PAPER REVIEW-

SPEECHT5: UNIFIED-MODAL ENCODER-DECODER PRE-TRAINING FOR SPOKEN LANGUAGE PROCESSING

JUNYI AO[1,2,*], RUI WANG[3,*], LONG ZHOU[4,*] CHENGYI WANG[4], SHUO REN[4], YU WU[4], SHUJIE LIU[4], TOM KO[1], QING LI[2], YU ZHANG[1,5], ZHIHUAWEI[3], YAO QIAN[4], JINYU LI[4], FURU WEI[4]

[1]DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING, SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

[2]DEPARTMENT OF COMPUTING, THE HONG KONG POLYTECHNIC UNIVERSITY

[3]DEPARTMENT OF COMPUTER SCIENCE AND TECHNOLOGY, TONGJI UNIVERSITY

[4]MICROSOFT

[5]PENG CHENG LABORATORY

Ao, J., Wang, R., Zhou, L., Liu, S., Ren, S., Wu, Y., Ko, T., Li, Q., Zhang, Y., Wei, Z., Qian, Y., Li, J., & Wei, F. (2021). SpeechT5: Unified-Modal Encoder-Decoder Pre-training for Spoken Language Processing. *ArXiv, abs/2110.07205*.

- **PROBLEM**:
  - Pre-training techniques in NLP & self-supervised speech representation learning suffers from two problems:

(1) Most of them learn the speech representation with only unlabeled speech data but ignore the importance of textual data to spoken language tasks (e.g., automatic speech recognition) which require the modality transformation.

(2) Most of these models solely rely on a pre-trained speech encoder for various downstream tasks, leaving the decoder not pre-trained for the sequence-to-sequence generation tasks. How to design a unified encoder-decoder model that can take advantage of both unlabeled speech and text data to improve various spoken language processing tasks is not well explored.

- **OBJECTIVE:**

Propose SpeechT5, a unified modal framework for learning joint contextual representations for speech and text data via a shared encoder-decoder structure.

RELATED WORK

Approach	Merits	Demerits
BERT, wav2vec 2.0, HuBERT (Large scale pre-training models)	Have strong capability of generalization and efficient usage of large-scale data.	Gear towards single-modal learning, hence they can only be used in either text or speech modeling.
Speech2vec, VoxCeleb (speech-language pre-training work)	Attempts to improve spoken language understanding tasks	These methods only focus on an encoder with task-specific layers for different tasks and do not pre-train a decoder for generation tasks such as speech synthesis or text generation.
Han et al., 2021; Ye et al., 2021; Tang et al., 2021a; Zheng et al., 2021; Tang et al., 2021b	Series of research work to investigate joint text and speech training	But they are mainly designed for speech to text tasks.

➢ Related Models

1) **T5** (Raffel et al., 2019)- "Text-to-Text Transfer Transformer"

The core idea of the T5 model - a unified framework for a variety of text-based language problems, is to treat every text processing problem as a "text-to-text" problem.

2) **Speech Chain** (Tjandra et al., 2020), which leverages the ASR model and TTS model to build a closed-loop machine speech chain to train models on the concatenation of both labeled and unlabeled data.

3) **SpeechNet** (Chen et al., 2021b), which designs a universal modularized model to perform multiple speech processing tasks with multi-task learning.

ABOVE MODELS **vs SpeechT5**

(1) SpeechT5 is a shared cross-modal encoder-decoder framework, whose input and output are speech or text through

multiple pre/post-nets.

(2) SpeechT5 attempts to pre-train and improve the universal model with large-scale unlabeled text and speech data.

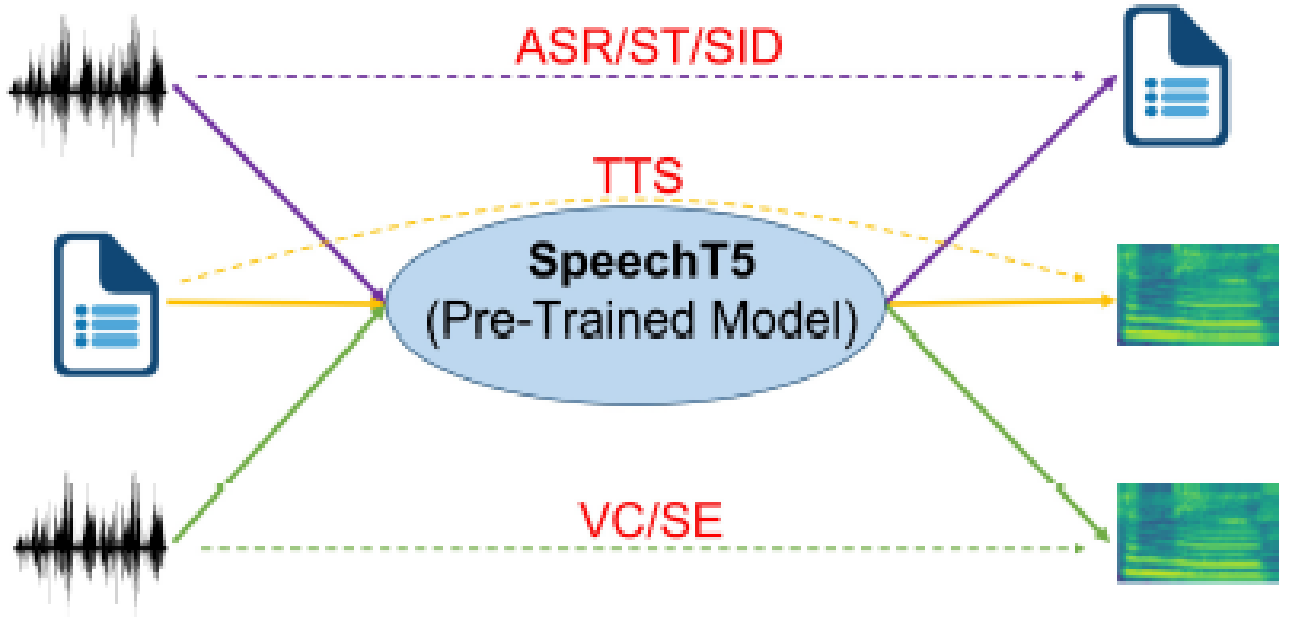
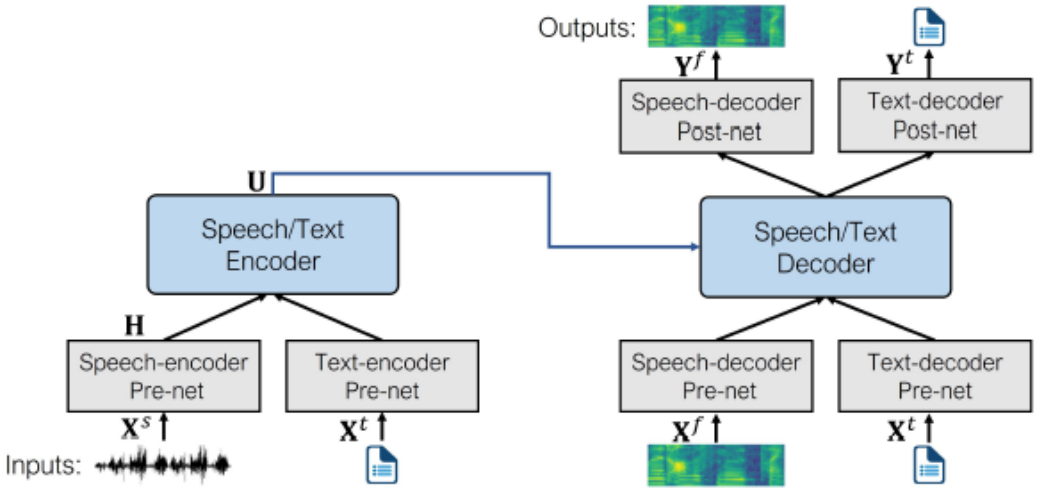


Figure 1: An illustration of the SpeechT5 framework, which treats spoken language processing tasks as a speech/text to speech/text format, including automatic speech recognition (ASR), speech translation (ST), speech identification (SID), text to speech (TTS), voice conversion (VC), and speech enhancement (SE).

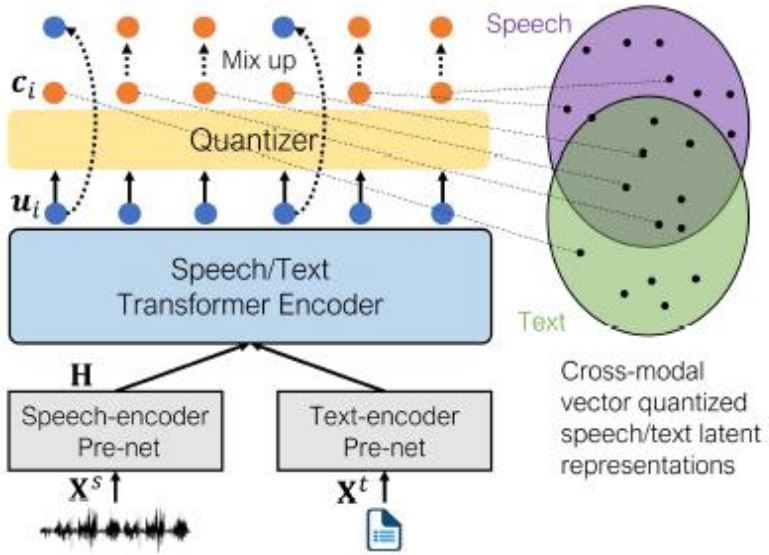
METHODS

Model Architecture

- Consists of an encoder-decoder module and six modal-specific pre/post nets.
- The pre-nets convert the input speech $X^s \in D^s$ or text $X^t \in D^t$ to a unified space of hidden representations and then feed them into the shared encoder-decoder to perform the sequence-to sequence conversion.
- Finally, the post-nets generate the output in the speech or text modality, based on the decoder output.



(a) The model architecture of SpeechT5



(b) The joint pre-training approach

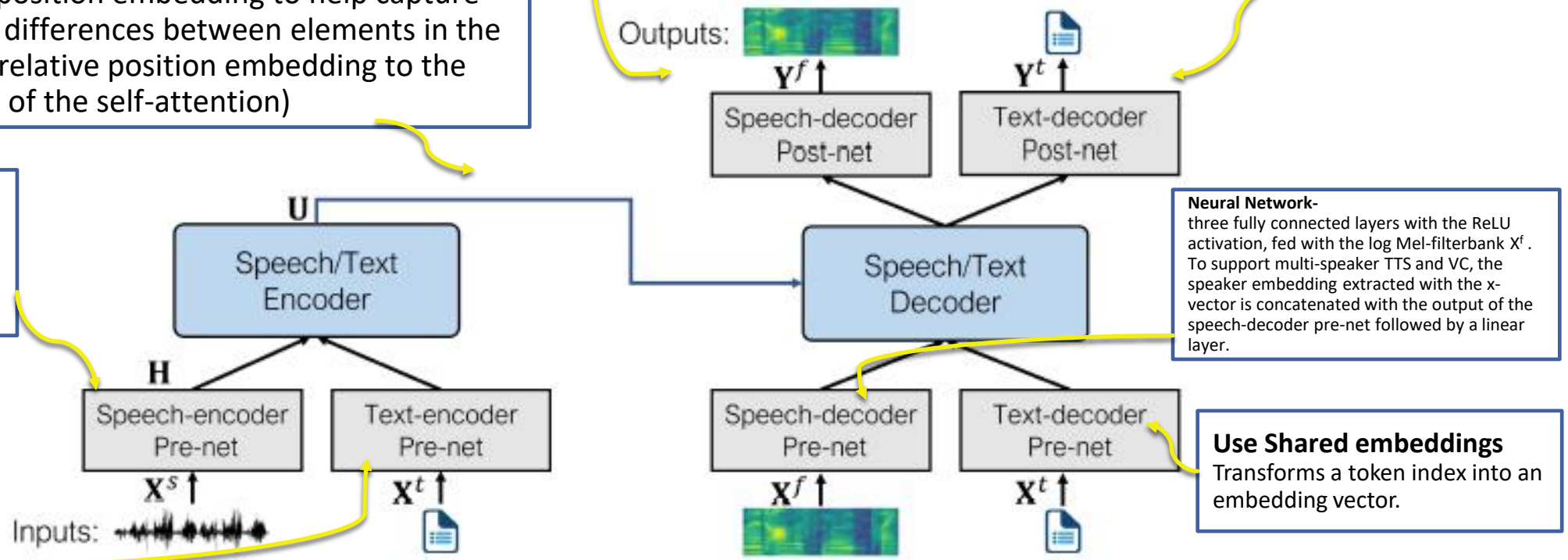
Encoder-Decoder Backbone -The Transformer encoder-decoder model (Vaswani et al., 2017) is used as the backbone network of SpeechT5.

- Employ the relative position embedding to help capture the relative position differences between elements in the input. (only add the relative position embedding to the dot-product weights of the self-attention)

1. The first module uses a linear layer fed with the decoder output to predict the log Melfilterbank, followed by five 1-dimensional convolutional layers to produce a residual to refine the predicted Yf.
2. **A linear module** to project the decoder output to a scalar for predicting the stop token.

Use Shared embeddings
Transforms the hidden state into the probability distribution of tokens, normalized by the softmax function.

Convolutional feature extractor of wav2vec 2.0
- to downsample raw waveform X^s and produce a sequence of a speech utterance $H = (h^1, \ldots, h_{Nh})$.



Neural Network-
three fully connected layers with the ReLU activation, fed with the log Mel-filterbank X^f. To support multi-speaker TTS and VC, the speaker embedding extracted with the x-vector is concatenated with the output of the speech-decoder pre-net followed by a linear layer.

Use Shared embeddings
Transforms a token index into an embedding vector.

Use Shared embeddings
Transforms a token index into an embedding vector.

Input/Output Representations - The model is fed with speech/text as the input and generates the corresponding output in the speech/text format.

- Split text into a sequence of characters $X^t = (x^t_1, \ldots, x^t_{Nt})$ as the input and output.
- For speech modality, the raw waveform $X^s = (x^s_1, \ldots, x^s_{Ns})$ is used as the input, and a sequence of the log Mel-filter bank features $X^f = (x^f_1, \ldots, x^f_{Nf})$ extracted from raw audio using librosa tool is adopted as the target output.
- A **vocoder** (Kong et al., 2020) is leveraged to generate the final waveform from the generated features.

Pre training

- Can be pre trained with large-scale collections of **unlabeled speech** and **text corpus**.

[1] Speech pre-training-SpeechT5 is trained with two types of tasks: bidirectional masked prediction and sequence-to-sequence generation.

I. **Bidirectional masked prediction-** this uses a masked language model similar to BERT. – produce masked speech utterance H ($\hat{H}$). *In Speech-encoder pre-net.*

   - *Transformer encoder* takes masked H as input & produces hidden representation U = (u_1,….,u_{Nh})
   - Calculate cross-entropy loss ($\mathcal{L}^{s}_{mlm}$)

II. **Sequence-to-sequence generation-** propose to reconstruct the original speech, given the randomly masked input as introduced in bidirectional masked prediction.

   - Following **seq2seq TTS** models (Li et al., 2019), generate predicted output Y^f , which is generated through the spee*ch-decoder pre-net, Transformer decoder, and speech-decoder post-net*, to be close to the original X^f by minimizing their L1-distance ($\mathcal{L}^{s}_{1}$) [speech-reconstruction loss]
   - Calculate the binary cross-entropy (BCE) loss ($\mathcal{L}^{s}_{bce}$) , a cross-entropy loss specific to the stop token when speech is decoded.

[2] Text Pre-Training – With unlabeled text data, reconstruct the model output Y^t to the original text X^t, using the corrupted text $\hat{H}^t$ as the input generated with a mask-based noising function.

- Following the text infilling approach in BART (Lewis et al., 2020), randomly sample 30% of text spans to mask, where the span length of text spans draws from a Poisson distribution (= 3:5), and each span is replaced with a single mask token.
- Optimizing to generate the original sequence with the maximum likelihood estimation ($\mathcal{L}_{mle}^{t}$)

[3] Joint Pre-Training-

- The **proposed joint pre-training** method can align the textual and acoustic information into a unified semantic space.
- To build a cross-modality mapping between speech and text, which is essential for tasks such as ASR and TTS, propose a ***cross-modal vector quantization method*** to learn representations capturing the modality-invariant information.
- Utilize vector quantized embeddings as a bridge to align the speech representation and text representation through a shared codebook.
- Calculate diversity loss ($\mathcal{L}_d$) L_d is used to encourage sharing more codes by maximizing the entropy of the averaged Softmax distribution.

- Final pre-training loss with unlabeled speech & text data can be formulated as:

$$\mathcal{L} = \mathcal{L}_{mlm}^{s} + \mathcal{L}_{1}^{s} + \mathcal{L}_{bce}^{s} + \mathcal{L}_{mle}^{t} + \gamma \mathcal{L}_{d}.$$

Fine tuning

- Fine-tuning the encoder-decoder backbone via the loss of the downstream task.- The goal is to measure the learning abilities of SpeechT5.
- Study the performance on a diverse set of downstream tasks such as ASR, TTS, ST, VC, SE, and SID.

- Example : ASR, the final model consists:
  - speech-encoder pre-net, encoder-decoder, text-decoder pre-net, and text-decoder post-net, which are initialized by SpeechT5 and **fine-tuned via the cross-entropy loss** on the corresponding training data.
- The baseline systems have the same architecture as SpeechT5, but the weights of the baseline encoder are initialized by the HuBERT BASE model (Hsu et al., 2021) if the input data of the downstream tasks is speech.
- It allows raw waveform as the model input and can provide a strong baseline.

EXPERIMENTS

- Pre-Training Setup
- All models are implemented in Fairseq - Facebook AI Research Sequence-to-Sequence Toolkit written in Python. https://github.com/pytorch/fairseq
- A medium-size encoder-decoder model (12 encoder blocks, 6 decoder blocks, model dimension 768, inner dimension 3072, and the number of attention heads is 12) is trained using LibriSpeech for speech pre-training and LibriLM for text pre-training.
- The above encoder setting is the same as that in wav2vec 2.0 BASE and HuBERT BASE.
- The speech-encoder pre-net contains 7 blocks of temporal convolutions, each of which is composed of 512 channels with strides (5; 2; 2; 2; 2; 2; 2) and kernel sizes (10; 3; 3; 3; 3; 2; 2). For the speech-decoder pre-net and post-net, we use the same setting as the pre-net and post-net in Shen et al. (2018) except that the number of channels of the post-net is 256.
- For textencoder/ decoder pre/post-net, a shared embedding layer with dimension 768 is used. For the vector quantization, we use two codebooks with 100 entries for the shared codebook module, resulting in a theoretical maximum of K = 104 code entries.
- For speech pre-training, used the full 960 hours of LibriSpeech audio.
- For text pre-training, used the normalized language model training text of LibriSpeech as unlabeled data, which contains 400M sentences. Optimize the model with Adam (Kingma and Ba, 2014) by warming up the learning rate for the first 8% of updates to a peak of 2×10^{-4}, which is linear decayed for the following updates. Pre-train the proposed SpeechT5 model on 32 V100 GPUs with a batch size of around 90s samples per GPU for speech and 12k tokens per GPU for text and set the update frequency to 2 for 500k steps.

EVALUATION:

Evaluation on ASR

- Fine-tune the ASR model with the LibriSpeech 100/960 hours data and train the language model(LM) with the LibriSpeech LM text data, which is used for shallow fusion. Besides the cross-entropy loss for the decoder, they have added an extra linear layer to calculate the connectionist temporal classification (CTC) loss on the top of the encoder so that they can apply the joint CTC/attention decoding, to boost the performance.
- Measure the performance of ASR by the *word error rate (WER).*

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**

Table 1: Results of ASR (speech to text) on the LibriSpeech dev and test sets when training on the 100 hours subset of LibriSpeech. † indicates that results are not reported in the corresponding paper and evaluated by ourselves.

- Without LM fusion, the baseline outperforms wav2vec 2.0 BASE and HuBERT BASE.
- SpeechT5 model achieves significant improvements on all settings compared to wav2vec 2.0 BASE, HuBERT BASE.
- when decoding with LM fusion, SpeechT5 obtains the lower WERs than wav2vec 2.0 BASE on all sets and achieves the state-of-the-art performance

Model	LM	dev-clean	dev-other	test-clean	test-other
wav2vec 2.0 BASE (Baevski et al., 2020)	-	3.2	8.9	3.4	8.5
Baseline (w/o CTC)	-	3.1	7.8	3.1	7.6
Baseline	-	2.8	7.6	2.8	7.4
SpeechT5 (w/o CTC)	-	2.8	7.6	3.1	7.3
SpeechT5	-	**2.5**	**7.4**	**2.7**	**7.1**
wav2vec 2.0 BASE (Baevski et al., 2020)	4-gram	2.0	5.9	2.6	6.1
wav2vec 2.0 BASE (Baevski et al., 2020)	Transf.	1.8	4.7	2.1	4.8
Baseline	Transf.	2.0	4.5	1.9	4.5
SpeechT5	Transf.	**1.8**	**4.3**	**1.9**	**4.4**

Table 11: WER of ASR when training on the 960 hours labeled data of LibriSpeech.

- SpeechT5 model achieves significant improvement even without LM fusion, and it performs comparable or even better than wav2vec 2.0 with LM fusion

Evaluation on TTS

- Fine-tune the pre-trained model on the 460-hours LibriTTS clean sets with the L_1 loss, BCE loss, and attention loss.
- Evaluate the *Naturalness* with the open-source NISQA-TTS (Mittag and Möller, 2020), the *Mean Option Score (MOS),* and the *Comparison Mean Option Score (CMOS)* by native speakers on the randomly selected 200 sentences with various lengths (no overlapping with training data).

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**

Table 3: Results of TTS (text to speech) on the LibriTTS.

Model	Naturalness
SpeechT5	2.79
w/o $\mathcal{L}^s_{mlm}$	2.91

Table 13: Comparisons between SpeechT5 and its variant without using $\mathcal{L}^s_{mlm}$.

Evaluation on ST

- Evaluate the ST task on the MUST-C dataset (Di Gangi et al., 2019), including English-German (EN-DE) and English-French (EN-FR) translation tasks.
- Use the default training setting of speech translation in Fairseq ST (Wang et al., 2020), and average the last 10 checkpoints and use a beam size of 5 for decoding.
- Translation results are evaluated with case-sensitive *BLEU (**bi**lingual **e**valuation **u**nderstudy)* (Papineni et al., 2002)

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**

Table 4: Results of ST (speech to text) on the MUST-C EN-DE and EN-FR.

Evaluation on VC

- Follow the many-to-many setting and utilize speech recordings of four speakers in the **CMU Arctic** (Kominek and Black, 2004), including clb, bdl, slt, and rms.
- Employ the average of *MCD (Mel-Cepstral Distortion)* and *WER* as the metrics for the VC task evaluation.
- Fine tuning with L_1 loss, BCE loss, and attention loss.

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**

Table 2: Results of VC (speech to speech) on the CMU Arctic. The bdl, clb, and slt denote three speakers.

Evaluation on SE

- SE is the task of removing background noise from a degraded speech signal and improving the intelligibility and the perceived quality of the signal.
- Used WHAM (WSJ0 Hipster Ambient Mixtures) dataset.
- Fine tuning with L_1 loss, BCE loss, and attention loss.

Model	WER
Ground Truth Speech	3.2%
Noisy Speech (Wichern et al., 2019)	76.1%
Baseline	10.9%
SpeechT5	**8.9%**

Table 5: Results of SE (speech to speech) on the WHAM!.

Evaluation on SID

- Convert SID, a multi-class classification task of classifying each utterance for its speaker identity, to a speech to text task by sequence-to-sequence model.
- Used VoxCeleb1 dataset.
- Fine tuning with cross entropy.

Model	ACC
SUPERB (Yang et al., 2021)	
wav2vec 2.0 BASE (Baevski et al., 2020)	75.18%
HuBERT BASE (Hsu et al., 2021)	81.42%
HuBERT LARGE (Hsu et al., 2021)	90.33%
SpeechNet (Chen et al., 2021b)	
Single Task	86.00%
Multi-Task with TTS	87.90%
Thin ResNet-34 (Chung et al., 2020)	89.00%
Ours	
Baseline	91.92%
SpeechT5	**96.49%**

Table 6: Results of SID (speech to text) on the VoxCeleb1. The SUPERB fine-tuning freezes the encoder.

Ablation Study

- To better understand why the proposed SpeechT5 model is effective, they investigate the influence of the pre-training methods by removing each of them independently.

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%

Table 7: Ablation study for the SpeechT5 model. Different variants of the SpeechT5 model, including the SpeechT5 model without speech pre-training (PT), text pre-training, joint pre-training method, or the bidirectional masked prediction loss, are evaluated on the ASR (test subsets with WER), VC (bdl to slt with MCD), and SID (test set with ACC) tasks.

ACHIEVEMENTS

- To the best of the knowledge, this is the first work to investigate a unified encoder-decoder framework for various spoken language processing tasks.
- Proposing the cross-modal vector quantization approach, which learns the implicit alignment between acoustic and textual representation with large-scale unlabeled speech and text data.
- Extensive experiments on spoken language processing tasks demonstrate the effectiveness and superiority of the proposed SpeechT5 model.

CONCLUSION

- SpeechT5 as a pre-trained encoder-decoder model for various spoken language processing tasks.
- The proposed unified encoder-decoder model can support generation tasks such as speech translation and voice conversion.
- Massive experiments show that SpeechT5 significantly outperforms all baselines in several spoken language processing tasks.

FURTHER WORKS

- Going to pre-train the SpeechT5 with a larger model and more unlabeled data.
- Extending the proposed SpeechT5 framework to address multilingual spoken language processing tasks.

Limitations of the study

- Unclear points on the computation of loss:
  - (a) speech reconstruction L1 loss (the reason for adding this loss component is unclear and no ablation study supports its usefulness)
  - (b) diversity loss - cross-modal objective to better align speech and text representations (it is unclear, from the paper whether or not this specific loss needs aligned speech-text data to be computed.

- Fine-tuning process
  - multi-modal fine tuning -> mSLAM

mSLAM benefits from multi-modal fine-tuning, further improving the quality of speech translation by directly leveraging text translation data during the fine-tuning process.