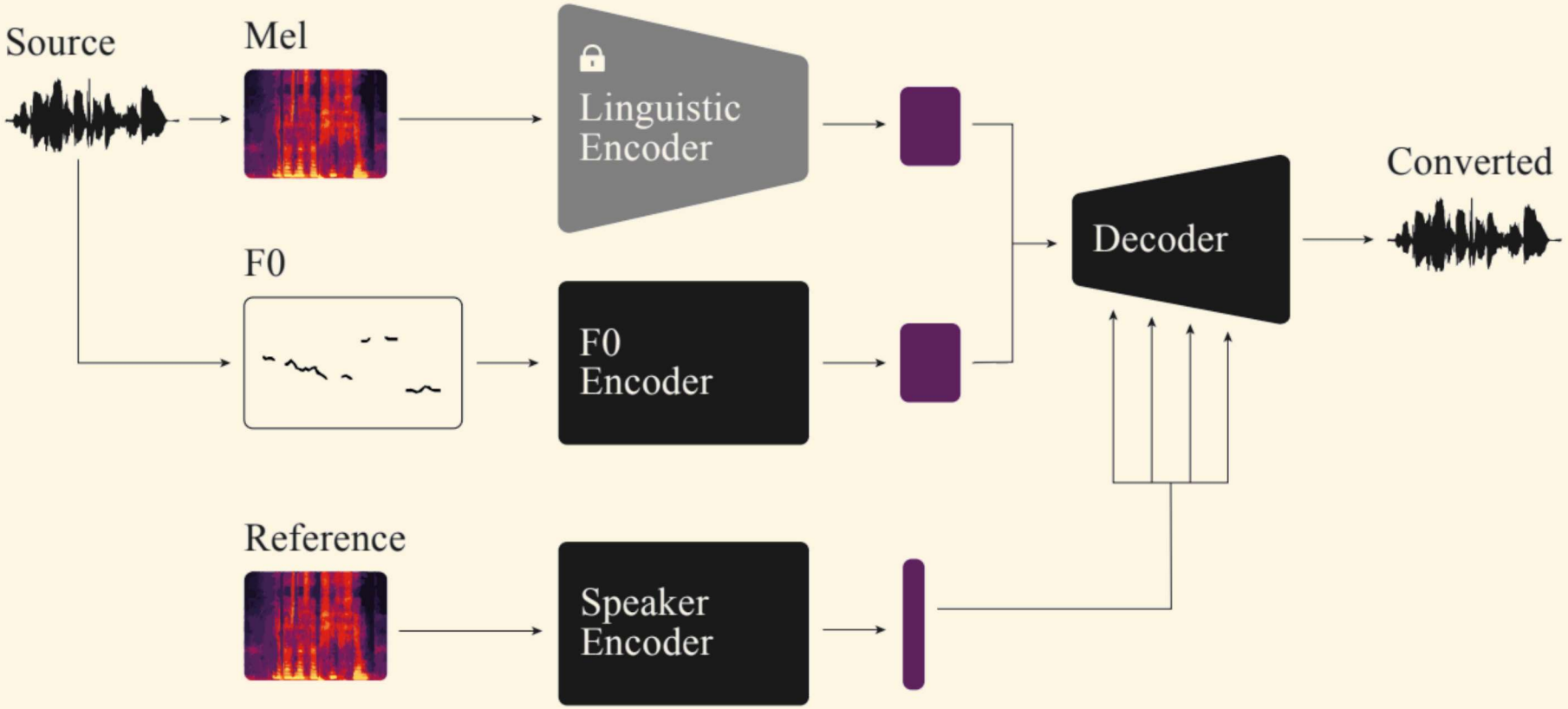
- The goal of voice conversion (VC) is to convert **input** voice to match the **target speaker**'s voice.
- The development of **any-to-any VC** systems is of particular interest.
- any-to-any conversion quality is **still inferior** to natural speech.

Contributions

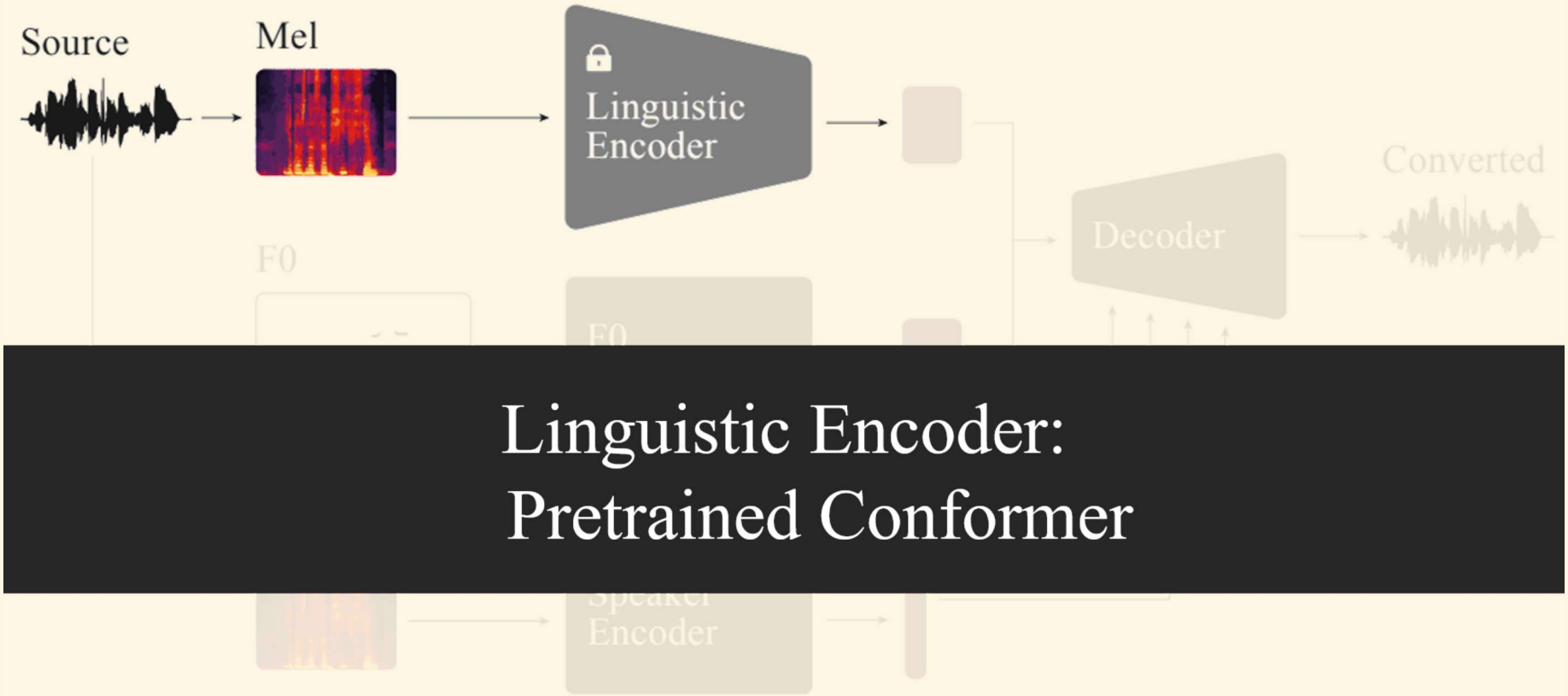
- Propose a **high-quality any-to-any VC architecture** that converts **ASR feature**, **pitch** and **speaker style** into waveforms.
- It is **better than other baselines** in terms of voice **quality**, **similarity** and content **consistency**.

HiFi-VC

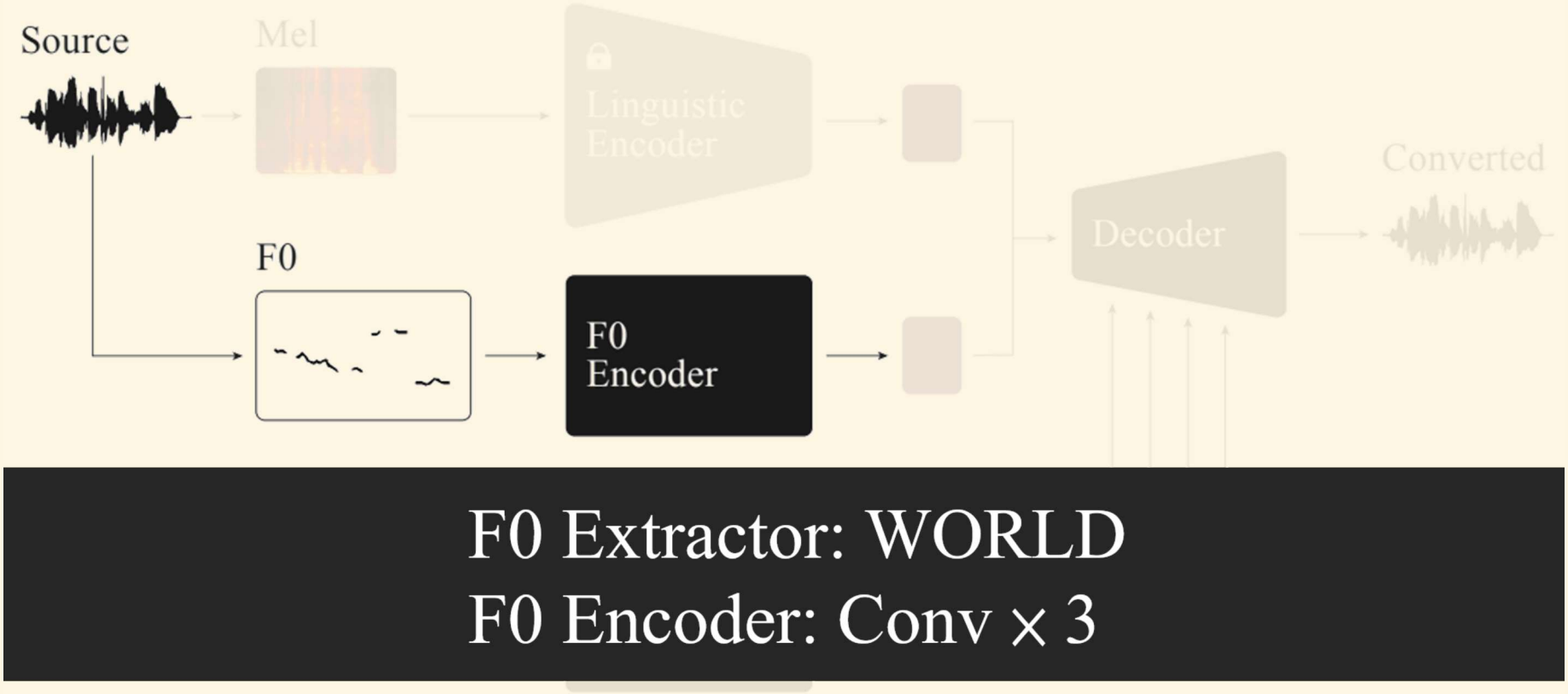

Source
Mel
Linguistic Encoder
F0
F0 Encoder
Reference
Speaker Encoder
Decoder
Converted
Discriminators
Reconstruction Loss
Speaker KL Loss



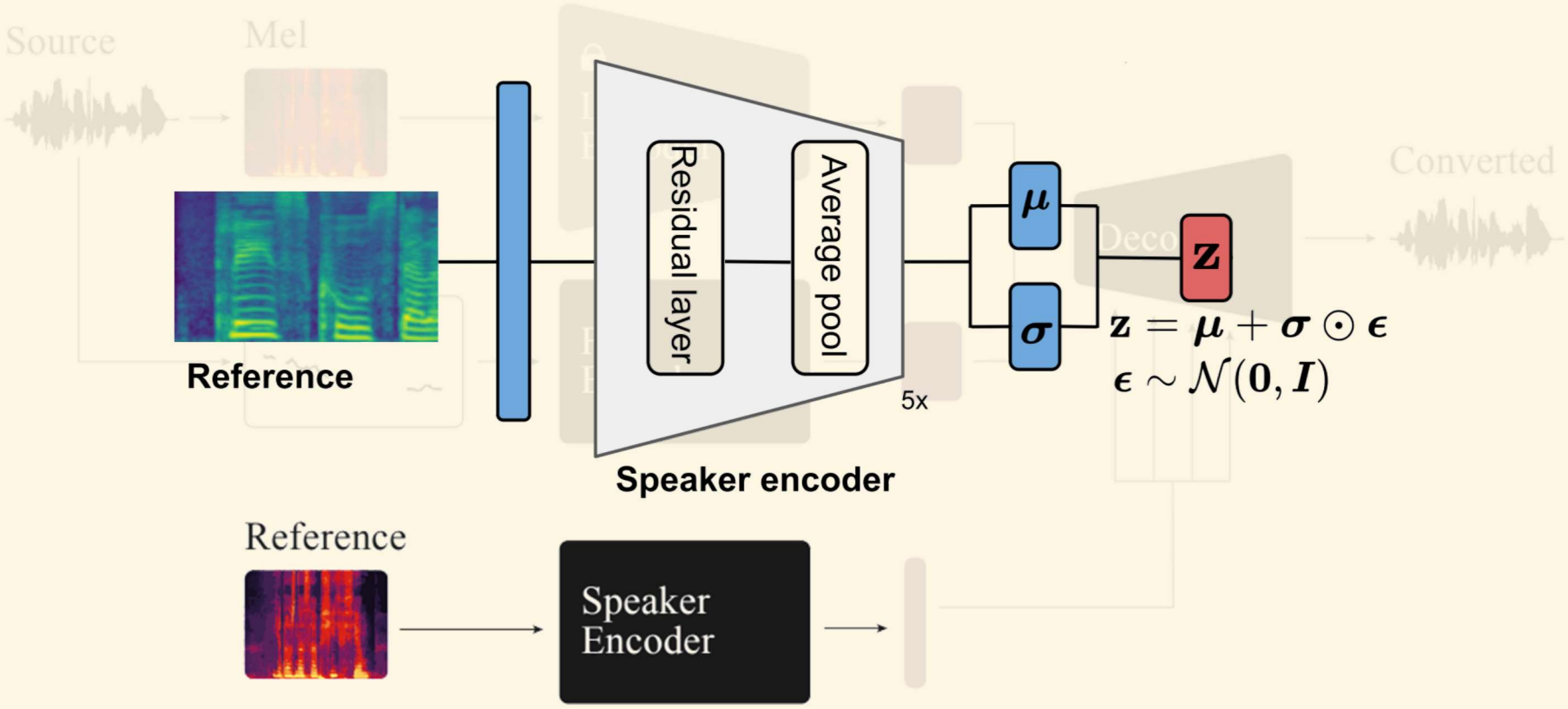
$$Voice = Content + Pitch + Speaker$$

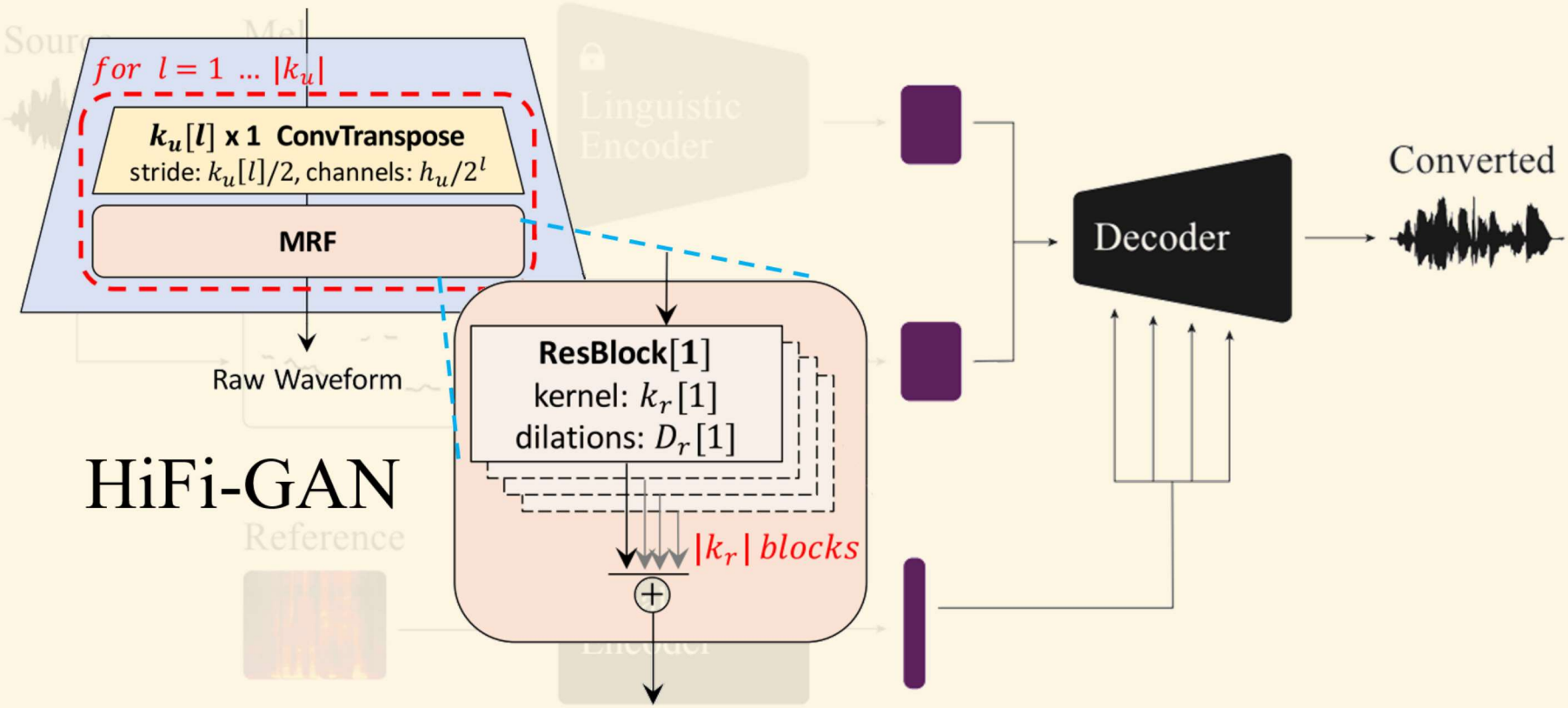


Linguistic Encoder: Pretrained Conformer



F0 Extractor: WORLD
F0 Encoder: Conv × 3


Source
Mel
Converted
Reference
Residual layer
Average pool
5x
Speaker encoder
μ
σ
$\mathbf{z}$
$\mathbf{z} = \boldsymbol{\mu} + \boldsymbol{\sigma} \odot \boldsymbol{\epsilon}$
$\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$
Reference
Speaker Encoder


for $l = 1 \ldots |k_u|$
$k_u[l]$ x 1 ConvTranspose
stride: $k_u[l]/2$, channels: $h_u/2^l$
MRF
Raw Waveform
ResBlock[1]
kernel: $k_r[1]$
dilations: $D_r[1]$
$|k_r|$ blocks
HiFi-GAN
Linguistic Encoder
Reference
Decoder
Converted

Discriminator Loss

$$\mathcal{L}_{advD}(s, \tilde{s}) = (D(s) - 1)^2 + D(\tilde{s})^2$$

- $\tilde{s}$ is the predicted waveform, s is the natural one.
- D is a discriminator network.

Predictor Loss

$$\mathcal{L}_{pred} = \lambda_{rec} \cdot \mathcal{L}_{rec} + \lambda_{advG} \cdot \mathcal{L}_{advG} + \lambda_{FM} \cdot \mathcal{L}_{FM} + \lambda_{spk} \cdot \mathcal{L}_{spk}$$

Predictor Loss (cont.)

Reconstruction Loss

$$\mathcal{L}_{rec}(s, \tilde{s}) = ||M_s - M_{\tilde{s}}||_1$$

- M_* is the Mel-spectorgam of s or $\tilde{s}$.

Predictor Loss (cont.)

Adversarial & Feature Matching Loss

$$\mathcal{L}_{advG}(\tilde{s}) = (D(\tilde{s}) - 1)^2$$

$$\mathcal{L}_{FM}(s, \tilde{s}) = \sum_{i=1}^{L} \frac{1}{N_i} ||D_i(s) - D_i(\tilde{s})||_1$$

- L is the number of layers.
- D_i is the feature of the i-th discriminator layer.
- N_i is the dimension of D_i.

Predictor Loss (cont.)

Regularization Loss

$$\mathcal{L}_{spk}(s) = D_{KL}(\mathcal{N}(\mu_s, \sigma_s)||\mathcal{N}(\mathbf{0}, \mathbf{I}))$$

Predictor Loss (cont.)

$$
\begin{aligned}
\mathcal{L}_{pred} =& \lambda_{rec} \cdot \mathcal{L}_{rec} + \lambda_{advG} \cdot \mathcal{L}_{advG} + \\
& \lambda_{FM} \cdot \mathcal{L}_{FM} + \lambda_{spk} \cdot \mathcal{L}_{spk} \\
=& 45 \cdot \mathcal{L}_{rec} + 1 \cdot \mathcal{L}_{advG} + \\
& 1 \cdot \mathcal{L}_{FM} + 0.01 \cdot \mathcal{L}_{spk}
\end{aligned}
$$

Experiments

Model	Many-to-Many Voice Quality ↑				
	F2F	F2M	M2M	M2F	Mean
Ground Truth	4.30	N/A	4.35	N/A	4.33
AutoVC[7]	2.22	2.14	2.27	2.15	2.20
VQMIVC[8]	3.93	3.69	3.74	3.78	3.78
NVC-Net[15]	3.73	3.17	3.71	3.35	3.49
HiFi-VC (Proposed)	**4.10**	**4.09**	**4.11**	**4.01**	**4.08**

Model	Any-to-Any Voice Quality ↑				
	F2F	F2M	M2M	M2F	Mean
Ground Truth	4.27	N/A	4.47	N/A	4.37
AutoVC[7]	2.08	1.61	1.64	2.03	1.84
VQMIVC[8]	3.64	3.73	3.67	3.70	3.69
NVC-Net[15]	3.68	3.41	3.64	3.42	3.54
HiFi-VC (Proposed)	**4.00**	**3.98**	**4.06**	**4.09**	**4.03**

Model	Many-to-Many Similarity ↑				
	F2F	F2M	M2M	M2F	Mean
Ground Truth	4.37	N/A	4.44	N/A	4.40
AutoVC[7]	1.59	1.66	1.73	1.47	1.61
VQMIVC[8]	2.97	3.10	3.19	2.97	3.06
NVC-Net[15]	3.91	3.79	3.83	3.71	3.81
HiFi-VC (Proposed)	**4.03**	**4.17**	**4.11**	**3.99**	**4.08**

Model	Any-to-Any Similarity ↑				
	F2F	F2M	M2M	M2F	Mean
Ground Truth	4.39	N/A	4.17	N/A	4.28
AutoVC[7]	1.59	1.66	1.73	1.47	1.61
VQMIVC[8]	1.96	2.23	2.22	1.95	2.09
NVC-Net[15]	2.22	1.82	1.82	2.06	1.98
HiFi-VC (Proposed)	**3.50**	**2.54**	**2.70**	**3.34**	**3.02**

Model	Many-to-Many		
	WER (%) ↓	CER (%) ↓	PCC ↑
AutoVC[7]	85.1	58.1	0.22
VQMIVC[8]	32.5	16.9	0.51
NVC-Net[15]	37.9	21.4	0.42
HiFi-VC (Proposed)	**14.6**	**6.4**	**0.60**

Model	Any-to-Any		
	WER (%) ↓	CER (%) ↓	PCC ↑
AutoVC[7]	95.4	67.6	0.12
VQMIVC[8]	32.2	16.7	0.55
NVC-Net[15]	32.5	16.5	0.12
HiFi-VC (Proposed)	**10.3**	**4.2**	**0.66**

Conclusion

- ASR feature + Speaker Embedder is great in any-to-any voice conversion task.
- Separation of vocoder and decoder is not necessary.

Thanks for your attention.

Q&A