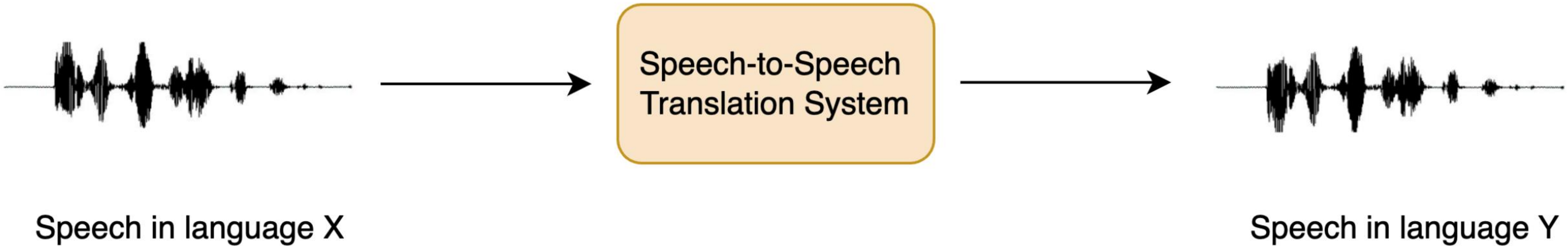
DIFFNORM: Self-Supervised Normalization for Non-autoregressive Speech-to-speech Translation

Weiting Tan, Jingyu Zhang, Lingfeng Shen, Daniel Khashabi, Philipp Koehn

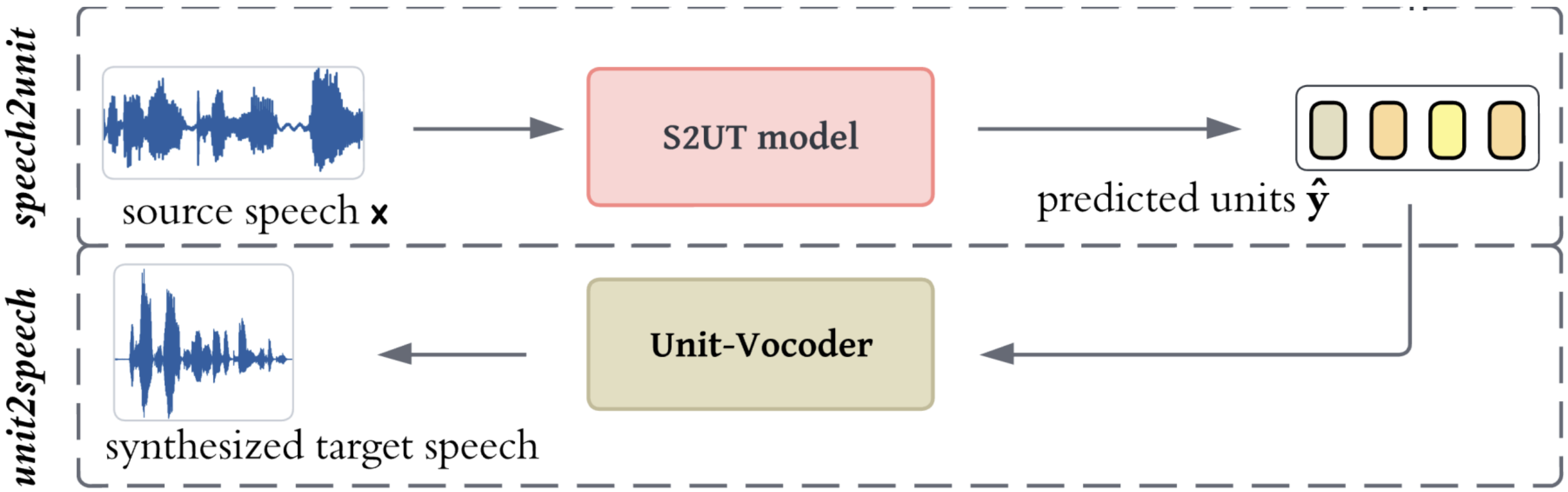


CENTER FOR LANGUAGE AND SPEECH PROCESSING
JOHNS HOPKINS UNIVERSITY

Speech-to-Speech Translation (S2ST)

Two stages

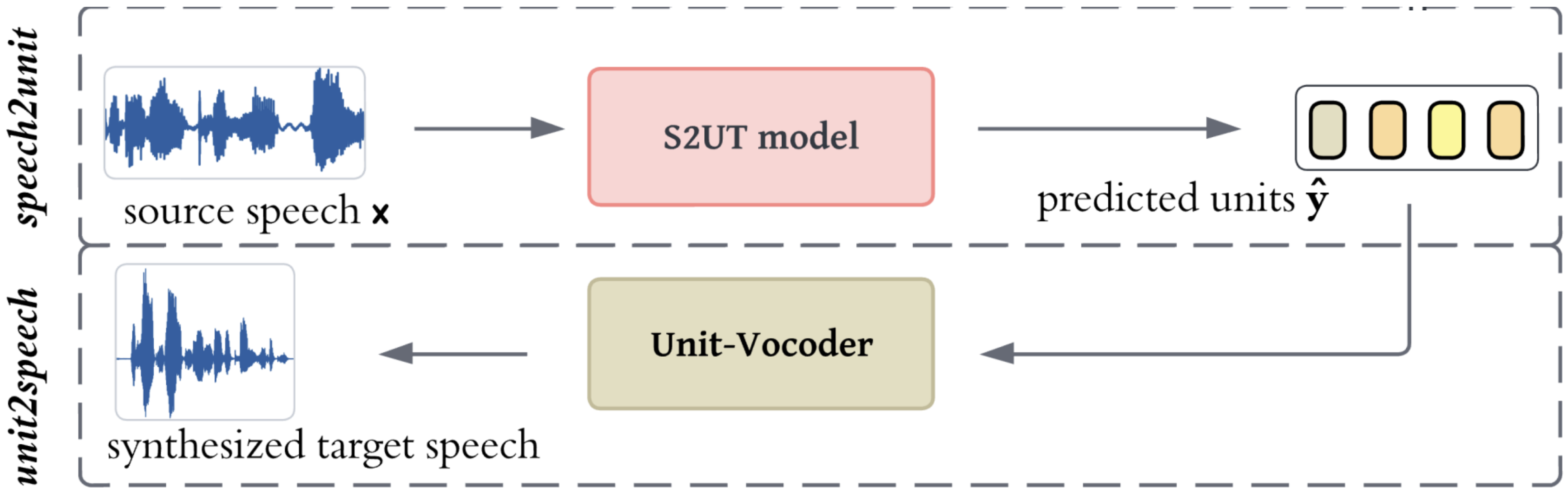
- S2UT (speech-to-unit translation): Convert source speech into target speech units
- Unit-Vocoder: Synthesize target speech from target speech units



Non-autoregressive Speech-to-Speech Translation

S2UT Model

- Transformer/Conformer-based
- Non-autoregressive Transformer (NAT): Masked-Predict Language Model



Non-autoregressive Speech-to-Speech Translation

S2UT Backbone: CMLM [1]

- Source encoded by Transformer/Conformer-Encoder
- Target units predicted by Transformer-Decoder non-autoregressively
  - Use Iterative Refinement during decoding
  - Tokens of all positions are predicted and tokens with TopK prob are kept for next iteration

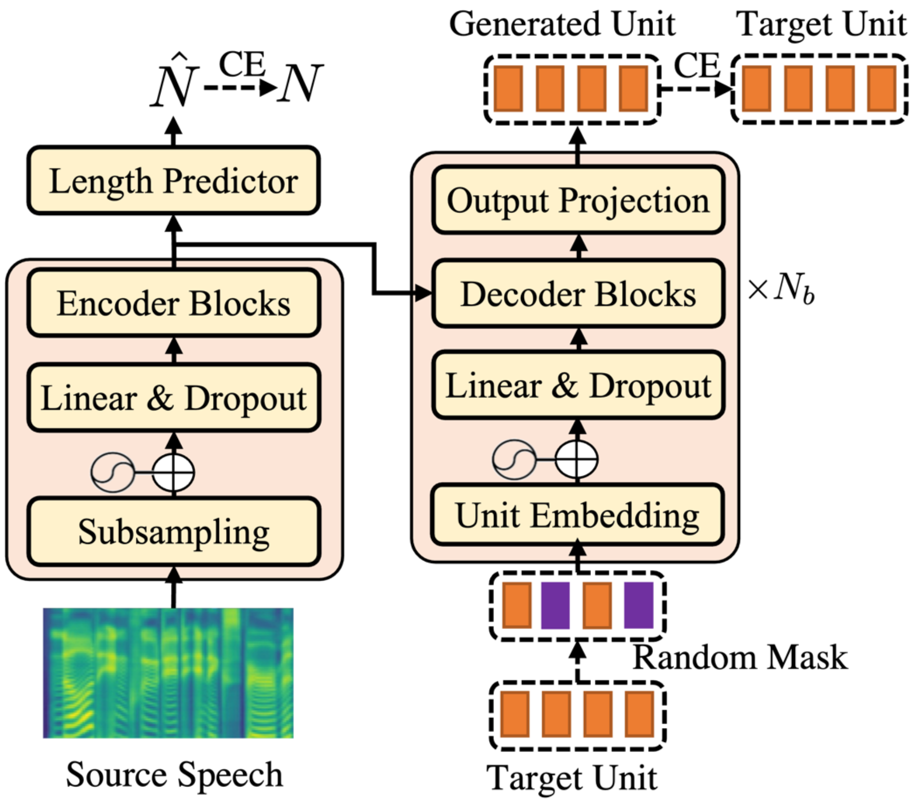


Figure from [2]

[1] Ghazvininejad et al., (2019). Mask-Predict: Parallel Decoding of Conditional Masked Language Models
[2] Huang et al., (2023). TRANSPEECH: SPEECH-TO-SPEECH TRANSLATION WITH BILATERAL PERTURBATION

Challenge in Non-autoregressive S2ST

Multi-modality Problem

- Acoustic: the same content can sound differently due to acoustic conditions
- Linguistic: Multiple correct translations exist for the same source speech

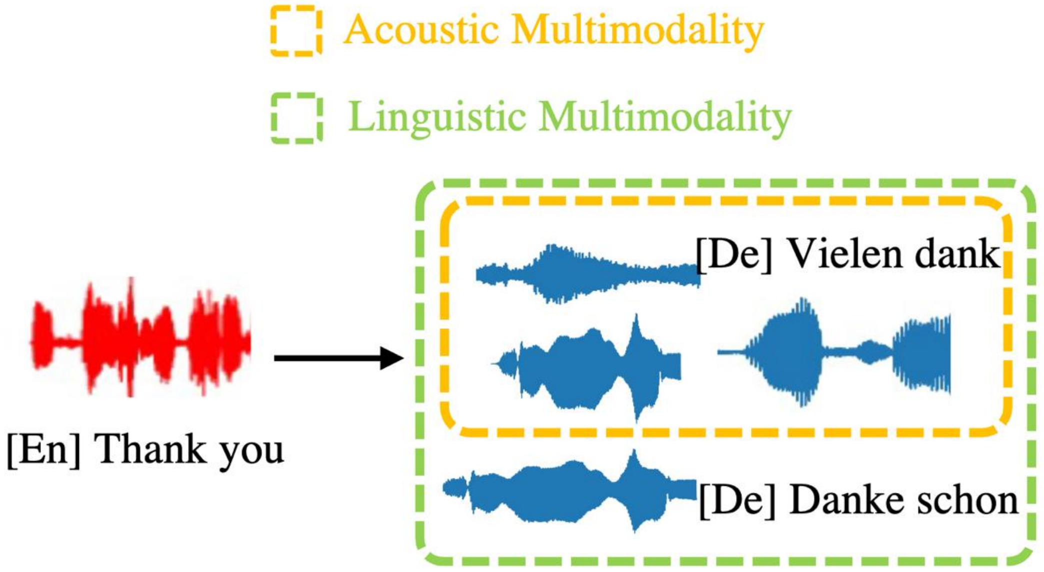
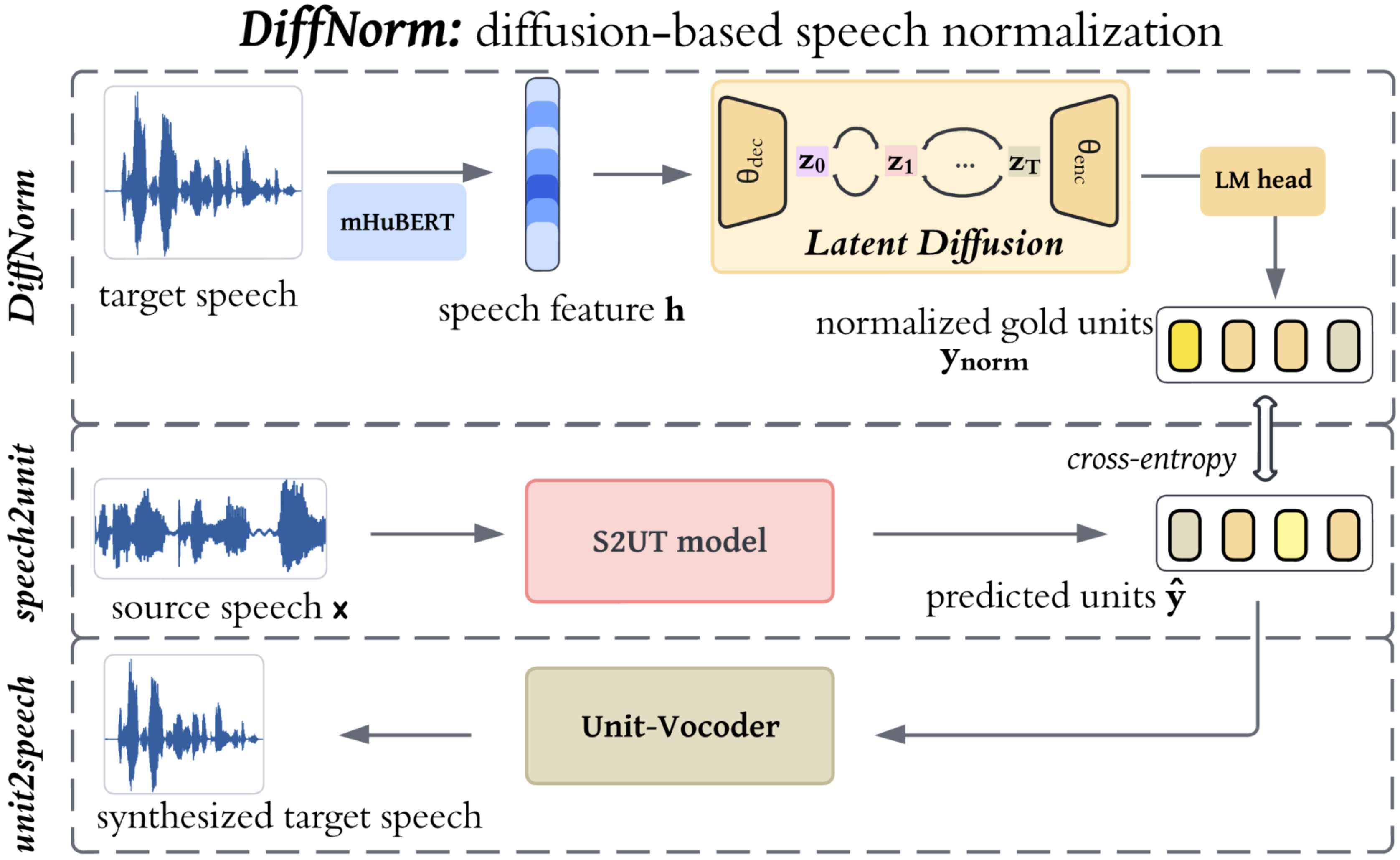


Figure from [1]

[1] Huang et al., (2023). TRANSPEECH: SPEECH-TO-SPEECH TRANSLATION WITH BILATERAL PERTURBATION

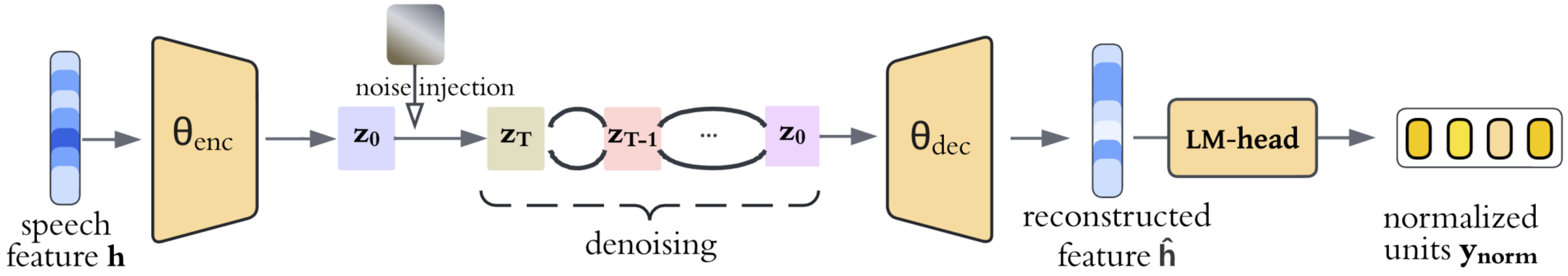
Strategy: Speech Normalization with Diffusion



Strategy: Speech Normalization with Diffusion

Construct Normalized Speech Units:

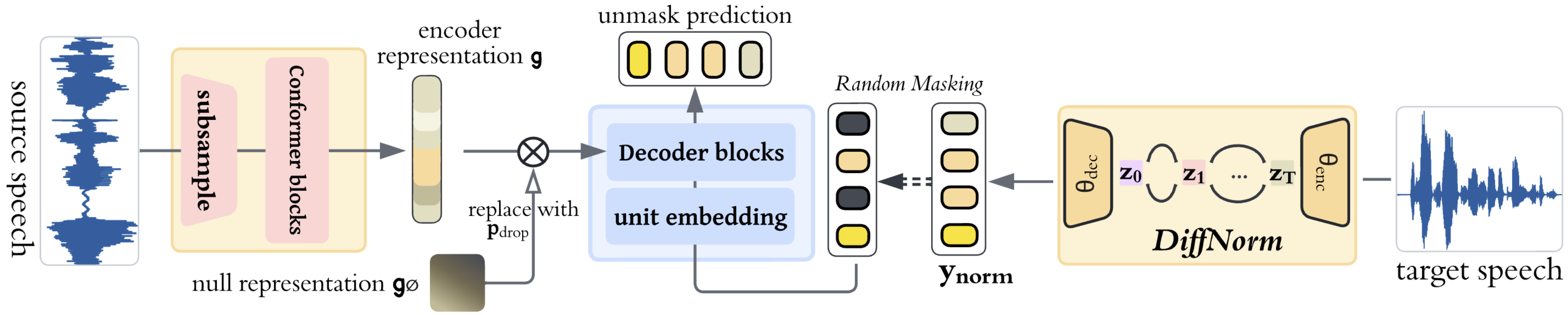
- Train VAE model on target speech feature
- Train Diffusion Model on VAE latents
- Units Construction:
  - Choose a start time T to inject noise into the clean latents (z_0 -> z_T)
  - Denoise with pre-trained Diffusion Model and reconstruct feature
  - Predict normalized speech units with reconstructed feature



CMLM with DiffNorm Units

Training with classifier-guidance (adapted from Diffusion to NAT)

- Randomly replace source representation with null representation
- Improve decoder’s iterative decoding quality, especially for long-sequences



Selected Experiment Results

ID	*System*	*Quality* ↑		*Inference Speed* ↑	
		En-Es	En-Fr	Speed	Speedup
Autoregressive					
1	Transformer† [30]	10.07	15.28	870	1.00×
2	Norm Transformer† [31]	12.98	15.93	870	1.00×
3	Conformer†	13.75	17.07	895	1.02×
Non-autoregressive Model					
4	CMLM	12.58	15.62	4651	5.34×
5	CMLM + BiP† [20]	12.62	16.97		
Our Improved Non-autoregressive Model					
6	CMLM + DiffNorm	18.96	17.27	4651	5.34×
7	CMLM + CG‡	17.06	16.89		
8	CMLM + DiffNorm + CG‡	**19.49**	**17.54**		

Table 2: Comparison of speech-to-speech models evaluated by quality (ASR-BLEU) and speed (units/seconds). Results with † are taken from the prior work [20]. ‡ We use $w = 0.5$ for CG. **Our NAT models achieve superior translation quality while maintaining their fast inference speed**.

Selected Experiment Results

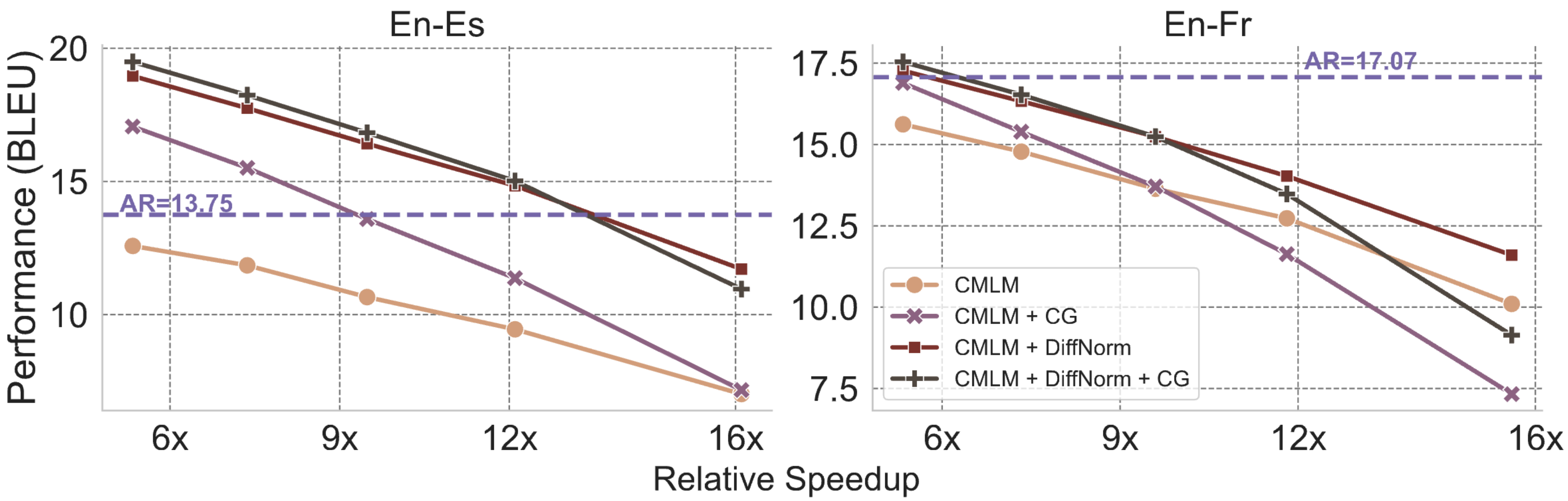


Figure 4: Trade-off between quality (ASR-BLEU) and latency for varying numbers of decoding iterations. Five markers correspond to {15, 10, 7, 5, 3} decoding iterations. Decreasing the number of iterations results in a decline in model performance, traded off for faster speedup. With DiffNorm and CG, **our S2UT model achieves a better quality-latency trade-off** than CMLM and outperforms a strong autoregressive baseline with large speedups.

THANK YOU!

Please feel free to reach out to me with questions/suggestions at wtan12@jhu.edu