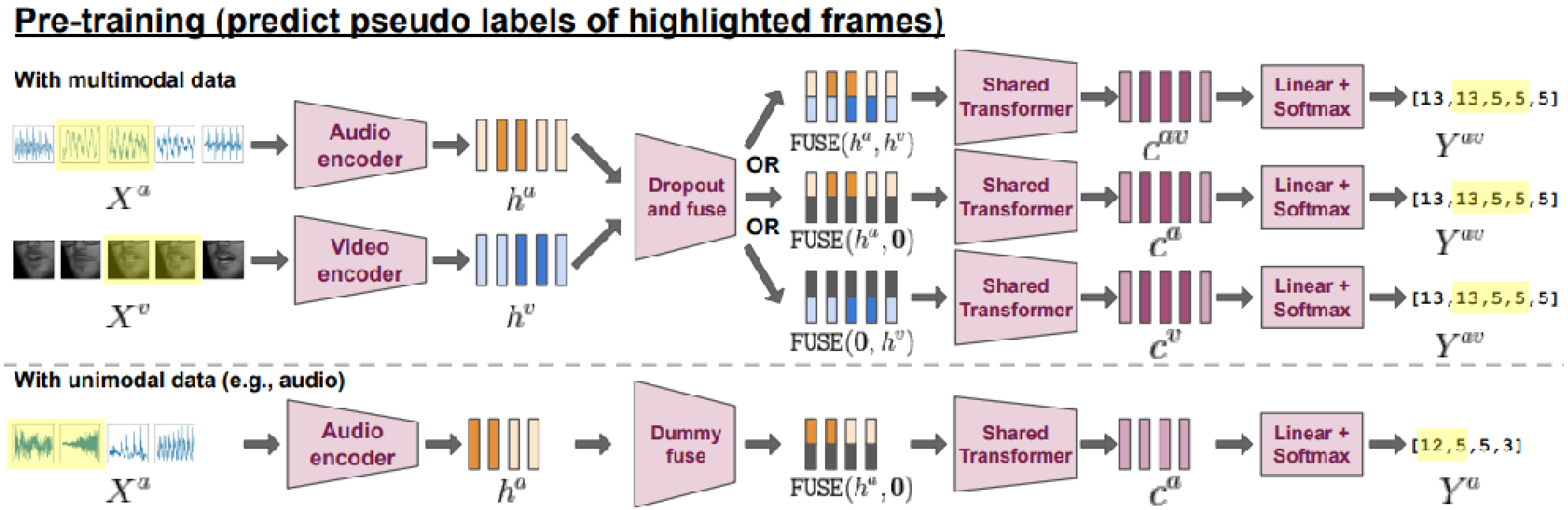
NIPS 2022

u-HuBERT

- Mask multimodal data in self-supervised learning, to learn the correlation among modalities



Hsu W N, Shi B. u-HuBERT: Unified Mixed-Modal Speech Pretraining And Zero-Shot Transfer to Unlabeled Modality[C]//Advances in Neural Information Processing Systems.

u-HuBERT

- Use one modality to fine-tune can obtain good performance with other modalities.

PT	PT mod-drop	FT mod	FT mod-drop	AV-WER Clean	AV-WER Noisy	A-WER Clean	A-WER Noisy	V-WER	Avg-WER
fine-tuned on 433h									
✗	n/a	AV	✗	3.8	17.2	28.2	87.6	83.4	44.0
✓	✗	AV	✗	1.3	4.8	21.4	52.6	42.3	24.5
✓	✓	AV	✗	1.2	5.2	1.7	25.5	32.4	13.2
✗	n/a	AV	✓	3.6	15.9	4.6	44.8	63.7	26.5
✓	✗	AV	✓	1.3	4.1	1.8	23.1	31.0	12.3
✓	✓	AV	✓	1.3	4.6	1.5	20.5	29.1	11.4
✗	n/a	A	n/a	✗	✗	4.0	37.3	✗	✗
✓	✗	A	n/a	1.5	18.0	1.6	20.9	96.8	27.8
✓	✓	A	n/a	1.3	4.6	1.4	19.3	31.6	11.6
✗	n/a	V	n/a	✗	✗	✗	✗	60.3	✗
✓	✗	V	n/a	11.3	21.8	80.3	97.7	28.0	47.8
✓	✓	V	n/a	2.1	5.1	2.3	20.9	28.7	11.8

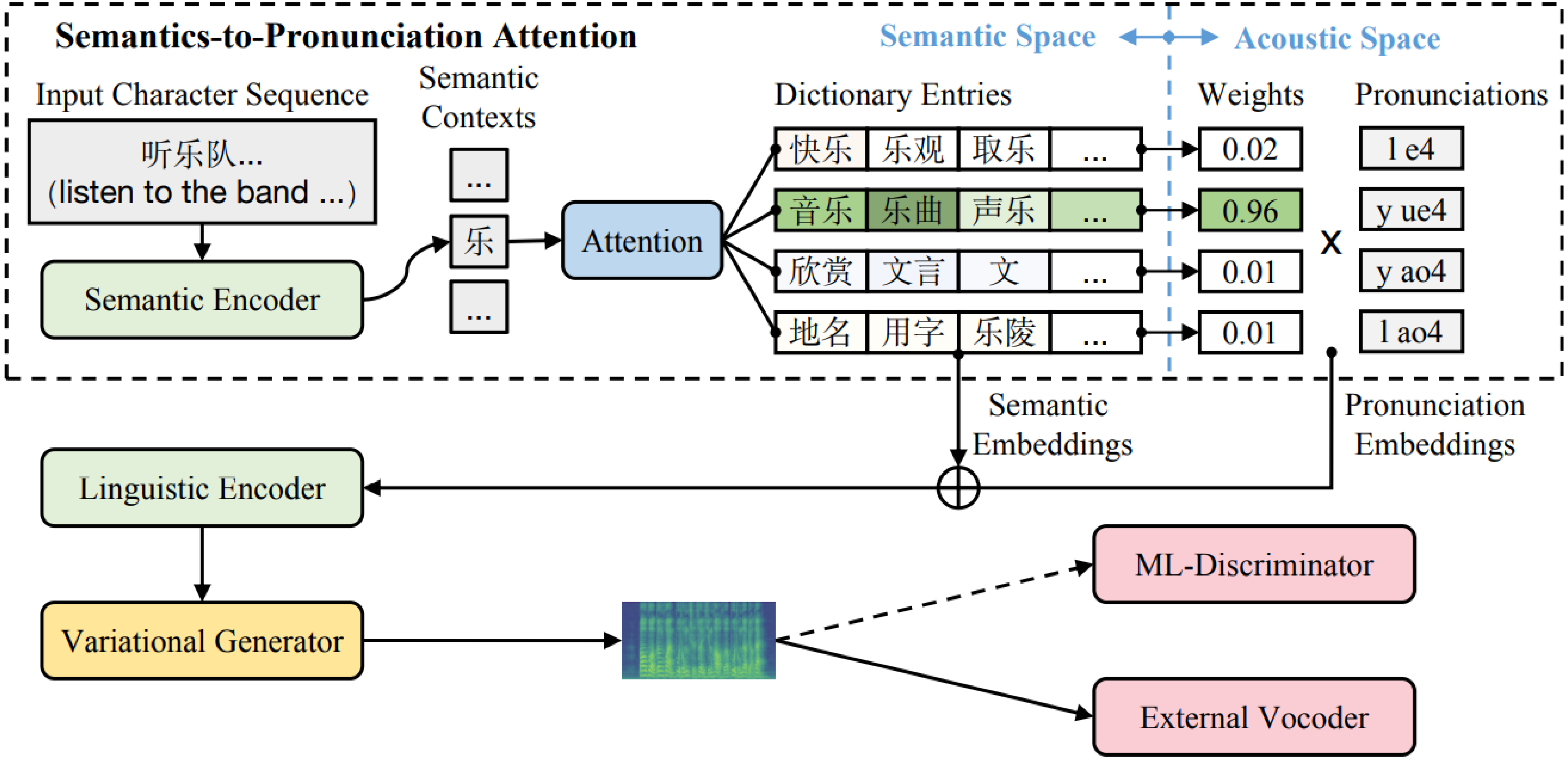
Dict-TTS

- Distinguish pronunciations of Chinese char, using extra knowledge

Character	Pronunciation	Definitions	Usages
乐:			
	<l è>	欢喜，快活；使人快乐的事情...	快乐。乐融融。其乐无穷。乐观。乐天。取乐。逗乐。快乐...
	<y uè>	声音，成调的声音。或姓氏...	音乐。声乐。乐池。乐音。乐曲（①音乐与歌曲；②伴奏...
	<y ào>	喜好、欣赏。用于文言文...	知者乐水，仁者乐山。
	<l ào>	地名用字。	河北省乐亭、山东省乐陵。

Jiang Z, Zhe S, Zhao Z, et al. Dict-TTS: Learning to Pronounce with Prior Dictionary Knowledge for Text-to-Speech[J]. arXiv preprint arXiv:2206.02147, 2022.

Dict-TTS


Semantics-to-Pronunciation Attention
Semantic Space
Acoustic Space
Input Character Sequence
听乐队...
(listen to the band ...)
Semantic Contexts
...
乐
...
Semantic Encoder
Attention
Dictionary Entries
快乐 乐观 取乐 ...
音乐 乐曲 声乐 ...
欣赏 文言 文 ...
地名 用字 乐陵 ...
Weights
0.02
0.96
0.01
0.01
X
Pronunciations
l e4
y ue4
y ao4
l ao4
Semantic Embeddings
Pronunciation Embeddings
Linguistic Encoder
Variational Generator
ML-Discriminator
External Vocoder

Dict-TTS

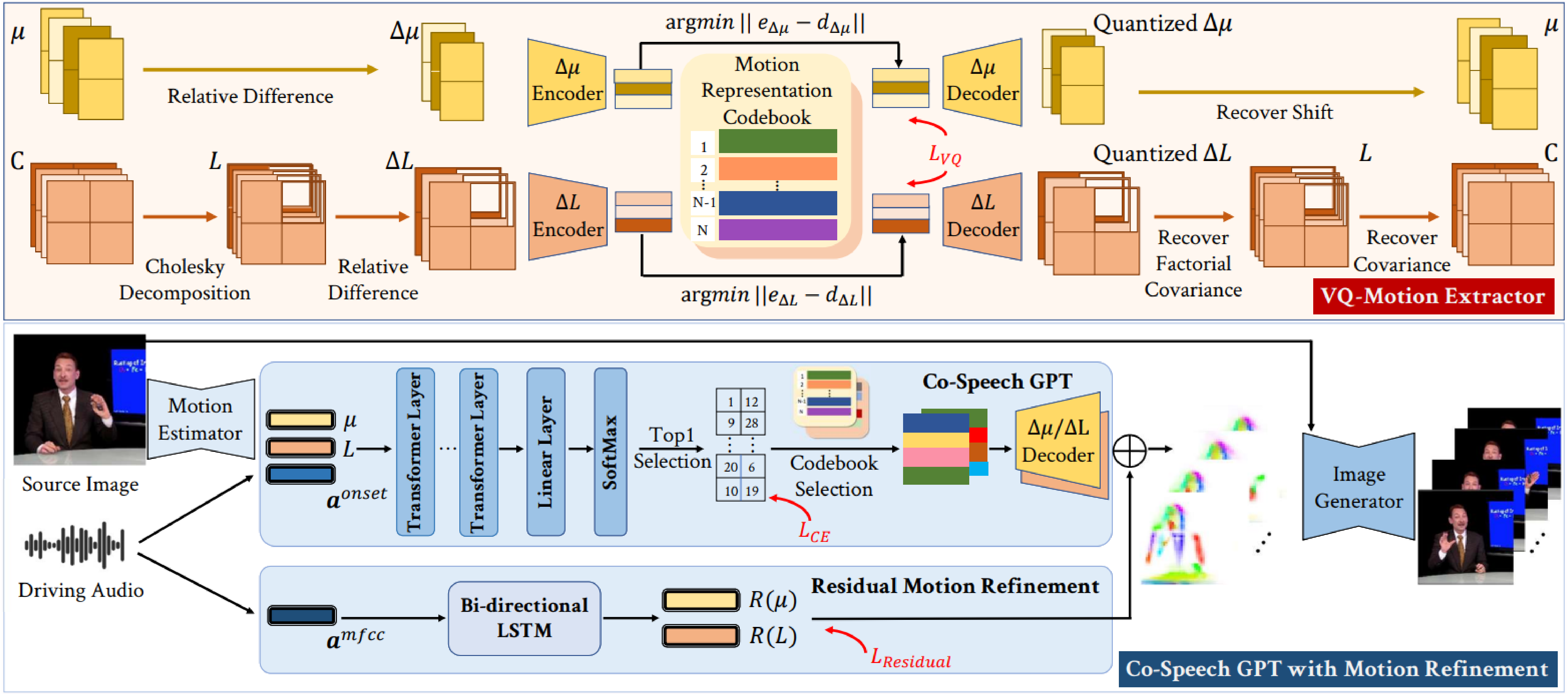
Table 1: The objective and subjective pronunciation accuracy comparisons. PER-O denotes phoneme error rate in the objective evaluation, PER-S denotes phoneme error rate in the subjective evaluation and SER-S denotes sentence error rate in the subjective evaluation.

Method	Biaobei			JSUT			Common Voice (HK)		
	PER-O	PER-S	SER-S	PER-O	PER-S	SER-S	PER-O	PER-S	SER-S
Character	-	3.73%	30.50%	-	13.78%	65.50%	-	1.89%	15.50%
Phoneme	2.78%	1.14%	7.00%	**1.55%**	**0.92%**	**4.25%**	-	1.45%	10.25%
Dict-TTS	**2.12%**	**1.08%**	**6.50%**	3.73%	2.57%	22.75%	-	**1.23%**	**9.75%**

Gesture generation

Figure 1: **Illustration of Problem Setting.** In this paper, we focus on audio-driven co-speech gesture video generation. Given an image with speech audio, we generate aligned speaker *image sequence*.

Liu X, Wu Q, Zhou H, et al. Audio-Driven Co-Speech Gesture Video Generation[J]. arXiv preprint arXiv:2212.02350, 2022.



μ
Relative Difference
$\Delta\mu$
$\Delta\mu$ Encoder
argmin $|| e_{\Delta\mu} - d_{\Delta\mu} ||$
Motion Representation Codebook
1
2
N-1
N
$\Delta\mu$ Decoder
Quantized $\Delta\mu$
Recover Shift
μ
C
L
ΔL
Cholesky Decomposition
Relative Difference
ΔL Encoder
L_{VQ}
ΔL Decoder
argmin $||e_{\Delta L} - d_{\Delta L}||$
Quantized ΔL
Recover Factorial Covariance
L
Recover Covariance
C
VQ-Motion Extractor
Source Image
Motion Estimator
μ
L
a^{onset}
Transformer Layer
Transformer Layer
Linear Layer
SoftMax
Top1 Selection
1 12
9 28
20 6
10 19
L_{CE}
Codebook Selection
Co-Speech GPT
$\Delta\mu/\Delta L$ Decoder
Image Generator
Driving Audio
a^{mfcc}
Bi-directional LSTM
$R(\mu)$
$R(L)$
Residual Motion Refinement
$L_{Residual}$
Co-Speech GPT with Motion Refinement

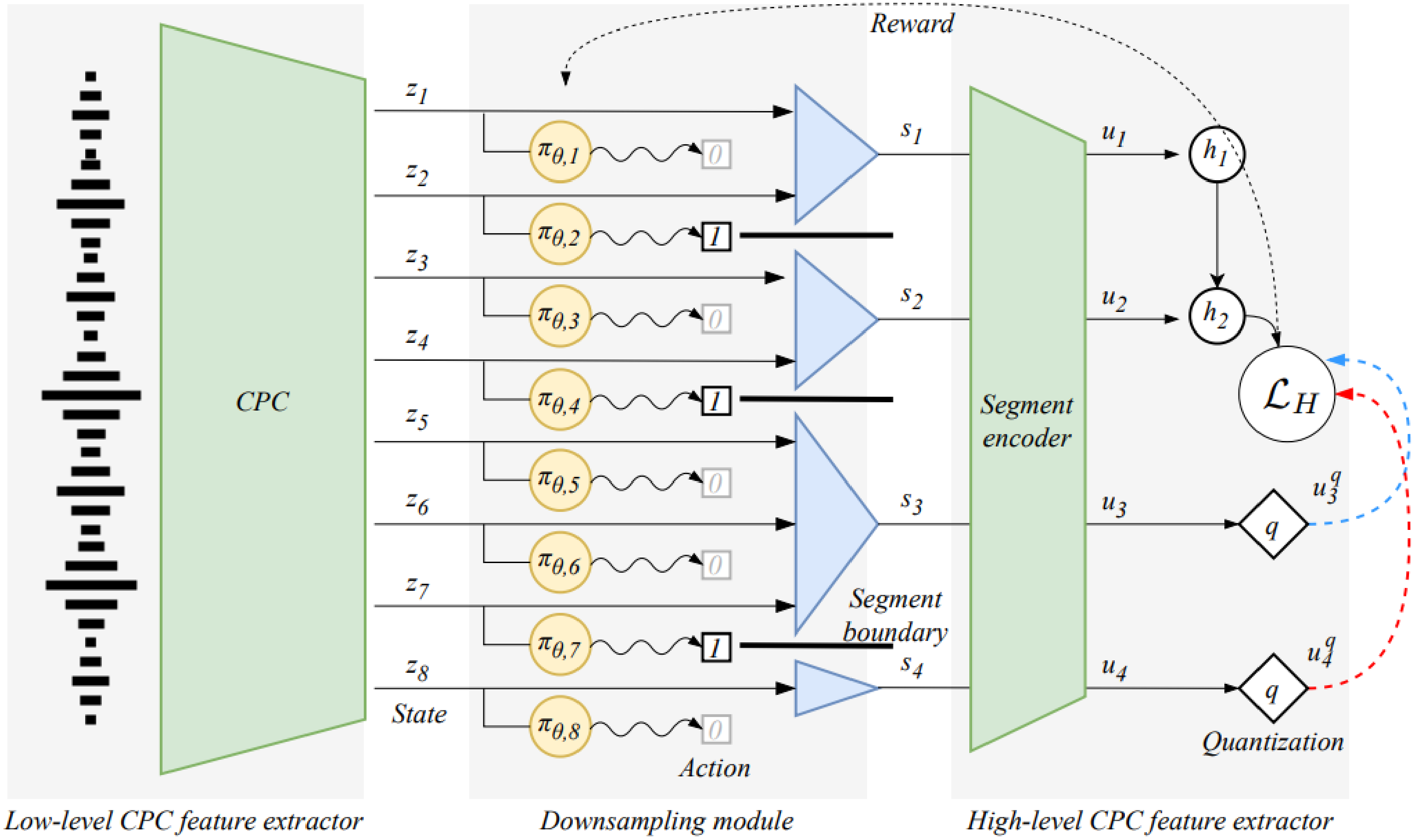
Reduce gap between streaming and non-streaming ASR

- In this paper, we argue that one main cause of such WER gap is that all the existing models are constrained to be locally normalized which makes them susceptible to label bias problem
- Using a non-normalized form w(c)/Z instead of p© for each frame.

context dep.	weight function		WER [%]	
	streaming	normalization	clean	other
1-gram	no	local	3.4	8.7
		global	3.3	8.4
	yes	local	7.0	17.4
		global	5.5	14.0
2-gram	no	local	2.8	6.7
		global	2.8	6.7
	yes	local	4.9	11.0
		global	3.8	9.5

Variani E, Wu K, Riley M, et al. Global Normalization for Streaming Speech Recognition in a Modular Framework[J]. arXiv preprint arXiv:2205.13674, 2022.

Two-level CPC with variable rate



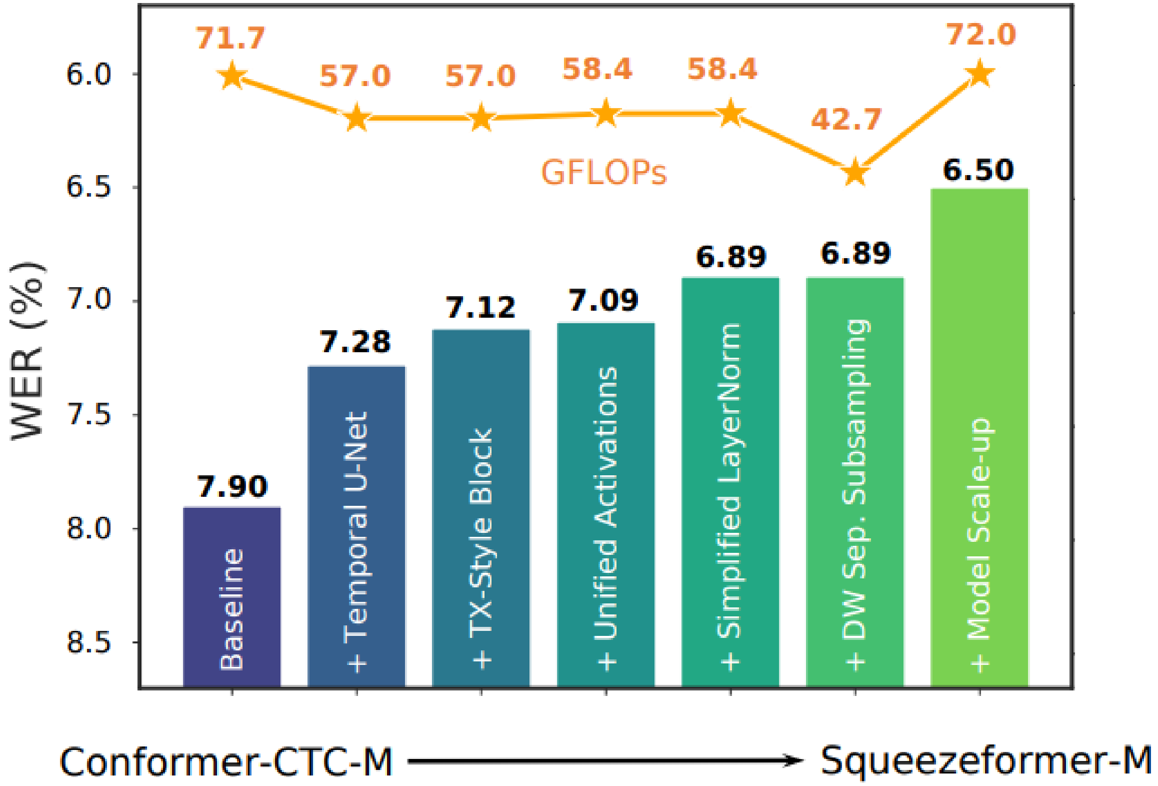
Cuervo S, Łańcucki A, Marxer R, et al. Variable-rate hierarchical CPC leads to acoustic unit discovery in speech[J]. arXiv preprint arXiv:2206.02211, 2022.

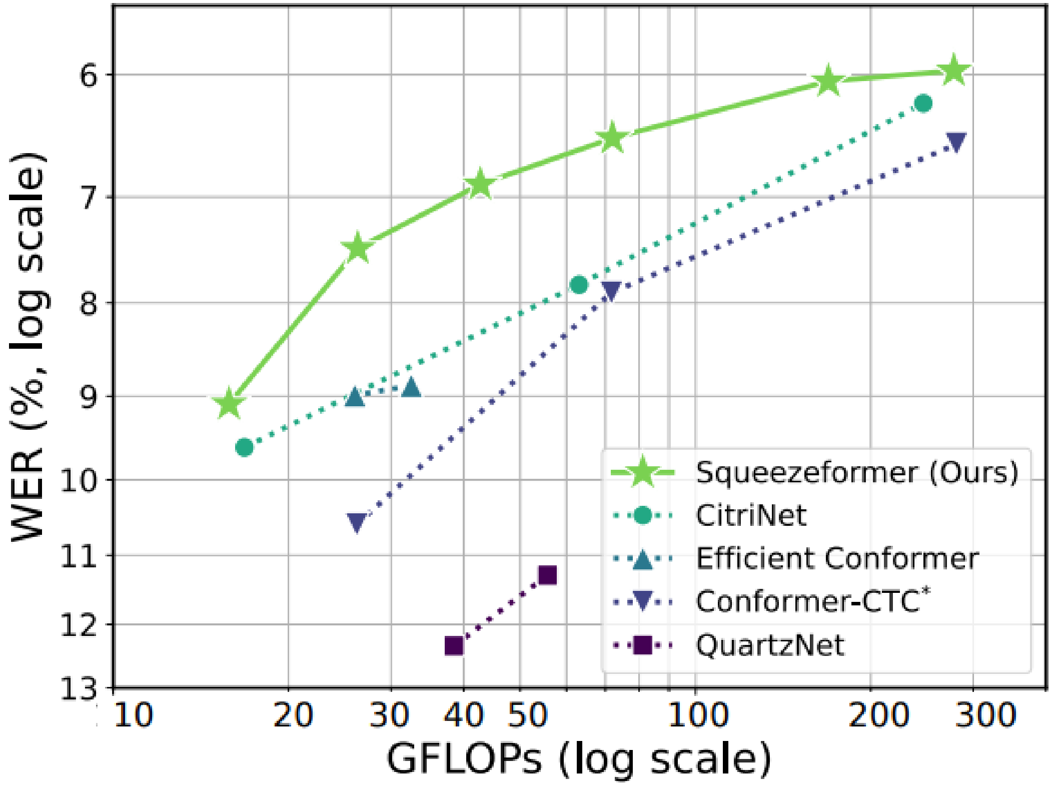
Two-level CPC with variable rate

- Using a boundary detector

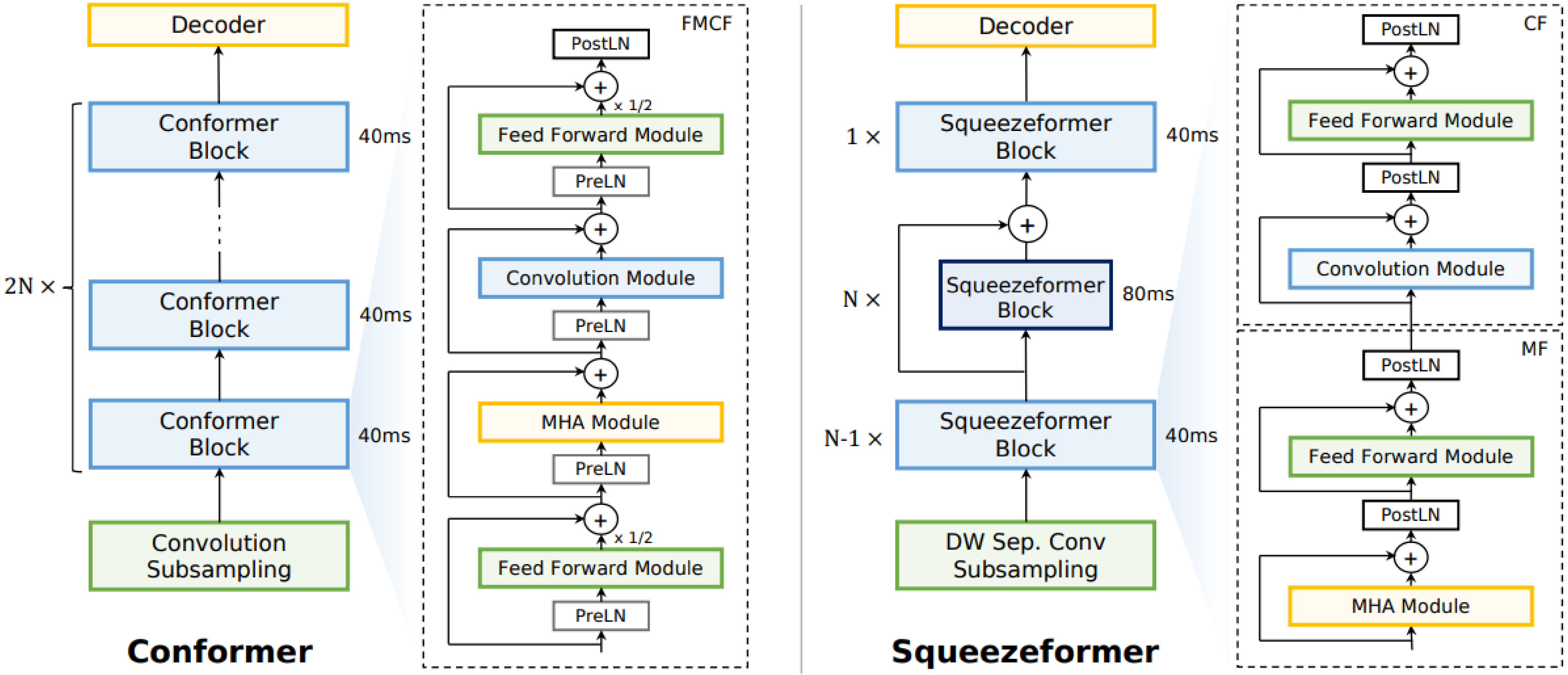
Architecture	Model	Frame accuracy ↑	Phone accuracy ↑	ABX within ↓	ABX across ↓
Single level	CPC [Rivière et al., 2020]	67.50	83.20	6.68	8.39
	ACPC [Chorowski et al., 2021]	68.60	83.33	5.37	7.09
Multi-level	Two-level CPC no downsampling	67.49	83.38	6.66	8.34
	SCPC [Bhati et al., 2021]	43.79	68.38	20.18	16.26
	Two-level CPC w. downsampling	67.92	83.39	6.66	8.32
	mACPC [Cuervo et al., 2022]	70.25	83.35	5.13	6.84
	Ours	**72.57**	**83.95**	**5.08**	**6.72**
	Downsampling (supervised)	71.01	84.70	5.07	6.68

SqueezeFormer: better than conformer





Kim S, Gholami A, Shaw A, et al. Squeezeformer: An Efficient Transformer for Automatic Speech Recognition[J]. arXiv preprint arXiv:2206.00888, 2022.

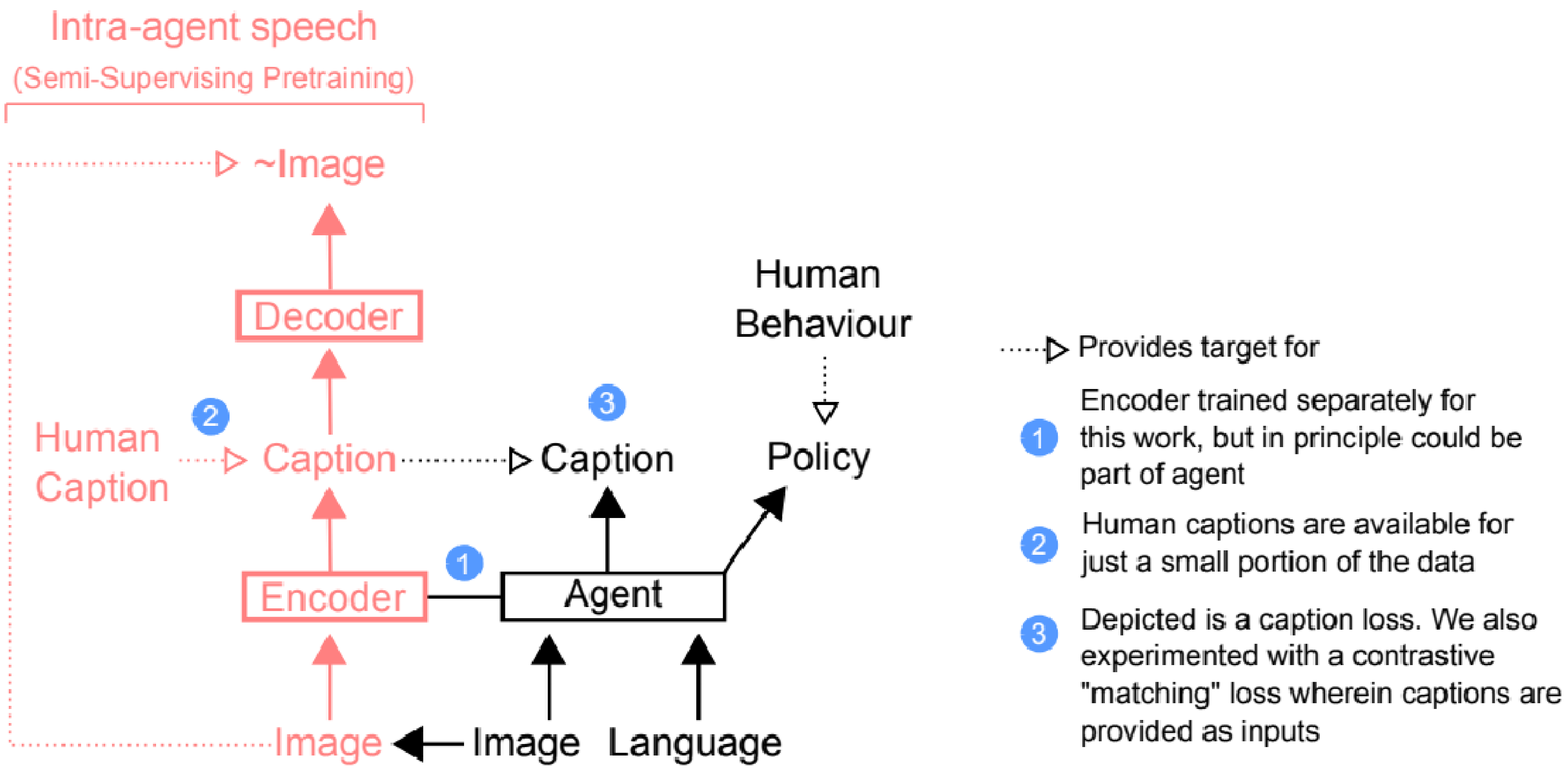


Decoder
Conformer Block
Conformer Block
Conformer Block
Convolution Subsampling
2N ×
40ms
40ms
40ms
FMCF
PostLN
x 1/2
Feed Forward Module
PreLN
Convolution Module
PreLN
MHA Module
PreLN
x 1/2
Feed Forward Module
PreLN
Conformer
Decoder
1 ×
Squeezeformer Block
40ms
N ×
Squeezeformer Block
80ms
N-1 ×
Squeezeformer Block
40ms
DW Sep. Conv Subsampling
Squeezeformer
CF
PostLN
Feed Forward Module
PostLN
Convolution Module
MF
PostLN
Feed Forward Module
PostLN
MHA Module

Result on librispeech

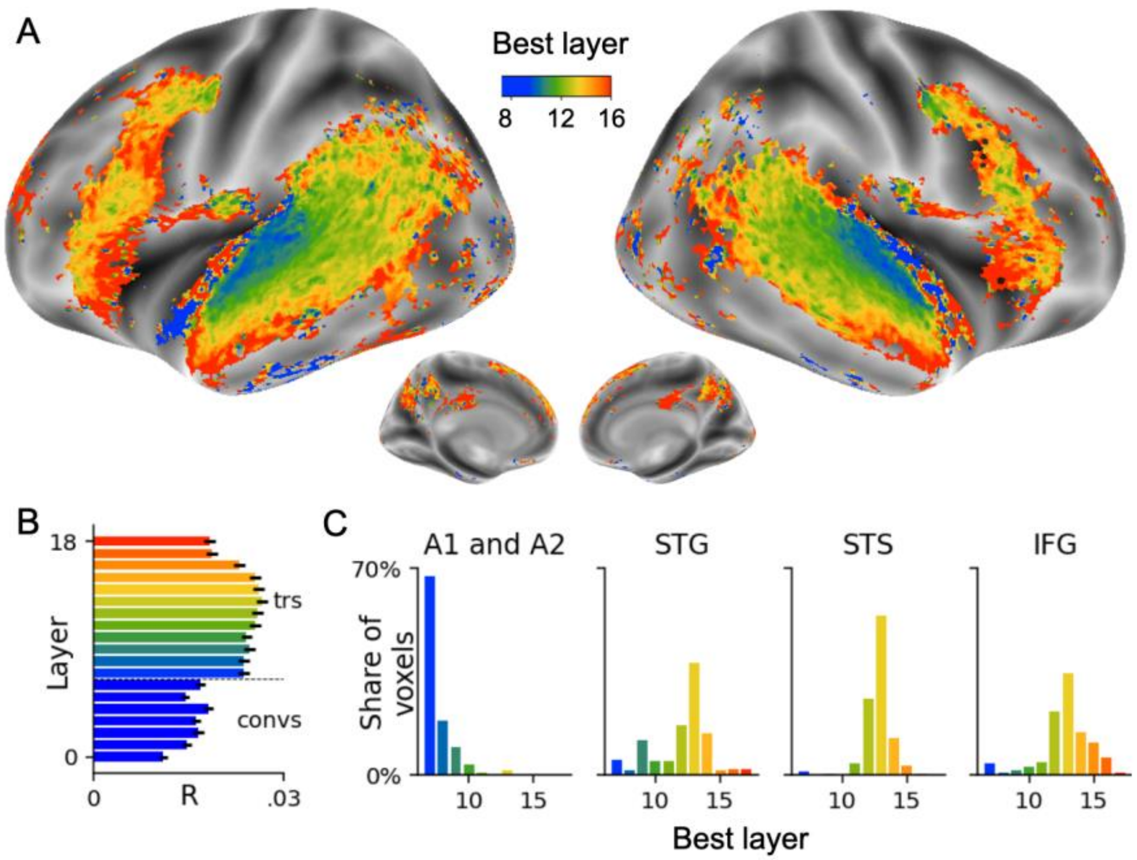
Model	Design change	test-clean	test-other	Params (M)	GFLOPs
Conformer-CTC-M	Baseline	3.20	7.90	27.4	71.7
	+ Temporal U-Net (§ 3.1.1)	2.97	7.28	27.5	57.0
	+ Transformer-style Block (§ 3.1.2)	2.93	7.12	27.5	57.0
	+ Unified activations (§ 3.2.1)	2.88	7.09	28.7	58.4
	+ Simplified LayerNorm (§ 3.2.2)	2.85	6.89	28.7	58.4
Squeezeformer-SM	+ DW sep. subsampling (§ 3.2.3)	2.79	6.89	28.2	42.7
Squeezeformer-M	+ Model scale-up (§ 3.2.3)	2.56	6.50	55.6	72.0

Self language to produce concept



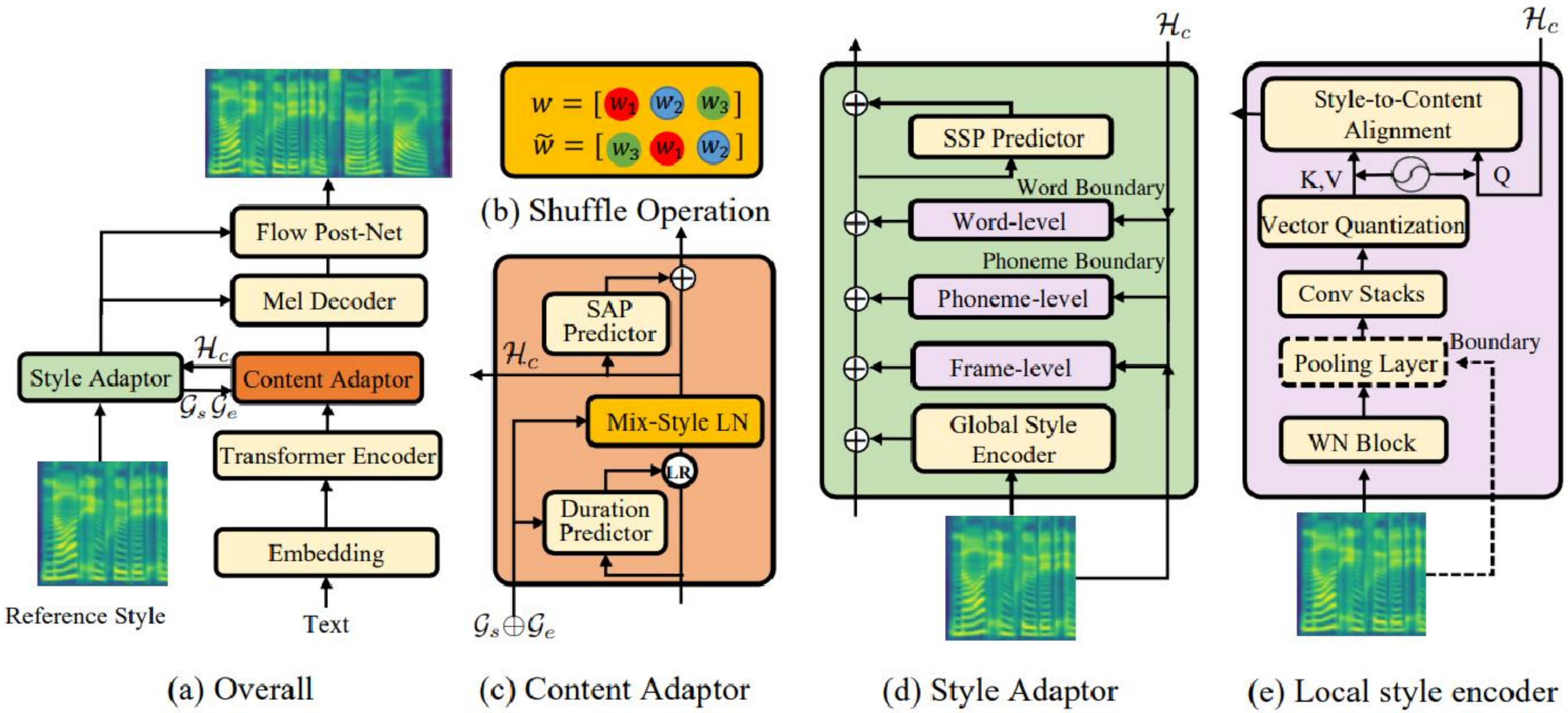
Yan C, Carnevale F, Georgiev P, et al. Intra-agent speech permits zero-shot task acquisition[J]. arXiv preprint arXiv:2206.03139, 2022.

Wav2.0 activations corresponds to brain activations



Millet J, Caucheteux C, Orhan P, et al. Toward a realistic model of speech processing in the brain with self-supervised learning[J]. arXiv preprint arXiv:2206.01685, 2022.

Multiple style encoder



Huang R, Ren Y, Liu J, et al. GenerSpeech: Towards Style Transfer for Generalizable Out-Of-Domain Text-to-Speech[C]//Advances in Neural Information Processing Systems.

Baseline	**Parallel**				**Non-Parallel**			
	7-point score	Perference (%)			7-point score	Perference (%)		
		X	Neutral	Y		X	Neutral	Y
Mellotron	1.51 ± 0.10	26%	14%	40%	1.62 ± 0.09	6%	28%	66%
FG-TransformerTTS	1.07 ± 0.14	22%	30%	48%	1.29 ± 0.10	34%	20%	46%
Expressive FS2	1.22 ± 0.12	30%	20%	50%	1.42 ± 0.11	24%	16%	60%
Meta-StyleSpeech	1.13 ± 0.09	26%	26%	48%	1.18 ± 0.12	14%	26%	60%
Styler	1.49 ± 0.10	18%	24%	58%	1.27 ± 0.09	20%	22%	58%

Binarizing connections on self-training model

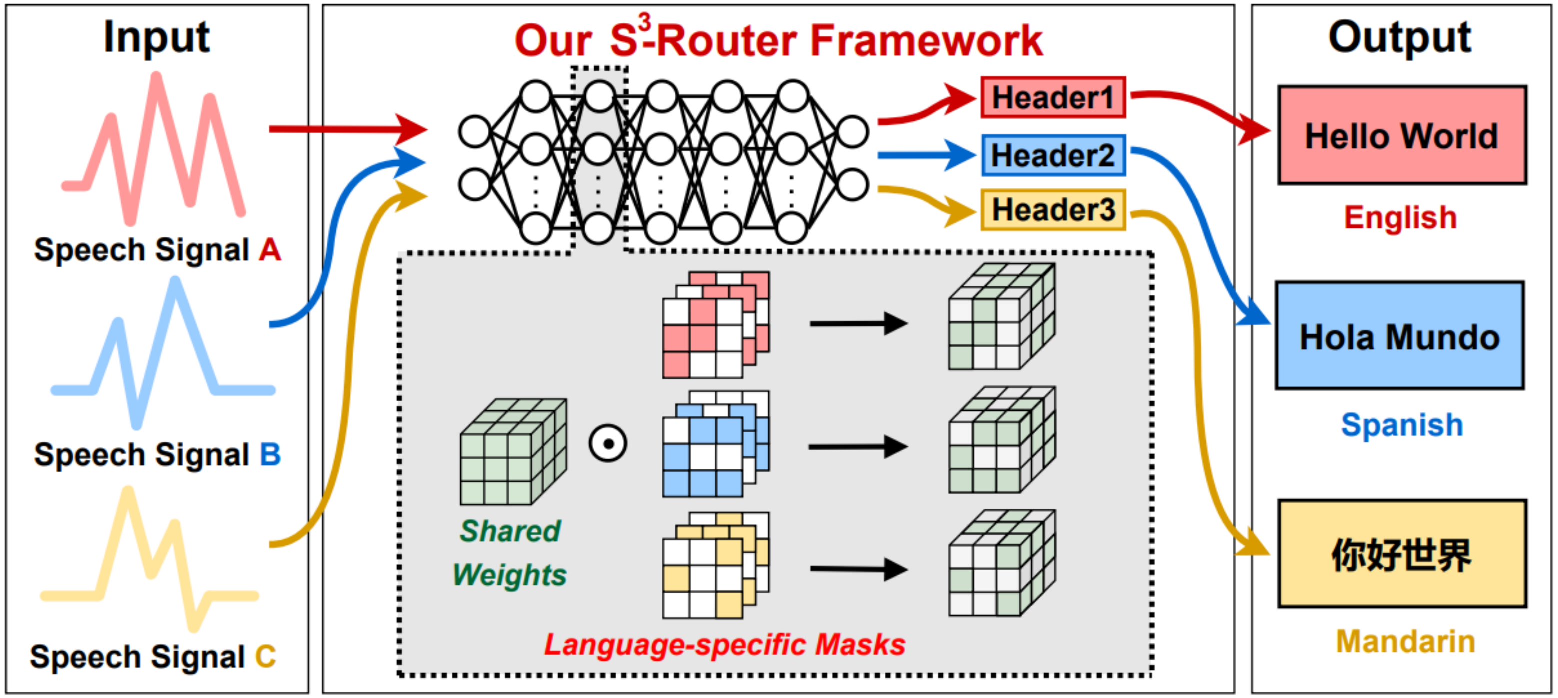


Figure 1: An overview of our S^3-Router framework, which receives multilingual speech signals denoted as A, B, and C here and then outputs the corresponding text transcript of predication, based on one *shared weight* model together with language-/task-specific *binary* masks.

- https://github.com/GATECH-EIC/S3-Router

- Fu Y, Zhang Y, Qian K, et al. Losses Can Be Blessings: Routing Self-Supervised Speech Representations Towards Efficient Multilingual and Multitask Speech Processing[J]. arXiv preprint arXiv:2211.01522, 2022.

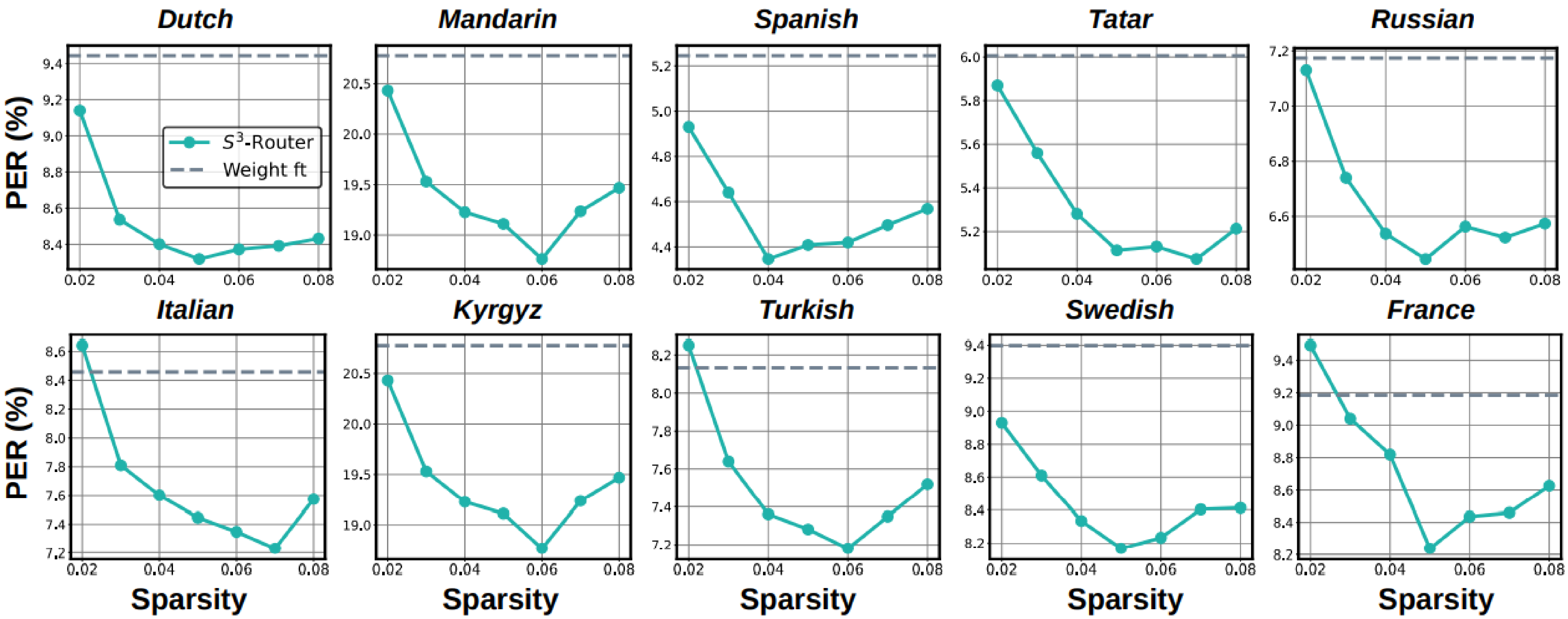


Figure 3: Benchmark our S^3-Router and weight finetuning on xlsr across 10 spoken languages.

Wav2Vec as additional code

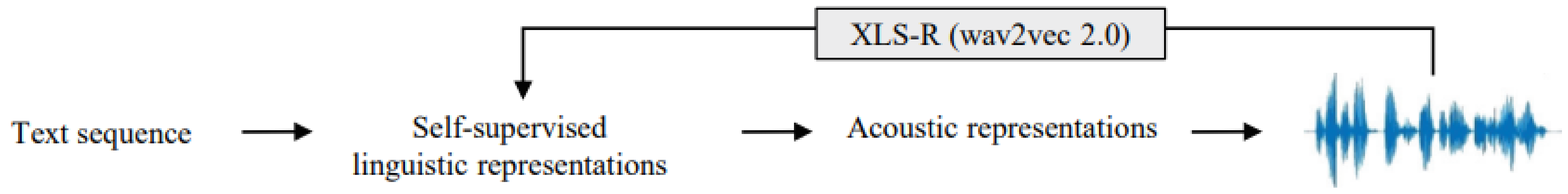


Figure 1: Hierarchical text-to-speech pipeline.

Lee S H, Kim S B, Lee J H, et al. HierSpeech: Bridging the Gap between Text and Speech by Hierarchical Variational Inference using Self-supervised Representations for Speech Synthesis[C]//Advances in Neural Information Processing Systems.

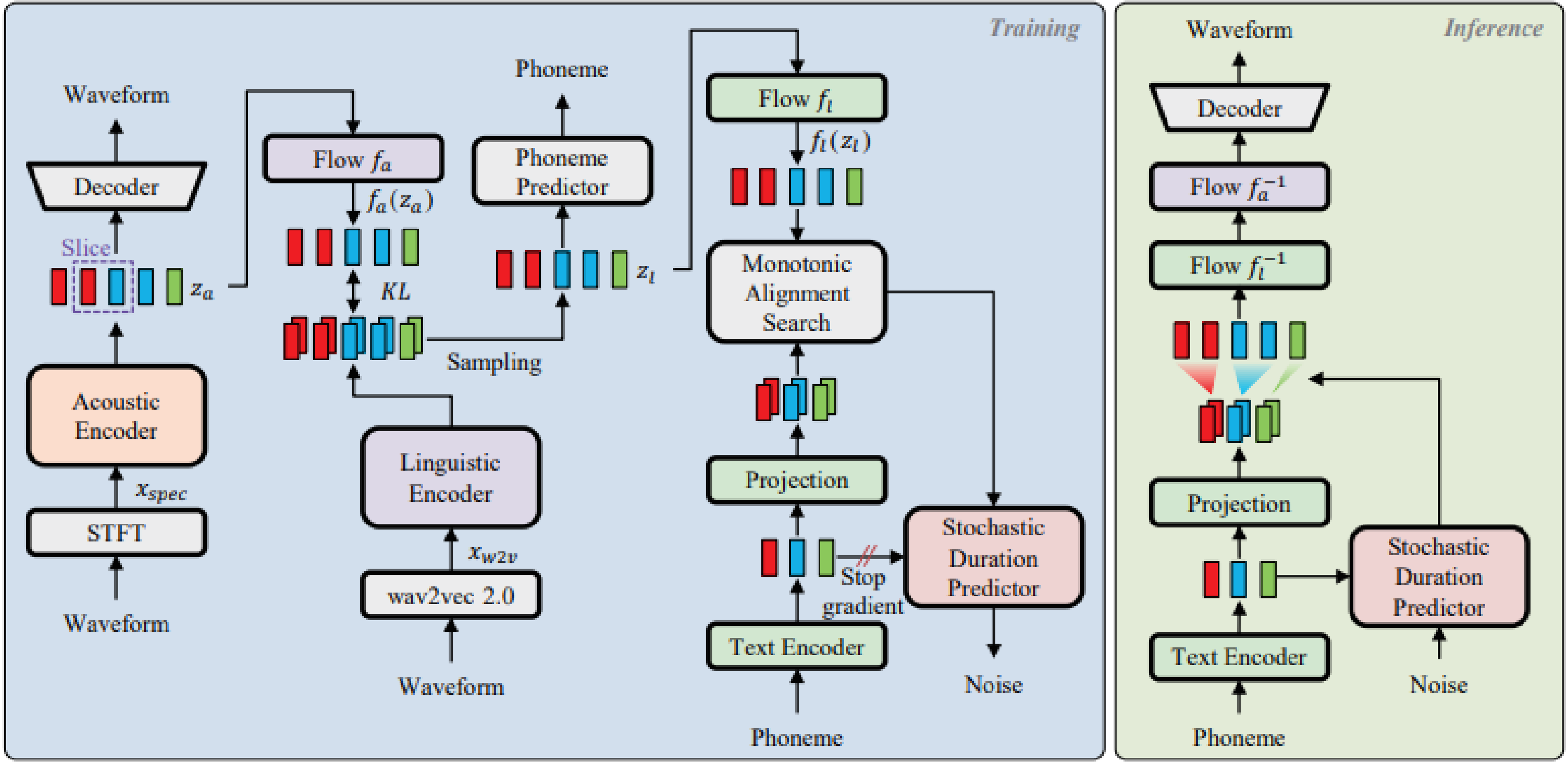


Figure 2: Overall framework of HierSpeech.