SpeechT5: Unified-Modal Encoder-Decoder Pre-Training for Spoken Language Processing

https://arxiv.org/abs/2110.07205

Hugging Face ML-4-Audio

https://hf.co/join/discord

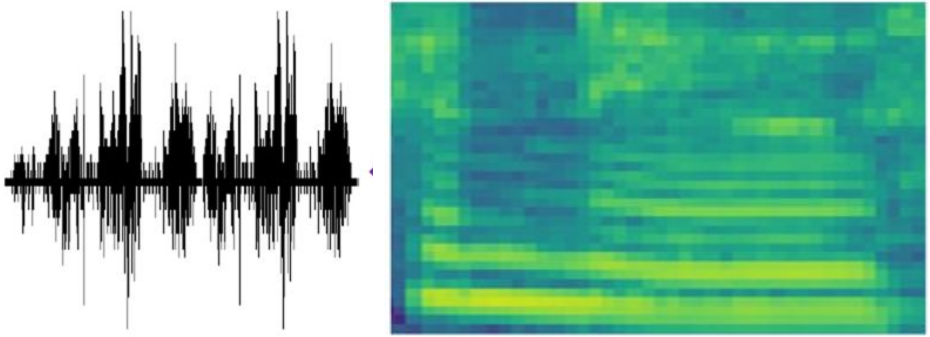
SpeechT5 Quick Facts

- Published in May 2022
- Developed by Microsoft Research Asia
- Inspired by T5 (Text-To-Text Transfer Transformer)
- First text-to-speech model added to 🤗 Transformers

Spoken Languages

Example: “This is a sentence”

Audio



Speech

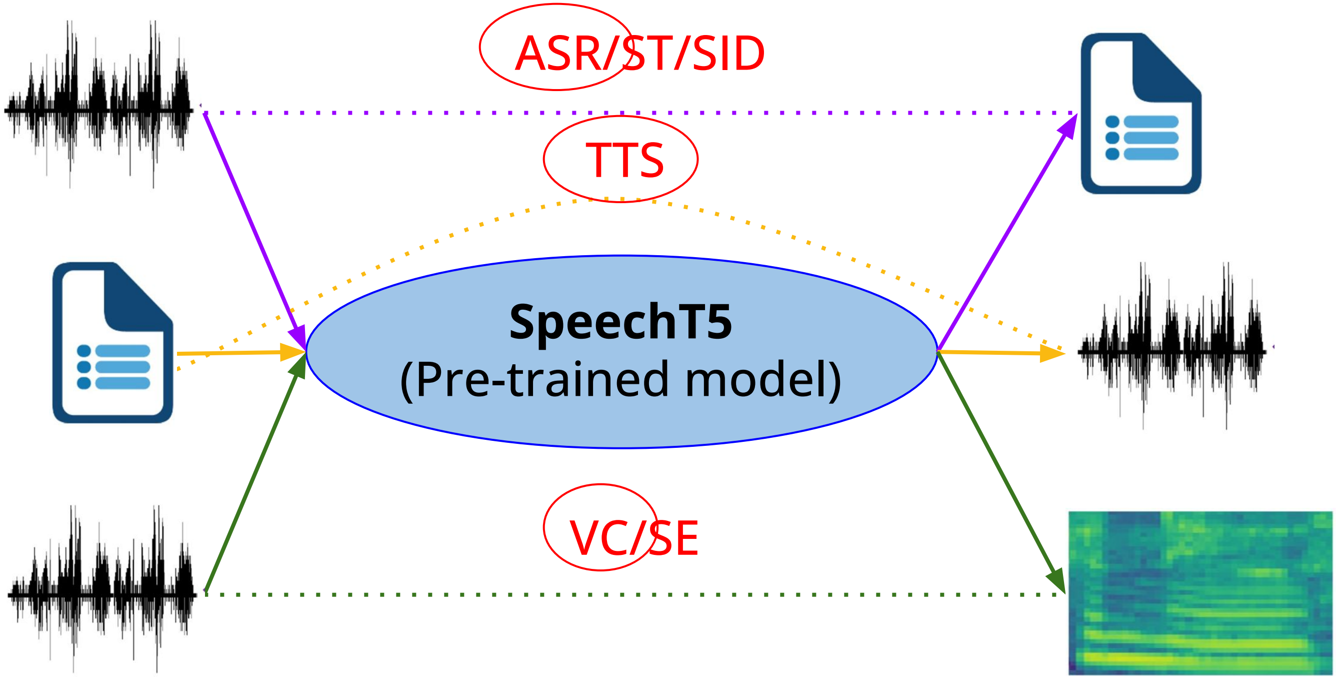
Symbolic

Text

Images taken from the original paper: https://arxiv.org/abs/2110.07205

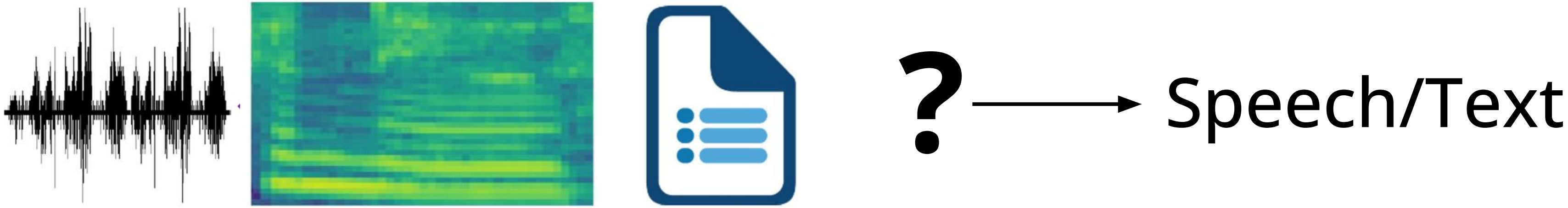
SpeechT5: Spoken Language Processing

- speech-to-text
- text-to-speech
- speech-to-speech



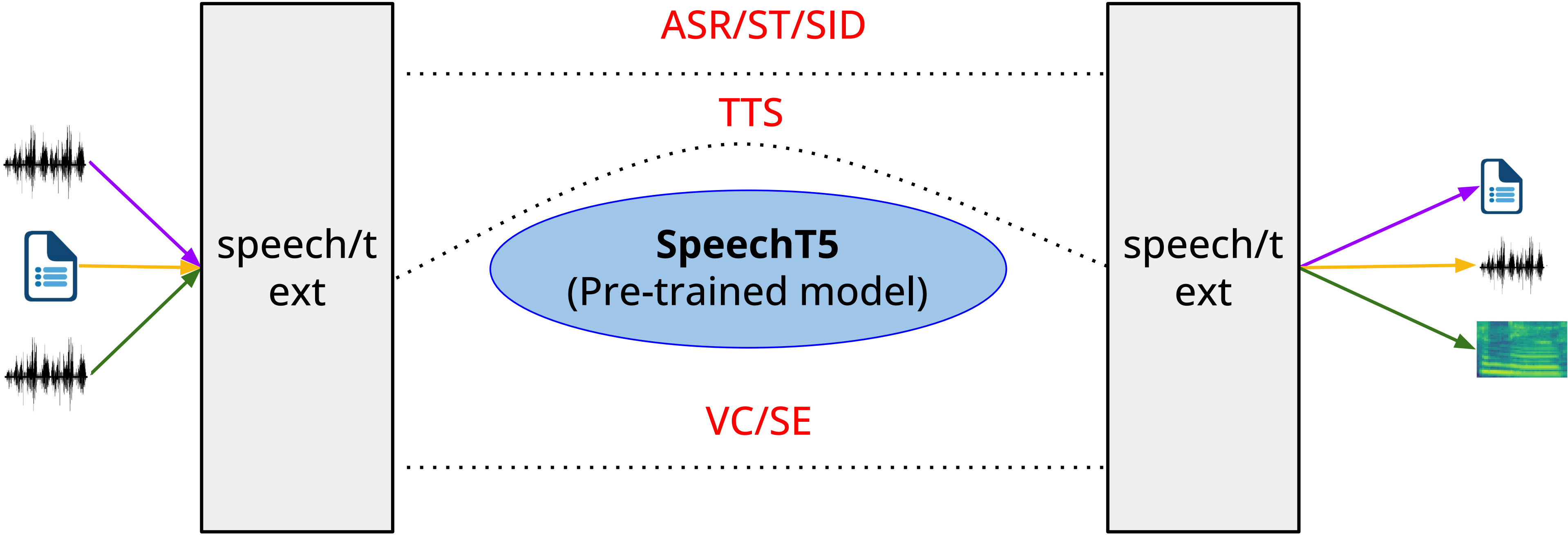
SpeechT5: Unified Modal Encode-Decoder Pre-Training

- Pre-trained models can significantly improve NLP tasks (ELMo, BERT).
- Two problems in previous speech pre-training works:
    - Learnt speech representation with only unlabeled speech data
    - Rely mainly only on a pre-trained speech encoder



? → Speech/Text

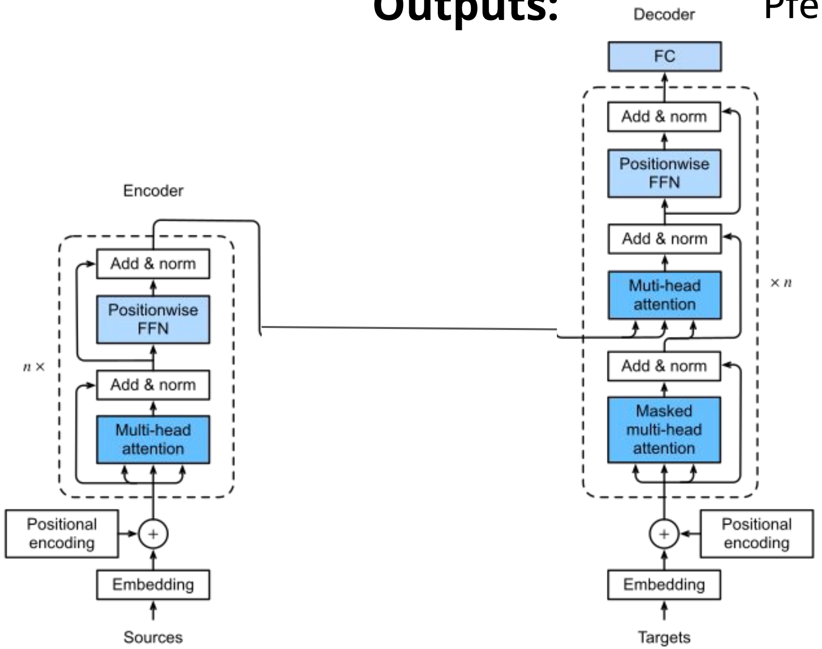
SpeechT5: Unified Modal Encode-Decoder Pre-Training



SpeechT5: Unified Modal Encode-Decoder Pre-Training

Outputs: Pfeil

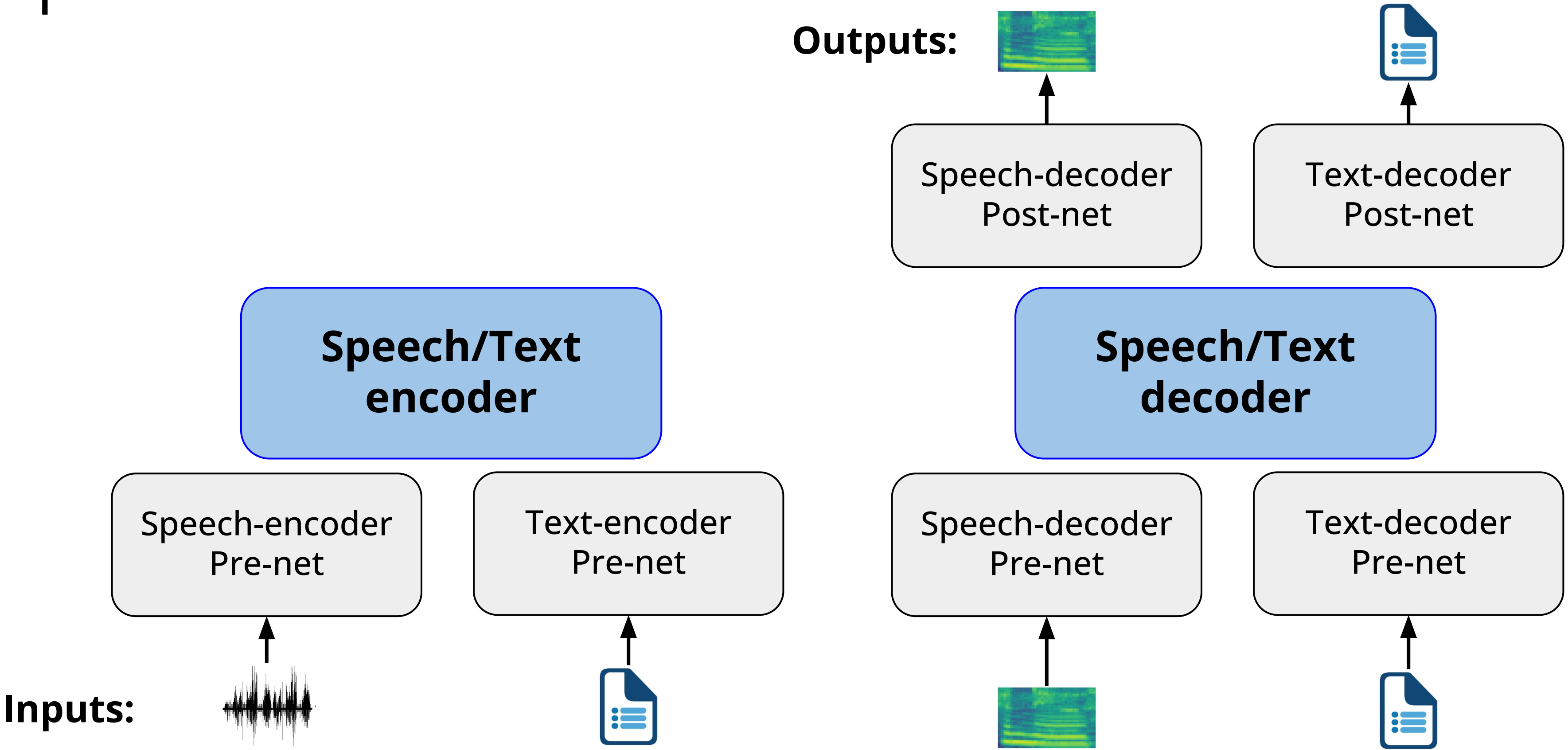
Speech/Text encoder



Speech/Text decoder

Inputs: Time flies like an arrow Die Zeigt fliegt wie ein

SpeechT5: Architecture



Outputs:

Speech-decoder Post-net

Text-decoder Post-net

Speech/Text encoder

Speech/Text decoder

Speech-encoder Pre-net

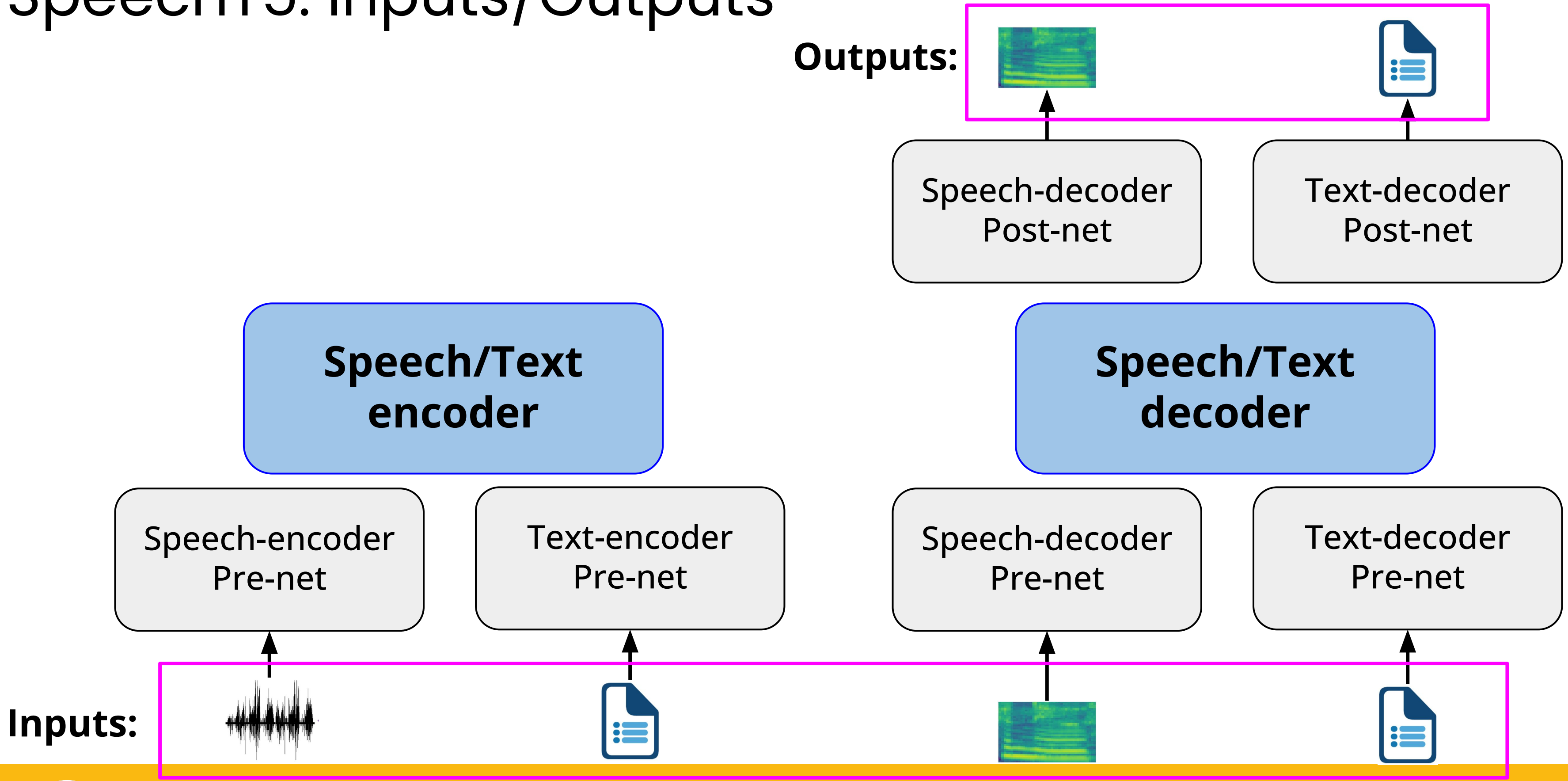
Text-encoder Pre-net

Speech-decoder Pre-net

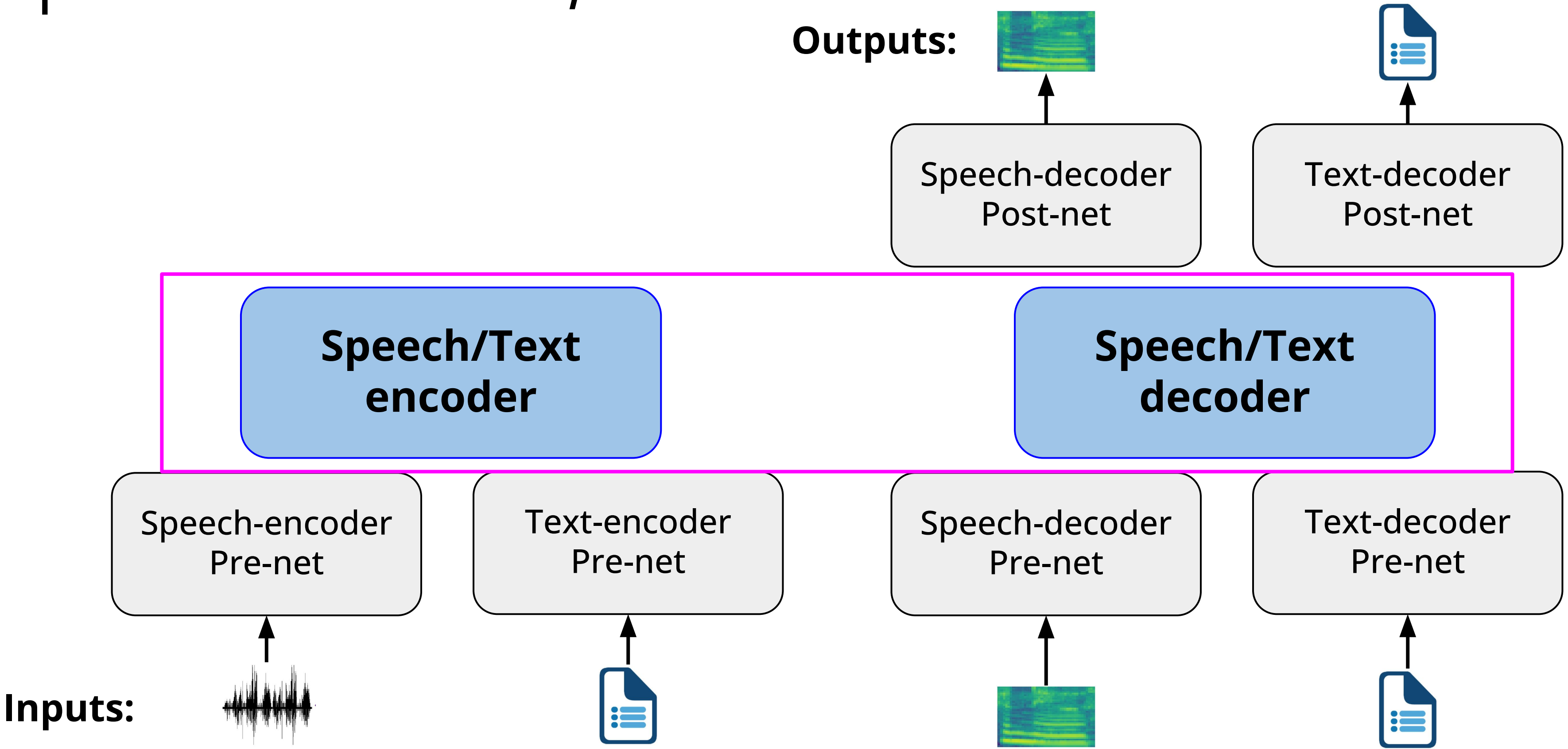
Text-decoder Pre-net

Inputs:

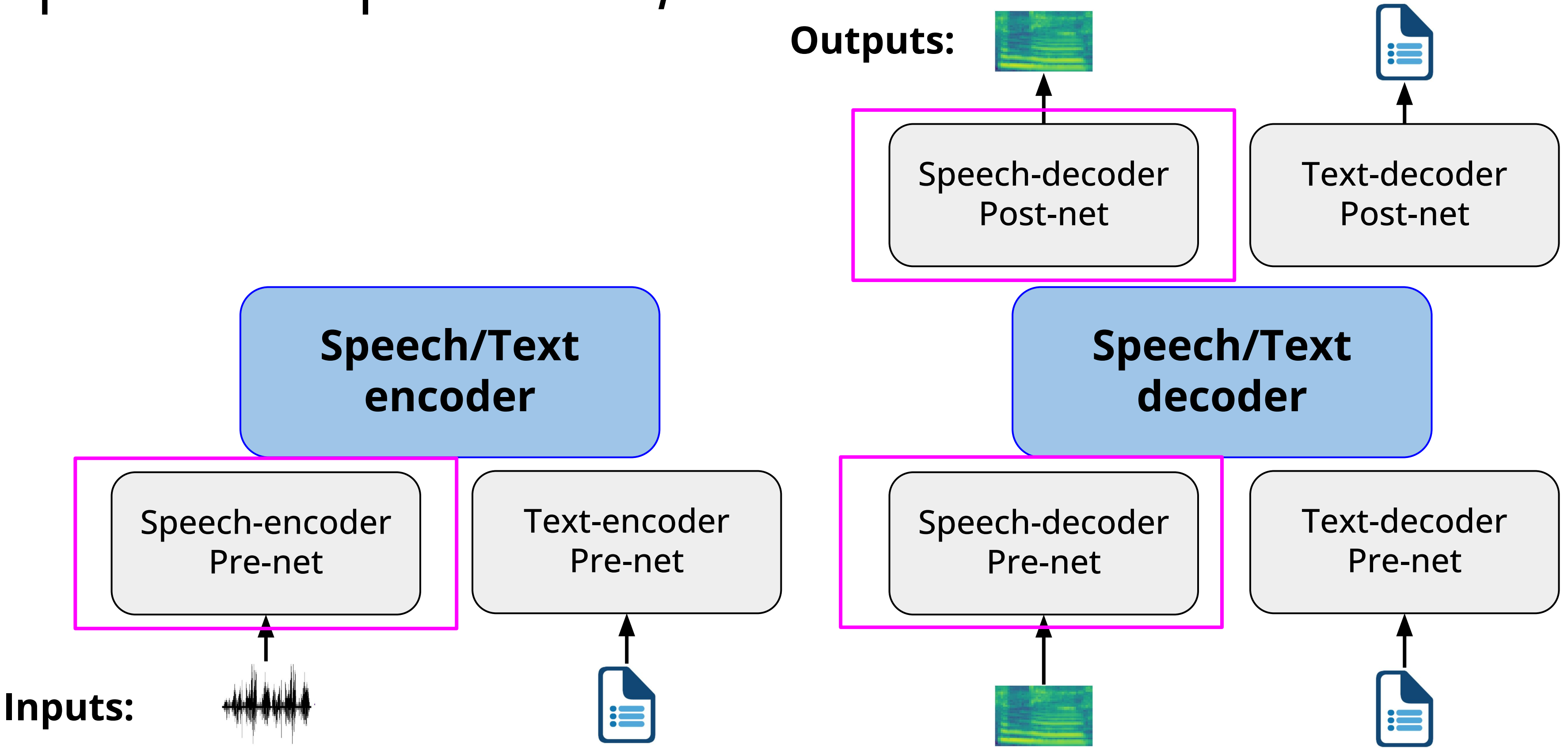
SpeechT5: Inputs/Outputs



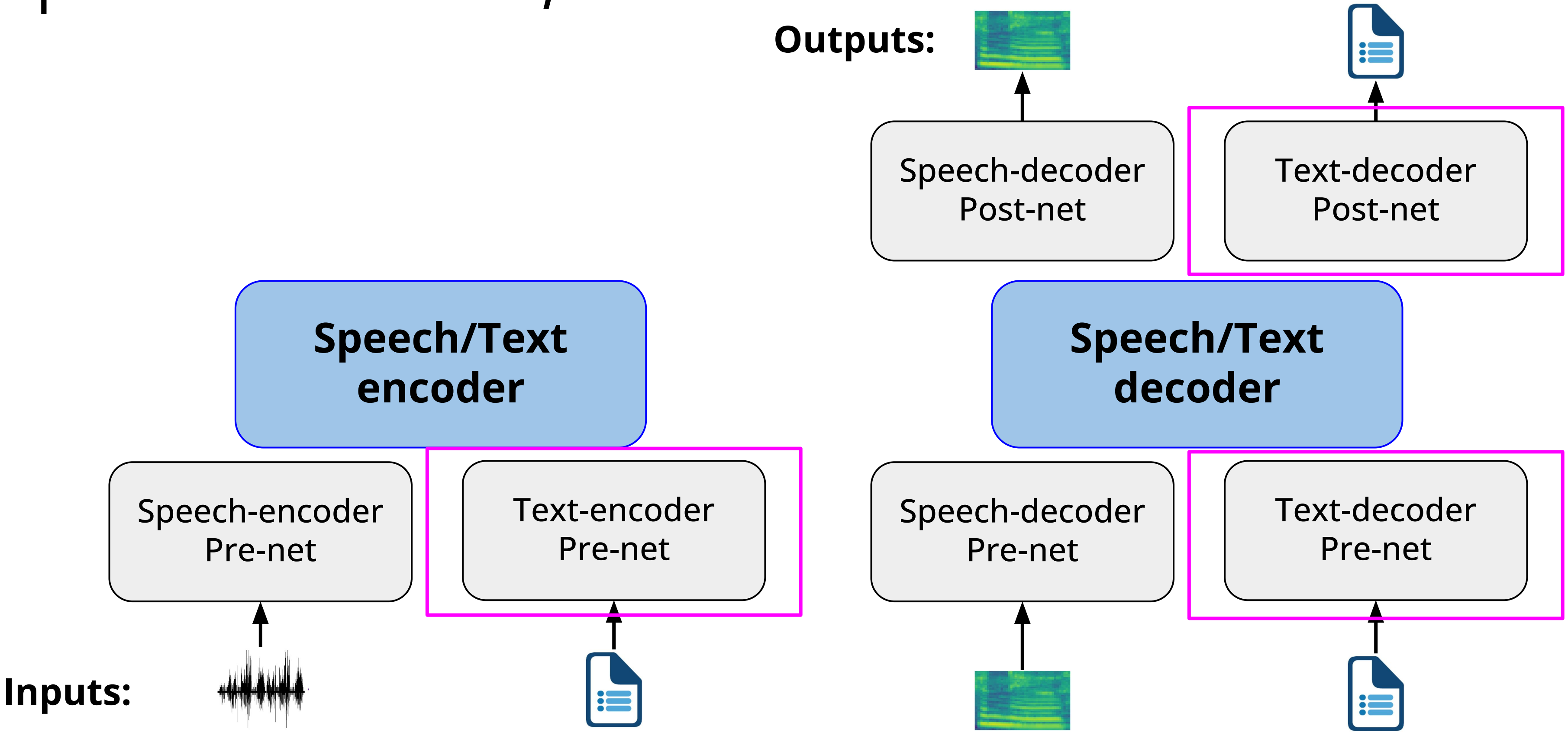
SpeechT5: Encoder/Decoder Backbone



SpeechT5: Speech Pre/Post-Net

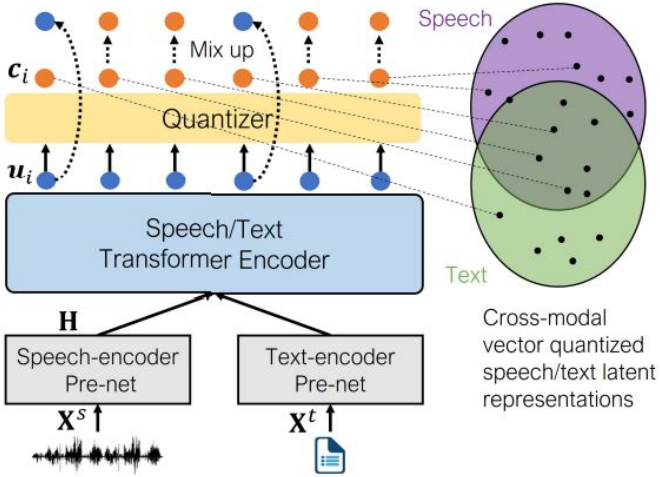


SpeechT5: Text Pre/Post-Net



SpeechT5: Pre-Training and Fine-Tuning

- Pre-training:
  - Speech pre-training
  - Text pre-training
  - Joint pre-training
- Fine-tuning

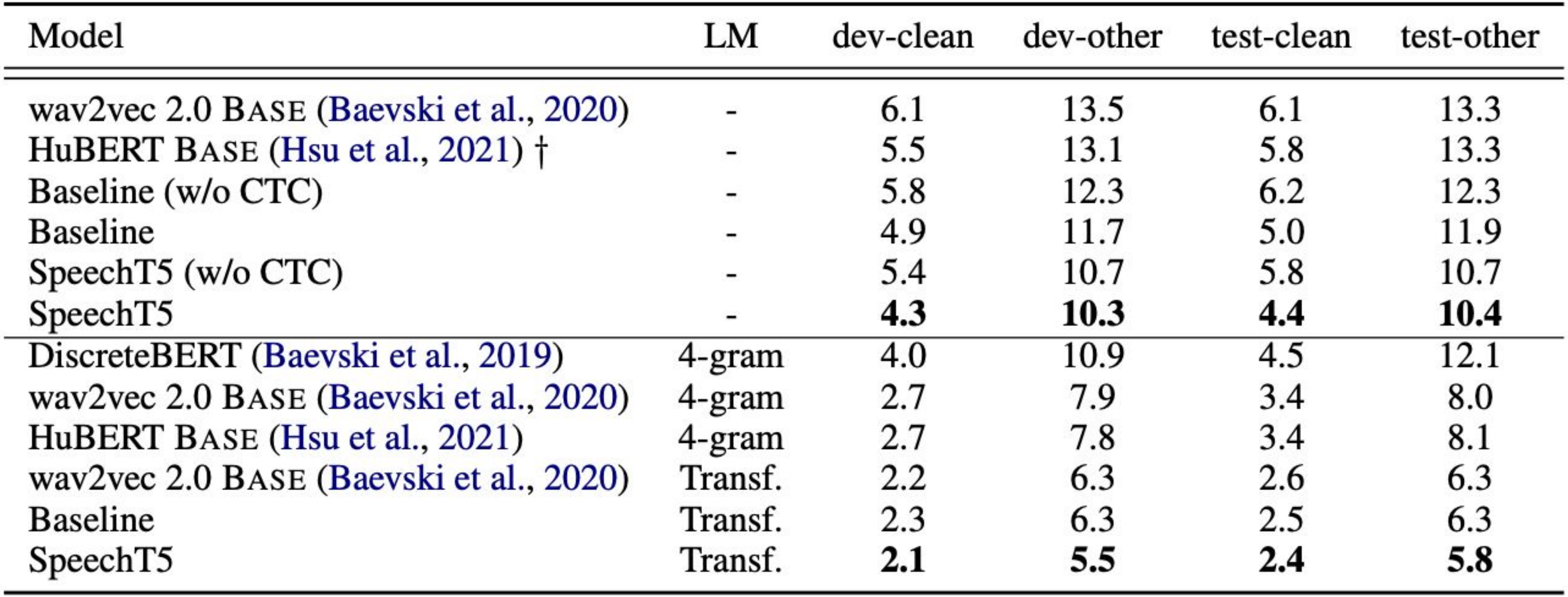

Speech
Mix up
c_i
Quantizer
u_i
Speech/Text Transformer Encoder
H
Speech-encoder Pre-net
Text-encoder Pre-net
X^s
X^t
Text
Cross-modal vector quantized speech/text latent representations

(b) The joint pre-training approach

SpeechT5: Architecture takeaways

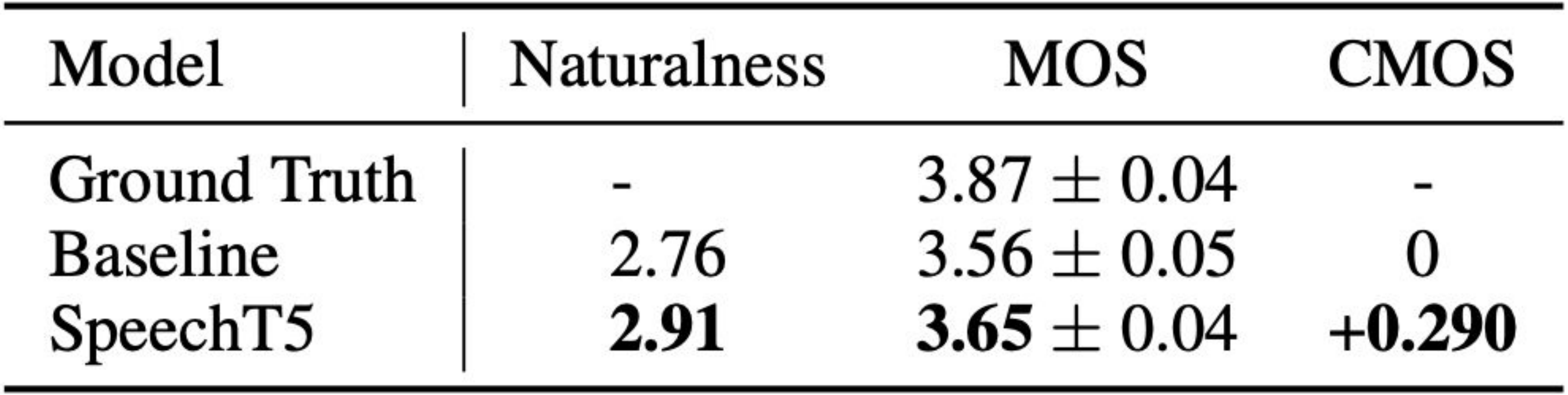
- Speech/text (unified space of hidden representations) as input/output of the model
- Encode-decoder backbone
- Six modal-specific pre/post-nets
- Tasks are fine-tuned from same initial weights, but final versions are quite different in the end

SpeechT5: Evaluation - ASR



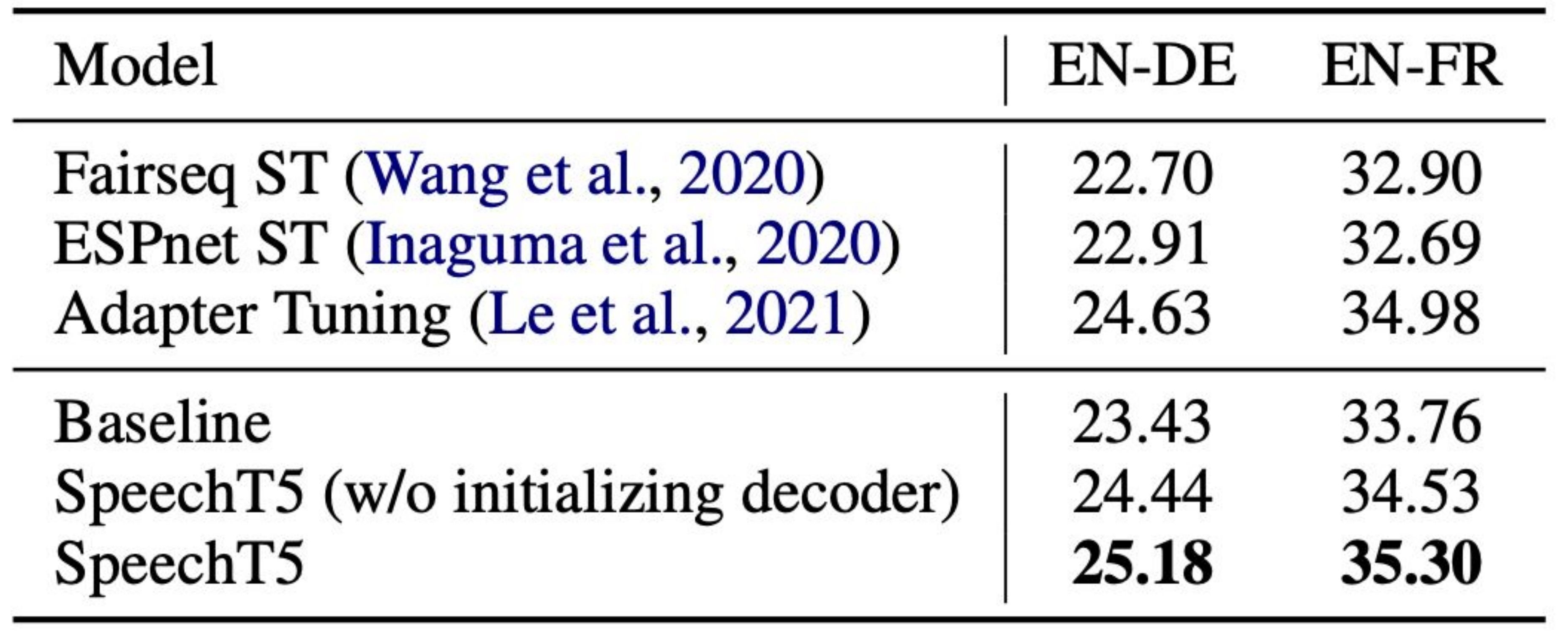
Model	LM	dev-clean	dev-other	test-clean	test-other
wav2vec 2.0 BASE (Baevski et al., 2020)	-	6.1	13.5	6.1	13.3
HuBERT BASE (Hsu et al., 2021) †	-	5.5	13.1	5.8	13.3
Baseline (w/o CTC)	-	5.8	12.3	6.2	12.3
Baseline	-	4.9	11.7	5.0	11.9
SpeechT5 (w/o CTC)	-	5.4	10.7	5.8	10.7
SpeechT5	-	**4.3**	**10.3**	**4.4**	**10.4**
DiscreteBERT (Baevski et al., 2019)	4-gram	4.0	10.9	4.5	12.1
wav2vec 2.0 BASE (Baevski et al., 2020)	4-gram	2.7	7.9	3.4	8.0
HuBERT BASE (Hsu et al., 2021)	4-gram	2.7	7.8	3.4	8.1
wav2vec 2.0 BASE (Baevski et al., 2020)	Transf.	2.2	6.3	2.6	6.3
Baseline	Transf.	2.3	6.3	2.5	6.3
SpeechT5	Transf.	**2.1**	**5.5**	**2.4**	**5.8**

SpeechT5: Evaluation - TTS



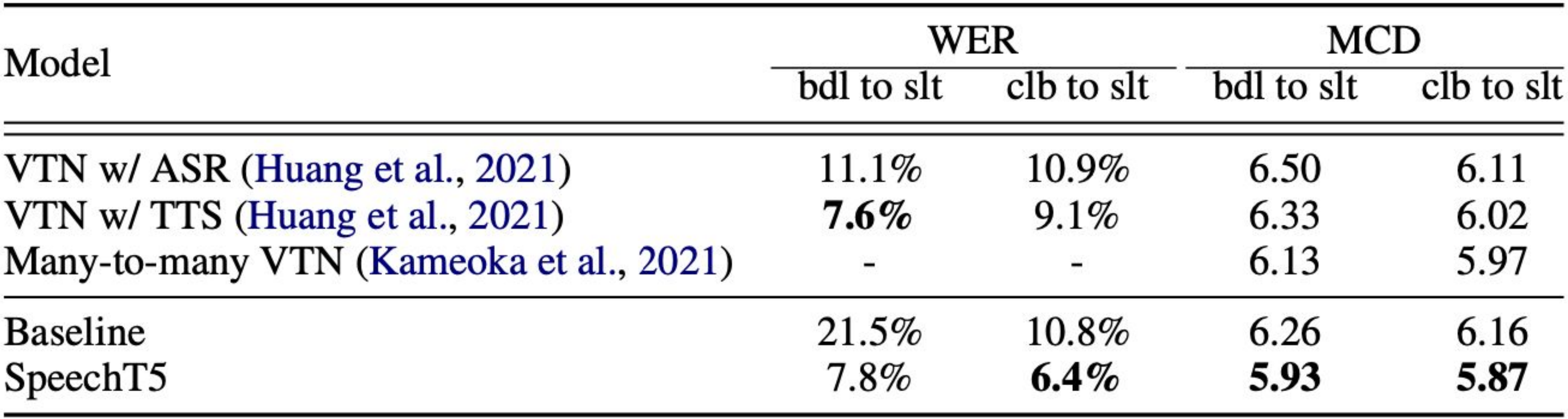
Model	Naturalness	MOS	CMOS
Ground Truth	-	3.87 ± 0.04	-
Baseline	2.76	3.56 ± 0.05	0
SpeechT5	**2.91**	**3.65** ± 0.04	**+0.290**

SpeechT5: Evaluation - ST



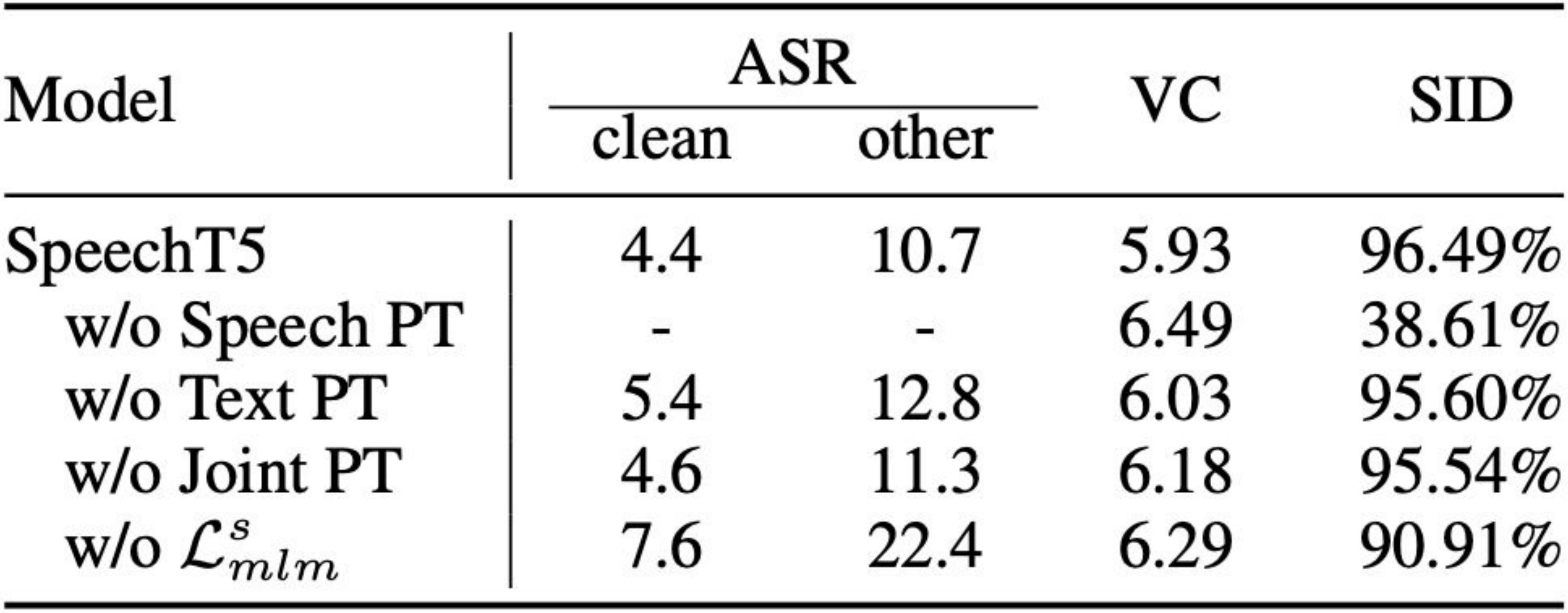
Model	EN-DE	EN-FR
Fairseq ST (Wang et al., 2020)	22.70	32.90
ESPnet ST (Inaguma et al., 2020)	22.91	32.69
Adapter Tuning (Le et al., 2021)	24.63	34.98
Baseline	23.43	33.76
SpeechT5 (w/o initializing decoder)	24.44	34.53
SpeechT5	**25.18**	**35.30**

SpeechT5: Evaluation - VC



Model	WER		MCD	
	bdl to slt	clb to slt	bdl to slt	clb to slt
VTN w/ ASR (Huang et al., 2021)	11.1%	10.9%	6.50	6.11
VTN w/ TTS (Huang et al., 2021)	**7.6%**	9.1%	6.33	6.02
Many-to-many VTN (Kameoka et al., 2021)	-	-	6.13	5.97
Baseline	21.5%	10.8%	6.26	6.16
SpeechT5	7.8%	**6.4%**	**5.93**	**5.87**

SpeechT5: Ablation study



Model	ASR		VC	SID
	clean	other		
SpeechT5	4.4	10.7	5.93	96.49%
w/o Speech PT	-	-	6.49	38.61%
w/o Text PT	5.4	12.8	6.03	95.60%
w/o Joint PT	4.6	11.3	6.18	95.54%
w/o $\mathcal{L}^{s}_{mlm}$	7.6	22.4	6.29	90.91%

SpeechT5: Conclusion

- Converting all Spoken language processing tasks into a speech/text to speech/text format works.
- Joint pre-training method utilizing cross-modal information works wonders.
- The unified encoder-decoder model works wonder on adjacent tasks like Speech Translation and Voice Conversion as well.

Next Steps

- Read the SpeechT5 release blogpost by Matthijs
- TTS demo on 🤗 Spaces - https://huggingface.co/spaces/Matthijs/speecht5-tts-demo
- VC demo on 🤗 Spaces - https://huggingface.co/spaces/Matthijs/speecht5-vc-demo
- Go! Create your own fancy demos and share on our discord channel (#ml-4-audio)

Thank You!