A^3T: Alignment-Aware Acoustic and Text Pretraining for Speech Synthesis and Editing

ICML 2022

He Bai[§], Renjie Zheng[†], Junkun Chen[‡], Xintong Li[†], Mingbo Ma[†], Liang Huang[‡†]

[§]University of Waterloo
[†]Baidu Research
[‡]Oregon State University

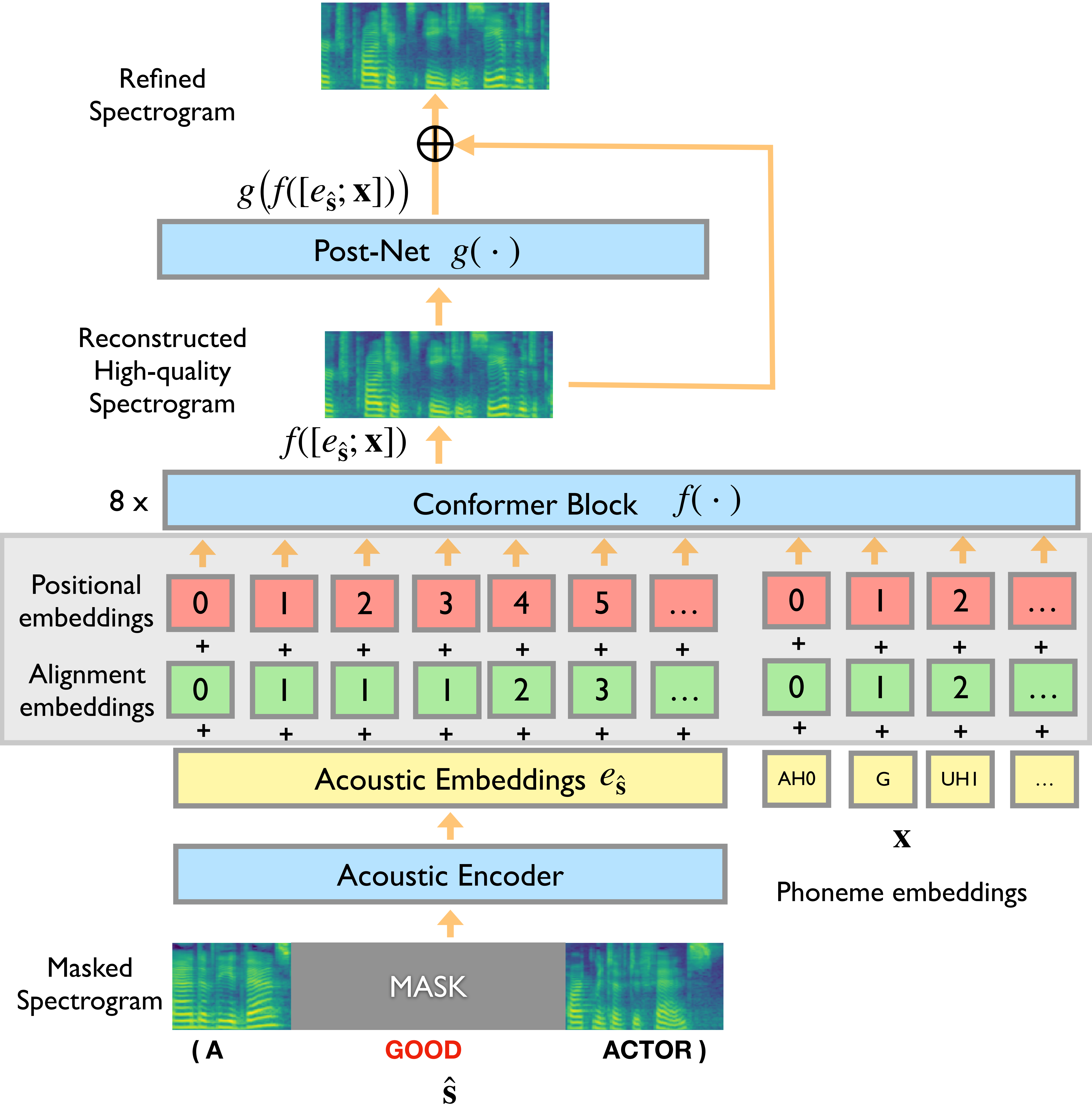
Introduction

- We propose an Alignment-Aware Acoustic and Text (A^3T) pretraining method for speech synthesis

- Without any further fine-tuning, our pre-trained model achieves the SOTA performance for speech editing.

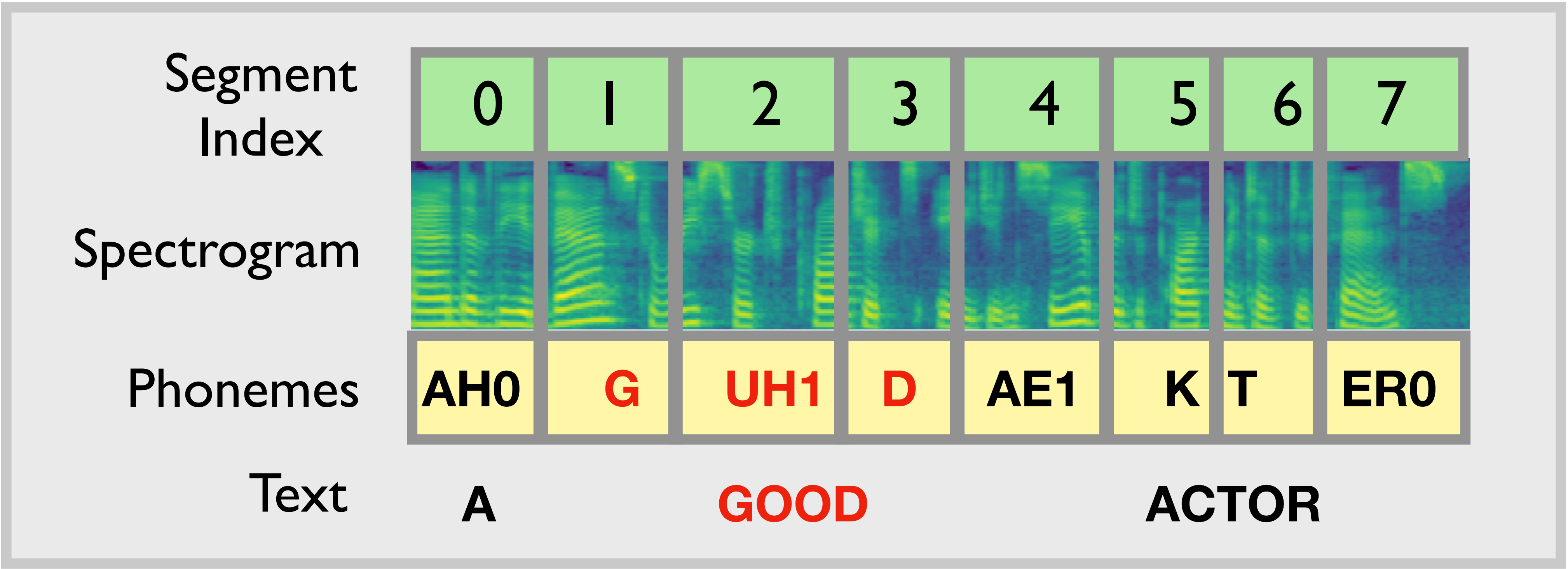
- Moreover, with our proposed Prompt-based Decoding, our pre-trained model can synthesis new speaker's speech without any speaker embedding.

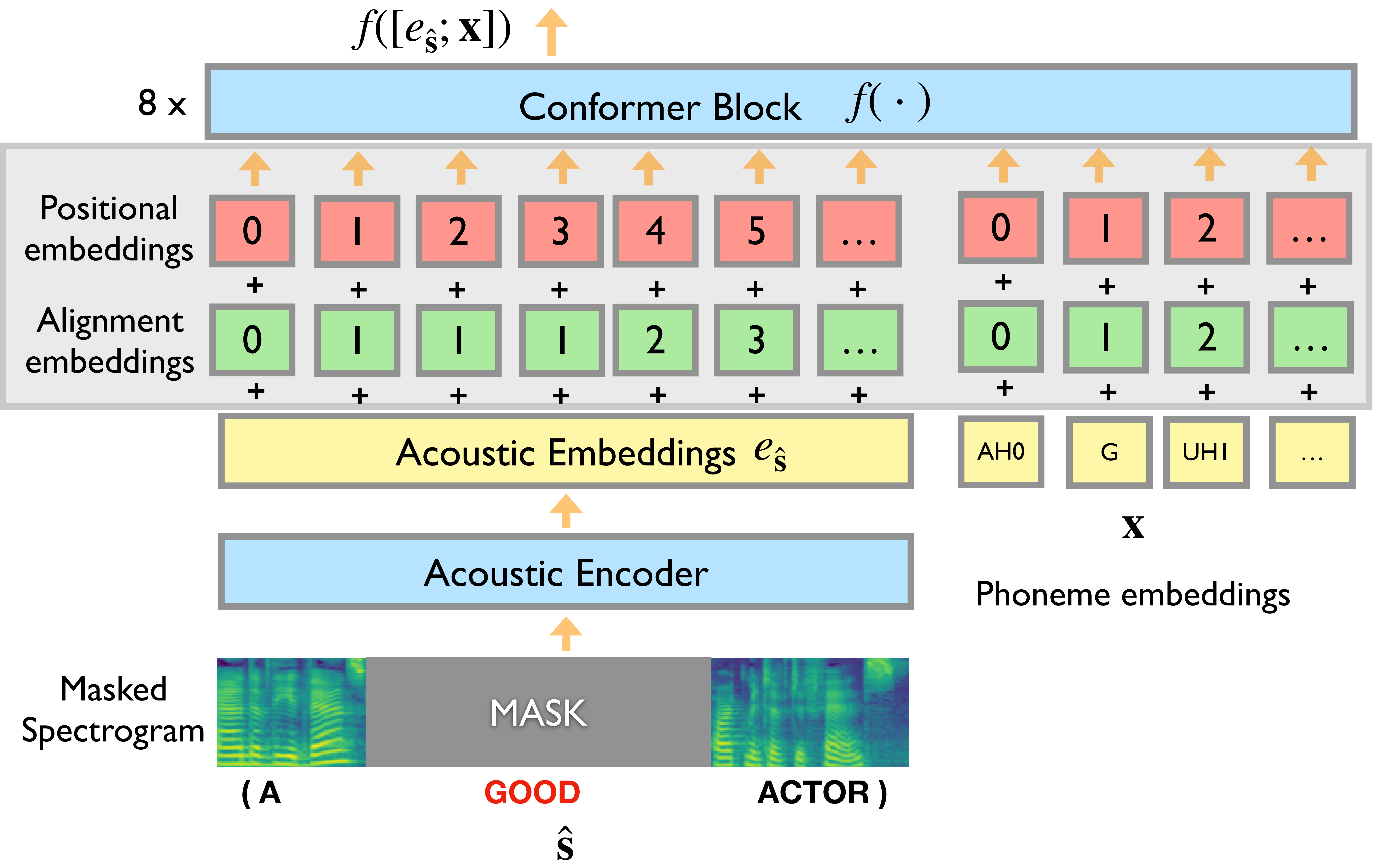
Our Model


Refined Spectrogram
$g\big(f([e_{\hat{\mathbf{s}}}; \mathbf{x}])\big)$
Post-Net $g(\cdot)$
Reconstructed High-quality Spectrogram
$f([e_{\hat{\mathbf{s}}}; \mathbf{x}])$
8 x
Conformer Block $f(\cdot)$
Positional embeddings
0 1 2 3 4 5 … 0 1 2 …
Alignment embeddings
0 1 1 1 2 3 … 0 1 2 …
Acoustic Embeddings $e_{\hat{\mathbf{s}}}$
AH0 G UH1 …
$\mathbf{x}$
Phoneme embeddings
Acoustic Encoder
Masked Spectrogram
MASK
(A GOOD ACTOR)
$\hat{\mathbf{s}}$

Our Model

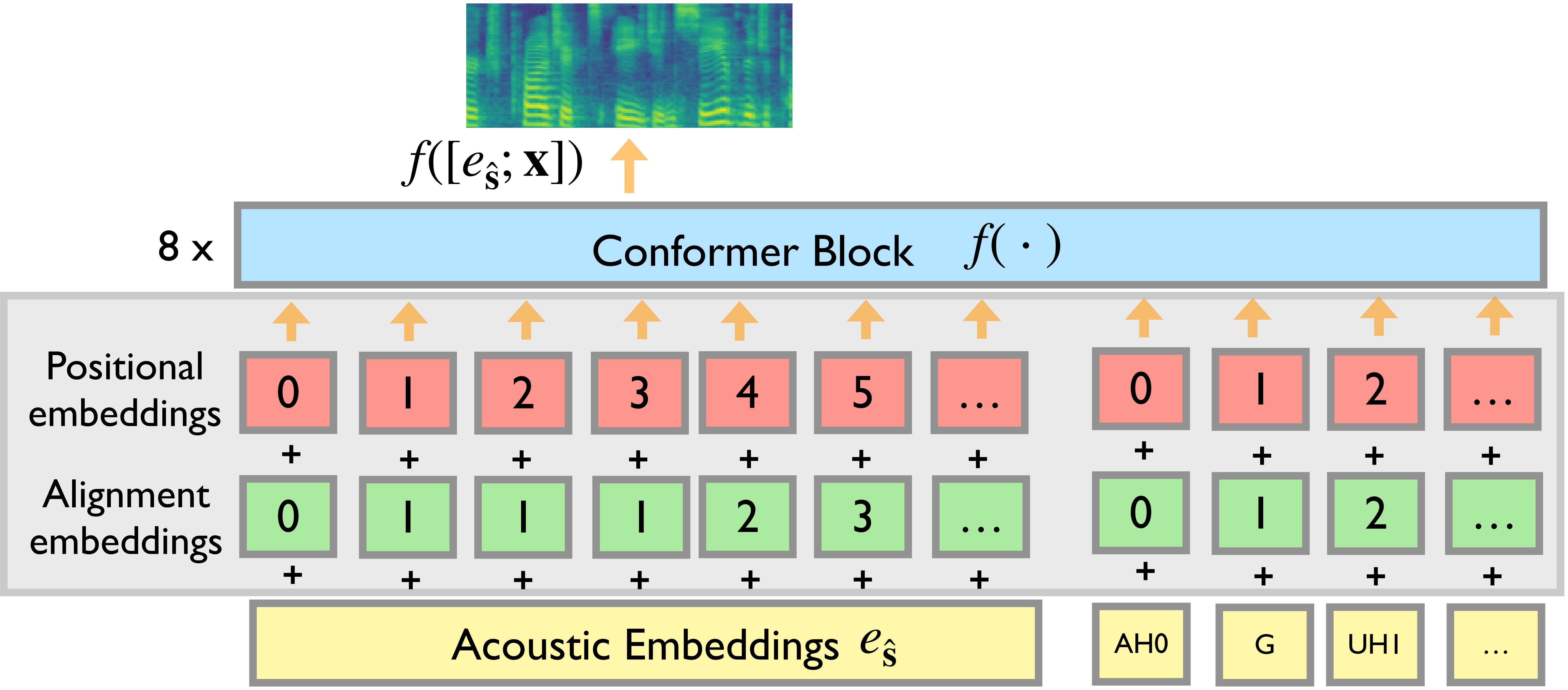
- Forced Alignment Preprocessing

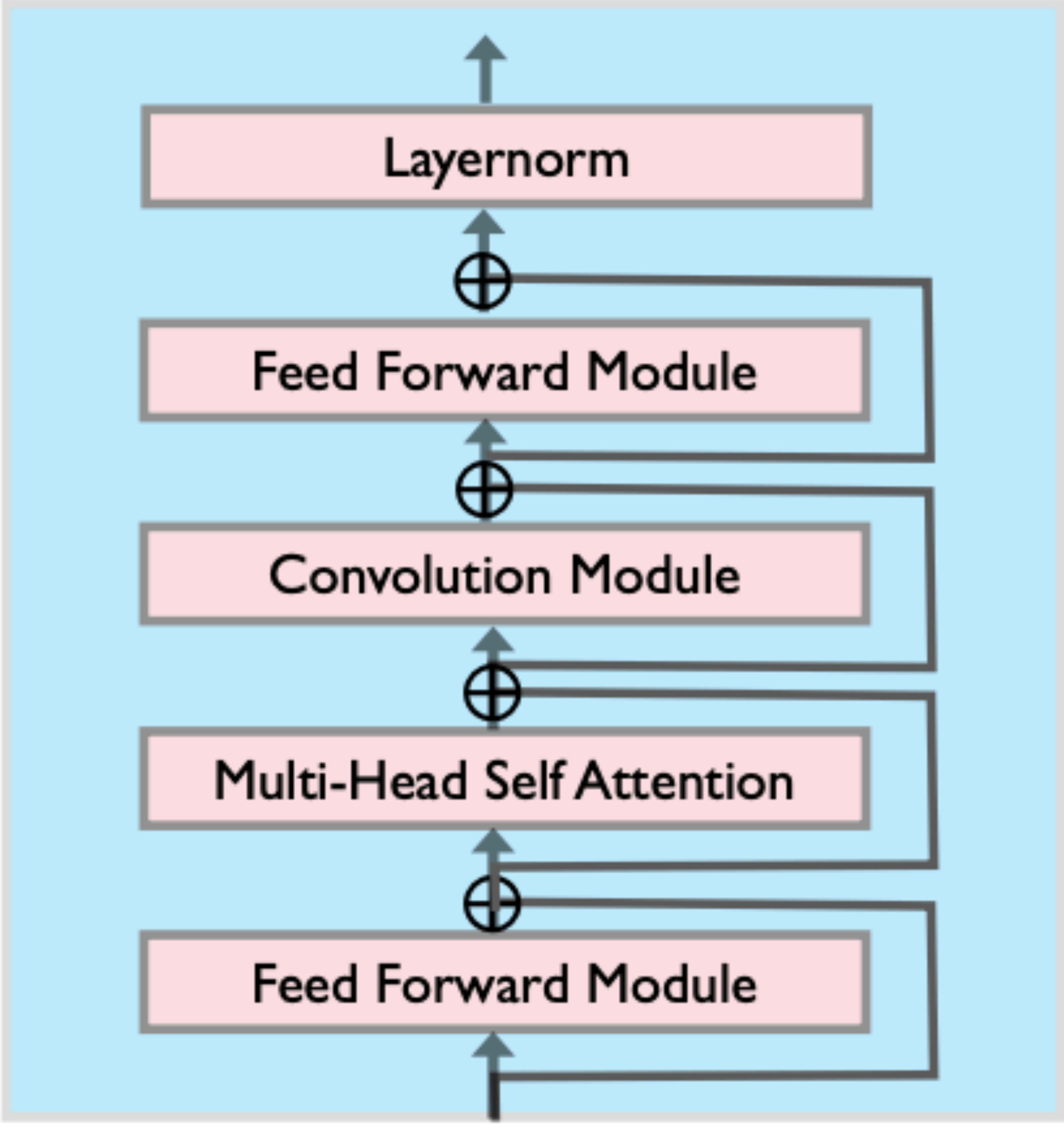




Our Model

- Conformer

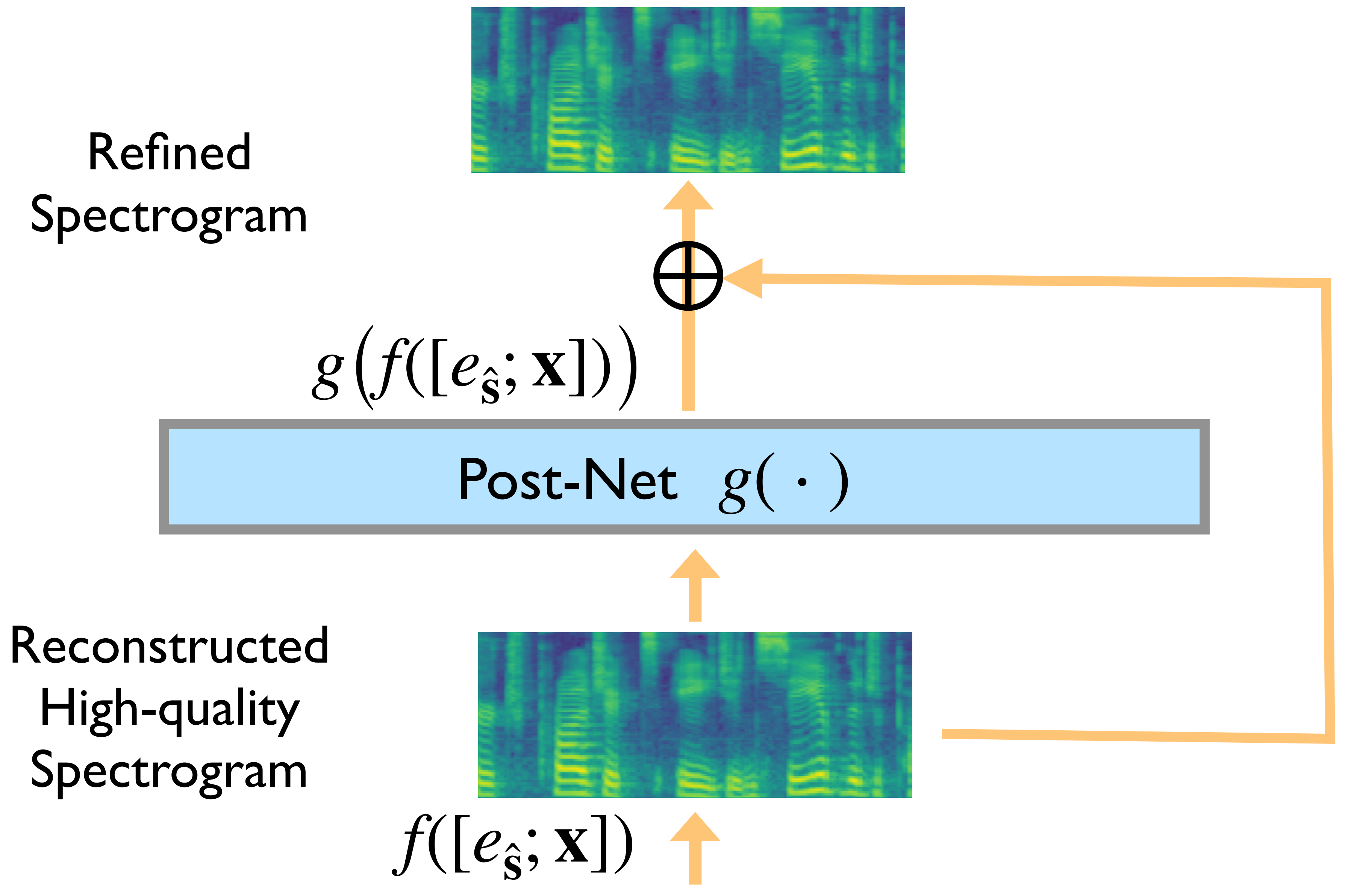


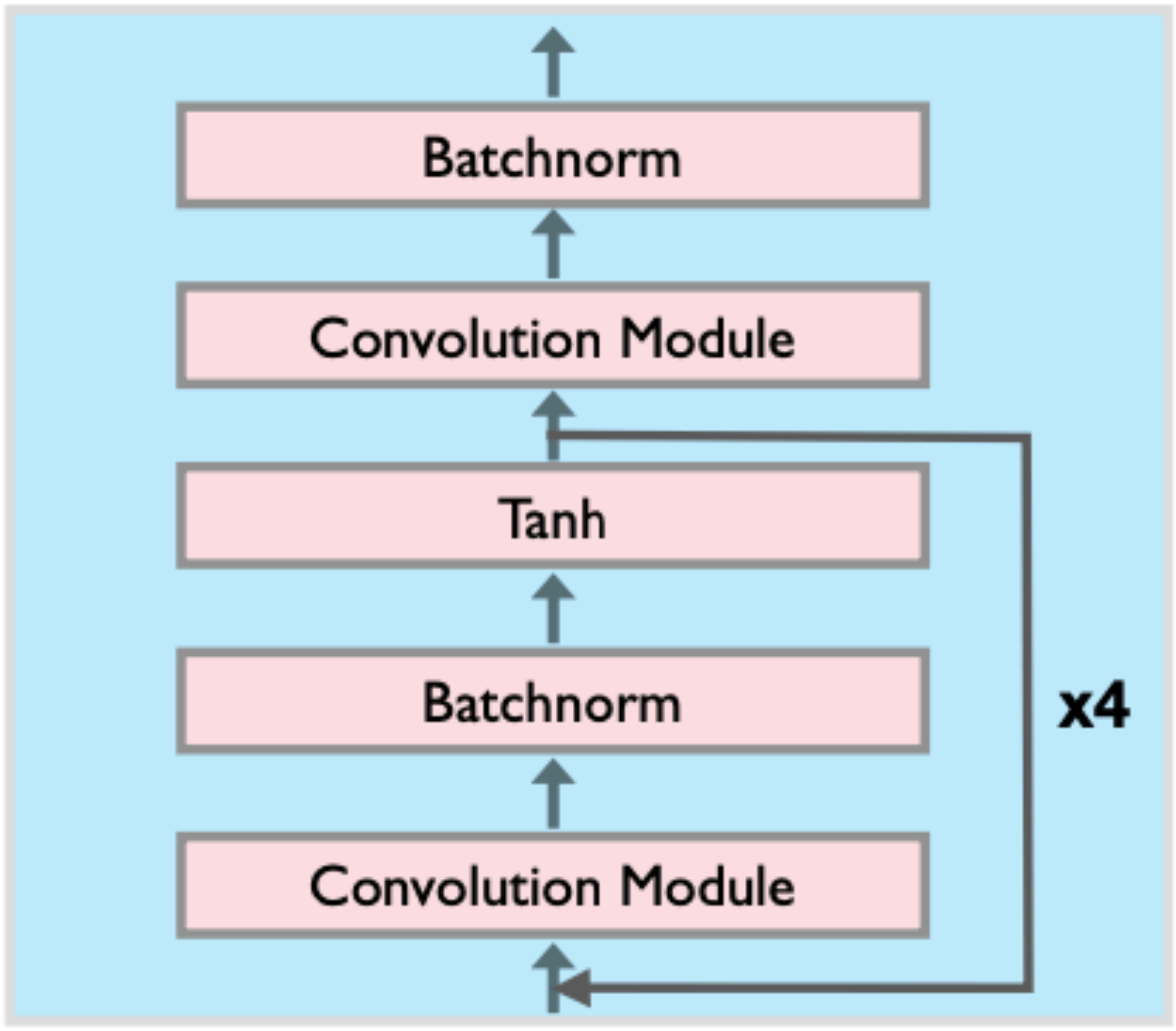


Gulati, Anmol, et al. "Conformer: Convolution-augmented transformer for speech recognition." *arXiv preprint arXiv:2005.08100* (2020).

Our Model

- PostNet and L1 loss





$$\ell_{\mathbf{s}}(D_{\mathbf{s},\mathbf{x}}) = \sum_{\langle \mathbf{s},\mathbf{x}\rangle \in D_{\mathbf{s},\mathbf{x}}} \| \underbrace{f([e_{\hat{\mathbf{s}}};\mathbf{x}]) + g\big(f([e_{\hat{\mathbf{s}}};\mathbf{x}])\big)}_{\text{refined spectrogram}} - \mathbf{s}\|_1 + \| \underbrace{f([e_{\hat{\mathbf{s}}};\mathbf{x}])}_{\text{reconstructed spectrogram}} - \mathbf{s}\|_1$$

Shen, Jonathan, et al. "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions." *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018.

Experiments

- 1. Ablation Study of Spectrogram Reconstruction
- 2. Speech Editing
- 3. Prompt-based Decoding for New speaker TTS (in-context learning)
- 4. For the fine-tuning experiments, please read our paper

Type	Name	# Speakers	# Samples	# Hours
TTS	LJSpeech	1	13K	24
TTS	VCTK	109	44K	44
TTS	LibriTTS	2,456	158K	586

Ablation Study

An example of ablation study in LJSpeech. Original text is “and of the Advanced Research Projects Agency of the Department of Defense”.

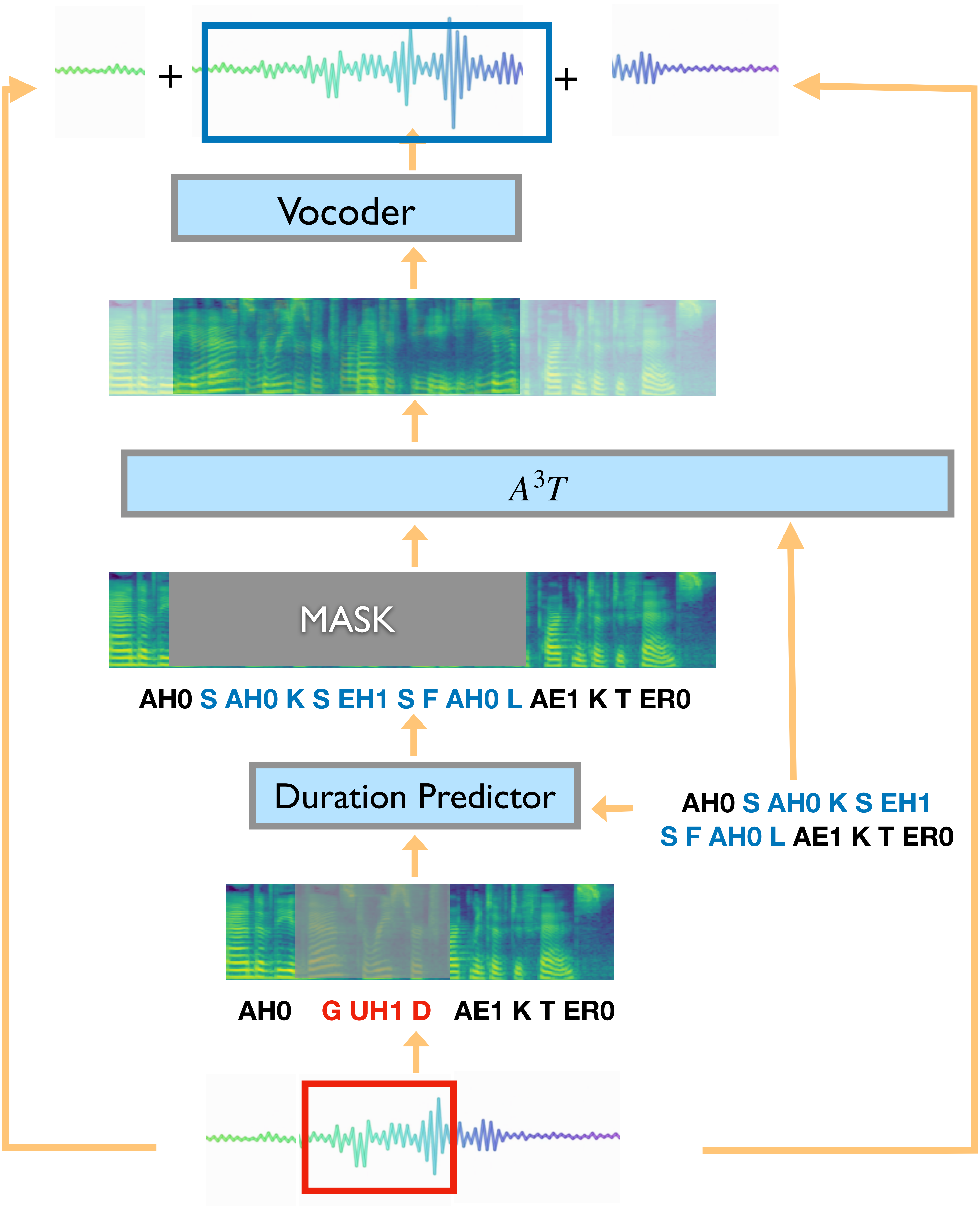
Ablation Study

Ablation MCD scores with LJSpeech dataset:

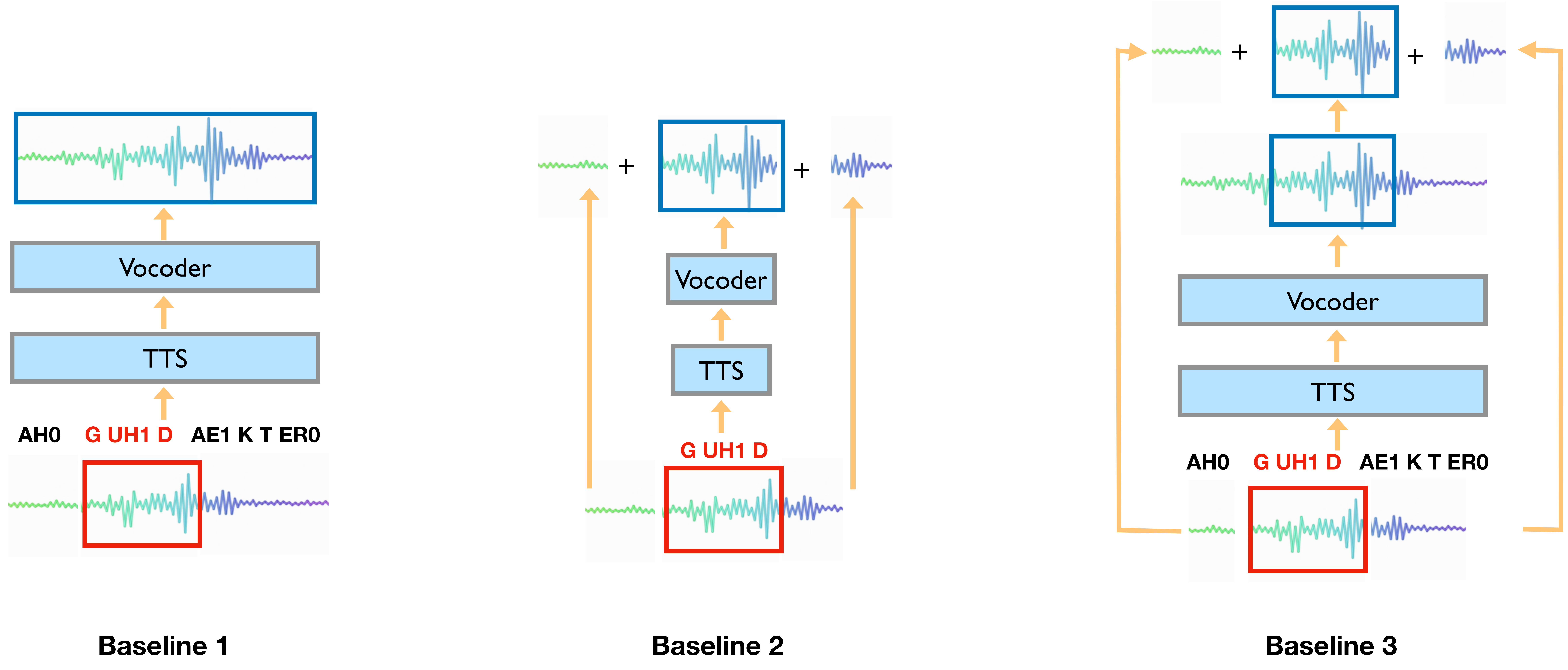
Example	**Model**	**MCD ↓**
Fig. 4(b)	A^3T	8.09
Fig. 4(c)	- Alignment Embeddings	10.73
Fig. 4(d)	- Conformer	12.43
Fig. 4(e)	- Post-Net	12.94
Fig. 4(f))	- L1 loss	11.55

A^3T for Speech Editing

we use FastSpeech 2 duration predictor



Speech Editing Baseline



Speech Editing Results

MCD scores:

Model	VCTK MCD ↓	LJSpeech MCD ↓
Baseline 1/3	10.66	10.32
Baseline 2	12.06	10.91
A^3T	**7.76**	**9.26**
w/o Alignment Emb.	11.37	10.30

MOS scores:

Model	Insert	Replace
Baseline 1	3.02 ± 0.20	2.64 ± 0.16
Baseline 2	2.89 ± 0.17	2.70 ± 0.16
Baseline 3	2.89 ± 0.17	2.44 ± 0.16
Tan et al. (2021)	3.50 ± 0.16	3.58 ± 0.16
A^3T	$\mathbf{3.53} \pm 0.17$	$\mathbf{3.65} \pm 0.15$
w/o Alignment Emb.	2.48 ± 0.21	1.98 ± 0.17

Speech Editing Examples 1 (Single Speaker)

Original	who responded to the unplanned event with dispatch.
Edited 1	unplanned → unexpected
Edited 2	unplanned event → unexpected question

Speech Editing Examples 2 (Multi-speaker)

Original

for that reason cover should not be given

Tan et al.

for that reason cover is impossible to be given

Ours

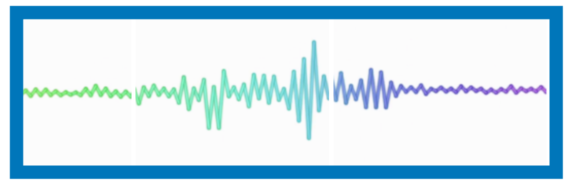
Tan et al.

for that theoretical and realistic reason cover should not be given

Ours

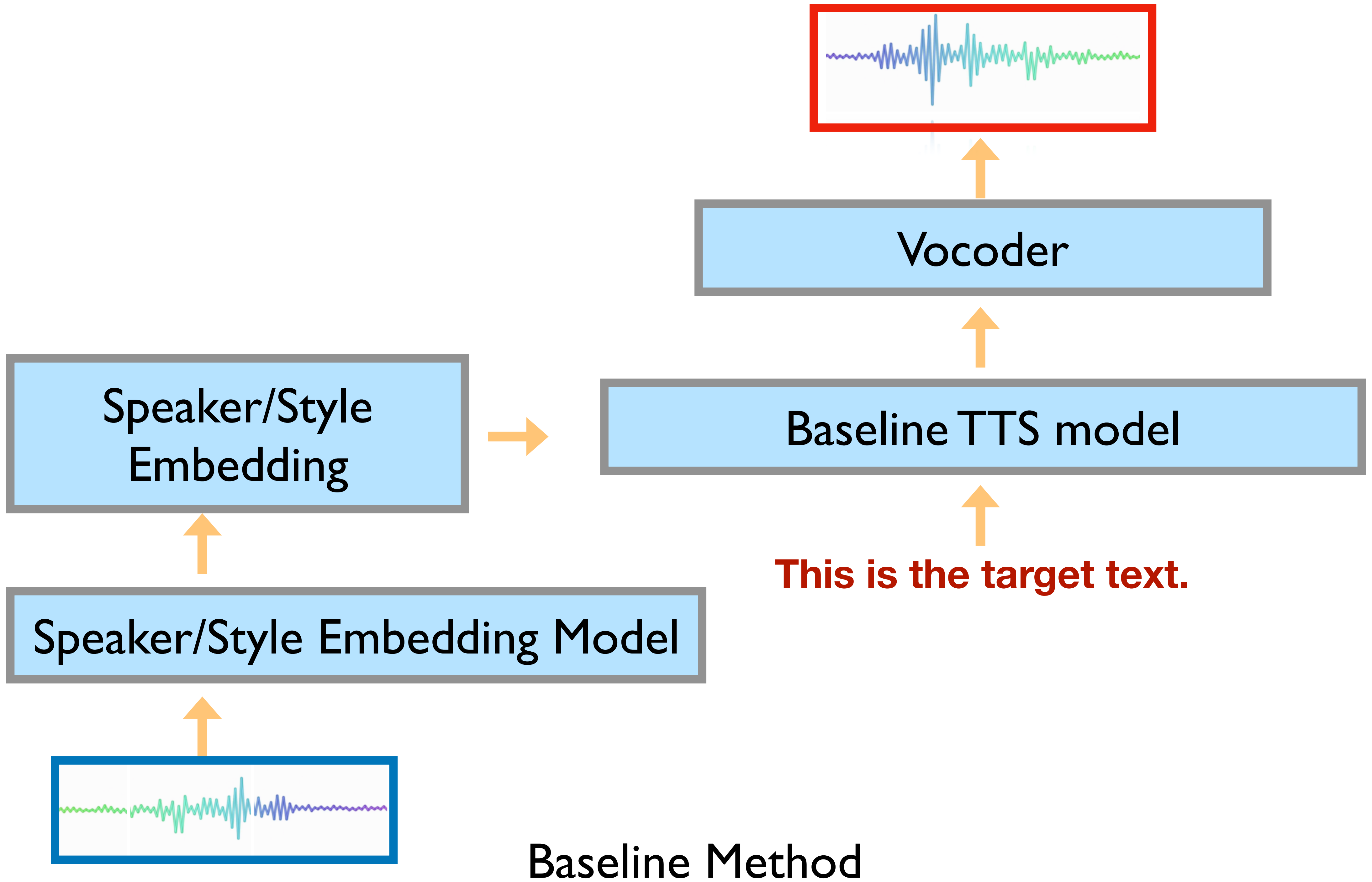
Prompt-based Decoding for Speech Synthesis

Input example:

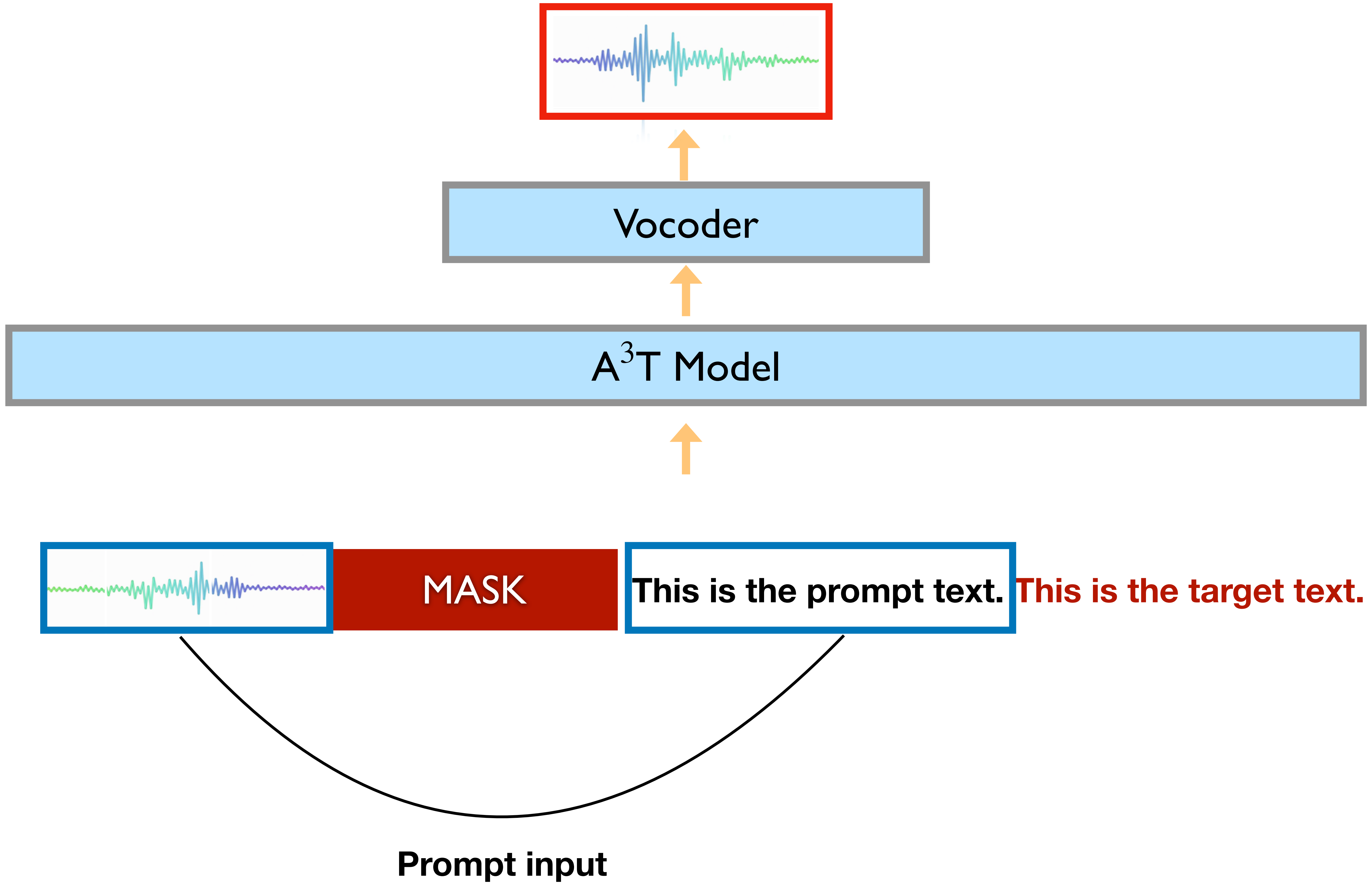
Unseen speaker's speech: 

Unseen speaker's Text: **This is the prompt text.**

Text we want to pronounce : **This is the target text.**



Baseline Method



Our Method

New Speaker Speech Synthesis Examples

Prompt

Ours

Baseline

Prompt

Ours

Baseline

In-context learning Example

Prompt

Ours

Conclusion

- This is the first pre-training method for speech synthesis, which can be used like GPT3, without any fine-tuning, to generate high quality speech and can benefit from the prompt in-context learning.
- Our model outperforms the SOTA speech editing system
- Our model can do new speaker TTS without any speaker embedding