SpeechUT: Bridging Speech and Text with Hidden-Unit for Encoder-Decoder Based Speech-Text Pre-training

Ziqiang Zhang[1,*], **Long Zhou**[2,†], **Junyi Ao**[3,*], **Shujie Liu**[2],
Lirong Dai[1], **Jinyu Li**[2], **Furu Wei**[2]
[1]University of Science and Technology of China
[2]Microsoft
[3]The Chinese University of Hong Kong, Shenzhen

Abstract

The rapid development of single-modal pre-training has prompted researchers to pay more attention to cross-modal pre-training methods. In this paper, we propose a unified-modal speech-unit-text pre-training model, SpeechUT, to connect the representations of a speech encoder and a text decoder with a shared unit encoder. Leveraging hidden-unit as an interface to align speech and text, we can decompose the speech-to-text model into a speech-to-unit model and a unit-to-text model, which can be jointly pre-trained with unpaired speech and text data respectively. Our proposed SpeechUT is fine-tuned and evaluated on automatic speech recognition (ASR) and speech translation (ST) tasks. Experimental results show that SpeechUT gets substantial improvements over strong baselines, and achieves state-of-the-art performance on both the LibriSpeech ASR and MuST-C ST tasks. To better understand the proposed SpeechUT, detailed analyses are conducted. The code and pre-trained models are available at `https://aka.ms/SpeechUT`.

1 Introduction

Self-supervised pre-training with large-scale unlabeled data obtains remarkable progress on various downstream tasks (Devlin et al., 2019; Radford et al., 2019; Dong et al., 2019; Baevski et al., 2020; Hsu et al., 2021; Chen et al., 2021). Specifically, pre-trained models, such as BERT (Devlin et al., 2019) and GPT (Radford et al., 2019), have extensively promoted the development of natural language processing (NLP). Researchers also develop many pre-trained speech models utilizing a mass of unlabeled audio data, e.g., wav2vec (Baevski et al., 2020) and HuBERT (Hsu et al., 2021). Although text and speech are two different modalities, they have a natural relationship because they can be viewed as two kinds of expressions of language. Hence, joint pre-training of speech and text has received increasing attention from the research community in recent years (Ao et al., 2022a; Bapna et al., 2021; Tang et al., 2022).

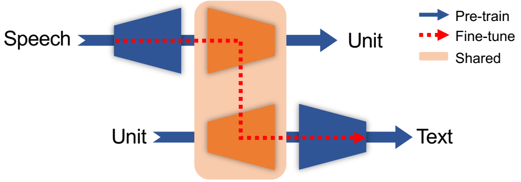


Figure 1: A high-level illustration of SpeechUT. After pre-trained with speech-to-unit and unit-to-text tasks (blue arrows), the model with a shared unit encoder enables speech-to-text tasks for fine-tuning (red arrow).

One line of speech-text joint pre-training builds a shared encoder to learn speech and text representation jointly, such as SLAM (Bapna et al., 2021), which needs a random initialization of the decoder parameter for fine-tuning an encoder-decoder model. Another line of studies, e.g., SpeechT5 (Ao et al., 2022a) and STPT (Tang et al., 2022), directly pre-trains an encoder-decoder model on speech and text corpus to boost the performance of automatic speech recognition (ASR) and speech translation (ST), leveraging unsupervised vector quantization (van den Oord et al., 2017) and supervised speech-text data to encourage the alignment of speech and text respectively. For these cross-modal speech-to-text models, a key problem is how to naturally connect the speech encoder and the text decoder.

Our preliminary observation shows that an intermediate hidden-unit representation (Hsu et al., 2021) can be regarded as the bridge between speech and text modalities, and it can provide a strong mapping relationship with both of them (see Appendix A). This inspires us to leverage hidden-unit as the semantic interface between the speech encoder and the text decoder in the encoder-decoder framework, and decompose the speech-to-text model into

[*] Work done during internship at Microsoft Research Asia.
[†] Corresponding author (lozhou@microsoft.com).

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 1663–1676
December 7-11, 2022 ©2022 Association for Computational Linguistics

a speech-to-unit model and a unit-to-text model, which can be pre-trained with unpaired speech and text data respectively, as shown in Figure 1.

In this paper, we propose a unified speech-unit-text pre-training method (**SpeechUT**), using hidden-unit representation as a bridge between the speech-encoder and the text-decoder. SpeechUT leverages three unsupervised pre-training tasks, including a speech-to-unit (S2U) task to model the mapping between speech and unit like HuBERT, masked unit modeling (MUM) task to learn better unit representation, and a unit-to-text (U2T) task to recover text from middle shared hidden-unit representation. To generate training data for S2U, MUM, and U2T, two off-line generators trained with a small amount of paired data (100h) are introduced to produce discrete unit sequences for large-scale unpaired speech and text. Experiments are conducted on two typical speech-to-text tasks, ASR and ST, followed by principal analysis to better understand the proposed method. The contributions of this paper are summarized as follows,

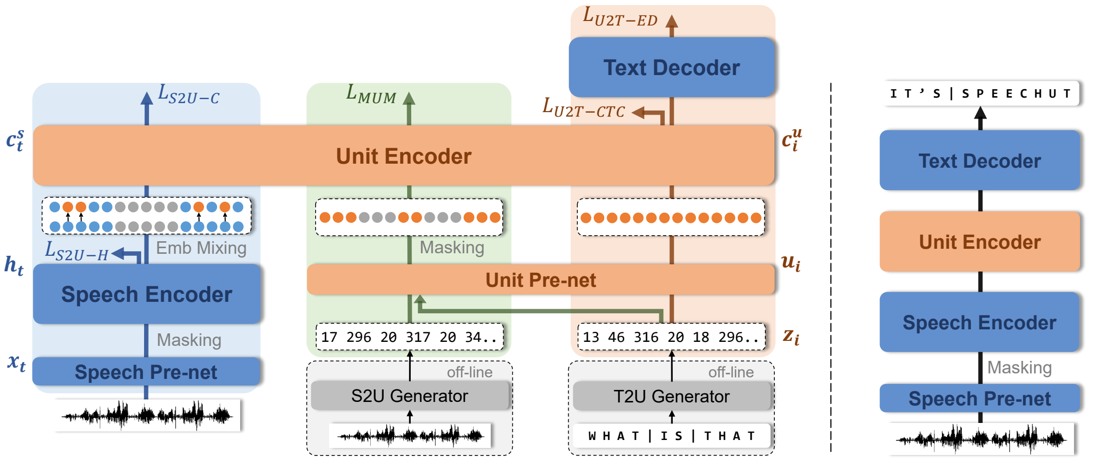
- We propose a unified speech-text pre-training method SpeechUT to bridge the speech encoder and the text decoder with hidden units.
- We decouple the speech-to-text model into speech-to-unit and unit-to-text models, to take advantage of a large amount of unpaired speech and text data for pre-training.
- Our proposed SpeechUT achieves state-of-the-art performance in downstream speech recognition and speech translation tasks.

2 Related Work

The proposed SpeechUT is built upon the Transformer encoder-decoder model (Vaswani et al., 2017) and relates to discrete speech representation learning and joint speech-text pre-training. We discuss these topics in the following.

Discrete Speech Representation Learning Discretizing continuous speech signals for speech representation learning has drawn substantial attention. Vq-wav2vec (Baevski et al., 2019) and wav2vec 2.0 (Baevski et al., 2020) attempt at discretizing speech signals into quantized units from a learnable codebook (van den Oord et al., 2017). PBERT (Wang et al., 2022a) instead uses phonemes as the discrete targets in a semi-supervised setting. SemFace (Ren et al., 2021) proposes to use language-independent vector quantized units as the semantic interface of encoder pre-training and decoder pre-training. Inspired by the masked language model in BERT (Devlin et al., 2019), HuBERT (Hsu et al., 2021) first introduces the masked speech prediction of hidden units to pre-train a universal speech model. Particularly, the hidden units can be clustered from log Mel-filterbank features or the hidden states of the previous pre-trained model. Recently, some studies explore leveraging the discrete hidden units to build speech-to-speech translation systems (Lee et al., 2021a,b), which first convert source speech into target units, then generate the target waveform from predicted units. However, our goal in this paper is to jointly pre-train speech and text with the hidden units as the intermediate bridge.

Joint Speech-Text Pre-Training Single-modal pre-trained models have achieved remarkable results in both natural language processing and spoken language processing, such as BERT (Vaswani et al., 2017), UniLM (Dong et al., 2019), XLNet (Yang et al., 2019), wav2vec 2.0 (Baevski et al., 2020), HuBERT (Hsu et al., 2021), and WavLM (Chen et al., 2021). Thanks to the rapid development of these single-modal pre-training works, researchers begin to pre-train a cross-modal model with both speech and text data (Chung et al., 2021b; Kim et al., 2021; Qian et al., 2021; Ao et al., 2022a; Bapna et al., 2021; Zhang et al., 2022b; Tang et al., 2022). One category of these works focuses on pre-training a unified encoder model for spoken language understanding (Chung et al., 2021b; Kim et al., 2021; Qian et al., 2021; Zhang et al., 2022a). In parallel to our work, SpeechLM (Zhang et al., 2022a) leverages two kinds of tokenizers to tokenize speech and text, and aims at unifying speech and text modalities into the same semantic space within one encoder model. When fine-tuning an encoder-decoder model, a randomly initialized decoder needs to be superimposed on the encoder for speech-to-text tasks (Bapna et al., 2021, 2022). Besides, Maestro (Chen et al., 2022) utilizes paired speech-text data to learn speech-text alignment through a modality-matching algorithm in RNN-T framework. Our proposed SpeechUT model is most related to encoder-decoder pre-trained models like SpeechT5 (Ao et al., 2022a) and STPT (Tang et al., 2022), in which speech and text are directly connected by a shared encoder. Unlike them, SpeechUT leverages hidden units (Hsu et al., 2021) as the bridge between the speech encoder and the



(a) Pre-training pipeline of SpeechUT

(b) Fine-tuning for ASR/ST

Figure 2: (a) The overall framework of SpeechUT, which is pre-trained with the speech-to-unit (S2U) task, the masked unit modeling (MUM) task and the unit-to-text (U2T) task jointly. The discrete units are extracted from off-line speech-to-unit (S2U) and text-to-unit (T2U) generators. (b) Fine-tuning is performed for speech-to-text tasks by cascading the speech encoder, the unit encoder, and the text decoder into an end-to-end model.

text decoder, decoupling the conventional model into two pre-trained speech-to-unit and unit-to-text models.

3 SpeechUT

Figure 2 shows the overall framework of SpeechUT, which leverages the unit representation as the bridge between speech and text. In this section, we will introduce the model architecture, pre-training, and fine-tuning methods.

3.1 Model Architecture

As illustrated in Figure 2(a), SpeechUT mainly contains a speech encoder, a unit encoder, and a text decoder. In addition, speech and unit pre-nets pre-process the input waveform and the text tokens into fixed-dimensional hidden states, respectively.

Speech/Unit Pre-nets The speech pre-net is a stack of 1-D convolutional layers with 512 channels and kernel sizes of [10,3,3,3,3,2,2]. The overall downsampling rate is 320. Given a 16K Hz speech waveform, the speech pre-net will convert it into a sequence of speech features, $\boldsymbol{X} = (x_1, x_2, \ldots, x_T)$, where T is the sequence length. The unit pre-net is a simple embedding layer which converts a sequence of unit tokens, $\boldsymbol{Z} = (z_1, z_2, \ldots, z_L)$, into latent embeddings, $\boldsymbol{U} = (u_1, u_2, \ldots, u_L)$, where L is the sequence length. The latent embeddings are then equipped with learned positional encodings.

Speech Encoder The speech encoder is a stack of Transformer layers (Vaswani et al., 2017) that transforms the local speech features $\boldsymbol{X}$ into contextualized speech hidden states, $\boldsymbol{H} = (h_1, h_2, \ldots, h_T)$.

Unit Encoder The unit encoder has the same architecture and layer numbers as the speech encoder. It is designed to align the speech hidden states $\boldsymbol{H}$ and the unit embeddings $\boldsymbol{U}$ into the same latent space. The unit encoder takes two types of input, $\boldsymbol{H}$ and $\boldsymbol{U}$, and outputs high-level contextualized representations, $\boldsymbol{C^s} = (c_1^s, c_2^s, \ldots, c_T^s)$, and $\boldsymbol{C^u} = (c_1^u, c_2^u, \ldots, c_L^u)$, respectively.

Text Decoder The text decoder is a Transformer decoder (Vaswani et al., 2017) consisting of a text embedding layer, stacked Transformer layers, and a text output layer. It is used to generate the target text sequence $\boldsymbol{Y} = (y_1, y_2, \ldots, y_{|\boldsymbol{Y}|})$ from left to right according to the output of the unit encoder.

3.2 Pre-Training Tasks

To pre-train the components of SpeechUT, we propose three pre-training tasks:

Speech-to-Unit (S2U) Task The speech-to-unit task is similar to HuBERT (Hsu et al., 2021), where the model needs to predict the units of the masked positions based on the non-mask regions in a speech sequence. Particularly, SpeechUT enables this prediction task for both the output of the speech encoder ($\boldsymbol{H}$) and the output of the unit encoder ($\boldsymbol{C^s}$),

$$\begin{aligned}\mathcal{L}_{S2U} &= \mathcal{L}_{S2U-H} + \mathcal{L}_{S2U-C} \\ &= -\sum_{t \in \mathcal{M}} \left(\log p\left(z_t|h_t\right) + \log p\left(z_t|c_t^s\right)\right)\end{aligned} \quad (1)$$

where $\mathcal{M}$ is a set of masked positions and z_t is the corresponding unit at position t. $p(.)$ computes the probabilities, i.e.,

$$p(z|h_t) = \frac{\exp(\cos(\boldsymbol{W}^s h_t, \boldsymbol{e}_z)/\tau)}{\sum_{z' \in \mathcal{Z}} \exp(\cos(\boldsymbol{W}^s h_t, \boldsymbol{e}_{z'})/\tau)} \quad (2)$$

$$p(z|c_t^s) = \text{softmax}\left(\boldsymbol{W}^u c_t^s\right) \quad (3)$$

where $\boldsymbol{W}^s$ and $\boldsymbol{W}^u$ are projection weights, τ is the temperature coefficient set to 0.1, and $\mathcal{Z}$ is the set of unit categories. cos(.) computes cosine similarity between two vectors following HuBERT (Hsu et al., 2021). Here $\boldsymbol{e}$ is a unit embedding matrix preserved by the speech encoder, it does not share parameters with the unit pre-net since HuBERT uses a lower embedding dimension.

Unit-to-Text (U2T) Task SpeechUT performs the unit-to-text task as a regular encoder-decoder based sequence-to-sequence task (Vaswani et al., 2017). The text sequence serves as the target and the corresponding generated unit sequence serves as the input. Conditioned on the output of the unit encoder, $\boldsymbol{C^u}$, the loss is formulated as

$$\mathcal{L}_{U2T-CE} = -\sum_{i=1}^{|\boldsymbol{Y}|} \log p(y_i|\boldsymbol{Y}_{<i}, \boldsymbol{C^u}) \quad (4)$$

where $\boldsymbol{Y} = (y_1, y_2, \ldots, y_{|\boldsymbol{Y}|})$ is the text sequence and $\boldsymbol{Y}_{<i}$ is its prefix from position 0 to position i. $p(.)$ is parameterized by a linear softmax layer.

Besides, to enhance the unit-to-text generation, following (Watanabe et al., 2017) we formulate a joint CTC (Graves et al., 2006) objective which directly predicts the target text sequence from the unit encoder,

$$\mathcal{L}_{U2T-CTC} = -\log p_{CTC}(\boldsymbol{Y}|\boldsymbol{C^u}) \quad (5)$$

$$\mathcal{L}_{U2T} = \mathcal{L}_{U2T-CE} + \mathcal{L}_{U2T-CTC} \quad (6)$$

where $p_{CTC}(.)$ is parameterized by a single 1-D convolutional layer with a kernel size of 2 and channel of 768, followed by a linear projection to the text vocabulary.

Masked Unit Modeling (MUM) Task Note that in S2U and U2T tasks, the unit serves as the target and the input, respectively. To enhance the unit-in, unit-out property, inspired by BERT (Vaswani et al., 2017) and HuBERT (Hsu et al., 2021), SpeechUT performs an additional masked unit modeling (MUM) task, with the training data combining all the units in S2U and U2T tasks. The unit encoder needs to predict the unit categories of the masked positions in a unit sequence, with loss formulated as

$$\mathcal{L}_{MUM} = -\sum_{i \in \mathcal{M}} \log p(z_i|c_i^u) \quad (7)$$

where $\mathcal{M}$ is a set of masked positions and the probability $p(.)$ is computed as

$$p(z|c_i^u) = \text{softmax}\left(\boldsymbol{W}^u c_i^u\right) \quad (8)$$

Multi-task Learning In the pre-training stage, SpeechUT performs multi-task pre-training with three tasks,

$$\mathcal{L} = \mathcal{L}_{S2U} + \lambda \mathcal{L}_{U2T} + \gamma \mathcal{L}_{MUM} \quad (9)$$

where λ and γ control the balance of losses. During multi-task learning, SpeechUT is expected to connect the speech encoder and the text decoder by the unit encoder. Thus the data could flow smoothly from the speech input end to the text output end even without consuming speech-text paired data.

3.3 Hidden-Unit Generation

Using these three tasks for pre-training, we need to construct three kinds of training data, the unit data, the speech-unit paired data, and the unit-text paired data. The unit data is the combination of the units in speech-unit and unit-text data. To get the latter two, we introduce two off-line unit generators, the speech-to-unit (S2U) generator and the text-to-unit (T2U) generator. The S2U generator could be any off-line unsupervised clustering model that discretizes the unlabeled speech sequences into the hidden units, e.g., the k-means model learned from HuBERT (Hsu et al., 2021). Besides, our T2U generator is a sequence-to-sequence model (Vaswani et al., 2017). As the units generated from the text should have the same style as the units generated from speech, we leverage a small amount of paired ASR data[1] to train the T2U generator.

[1] A small amount of ASR data is enough to train the T2U generator (see Appendix A).

Specifically, we generate the units from speech for a small paired dataset using the S2U generator, and then remove the repetitive units of adjacent frames to get *reduced* units (Ao et al., 2022b). The *reduced* units and the corresponding transcription form the training data for the T2U generator. With the trained T2U generator, large-scale unpaired text corpora can be converted to a large unit-text paired corpus for the U2T pre-training task.

3.4 Embedding Mixing Mechanism

Multi-task learning assumes the representations of different modalities are aligned to the same latent space (Wang et al., 2022c). However, we found that the unit encoder always performs two individual tasks for speech and unit without providing explicit alignment information between them. To better align the speech and unit representations in the unit encoder, we adopt a simple embedding mixing mechanism for S2U task, which is to mix the embeddings of two modalities in one sequence. Since each unlabeled speech sequence has the generated units at each time position, we randomly replace a portion of speech hidden states h_t in the sequence with the corresponding unit embeddings u_t, i.e.,

$$h_t' = \begin{cases} u_t & t \in \mathcal{R} - \mathcal{M} \\ h_t & \text{otherwise} \end{cases} \tag{10}$$

where $\mathcal{M}$ is the masked positions in Eqn. (1) and $\mathcal{R}$ is a set of randomly selected positions. $\mathcal{R} - \mathcal{M}$ means the embedding mixing is restricted by only operating on the non-mask positions. Different from previous work (Chen et al., 2022; Wang et al., 2022c; Fang et al., 2022), which rely on force-aligned phoneme or word labels, SpeechUT uses units for mixing, making it available for full of unlabeled data.

3.5 Fine-Tuning for ASR and ST

After pre-training, we drop the unit pre-net and stack the speech encoder, the unit encoder, and the text decoder into a complete sequence-to-sequence model, which can be fine-tuned for any speech-to-text task, such as ASR and ST. Note that all modules have been pre-trained, including the text output layer, and no new parameters are introduced in the fine-tuning stage.

4 Experiments

4.1 Dataset

We conducted experiments individually for the ASR task on English and ST tasks in three directions: English (En) to German (De), Spanish (Es), and French (Fr). For ASR pre-training, the S2U task uses unlabeled speech data from LibriSpeech (Panayotov et al., 2015) and LibriLight (Kahn et al., 2020), which contain about 960 and 60,000 hours of speech respectively. U2T task uses text from LibriSpeech LM Corpus[2], containing about 40M sentences. MUM task uses the combination of units generated from the speech and the text.

For ST pre-training, the S2U task uses unlabeled speech data from LibriSpeech and MuST-C (Di Gangi et al., 2019). The latter contains hundreds of hours of speech (see Appendix B). U2T task only optimizes $\mathcal{L}_{U2T-ED}$ and uses the paired machine translation (MT) data from WMT datasets, where the English-side text is used to generate units, and the target-side text serves as the target of the text decoder. WMT contains about 4.6M, 15M and 40M paired sentences for En-{De[3],Es[4],Fr[5]}, respectively. MUM task also uses the combination of units from two sources.

The T2U generator is trained on LibriSpeech 100 hours subset (`train-clean-100`) and used for both ASR and ST pre-training. For downstream tasks, we use LibriSpeech 100 and 960 hours training set for ASR fine-tuning and MuST-C En-De/Es/Fr train sets for ST fine-tuning. More details about the dataset and the text pre-processing can be found in Appendix B.

4.2 Model Configuration

SpeechUT The base model consists of 6 Transformer layers with relative positional attention bias (Shaw et al., 2018) for all encoder/decoders. The model dimension is 768 and the FFN dimension is 3072. The large model scales up to 12 Transformer layers for the speech/unit encoder with the model dimension of 1024 and the FFN dimension of 4096, and 12 Transformer layers for the text decoder without changing model dimensions. We use the character vocabulary for ASR tasks and 10k SentencePiece (Kudo and Richardson, 2018) for ST tasks. CTC prediction head is not applied for

[2] http://www.openslr.org/11/

[3] https://www.statmt.org/wmt16/

[4] https://www.statmt.org/wmt13/

[5] https://www.statmt.org/wmt14/

Model	Size	Pre-training Data			WER (↓) Without LM		WER (↓) With LM		
		Speech	Paired	Text	test-clean	test-other	LM	test-clean	test-other
960h hours pre-trained									
wav2vec 2.0 (Baevski et al., 2020)	Base (0.1B)	960h	-	-	6.1	13.3	4-gram	3.4	8.0
HuBERT (Hsu et al., 2021)	Base (0.1B)	960h	-	-	6.3	13.2	4-gram	3.4	8.1
WavLM (Chen et al., 2021)	Base (0.1B)	960h	-	-	5.7	12.0	4-gram	3.4	7.7
ILS-SSL (Wang et al., 2022b)	Base (0.1B)	960h	-	-	4.7	10.1	4-gram	3.0	6.9
data2vec (Baevski et al., 2022)	Base (0.1B)	960h	-	-	4.2*	9.7*	4-gram	2.8	6.8
PBERT (Wang et al., 2022a)	Base (0.15B)	960h	100h†	-	4.7	10.7	4-gram	3.1	7.3
SpeechT5 (Ao et al., 2022a)	Base (0.15B)	960h	-	40M	4.4	10.4	Transf.	2.4	5.8
Speech2C (Ao et al., 2022b)	Base (0.15B)	960h	-	-	4.3	9.0	Transf.	2.4	5.2
Wav2seq (Wu et al., 2022)	Base (0.15B)	960h	-	-	-	11.2	-	-	-
wav2vec 2.0 (Baevski et al., 2020)	**Large** (0.3B)	960h	-	-	4.7	9.0	Transf.	2.3	5.0
Baseline (Ours)	Base (0.15B)	960h	-	40M	3.8	8.0	Transf.	2.3	5.1
SpeechUT (Ours)	Base (0.15B)	960h	100h†	40M	**2.7**	**6.8**	Transf.	**2.0**	**4.5**
60kh hours pre-trained									
wav2vec 2.0 (Baevski et al., 2020)	**Large** (0.3B)	**60kh**	-	-	3.1	6.3	Transf.	2.0	4.0
HuBERT (Hsu et al., 2021)	**Large** (0.3B)	**60kh**	-	-	-	-	Transf.	2.1	3.9
WavLM (Chen et al., 2021)	**Large** (0.3B)	**94kh**	-	-	-	-	Transf.	2.1	4.0
ILS-SSL (Wang et al., 2022b)	**Large** (0.3B)	**60kh**	-	-	2.9	5.8	Transf.	2.0	4.0
STPT (Tang et al., 2022)	Base (0.16B)	**60kh**	100h	40M	3.5	7.2	-	-	-
SpeechUT (Ours)	**Large** (0.38B)	**60kh**	100h†	40M	**2.2**	**4.5**	Transf.	**1.9**	**3.6**

Table 1: ASR performance on 100-hour LibriSpeech benchmark. Speech/Paired/Text indicates the unlabeled speech data, the paired ASR data, and the unpaired text data respectively. * indicates our reproduction results, and † indicates the data is not directly used for pre-training.

ST tasks. The total parameter size is about 156M for the ASR base model, 162M for the ST base model, and 380M for ASR large model.

Baseline For comparison, we also implement a baseline with similar architecture but without using units as an intermediate modality. The baseline combines the Speech2C (Ao et al., 2022b) task with the BART (Lewis et al., 2020) task to perform multi-task pre-training. Specifically, Speech2C takes speech as input and predicts the corresponding units at the decoder. BART takes the corrupted character-level text sequence as input and predicts the complete sequence at the decoder. The baseline consists of a shared 12-layer encoder and a shared 6-layer decoder. The model size is the same as the SpeechUT base model.

Unit Generators The S2U generator is a k-means model with 500 classes learned from the released HuBERT base model (Hsu et al., 2021). The T2U generator has 6 Transformer layers for both the encoder and the decoder, the model dimension is 768 and the FFN dimension is 3072.

4.3 Training Details

All the experiments are conducted in Fairseq (Ott et al., 2019). The loss weights (λ, γ) are set to $(0.1, 0.5)$ for ASR pre-training and $(1.0, 0.5)$ for ST pre-training. Before each optimization step, the model simultaneously consumes batched data from 3 tasks and accumulates their gradients. The masking in S2U and MUM tasks follows the same configuration with HuBERT (Hsu et al., 2021), with the mask probability of 8% and the mask length of 10. The selection of $\mathcal{R}$ also follows the masking strategy, but with the probability of 4% and the window length of 5.

For ASR fine-tuning, we keep the pre-trained CTC prediction head and tune a CTC/Attention multi-task model (Watanabe et al., 2017) with the CTC weight of 0.5. For ST fine-tuning, only encoder-decoder loss is optimized. More details about pre-training and fine-tuning can be found in Appendix C.

4.4 Evaluation on Speech Recognition

We evaluate the performance by the word error rate (WER) computed on LibriSpeech `test-clean` and `test-other` sets. We also leverage an external Transformer language model (Transf. LM) for shallow fusion (Gulcehre et al., 2015). The LM has a similar size to that used in the previous works and is trained on LibriSpeech LM Corpus (see Appendix D). The results are summarized in Table 1, compared with several previous self-supervised approaches, including encoder-based models like wav2vec 2.0 (Baevski et al., 2020), HuBERT (Hsu et al., 2021), data2vec (Baevski et al., 2022), and

Models	Sizes	Pre-training Data			Fine-tuning BLEU (↑)		
		Speech (h)	ASR (h)	MT (#utt)	En-De	En-Es	En-Fr
FAT-ST (Zheng et al., 2021)	-	3.7k	1.4∼1.5k	1.9∼2.0M	25.5	30.8	-
SATE (Xu et al., 2021a)	-	-	1.4k	18M	28.1	-	-
STEMM (Fang et al., 2022)	-	960	408∼504	4.6∼40M	28.7	31.0	37.4
ConST (Ye et al., 2022)	0.15B	960	408∼504	4.6∼40M	28.3	32.0	38.3
STPT (Tang et al., 2022)	0.16B	60k	408∼504	4.6∼40M	29.2[6]	33.1	39.7
SpeechUT (Ours)	0.16B	1.4∼1.5k	100†	4.6∼40M	**30.1**	**33.6**	**41.4**

Table 2: ST performance on MuST-C dataset. Speech/ASR/MT indicates auxiliary unlabeled speech data, ASR data, and MT data. † indicates the data is not directly used for pre-training.

PBERT (Wang et al., 2022a), and encoder-decoder models like SpeechT5 (Ao et al., 2022a), Speech2C (Ao et al., 2022b), and STPT (Tang et al., 2022).

Table 1 shows that SpeechUT outperforms all the encoder-based models by a large margin. Our base model even behaves better than the large model of wav2vec 2.0 with 960 hours of pre-training data. SpeechUT also outperforms all the previous encoder-decoder speech-text pre-trained models, including SpeechT5, STPT, and our baseline, achieving a new state-of-the-art performance on the `train-clean-100` set. Moreover, the SpeechUT Large with LM gets the WER of 1.9 and 3.6 on `test-clean` and `test-other` sets. Due to the space limitation, the results using 960 hours of training data are given in Appendix E.

4.5 Evaluation on Speech Translation

We evaluate the proposed SpeechUT on En-{De, Es, Fr} language pairs. The results are shown in Table 2, with a comparison to recent state-of-the-art approaches, such as ConST (Ye et al., 2022) and STPT (Tang et al., 2022). For convenience, we reuse the off-line T2U generator trained on LibriSpeech, which is inevitably related to external 100-hour ASR data. But we do not use any ASR labels of MuST-C as all the previous works do, which is much more than 100 hours. As shown in Table 2, our SpeechUT achieves the performance of 30.1, 33.6, and 41.4 BLEU scores on En-De, En-Es, and En-Fr, respectively, demonstrating the superiority of SpeechUT over previous works. Specifically, SpeechUT outperforms the previous state-of-the-art methods by at most +1.7 BLEU (En-Fr) with significantly less pre-training data.

5 Analysis & Discussion

5.1 Ablation Study

$\mathcal{L}_{S2U}$	$\mathcal{L}_{U2T}$	$\mathcal{L}_{MUM}$	Mix	dev-c	dev-o
✓	✓	✓	✓	2.5	7.0
✓	$w\backslash o$ CTC	✓	✓	2.6	7.2
✓	$w\backslash o$ CTC	–	✓	2.7	7.3
✓	$w\backslash o$ CTC	–	–	3.0	7.9

Table 3: Ablation study. The performance is evaluated by WER on `dev-clean` and `dev-other` set after fine-tuning on `train-clean-100` set.

To better understand the effect of each component of SpeechUT, we pre-train different models in the absence of different tasks as well as the embedding mixing mechanism. Specifically, these models are pre-trained on 960 hours of speech and fine-tuned on `train-clean-100`. The results are listed in Table 3. First, the embedding mixing mechanism has the biggest impact, as the absence leads to the biggest degeneration of WER, which demonstrates its importance and effectiveness. Second, it can be noticed that the CTC loss, as a part of the U2T task, has a minor influence (0.1~0.2 WER) on the fine-tuning performance. Finally, while the MUM loss has the minimum effect, we speculate that the U2T task has already modeled the unit well.

Model	Size	Dev		Test	
		clean	other	clean	other
Self-training (Xu et al., 2021b)	300M	2.2	4.6	2.4	5.0
Semi-supervised pre-training	156M	2.4	4.9	2.5	5.1
SpeechUT	156M	**1.6**	4.5	2.0	4.5
SpeechUT + *Self-training*	156M	1.7	**4.0**	**1.9**	**4.2**

Table 4: ASR performance (WER) using LibriSpeech 100-hour supervised and 860-hour unsupervised speech data. LM is used for decoding.

5.2 Effect of Paired Data Usage

SpeechUT employs a small amount of paired ASR data to train the T2U generator. Here, we analyze and verify our method on using the ASR

[6] En-De result is from their released code.

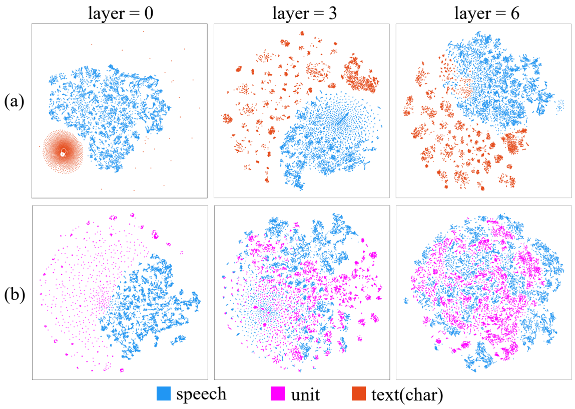


Figure 3: 2-D illustration of the token-level representations of the unit encoder. (a) Semi-supervised pre-trained model; (b) SpeechUT.

data compared with the other two methods, including 1) self-training (Xu et al., 2021b) and 2) semi-supervised pre-training which combines supervised and unsupervised learning in a multi-task pre-training process like mSLAM (Bapna et al., 2022). Since the ASR result in this setting is not reported in mSLAM, we implemented the semi-supervised pre-training based on encoder-decoder model[7]. Experimental results in Table 4 show our model behaves better than the self-training and semi-supervised pre-training. Moreover, our method has the following advantages: (1) instead of using pseudo text in self-training, SpeechUT uses real text data for the decoder pre-training; (2) simply semi-supervised pre-training only learns speech-text alignment within a small amount of paired data, while SpeechUT could align large-scale unpaired speech and text with units as a bridge; (3) SpeechUT is also complementary to self-training, achieving further performance improvement, as shown in Table 4.

5.3 Is the Encoder Getting Better?

The U2T and MUM tasks can pre-train the unit encoder with the generated unit and unit-text data. Here, we attempt at evaluating the effect of U2T and MUM losses on the encoder and judging whether they can help the encoder learn better. We first fine-tune a CTC model based on SpeechUT encoder, which obtains the 3.8 and 9.7 WER on `dev-clean` and `test-other` sets as shown in Table 5. Second, we pre-train an encoder without the text decoder and U2T task, and pre-train another encoder model by further removing the MUM task, whose results are summarized in the last two lines of Table 5. The evaluation demonstrates that our joint speech-unit-text pre-training method can still boost the performance of the encoder-only model, which means the encoder itself also learns better with U2T and MUM tasks.

Pre-trained model	dev-clean	dev-other
SpeechUT ($w\backslash o$ decoder)	3.8	9.7
- $w\backslash o$ $\mathcal{L}_{U2T}$	4.3	10.3
- $w\backslash o$ $\mathcal{L}_{U2T}$, $\mathcal{L}_{MUM}$	4.5	10.7

Table 5: ASR performance (WER) of encoder-only CTC models. SpeechUT ($w\backslash o$ decoder) is fine-tuned by discarding the pre-trained decoder, other models are pre-trained & fine-tuned without decoders.

Total	Vowels	Consonants	Silence
85.4%	79.6%	85.5%	96.7%

Table 6: Proportion where the paired speech and unit representations agree to the same phonemes.

5.4 Are the Speech and the Unit Aligned?

SpeechUT aims to align the representations of speech and unit using the unit encoder as a bridge, so that information can flow smoothly from the speech end to the text end. To verify this, we first demonstrate the alignment by validating the data distribution, i.e., the speech representation and the unit representation should follow the same distribution if they are aligned. Figure 3 plots the token-level representations of different layers of the unit encoder (layer=0 indices the inputs). The data are sampled from unpaired speech and unit sequences from LibriSpeech `dev-clean` set. T-SNE (Van der Maaten and Hinton, 2008) is performed to reduce the dimension to 2D. Figure 3(a) shows that the representations are divided into two distinct regions for speech and text respectively in the semi-supervised pre-trained model, which means the model processes the two kinds of inputs independently, leading to no alignment between speech and text. While, SpeechUT shows another behavior as shown in Figure 3(b), where the hidden states of the two modalities are mapped to the same distribution as the layer increases.

We further validate the alignment by a linear phoneme classifier. Specifically, we train the linear

[7] Specifically, it combines the speech-to-unit task (Eqn. (1)), the text-to-text BART (Lewis et al., 2020) task and the supervised CTC/Attention ASR task jointly for pre-training.

phoneme classifier using fixed speech representations extracted from the 6-th layer of the unit encoder paired with frame-level phoneme labels[8] on `train-clean-100` set. The classifier is then tested by the unit inputs on `dev-clean` set to see whether they predict the same phonemes with their paired speech inputs. Table 6 shows that SpeechUT is able to align the most portion of units (about 85%) with speech, where an interesting phenomenon shows that the alignment varies distinctly with respect to different kinds of phonemes.

6 Conclusion

In this paper, we propose SpeechUT, a unified-modal speech-unit-text pre-training model, which bridges the modality gap between speech and text representation with hidden units. By pre-training with the speech-to-unit task, masked unit modeling task, and unit-to-text task, SpeechUT significantly outperforms strong baselines as well as previous works and achieves state-of-the-art performance on downstream speech recognition and speech translation tasks. In the future, we are interested in removing the dependence on a small amount of paired ASR data before pre-training, and extending SpeechUT to a multilingual model.

Limitations

While the proposed SpeechUT model leverages hidden-unit representation as the bridge between speech and text, and obtains significant improvement over previous works, it still has some limitations: (1) the current method is a semi-supervised pre-training method, where the T2U generator needs paired ASR data to train, and takes external time to generate the units from the text; (2) the proposed SpeechUT only supports speech-to-text tasks, and it would be nice to able to help text-to-speech and speech-to-speech tasks; (3) we have to pre-train an independent model for each translation pair in the current method, which is time-consuming and resource-consuming; (4) the effectiveness of applying SpeechUT to other speech domains (e.g. child speech, accented speech) needs to be further investigated.

Ethics Statement

This work presents SpeechUT, a pre-trained model for speech recognition and speech translation. We evaluate our methods on standard benchmarks of the research community. The datasets used in this study contain LibriSpeech, LibriLight, LibriSpeech LM Corpus, and MuST-C. They are all public datasets and are widely used in the research community.

Acknowledgements

This work is supported by the “Strategic Priority Research Program” of the Chinese Academy of Sciences, Grant No. XDC08010200.

References

Junyi Ao, Rui Wang, Long Zhou, Chengyi Wang, Shuo Ren, Yu Wu, Shujie Liu, Tom Ko, Qing Li, Yu Zhang, Zhihua Wei, Yao Qian, Jinyu Li, and Furu Wei. 2022a. SpeechT5: Unified-modal encoder-decoder pre-training for spoken language processing. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5723–5738, Dublin, Ireland. Association for Computational Linguistics.

Junyi Ao, Ziqiang Zhang, Long Zhou, Shujie Liu, Haizhou Li, Tom Ko, Lirong Dai, Jinyu Li, Yao Qian, and Furu Wei. 2022b. Pre-training transformer decoder for end-to-end asr model with unpaired speech data. *arXiv preprint arXiv:2203.17113*.

Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. 2022. Data2vec: A general framework for self-supervised learning in speech, vision and language. *arXiv preprint arXiv:2202.03555*.

Alexei Baevski, Steffen Schneider, and Michael Auli. 2019. vq-wav2vec: Self-supervised learning of discrete speech representations. *arXiv preprint arXiv:1910.05453*.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Proceedings of the 34th Conference on Neural Information Processing Systems*, volume 33, pages 12449–12460.

Ankur Bapna, Colin Cherry, Yu Zhang, Ye Jia, Melvin Johnson, Yong Cheng, Simran Khanuja, Jason Riesa, and Alexis Conneau. 2022. mslam: Massively multilingual joint pre-training for speech and text. *arXiv preprint arXiv:2202.01374*.

Ankur Bapna, Yu-an Chung, Nan Wu, Anmol Gulati, Ye Jia, Jonathan H Clark, Melvin Johnson, Jason Riesa, Alexis Conneau, and Yu Zhang. 2021. Slam:

[8] `http://www.kaldi-asr.org/downloads/build/6/trunk/egs/librispeech/`. The phoneme inventory size is 42, including 15 vowels, 24 consonants, and 3 silence phones.

A unified encoder for speech and language modeling via speech-text joint pre-training. *arXiv preprint arXiv:2110.10329*.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, and Furu Wei. 2021. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *arXiv preprint arXiv:2110.13900*.

Zhehuai Chen, Yu Zhang, Andrew Rosenberg, Bhuvana Ramabhadran, Pedro Moreno, Ankur Bapna, and Heiga Zen. 2022. Maestro: Matched speech text representations through modality matching. *arXiv preprint arXiv:2204.03409*.

Yu-An Chung, Yu Zhang, Wei Han, Chung-Cheng Chiu, James Qin, Ruoming Pang, and Yonghui Wu. 2021a. w2v-bert: Combining contrastive learning and masked language modeling for self-supervised speech pre-training. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 244–250.

Yu-An Chung, Chenguang Zhu, and Michael Zeng. 2021b. SPLAT: Speech-language joint pre-training for spoken language understanding. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1897–1907.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Mattia A. Di Gangi, Roldano Cattoni, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2019. MuST-C: a Multilingual Speech Translation Corpus. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2012–2017.

Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2019. Unified language model pre-training for natural language understanding and generation. In *Proceedings of the 33rd Conference on Neural Information Processing Systems*, volume 32, pages 13063–13075.

Qingkai Fang, Rong Ye, Lei Li, Yang Feng, and Mingxuan Wang. 2022. STEMM: Self-learning with speech-text manifold mixup for speech translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7050–7062, Dublin, Ireland. Association for Computational Linguistics.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. ICML '06, page 369–376, New York, NY, USA. Association for Computing Machinery.

Caglar Gulcehre, Orhan Firat, Kelvin Xu, Kyunghyun Cho, Loic Barrault, Huei-Chi Lin, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2015. On using monolingual corpora in neural machine translation. *arXiv preprint arXiv:1503.03535*.

Takaaki Hori, Shinji Watanabe, and John Hershey. 2017. Joint CTC/attention decoding for end-to-end speech recognition. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 518–529, Vancouver, Canada. Association for Computational Linguistics.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460.

Jacob Kahn, Morgane Rivière, Weiyi Zheng, Evgeny Kharitonov, Qiantong Xu, Pierre-Emmanuel Mazaré, Julien Karadayi, Vitaliy Liptchinsky, Ronan Collobert, Christian Fuegen, et al. 2020. Libri-light: A benchmark for asr with limited or no supervision. In *Proceedings of the 2020 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 7669–7673. IEEE.

Minjeong Kim, Gyuwan Kim, Sang-Woo Lee, and Jung-Woo Ha. 2021. St-bert: Cross-modal language model pre-training for end-to-end spoken language understanding. In *Proceedings of the 2021 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 7478–7482.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Ann Lee, Peng-Jen Chen, Changhan Wang, Jiatao Gu, Xutai Ma, Adam Polyak, Yossi Adi, Qing He, Yun Tang, Juan Pino, et al. 2021a. Direct speech-to-speech translation with discrete units. *arXiv preprint arXiv:2107.05604*.

Ann Lee, Hongyu Gong, Paul-Ambroise Duquenne, Holger Schwenk, Peng-Jen Chen, Changhan Wang, Sravya Popuri, Juan Pino, Jiatao Gu, and Wei-Ning Hsu. 2021b. Textless speech-to-speech translation on real data. *arXiv preprint arXiv:2112.08352*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an asr corpus based on public domain audio books. In *Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 5206–5210. IEEE.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Yao Qian, Ximo Bianv, Yu Shi, Naoyuki Kanda, Leo Shen, Zhen Xiao, and Michael Zeng. 2021. Speech-language pre-training for end-to-end spoken language understanding. In *Proceedings of the 2021 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 7458–7462.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Shuo Ren, Long Zhou, Shujie Liu, Furu Wei, Ming Zhou, and Shuai Ma. 2021. Semface: Pre-training encoder and decoder with a semantic interface for neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4518–4527.

Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. 2018. Self-attention with relative position representations. pages 464–468.

Gabriel Synnaeve, Qiantong Xu, Jacob Kahn, Tatiana Likhomanenko, Edouard Grave, Vineel Pratap, Anuroop Sriram, Vitaliy Liptchinsky, and Ronan Collobert. 2019. End-to-end asr: from supervised to semi-supervised learning with modern architectures. *arXiv preprint arXiv:1911.08460*.

Yun Tang, Hongyu Gong, Ning Dong, Changhan Wang, Wei-Ning Hsu, Jiatao Gu, Alexei Baevski, Xian Li, Abdelrahman Mohamed, Michael Auli, and Juan Pino. 2022. Unified speech-text pre-training for speech translation and recognition. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1488–1499, Dublin, Ireland. Association for Computational Linguistics.

Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. 2017. Neural discrete representation learning. In *Proceedings of the 31st Conference on Neural Information Processing Systems*, volume 30.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st Conference on Neural Information Processing Systems*, volume 30, pages 6000–6010.

Chengyi Wang, Yiming Wang, Yu Wu, Sanyuan Chen, Jinyu Li, Shujie Liu, and Furu Wei. 2022a. Supervision-guided codebooks for masked prediction in speech pre-training. *arXiv preprint arXiv:2206.10125*.

Chengyi Wang, Yu Wu, Sanyuan Chen, Shujie Liu, Jinyu Li, Yao Qian, and Zhenglu Yang. 2022b. Improving self-supervised learning for speech recognition with intermediate layer supervision. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7092–7096.

Wei Wang, Shuo Ren, Yao Qian, Shujie Liu, Yu Shi, Yanmin Qian, and Michael Zeng. 2022c. Optimizing alignment of speech and language latent spaces for end-to-end speech recognition and understanding. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7802–7806.

Shinji Watanabe, Takaaki Hori, Suyoun Kim, John R. Hershey, and Tomoki Hayashi. 2017. Hybrid ctc/attention architecture for end-to-end speech recognition. *IEEE Journal of Selected Topics in Signal Processing*, 11(8):1240–1253.

Felix Wu, Kwangyoun Kim, Shinji Watanabe, Kyu Han, Ryan McDonald, Kilian Q Weinberger, and Yoav Artzi. 2022. Wav2seq: Pre-training speech-to-text encoder-decoder models using pseudo languages. *arXiv preprint arXiv:2205.01086*.

Chen Xu, Bojie Hu, Yanyang Li, Yuhao Zhang, Shen Huang, Qi Ju, Tong Xiao, and Jingbo Zhu. 2021a. Stacked acoustic-and-textual encoding: Integrating the pre-trained models into speech translation encoders. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2619–2630, Online. Association for Computational Linguistics.

Qiantong Xu, Alexei Baevski, Tatiana Likhomanenko, Paden Tomasello, Alexis Conneau, Ronan Collobert, Gabriel Synnaeve, and Michael Auli. 2021b. Self-training and pre-training are complementary for speech recognition. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3030–3034.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Proceedings of the 33rd Conference on Neural Information Processing Systems*, 32.

Rong Ye, Mingxuan Wang, and Lei Li. 2022. Cross-modal contrastive learning for speech translation. *arXiv preprint arXiv:2205.02444*.

Yu Zhang, James Qin, Daniel S Park, Wei Han, Chung-Cheng Chiu, Ruoming Pang, Quoc V Le, and Yonghui Wu. 2020. Pushing the limits of semi-supervised learning for automatic speech recognition. *arXiv preprint arXiv:2010.10504*.

Ziqiang Zhang, Sanyuan Chen, Long Zhou, Yu Wu, Shuo Ren, Shujie Liu, Zhuoyuan Yao, Xun Gong, Lirong Dai, Jinyu Li, and Furu Wei. 2022a. Speechlm: Enhanced speech pre-training with unpaired textual data. *arXiv preprint arXiv:2209.15329*.

Ziqiang Zhang, Long Zhou, Junyi Ao, Shujie Liu, Furu Wei, and Jinyu Li. 2022b. The yitrans end-to-end speech translation system for iwslt 2022 offline shared task. *arXiv preprint arXiv:2206.05777*.

Renjie Zheng, Junkun Chen, Mingbo Ma, and Liang Huang. 2021. Fused acoustic and text encoding for multimodal bilingual pretraining and speech translation. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 12736–12746. PMLR.

A Preliminary evaluation on hidden units

As a preliminary validation of using hidden units as an intermediate modality between speech and text, we 1) cascade a speech-to-unit model with a unit-to-text model for speech-to-text evaluation, and 2) cascade a text-to-unit model with a unit-to-text model for text-to-text evaluation.

The speech-to-unit model is a k-means model learned from the released HuBERT Base model, which is exactly the S2U generator introduced in Section 3.3. The text-to-unit model is an encoder-decoder based sequence-to-sequence model trained on *(speech-generated units, text)* paired data of LibriSpeech `train-clean-100`, which is exactly the T2U generator introduced in Section 3.3. The unit-to-text model has the same architecture with the text-to-unit model but with reversed inputs/outputs. The training data of the unit-to-text model comes from (1) `train-clean-100` subset like training text-to-unit model, and (2) pseudo unit-text data, in which we use text-to-unit model to generate pseudo unit from text data in LibriSpeech LM corpus.

The detailed results are listed in Table 7. Although units lose some information (e.g., the repetitive frames are merged) of speech, it still achieves low WER (7.3/18.1) compared to the oracle speech-to-text (CTC) model. On the other hand, cascading T2U and U2T models, which means translating text into units and then translating back, also achieves low WER. These results indicate the units produced by the off-line S2U/T2U generators remain the main linguistic information of both speech and text, thus working as a bridge between the two modalities.

Model	Dev		Test	
	clean	other	clean	other
Directly fine-tune CTC model from HuBERT				
S2T	-	-	6.3	13.2
Cascade two off-line models				
S2U → U2T	6.9	17.9	7.3	18.1
T2U → U2T	5.1	3.1	4.5	5.5

Table 7: WER between the true text and the generated text by different models.

B Data statistics

All the data used in our experiments are listed in Table 8. For LibriSpeech LM data, the text is directly processed into characters and sent to the T2U generator. For WMT data which is much noisier, we normalize the English-side text by removing punctuation and converting digits to spoken words before sending them to the T2U generator. We only keep the samples shorter than 250 words. When generating units from text, we filter out a few portions (about 15%) of data by thresholding the token-averaged decoding likelihood. The threshold is set to -0.666.

C Training details

Pre-training For the base model, the pre-training is conducted on 32 V100 GPUs with the update frequency of 1. The max-tokens of S2U, U2T, and MUM tasks on each GPU are 1,400,000 (87.5 seconds), 3,000, and 3,000, respectively. We use Adam optimizer. The maximum learning rate is $5e-4$ and increases linearly in the first 32K steps, then decays linearly to zero in the total 400k steps. All modules are randomly initialized before pre-training. The pre-training takes about 3 days.

For the large model, the pre-training is conducted on 64 V100 GPUs with the update frequency of 2. The max-tokens are set to 900,000 (56.25 seconds), 2000, and 2000 for S2U, U2T, and MUM tasks respectively. Other optimization configurations are the same as that of the base model. The pre-training takes about 12 days.

ASR fine-tuning Due to the limitation of the GPU memory, the max tokens are set to 1,300,000 (81.25 seconds). During fine-tuning, the speech masking probability is set to 5%. We use the tri-stage learning-rate scheduler with (warm-up, hold, decay) periods of $(10\%, 40\%, 50\%)$. The maximum learning rate is set to $1e-5$. The base model is fine-tuned on 8 GPUs with the update frequency of 2 for 40k steps. The large model is fine-tuned on 8 GPUs with the update frequency of 3 for 80k steps.

ST fine-tuning The max-tokens are set to 800,000 (50 seconds) due to the GPU memory and we drop the training samples longer than it. The speech masking probability is set to 5%. The label smoothing is set to 0.1. The learning rate increases linearly to $3e-5$ in the first 5K steps, then decays linearly to zero in total 50k steps. Models are fine-tuned on 8 GPUs with the update frequency of 4.

Task		Pre-training Data						T2U training data		Fine-tuning Data			
		Unlabeled Speech		Unpaired Text		MT		ASR		ASR		ST	
		name	#hour	name	#utt	name	#utt	name	#hour	name	#hour	name	#hour
ASR	Base	LS	960	LS LM	40M	-	-	LS	100	LS	100	-	-
	Large	LL	60k	LS LM	40M	-	-	LS	100	LS	100/960	-	-
ST	En-De	LS, MuST-C	1.4k	-	-	WMT16	4.6M	LS	100	-	-	MuST-C	408
	En-Es	LS, MuST-C	1.5k	-	-	WMT13	15.2M	LS	100	-	-	MuST-C	504
	En-Fr	LS, MuST-C	1.5k	-	-	WMT14	40.8M	LS	100	-	-	MuST-C	492

Table 8: Statistics of datasets used in experiments. LS: LibriSpeech, LL: LibriLight.

Model	Size	Pre-training Data			WER (↓) Without LM		WER (↓) With LM		
		Speech	Paired	Text	test-clean	test-other	LM	test-clean	test-other
wav2vec 2.0 (Baevski et al., 2020)	Large (0.3B)	60kh	-	-	2.2	4.5	Transf.	1.8	3.3
HuBERT (Hsu et al., 2021)	Large (0.3B)	60kh	-	-	-	-	Transf.	1.9	3.3
WavLM (Chen et al., 2021)	Large (0.3B)	94kh	-	-	-	-	Transf.	1.8	3.2
ILS-SSL (Wang et al., 2022b)	Large (0.3B)	60kh	-	-	1.9	3.8	Transf.	1.8	3.2
STPT (Tang et al., 2022)	Base (0.16B)	60kh	960h	40M	2.1	4.6	Unknown	2.1	4.5
w2v-Conformer (Zhang et al., 2020)	X-Large (0.6B)	60kh	-	-	1.7	3.5	LSTM.	1.5	3.2
w2v-Conformer (Zhang et al., 2020)	XX-Large (1.0B)	60kh	-	-	1.6	3.3	LSTM.	1.5	3.1
w2v-BERT (Chung et al., 2021a)	X-Large (0.6B)	60kh	-	-	1.5	2.9	LSTM.	1.5	2.8
w2v-BERT (Chung et al., 2021a)	XX-Large (1.0B)	60kh	-	-	1.5	2.8	LSTM.	1.5	2.7
SLAM (Bapna et al., 2021)	X-Large (0.6B)	60kh	960h	mC4-En	1.6	3.1	-	-	-
Maestro (Chen et al., 2022)	X-Large (0.6B)	60kh	∼5kh	54M	1.5	2.8	Conf.	1.5	2.7
SpeechUT (Ours)	Large (0.38B)	60kh	100h†	40M	1.6	3.6	Transf.	1.6	3.0

Table 9: ASR performance on 960-hour LibriSpeech benchmark. Speech/Paired/Text indicates the unlabeled speech data, the paired ASR data, and the unpaired text data respectively. † indicates the data is not directly used for pre-training. Transf./LSTM./Conf. indicate the Transformer/LSTM/Conformer language models.

D Inference details

ASR inference We select the model with the highest accuracy on dev-other set as the final model and apply the joint CTC/ED decoding (Hori et al., 2017). We also use a character-level Transformer language model (LM) for shallow fusion (Gulcehre et al., 2015), which is provided by Ao et al. (2022a) [9]. According to Ao et al. (2022a), the LM has a similar or higher word-level perplexity (means worse) than the Transformer LM used in the previous works in Table 1, the latter is provided by Synnaeve et al. (2019). During decoding, the beam size is set to 30 with LM fusion and 10 without it. The ED weight, CTC weight and the LM weight are set to (0.7, 0.3, 0.7) and (0.8, 0.2, 0) respectively after searching on dev-other set.

ST inference We average the parameters of the last 10 checkpoints for inference. The decoding beam is 10. We report the case-sensitive detokenized BLEU (Papineni et al., 2002) on tst-COMMON set.

E ASR results on 960-hour dataset

Table 9 lists the results of the SpeechUT Large model fine-tuned on full LibriSpeech 960 hours of ASR data, compared with several previous self-supervised methods. SpeechUT Large outperforms the previous works in the large model setting like wav2vec 2.0 (Baevski et al., 2020), HuBERT (Hsu et al., 2021), WavLM (Chen et al., 2021), and ILS-SSL (Wang et al., 2022b). While, w2v-Conformer (Zhang et al., 2020), w2v-BERT (Chung et al., 2021a), SLAM (Bapna et al., 2021), and Maestro (Chen et al., 2022) use much larger models with Conformer blocks and/or more speech/text data, which are beyond fair comparison.

[9] https://github.com/microsoft/SpeechT5