SPEECHLM: ENHANCED SPEECH PRE-TRAINING WITH UNPAIRED TEXTUAL DATA

Anonymous authors
Paper under double-blind review

ABSTRACT

How to boost speech pre-training with textual data is an unsolved problem due to the fact that speech and text are very different modalities with distinct characteristics. In this paper, we propose a cross-modal **Speech** and **L**anguage **M**odel (**SpeechLM**) to explicitly align speech and text pre-training with a pre-defined unified discrete representation. Specifically, we introduce two alternative discrete tokenizers to bridge the speech and text modalities, including phoneme-unit and hidden-unit tokenizers, which can be trained using a small amount of paired speech-text data. Based on the trained tokenizers, we convert the unlabeled speech and text data into tokens of phoneme units or hidden units. The pre-training objective is designed to unify the speech and the text into the same discrete semantic space with a unified Transformer network. Leveraging only 10K text sentences, our SpeechLM gets a 16% relative WER reduction over previous models in the Base setting (from 6.8 to 5.7) on the public LibriSpeech ASR benchmark. Moreover, SpeechLM with fewer parameters even outperforms previous works on CoVoST-2 speech translation tasks. We also evaluate our SpeechLM on various spoken language processing tasks under the universal representation evaluation framework SUPERB, demonstrating significant improvements on content-related tasks. Our code and models are available at `https://anonymous`.

1 INTRODUCTION

Speech and text are two important carriers of human communication, and they can be converted into each other through speech recognition and synthesis systems. In past years, the unimodal self-supervised representation learning has been well explored in natural language (Devlin et al., 2019; Dong et al., 2019) and speech (Schneider et al., 2019; Hsu et al., 2021). According to neuroscience, humans first pre-process speech and text with different cortices, and then extract the meaning with the same area, called the Wernicke-Geschwind area (Tremblay & Dick, 2016). Motivated by this, it is a very promising direction to design two pre-nets and a unified representation space (similar to the Wernicke area) so that the speech model would benefit greatly from text modality.

In terms of joint speech-text modeling, most approaches employ a speech encoder and a text encoder to map the speech and text inputs to hidden states, based on which, a shared encoder is used to learn cross-modality content information (Bapna et al., 2021; 2022; Chen et al., 2022b). To align the speech and text modalities, two alignment losses (TLM and STM) in SLAM (Bapna et al., 2021) are introduced with supervised ASR data. Extending SLAM to the multilingual scenario, mSLAM (Bapna et al., 2022) introduces CTC losses and uses SpanBERT (Joshi et al., 2020) to replace the BERT objective for pre-training on character-level text. Based on the RNN-T framework, Maestro (Chen et al., 2022b) learns shared representations with modality matching, duration prediction, and sequence alignment. Almost all previous work follows the same structure with a speech/text encoder and a shared encoder, however, the interface between the speech encoder and the text encoder is not well studied, which probably leads to the outputs of the two encoders in different spaces, and suffers from transfer interference and capacity dilution for the shared encoder (Bapna et al., 2021).

In this paper, we aim at bridging speech and text modalities via a well-defined interface, with which the model can easily benefit from additional textual data. We argue that such a interface should provide a shared semantic space for both speech and text, and preferably have strong interpretability and learnability. To this end, we explore two alternative representation spaces satisfying the above

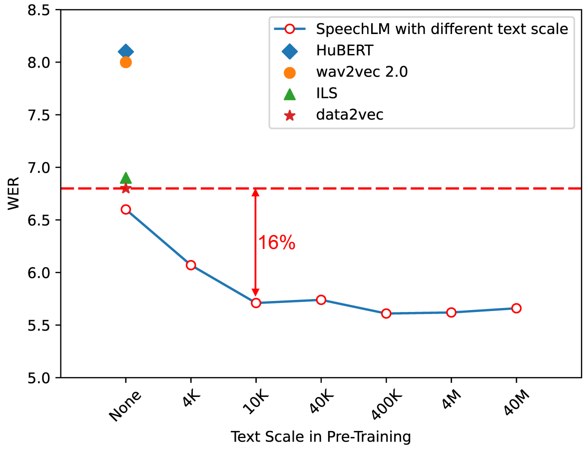

Figure 1: Automatic speech recognition (ASR) performance of different models on LibriSpeech `test-other` set in the Base setting, including HuBERT (Hsu et al., 2021), wav2vec 2.0 (Baevski et al., 2020b), ILS-SSL (Wang et al., 2022b), data2vec (Baevski et al., 2022), and the proposed SpeechLM pre-trained with different amount of text data. All models are in base size (0.1M parameters), and are pre-trained with 960h LibriSpeech data and fine-tuned on its 100h subset.

characteristics of the interface, which are based on phoneme units and hidden units. With them, we can convert the speech and text to a shared intermediate modality (phoneme/hidden units), and decouple the joint speech-text modeling into two sub-modules, speech/text to phoneme/hidden-unit learning components, and a unified unit representation learning component. More specifically, we employ a **Speech Transformer** encoder and an embedding layer to convert unlabeled speech and text into unit representations, respectively. After that, we leverage a functionally defined **Shared Transformer** encoder to jointly model the unit representations for both speech and text.

To prepare the discrete unit tokens for model pre-training, we introduce two discrete tokenizers, including **phoneme-unit tokenizer** and **hidden-unit tokenizer**. For the phoneme-unit tokenizer, we use a hybrid ASR model to transcribe unlabeled speech sequences to frame-level phoneme units, and convert unpaired text by the lexicon. For the hidden-unit tokenizer, we use the HuBERT-based k-means model to cluster speech into hidden units, and utilize a non-autoregressive model to transform the unlabeled text into hidden units. All tokenizer models are obtained with unsupervised data or a small amount of ASR data, and are used offline before pre-training. By converting to the shared phoneme/hidden-unit tokens, both unpaired speech and unpaired text data can be leveraged as training data for model pre-training.

Based on tokenized phoneme/hidden units, we propose two pre-training tasks. One is **Unit-based Masked Language Modeling (UMLM)** trying to predict the unit tokens from the masked speech. The other one is **Unit-based Connectionist Temporal Classification (UCTC)** task, aiming at reconstructing the whole text sequences from the masked unit sequences. To better align the representations of speech and text, we adopt a **Random Swapping Mechanism** for the UMLM task, which swaps the dense outputs from the Speech Transformer and the corresponding discrete unit embeddings before feeding into the Shared Transformer.

We evaluate the proposed model, namely **SpeechLM** which stands for **Speech** and **L**anguage **M**odel, on various spoken language processing tasks. Experimental results on LibriSpeech (Panayotov et al., 2015) ASR tasks show that the proposed SpeechLM, pre-trained with unlabeled speech and text, obtains a substantial WER reduction than previous single-modal pre-trained models, such as HuBERT (Hsu et al., 2021) and data2vec (Baevski et al., 2022). Particularly, under the standard LibriSpeech 960h-100h base setting, SpeechLM achieves the WER of 5.7 with 4-gram LM, which is a relative 30% and 16% reduction over HuBERT (8.1) and data2vec (6.8), respectively. Even with only little text data, e.g., 10K sentences, SpeechLM can obtain a substantial improvement and achieve superior performance, as shown in Figure 1. Experiments on CoVoST-2 (Wang et al., 2020) speech translation tasks with four language directions demonstrate that SpeechLM can be well generalized to other speech-to-text generation tasks. More evaluation on SUPERB (Yang et al., 2021) with 12 tasks shows that SpeechLM achieves better results compared to baselines, especially for content-related tasks. Further analysis demonstrates that SpeechLM can map both speech and text into a shared representation space well.

The contributions of this paper are summarized as follows. 1) We propose a simple yet effective speech representation model SpeechLM, consisting of a Speech Transformer and a Shared Transformer, with two simple and clear learning objectives, unit-based MLM and CTC tasks. 2) To the best of our knowledge, this is the first work to investigate phoneme/hidden-unit representation as the shared space of speech and text modalities. To achieve this goal, two discrete tokenizers are introduced to generate units from speech or text, and the random swapping mechanism is proposed to enhance the alignment of unit representations. 3) Experimental results demonstrate that SpeechLM outperforms its speech-only counterparts on various spoken language tasks. Even with only 10K pre-training text data, SpeechLM can get a 16% relative WER reduction over strong baselines on LibriSpeech ASR tasks.

2 BACKGROUND

Predictive Representation Learning for Speech Unlike natural language processing (NLP), speech signals are continuous, making it not straightforward to find the predictive labels for pre-training. To tackle this issue, a tokenizer, also referred to as a quantizer, is required to map continuous speech features into discrete tokens (Baevski et al., 2020a; Hsu et al., 2021; Chung et al., 2021). HuBERT (Hsu et al., 2021) is the pioneer in the exploration of predictive speech representation learning (SSL), which utilizes a k-means model on the middle layer of Transformer as the tokenizer to convert speech into discrete tokens. Chung et al. (2021) tries to combine a contrastive loss and a masked prediction loss in a self-supervised speech representation learning framework. In addition to the unsupervised tokenizers, Wang et al. (2022a) proposes a supervision-guided tokenizer, which is an acoustic model trained on limited labeled data, and can generate frame-level aligned phonemes as the predictive targets for SSL. In contrast, our goal is to take advantage of textual data to improve speech representation learning, which is a straightforward idea, especially for cross-modal tasks.

Joint Speech-Text Modeling With the rapid development of unimodal pre-training in speech and natural language processing (Devlin et al., 2019; Hsu et al., 2021), joint speech-text pre-training obtains more and more attention from research and industrial communities (Kim et al., 2021; Qian et al., 2021; Bapna et al., 2021; Ao et al., 2022; Tang et al., 2022; Zhang et al., 2022). Most previous studies (Kim et al., 2021; Qian et al., 2021; Ao et al., 2022) let speech and text share some parameters of a neural network in pre-training, however, the speech and the text are not guaranteed to lie in the same space, suffering from transfer interference and capacity dilution (Bapna et al., 2021). To alleviate this issue, SLAM (Bapna et al., 2021) and mSLAM (Bapna et al., 2022), which are most related work to our SpeechLM, leverage extra supervised speech-to-text tasks to enhance the speech-text alignment. However, these approaches still leave unpaired speech and text data modeled separately by using different types of input and pre-training targets. Our work is also related to MAESTRO (Chen et al., 2022b), which learns shared representations from speech and text modalities with a modality matching algorithm in RNN-T framework, but the modality matching could be only performed on paired speech-text data. Unlike mSLAM and MAESTRO, we utilize trained tokenizers to convert all unpaired speech and text into the same space and eliminate the influence of modal difference, so that the two modalities can interact naturally via the shared interface during the pre-training.

3 METHODS

Given unpaired speech and text data, SpeechLM is pre-trained to learn a unified representation of speech and text modalities with the help of offline discrete tokenizers. In this section, we will present the overall framework of SpeechLM, as well as the pre-training procedures and the tokenizers.

3.1 PHONEME/HIDDEN UNIT AS THE BRIDGE

Speech and language are two different modalities with different characteristics. We explore bridging speech and text pre-training with an explicitly defined discrete representation, where speech and text could be tokenized into a shared discrete space easily. Leveraging phoneme/hidden units as the bridge between speech and text has the following advantages: First, it is easier to separately align speech and text into a shared intermediate representation than to align them directly. Second, we can make full use of additional unpaired data to improve the alignment; Thirdly, we can leverage more fine-grained alignment information, i.e., at the frame level, to facilitate joint modeling.

To achieve this goal, we implement two tokenizers for both speech and text, a phoneme-unit tokenizer and a hidden-unit tokenizer, which will be described in detail in Section 3.4. The former aims

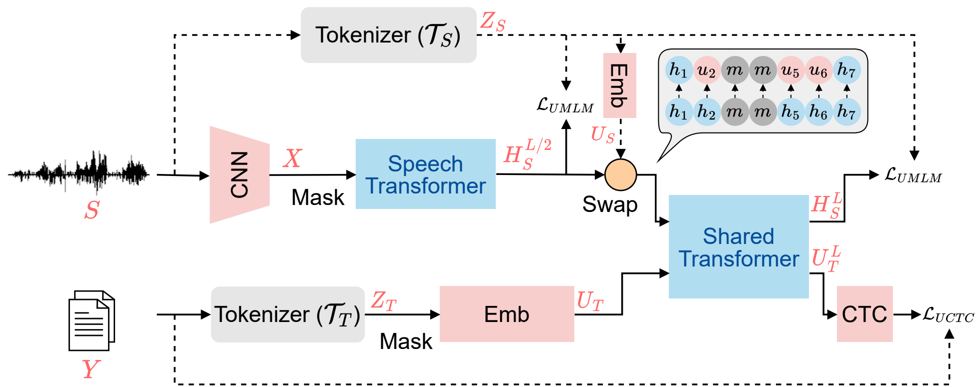

Figure 2: SpeechLM pre-training framework, which consists of a Speech Transformer and a Shared Transformer, and equips with discrete tokenizers.

to convert speech and text into the phoneme space, while the latter converts them into an acoustic clustering space. Given a speech sample $\boldsymbol{S}$ or a text sample $\boldsymbol{Y}$, a tokenizer ($\mathcal{T}_S$ for speech, $\mathcal{T}_T$ for text) yields a sequence of discrete units $\mathbf{Z}$, formulated as

$$\mathbf{Z}_S = (z_{S1}, \ldots, z_{SM}) = \mathcal{T}_S(S) \text{ or } \mathbf{Z}_T = (z_{T1}, \ldots, z_{TN}) = \mathcal{T}_T(Y) \tag{1}$$

where M and N are the lengths of the generated discrete unit sequences from speech and text, respectively.

3.2 Model Architecture

SpeechLM consists of a Speech Transformer and a Shared Transformer, which are enhanced with the random swapping mechanism, as illustrated in Figure 2. Next, we will introduce the main modules with the input of unpaired speech $\boldsymbol{S}$ and text $\boldsymbol{Y}$.

Speech Transformer Following HuBERT (Hsu et al., 2021), we use a standard Transformer (Vaswani et al., 2017) as the backbone of the Speech Transformer, equipped with relative position embedding (Shaw et al., 2018). A speech waveform $\boldsymbol{S}$ is first processed into a sequence of speech features $\boldsymbol{X} = (x_1, x_2, \ldots, x_M)$ by a stack of 1-D convolutional layers. We follow HuBERT to mask the speech feature $\boldsymbol{X}$ with the mask probability of 8% and the mask length of 10. Then the masked features, $\hat{\boldsymbol{X}}$, are fed into the Speech Transformer for higher level representations $\boldsymbol{H}_S^l = \text{Transformer}(\boldsymbol{H}_S^{l-1})$, where l means the layer and $\boldsymbol{H}_S^0 = \hat{\boldsymbol{X}}$ indicates the input.

Shared Transformer The Shared Transformer has the same architecture with the Speech Transformer and handles two types of input with respect to speech and text. The first one is the previous output of the Speech Transformer, $\boldsymbol{H}_S^{L/2} = (h_{S_1}^{L/2}, ..., h_{S_M}^{L/2})$, where L is the total number of Transformer layers. It is continuously processed by the Shared Transformer into $\boldsymbol{H}_S^L$. The second one is the embedding sequence $\boldsymbol{U}_T = (u_{T1}, \ldots, u_{TN})$, which is derived from the tokenized units $\boldsymbol{Z}_T$ by an embedding layer. It is then processed by the Shared Transformer into $\boldsymbol{U}_T^L$, where L indicates the last Transformer layer. Consequently, $\boldsymbol{H}_S^L$ and $\boldsymbol{U}_T^L$ are used as the encoded representations for speech and text. For textual representations, we further employ a CTC layer (Graves et al., 2006) that converts $\boldsymbol{U}_T^L$ to character-level representations.

Random Swapping Mechanism To better align the speech and textual representations into shared latent space at the early layer of the Shared Transformer, we introduce a random swapping mechanism. Note that each speech sequence can be tokenized into discrete units, we can randomly select some time positions (denoted as $i \in \mathcal{R}$) from a speech sequence and replace each $h_{S_i}^{L/2}$ with the corresponding unit embedding u_{Si}, which is derived from speech units z_{Si} by the same embedding layer mentioned above. To avoid information leakage, the swapping positions ($\mathcal{R}$) are only selected within unmasked regions of speech sequence. In this way, we can shuffle two modalities into one sequence and the model can treat them equally.

3.3 PRE-TRAINING TASKS

SpeechLM is jointly optimized by a unit-based masked language modeling task with unlabeled speech data and a unit-based connectionist temporal classification task with unlabeled text data.

Unit-based Masked Language Modeling (UMLM) The unit-based masked language modeling task is designed for speech pre-training, like HuBERT (Hsu et al., 2021) and ILS-SSL (Wang et al., 2022b). Given l-layer speech representations $\boldsymbol{H}_S^l = (h_{S1}^l, \dots, h_{SM}^l)$, UMLM tries to predict the corresponding tokenized units $\mathbf{Z}_S = (z_{S1}, \dots, z_{SM})$ at the masked positions. The probability of the predicted unit at position i is calculated with

$$p(z|h_{Si}^l) = \frac{\exp(\cos(\boldsymbol{W}h_{Si}^l, \boldsymbol{e}(z))/\tau)}{\sum_{z' \in \mathcal{Z}} \exp(\cos(\boldsymbol{W}h_{Si}^l, \boldsymbol{e}(z'))/\tau)} \tag{2}$$

where $\boldsymbol{W}$ is a projection matrix, $\boldsymbol{e}(.)$ is an embedding matrix, τ = 0.1 is the temperature coefficient, and $\mathcal{Z}$ is the set of phoneme/hidden-unit categories. Similar to ILS-SSL, the UMLM loss is computed on both the Speech Transformer ($\boldsymbol{H}_S^{L/2}$) and the Shared Transformer ($\boldsymbol{H}_S^L$), with the loss formulated as,

$$\mathcal{L}_{\mathrm{UMLM}} = -\sum_{i \in \mathcal{M}} \left(\log p(z_{Si}|h_{S_i}^{L/2}) + \log p(z_{Si}|h_{Si}^L) \right) \tag{3}$$

where z_{Si} is the corresponding speech unit at position i and $\mathcal{M}$ is a set of masked positions.

Unit-based Connectionist Temporal Classification (UCTC) Connectionist temporal classification (CTC) (Graves et al., 2006) is first proposed to address the sequence label problem where the output is shorter than the unsegmented input sequences. Here, we take the phoneme-unit or hidden-unit sequences $\boldsymbol{U}_T$ tokenized and upsampled from the unlabeled text as the input, and recognize the original text through the Shared Transformer and CTC layer. The input sequence is masked in the same way as the input of the speech signal. Given a text sequence $\boldsymbol{Y}$, the unit-based CTC loss is calculated as,

$$\mathcal{L}_{\mathrm{UCTC}} = -\log p_{CTC}(\boldsymbol{Y}|\boldsymbol{U}_T^L) \tag{4}$$

where $p_{CTC}(\cdot)$ is modeled by the CTC layer, whose goal is to transform the encoded unit representation $\boldsymbol{U}_T^L$ into the target characters $\boldsymbol{Y}$.

By taking advantage of unlabeled speech and text data, SpeechLM performs multi-task pre-training with UMUM and UCTC tasks, $\mathcal{L} = \mathcal{L}_{\mathrm{UMUM}} + \lambda\mathcal{L}_{\mathrm{UCTC}}$, where λ is used to control the weight of two losses. Through joint optimization and the random swapping mechanism, SpeechLM is expected to align speech and text into a unified representation.

3.4 UNIFIED TOKENIZERS

Figure 3 shows the overview of the proposed phoneme-unit tokenizer and hidden-unit tokenizer. Besides, the tokenizers are offline models, which are used to pre-process the unlabeled speech and text data before the pre-training.

Phoneme-Unit Tokenizer Inspired by PBERT (Wang et al., 2022a), which leverages phoneme labels as the pre-training targets, we introduce the phoneme-unit tokenizer ($\mathcal{T}^P$) to discretize speech signals ($\mathcal{T}_S^P$) as well as text samples ($\mathcal{T}_T^P$). For speech data, the tokenizer is composed of an acoustic model, whose goal is to convert acoustic features into a posterior distribution of phoneme units through a weight finite-sate transducer (WFST) based decoder (Mohri et al., 2002). We implement it using the open-source Kaldi toolkit[1] with a small amount of paired ASR data and language model (LM) data, with details described in Appendix A.1 due to space limitation. For text data, we can directly convert words into phonemes by looking up the provided lexicons. We further upsample the phoneme sequences of text by randomly repeating each phoneme many times to make sure they have similar lengths to the phoneme sequences of speech.

[1]https://github.com/kaldi-asr/kaldi

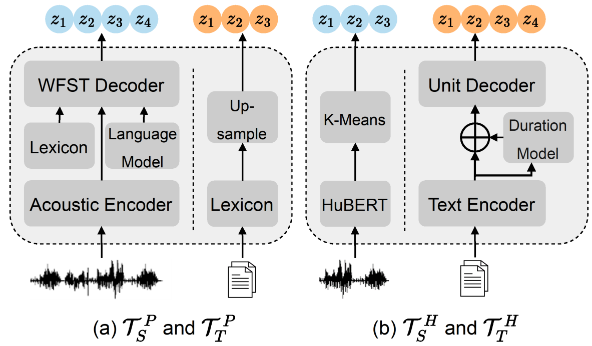

Figure 3: Two alternative tokenizers for speech and text. (a) Phone-unit; (b) Hidden-unit tokenizer.

Hidden-Unit Tokenizer We follow HuBERT to tokenize speech into hidden units with a k-means cluster model, where the clustering feature is the intermediate hidden states of the 2nd round HuBERT model. Inspired by Zhang et al. (2022), to tokenize text data into the same hidden-unit space, we propose a non-autoregressive text to hidden-unit model ($\mathcal{T}_T^H$), which is based on FastSpeech (Ren et al., 2019). The model consists of a text encoder, duration model, and a unit decoder, as shown in Figure 3 (b). $\mathcal{T}_T^H$ is trained with a small amount of text-to-unit pairs from ASR data, where the text side is the phoneme transcriptions with phoneme's durations, and the units are tokenized from the corresponding speech by $\mathcal{T}_S^H$. At inference time, $\mathcal{T}_T^H$ only consumes non-aligned phoneme sequences converted from raw text since the duration is automatically estimated.

4 Experiment

We evaluate SpeechLM models on various spoken language tasks, including automatic speech recognition (ASR), speech translation (ST), and the universal representation evaluation SUPERB (Yang et al., 2021).

4.1 Data

We use unlabeled speech data from LibriSpeech (Panayotov et al., 2015) and LibriLight (Kahn et al., 2020) to pre-train **Base** and **Large** models respectively. LibriSpeech contains 960 hours of labeled speech where the labels are not used in pre-training. LibriLight has about 60,000 hours of unlabeled speech in the same domain as LibriSpeech. The unpaired text data are from LibriSpeech LM corpus[2], containing about 40M English sentences. The paired data for optimizing the tokenizers are the full LibriSpeech data in the **Large** setting and the 100-hour subset (`train-clean-100`) in the **Base** setting. For downstream tasks, we use LibriSpeech for ASR evaluation, and four translation directions of CoVoST-2 (Wang et al., 2020) for ST evaluation. For all tasks of SUPERB evaluation, the data details can be found in Yang et al. (2021).

4.2 Pre-Training Setup

The network architecture of SpeechLM follows that of HuBERT (Hsu et al., 2021) for a fair comparison. Specifically, the **Base** model consists of L=12 Transformer layers where both the Speech Transformer and the Shared Transformer have 6 layers. The **Large** model doubles the number of Transformer layers. The convolutional layers downsample the input waveform to a frame rate of 20ms. The CTC layer consists of a single 1-D convolutional layer followed by a linear layer, which outputs the probabilities of text characters All models are pre-trained on 32 GPUs for 400K steps. To align with HuBERT, the update frequency is set to 4 for **Large** models to simulate 128 GPUs. The batch size for the **Base** model is 4375 tokens after down(up)-sampling for both speech and text input, and for the **Large** model it is set to 2800. The text loss ($\mathcal{L}_{UCTC}$) is weighted by 0.1[3]. More details about the model configuration and training details can be found in Appendix A.3.

4.3 Evaluation on Automatic Speech Recognition

We first verify the pre-trained SpeechLM on ASR tasks, where the Speech Transformer, the Shared Transformer, and the CTC head are fine-tuned with a speech-to-text CTC loss. Base models are fine-tuned on the `train-clean-100` subset and Large models are fine-tuned on the full LibriSpeech.

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

[3]The effect of different weights (λ) is reported in Appendix A.2

Model	Size	Pre-training Data			WER (↓) w/o LM		WER (↓) w/ LM		
		Speech	Paired	Text	test-clean	test-other	LM	test-clean	test-other
100h fine-tuned									
Wav2vec 2.0 (Baevski et al., 2020b)	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., 2022a)	Base (0.1B)	960h	-	-	5.7	12.0	4-gram	3.4	7.7
PBERT (Wang et al., 2022b)	Base (0.1B)	960h	100h	-	4.2	9.5	4-gram	3.1	7.2
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
SpeechLM-H	Base (0.1B)	960h	100h	40M	3.9	9.0	4-gram	2.7	6.4
SpeechLM-P	Base (0.1B)	960h	100h	40M	**3.3**	**7.3**	4-gram	**2.5**	**5.7**
960h fine-tuned									
Wav2vec 2.0 (Baevski et al., 2020b)	Large (0.3B)	60kh	-	-	2.2	4.5	Transf.	1.8	3.3
HuBERT (Hsu et al., 2021)	Large (0.3B)	60kh	-	-	2.1^*	4.2^*	Transf.	1.9	3.3
WavLM (Chen et al., 2022a)	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
SpeechLM-P	Large (0.3B)	60kh	960h	40M	**1.9**	**3.6**	Transf.	**1.8**	**3.2**

Table 1: ASR performance (WER) of different pre-trained models on the LibriSpeech benchmark. **Speech/Text** indicates the unpaired speech and text data, **Paired** indicates the paired ASR data for building tokenizers instead of directly used in pre-training. * indicates our reproduction results.

We measure the quality of ASR by the word error rate (WER) evaluated on the standard LibriSpeech `test-clean/other` sets. Table 1 shows that in the Base setting, SpeechLM significantly outperforms previous models, such as wav2vec 2.0 (Baevski et al., 2020b), HuBERT (Hsu et al., 2021), and data2vec (Baevski et al., 2022). Particularly, the proposed SpeechLM obtains 30% and 16% relative WER reductions over HuBERT and data2vec on `test-other` set, respectively. Furthermore, our SpeechLM Large model achieves competitive or even better performance than previous work.[4]

4.4 Evaluation on Speech Translation

We then evaluate SpeechLM on speech-to-text translation tasks. Following Wang et al. (2021a), we use four language directions from English to German (de), Catalan (ca), Arabic (ar), and Turkish (tr) in CoVoST-2 (Wang et al., 2020). When fine-tuning, the pre-trained model serves as the encoder, followed by a randomly initialized decoder consisting of 6 Transformer layers with the model dimension of 768. We use character vocabulary for target languages in all translation tasks, and report the case-sensitive detokenized BLEU (Papineni et al., 2002) on the test set. The results are shown in Table 2, including the baselines that are fine-tuned from other pre-trained models, without using an external language model. Results show that by boosting the quality of speech representation learning and reducing the burden of cross-modal conversion, SpeechLM-H and SpeechLM-P achieve comparable results in the Base setting, and 2.4 BLEU improvement over HuBERT Base. Moreover, the proposed SpeechLM Large model significantly outperforms previous work, such as wav2vec 2.0 Large and SLAM X-Large (Bapna et al., 2021).

Pre-trained Model	Size (Encoder)	en-de	en-ca	en-ar	en-tr	avg
Pre-ASR (Wang et al., 2020)	-	16.3	21.8	12.1	10.0	15.1
HuBERT (Hsu et al., 2021) *	Base (0.1B)	21.6	28.4	15.9	14.4	20.1
SpeechLM-H	Base (0.1B)	23.8	30.9	17.9	16.1	22.2
SpeechLM-P	Base (0.1B)	**24.2**	**31.2**	**18.3**	**16.2**	**22.5**
wav2vec 2.0 (Wang et al., 2021a)	Large (0.3B)	23.8	32.4	17.4	15.4	22.3
SLAM (Bapna et al., 2021)	X-Large (0.6B)	27.2	33.3	18.5	16.8	24.0
SLAM→w2v-bert (Bapna et al., 2021)	X-Large (0.6B)	27.1	34.2	21.2	17.5	25.0
SpeechLM-P	Large (0.3B)	**27.6**	**35.9**	**21.7**	**19.5**	**26.2**

Table 2: BLEU scores on four translation tasks of CoVoST-2, comparing SpeechLM with previous self-supervised models. * indicates our reproduction results.

[4]SLAM and MAESTRO use 2× model size, larger amount of paired data, or different inference framework (e.g., RNN-T in MAESTRO), whose results are not comparable with the results in Table 1.

Method	#Params	Corpus	Speaker			Content					Semantics				ParaL
			SID	ASV	SD	PR	ASR	OOD-ASR	KS	QbE	ST	IC	SF		ER
			Acc ↑	EER ↓	DER ↓	PER ↓	WER ↓	WER ↓	Acc ↑	MTWV ↑	BLEU ↑	Acc ↑	F1 ↑	CER ↓	Acc ↑
FBANK	0	-	8.5E-4	9.56	10.05	82.01	23.18	63.58	8.63	0.0058	2.32	9.10	69.64	52.94	35.39
PASE+ (Ravanelli et al., 2020)	7.83M	LS 50 hr	37.99	11.61	8.68	58.87	25.11	61.56	82.54	0.0072	3.16	29.82	62.14	60.17	57.86
APC (Chung et al., 2019)	4.11M	LS 360 hr	60.42	8.56	10.53	41.98	21.28	63.12	91.01	0.0310	5.95	74.69	70.46	50.89	59.33
VQ-APC (Chung et al., 2020)	4.63M	LS 360 hr	60.15	8.72	10.45	41.08	21.20	63.56	91.11	0.0251	4.23	74.48	68.53	52.91	59.66
NPC (Liu et al., 2020a)	19.38M	LS 360 hr	55.92	9.40	9.34	43.81	20.20	61.66	88.96	0.0246	4.32	69.44	72.79	48.44	59.08
Mockingjay (Liu et al., 2020c)	85.12M	LS 360 hr	32.29	11.66	10.54	70.19	22.82	65.27	83.67	6.6E-04	4.45	34.33	61.59	58.89	50.28
TERA (Liu et al., 2020b)	21.33M	LS 960 hr	57.57	15.89	9.96	49.17	18.17	58.49	89.48	0.0013	5.66	58.42	67.50	54.17	56.27
DeCoAR 2.0 (Ling & Liu, 2020)	89.84M	LS 960 hr	74.42	7.16	6.59	14.93	13.02	53.62	94.48	0.0406	9.94	90.80	83.28	34.73	62.47
modified CPC (Rivière et al., 2020)	1.84M	LL 60k hr	39.63	12.86	10.38	42.54	20.18	62.54	91.88	0.0326	4.82	64.09	71.19	49.91	60.96
wav2vec (Schneider et al., 2019)	32.54M	LS 960 hr	56.56	7.99	9.9	31.58	15.86	55.86	95.59	0.0485	6.61	84.92	76.37	43.71	59.79
vq-wav2vec (Baevski et al., 2020a)	34.15M	LS 960 hr	38.80	10.38	9.93	33.48	17.71	60.66	93.38	0.0410	5.66	85.68	77.68	41.54	58.24
Wav2vec 2.0 Base (Baevski et al., 2020b)	95.04M	LS 960 hr	75.18	6.02	6.08	5.74	6.43	46.95	96.23	0.0233	14.81	92.35	88.30	24.77	63.43
HuBERT Base (Hsu et al., 2021)	94.68M	LS 960 hr	81.42	5.11	5.88	5.41	6.42	46.69	96.30	0.0736	15.53	98.34	88.53	25.20	64.92
WavLM Base (Chen et al., 2022a)	94.70M	LS 960 hr	84.51	4.69	**4.55**	4.84	6.21	**42.81**	**96.79**	**0.0870**	20.74	**98.63**	89.38	22.86	**65.94**
SpeechLM-H Base	94.70M	LS 960 hr	75.12	6.76	6.48	4.2	5.56	45.78	96.04	0.0526	20.72	97.6	88.76	23.49	63.31
SpeechLM-P Base	94.70M	LS 960 hr	75.11	6.13	6.56	**2.43**	**4.22**	47.22	94.61	0.0458	**22.79**	98.6	**89.4**	**22.36**	62.09

Table 3: Universal speech representation evaluation on the SUPERB benchmark with 12 tasks.

4.5 Universal Representation Evaluation

We further evaluate our SpeechLM models on SUPERB (Yang et al., 2021), which is designed to provide a standard and comprehensive testbed for pre-trained models on various speech tasks, including Speaker Identification (SID), Automatic Speaker Verification (ASV), Speaker Diarization (SD), Phoneme Recognition (PR), Automatic Speech Recognition (ASR), Out-Of-Domain Automatic Speech Recognition (OOD-ASR), Keyword Spotting (KS), Query by Example Spoken Term Detection (QbE), Speech Translation (ST), Intent Classification (IC), Slot Filling (SF), Emotion Recognition (ER). These tasks can be grouped into five aspects of speech: content, speaker, semantics, and paralinguistics (ParaL). Table 3 shows the universal speech representation evaluation results. Compared to the previous self-supervised learning methods, SpeechLM achieves better performance on several content-related and semantic-related tasks, such as PR, ASR, ST, and SF. Particularly, our proposed SpeechLM-P model obtains 50% and 32% relative WER reductions on PR and ASR tasks. Meanwhile, we can observe performance degradation for the speaker and paralinguistics-related tasks, especially when we use the phoneme-unit tokenizer. It indicates that with our joint speech and text pre-training method, the model learns more about extracting the content-related information while discarding the other aspects of speech signals.

4.6 Analysis

To better understand the effectiveness of SpeechLM, we conduct several experiments to investigate its main components, such as the random swapping mechanism, the comparison of two tokenizers, the comparison with vanilla multi-task learning, the amounts of unpaired text data, and further visualization analysis.

#	Model	Size	Pre-training Data			WER (↓) w/o LM		WER (↓) w/ LM		
			Speech	Paired	Text	test-clean	test-other	LM	test-clean	test-other
1	SpeechLM-P	Base (0.1B)	960h	100h	40M	3.3	7.3	4-gram	2.5	5.7
2	SpeechLM-P w/o swapping	Base (0.1B)	960h	100h	40M	4.1	8.5	4-gram	2.7	6.0
3	SpeechLM-P w/o text pre-training	Base (0.1B)	960h	-	-	4.1	9.2	4-gram	2.8	6.6
4	SpeechLM-H w/o text pre-training	Base (0.1B)	960h	-	-	4.7	10.1	4-gram	3.0	6.9
5	SpeechLM-H	Base (0.1B)	960h	100h	40M	3.9	9.0	4-gram	2.7	6.4
6	SpeechLM w/o paired data	Base (0.1B)	960h	-	40M	4.5	9.9	4-gram	3.0	6.9
7	#3 + ASR task	Base (0.1B)	960h	100h	-	4.1	8.6	4-gram	2.9	6.4
8	#1 + ASR task	Base (0.1B)	960h	100h	40M	3.3	7.3	4-gram	2.6	5.6

Table 4: Ablation study on 100-hour LibriSpeech benchmark. The paired data are used for the tokenizers (#1, #2, #5, #8) and/or multi-task pre-training (#7, #8).

Effect of Random Swapping Mechanism The proposed random swapping mechanism is the key component of SpeechLM to align the speech and text modalities in the same space. Here, we explore its effectiveness by removing it. As shown in lines 1-2 of Table 4, without the random swapping mechanism, the performance of SpeechLM declines dramatically from 7.3 WER to 8.5 WER in `test-other` set without LM, and it confirms our suspicions.

Comparison of Two Tokenizers To further compare the influence of the two tokenizers, we pre-train two models with only speech data, with results shown in lines 3-4 of Table 4. Experiments show that phoneme-unit tokenizer is superior to hidden-unit tokenizer for ASR tasks, which may be because the phoneme units have better semantic relevance with speech than hidden units. Moreover,

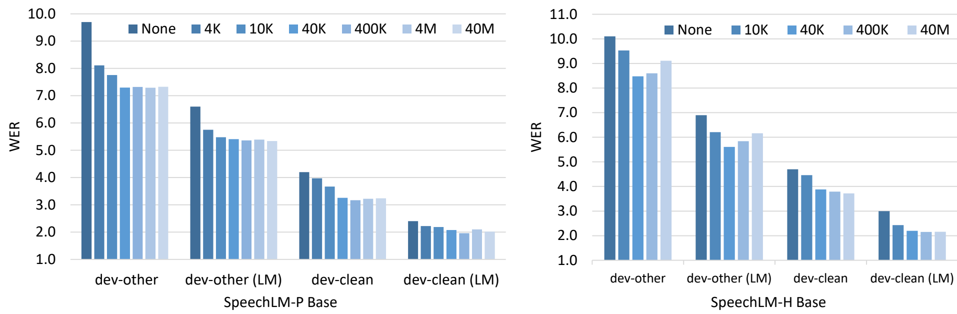

Figure 4: ASR performance with different amounts of text data on 100-hour LibriSpeech.

we explore whether we can obtain improvement by not relying on paired speech-text data for training tokenizers. We conduct an experiment (line 6) in which the speech side predicts the HuBERT hidden units in the mask regions, while the text side is trained with masked phoneme-to-character CTC loss. Compared to the results of SpeechLM-H with pair data (line 5), the performance is degraded drastically, indicating paired data is mandatory to bridge the modality gap.

Comparison with Multi-Task Learning Since we leverage a small amount of supervised data for pre-training, a baseline is joint supervised and unsupervised learning, which has been demonstrated effective in previous work (Wang et al., 2021b). Based on SpeechLM-P w/o text pre-training (line 3) in Table 4, line 7 shows multi-task learning with an extra ASR task (CTC loss on 100h data) can get a considerable improvement, as the WER decreases from 9.2 to 8.6. Furthermore, we pre-train the SpeechLM-P model jointly with an extra ASR task, and the result in the last line of Table 4 demonstrates that multi-task learning with ASR loss can hardly bring further performance gains.

Effect of Text Data Size Since the text corpus contains up to 40M sentences which is much larger than the number of speech samples (960-hour Librispeech contains about 30K sentences), we conduct experiments to analyze the effect of text data size for pre-training, by randomly sampling subsets from the original text corpus. Surprisingly, Figure 4 shows that the performance does not degrade until the text data are reduced to 10K sentences. We speculate that, such a small amount of text data might not be sufficient to learn a good language model to extract complex content information (Devlin et al., 2019), however, it could be sufficient to build a simple lexicon, i.e., the conversion rules from phoneme/hidden units to characters, which might be the main reason to help SpeechLM pre-training. Further exploration of this will be our future work.

Visualization Analysis Figure 5 illustrates the data distributions from different layers of the Shared Transformer in the SpeechLM-P Base model. The dimension is reduced to 2-D by T-SNE (Van der Maaten & Hinton, 2008). Data points are randomly sampled from unpaired speech and text samples from LibriSpeech `dev-clean` set. Layer=0 denotes the speech and text data flows before random swapping. It is shown that as the layer increases, SpeechLM is able to align speech and text representations into a shared space.

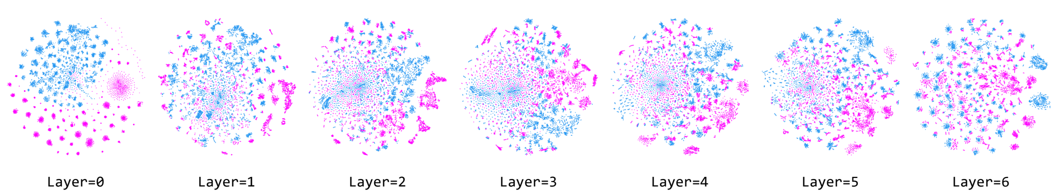

Figure 5: Layer-wise visualization of the Shared Transformer in SpeechLM-P Base. Frame-wise representations of unpaired speech (blue) and phonemes (red) are present.

5 Conclusion

In this work, we present SpeechLM, a text augmented speech pre-trained model, which achieves competitive performance on various spoken language tasks, such as automatic speech recognition and speech translation. To make full use of unpaired data, we propose two alternative discrete tokenizers based on phoneme units and hidden units to tokenize speech and text into the same semantic space. With the shared interface, SpeechLM can learn better speech representations with the help of text modality. Quantitative and qualitative analyses demonstrate the superiority and effectiveness of the proposed method. For future work, we would like to advance the work by deeply integrating the language model ability and extending to natural language tasks.

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. 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)*, pp. 5723–5738, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.393. URL `https://aclanthology.org/2022.acl-long.393`.

Alexei Baevski, Steffen Schneider, and Michael Auli. vq-wav2vec: Self-supervised learning of discrete speech representations. In *International Conference on Learning Representations (ICLR)*, 2020a.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020b.

Alexei Baevski, Wei-Ning Hsu, Alexis CONNEAU, and Michael Auli. Unsupervised speech recognition. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 27826–27839. Curran Associates, Inc., 2021. URL `https://proceedings.neurips.cc/paper/2021/file/ea159dc9788ffac311592613b7f71fbb-Paper.pdf`.

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

Ankur Bapna, Yu-an Chung, Nan Wu, Anmol Gulati, Ye Jia, Jonathan H Clark, Melvin Johnson, Jason Riesa, Alexis Conneau, and Yu Zhang. Slam: A unified encoder for speech and language modeling via speech-text joint pre-training. *arXiv preprint arXiv:2110.10329*, 2021.

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

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 2022a.

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

Yu-An Chung, Wei-Ning Hsu, Hao Tang, and James Glass. An Unsupervised Autoregressive Model for Speech Representation Learning. In *Interspeech*, pp. 146–150, 2019.

Yu-An Chung, Hao Tang, and James Glass. Vector-quantized autoregressive predictive coding. In *Interspeech*, pp. 3760–3764, 2020.

Yu-An Chung, Yu Zhang, Wei Han, Chung-Cheng Chiu, James Qin, Ruoming Pang, and Yonghui Wu. 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)*, pp. 244–250. IEEE, 2021.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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)*, pp. 4171–4186, June 2019. doi: 10.18653/v1/N19-1423.

Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. Unified language model pre-training for natural language understanding and generation. In *Proceedings of the 33rd Conference on Neural Information Processing Systems*, volume 32, pp. 13063–13075, 2019. URL https://proceedings.neurips.cc/paper/2019/file/c20bb2d9a50d5ac1f713f8b34d9aac5a-Paper.pdf.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. ICML '06, pp. 369–376, New York, NY, USA, 2006. Association for Computing Machinery. ISBN 1595933832. doi: 10.1145/1143844.1143891. URL https://doi.org/10.1145/1143844.1143891.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460, 2021. doi: 10.1109/TASLP.2021.3122291. URL https://ieeexplore.ieee.org/document/9585401.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S Weld, Luke Zettlemoyer, and Omer Levy. Spanbert: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77, 2020.

Jacob Kahn, Morgane Rivière, Weiyi Zheng, Evgeny Kharitonov, Qiantong Xu, Pierre-Emmanuel Mazaré, Julien Karadayi, Vitaliy Liptchinsky, Ronan Collobert, Christian Fuegen, et al. 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*, pp. 7669–7673. IEEE, 2020. doi: 10.1109/ICASSP40776.2020.9052942.

Minjeong Kim, Gyuwan Kim, Sang-Woo Lee, and Jung-Woo Ha. 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*, pp. 7478–7482, 2021. doi: 10.1109/ICASSP39728.2021.9414558.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Shaoshi Ling and Yuzong Liu. DeCoAR 2.0: Deep contextualized acoustic representations with vector quantization. *arXiv preprint arXiv:2012.06659*, 2020.

Alexander H Liu, Yu-An Chung, and James Glass. Non-autoregressive predictive coding for learning speech representations from local dependencies. *arXiv preprint arXiv:2011.00406*, 2020a.

Andy T Liu, Shang-Wen Li, and Hung-yi Lee. Tera: Self-supervised learning of transformer encoder representation for speech. *arXiv preprint arXiv:2007.06028*, 2020b.

Andy T. Liu, Shu-wen Yang, Po-Han Chi, Po-chun Hsu, and Hung-yi Lee. Mockingjay: Unsupervised speech representation learning with deep bidirectional transformer encoders. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020c.

Mehryar Mohri, Fernando Pereira, and Michael Riley. Weighted finite-state transducers in speech recognition. *Computer Speech & Language*, 16(1):69–88, 2002.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: an asr corpus based on public domain audio books. In *Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 5206–5210. IEEE, 2015. doi: 10.1109/ICASSP.2015.7178964.

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

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely. The kaldi speech recognition toolkit. IEEE Signal Processing Society, 2011. URL http://infoscience.epfl.ch/record/192584. IEEE Catalog No.: CFP11SRW-USB.

Yao Qian, Ximo Bianv, Yu Shi, Naoyuki Kanda, Leo Shen, Zhen Xiao, and Michael Zeng. 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*, pp. 7458–7462, 2021. doi: 10.1109/ICASSP39728.2021.9414900.

Mirco Ravanelli, Jianyuan Zhong, Santiago Pascual, Pawel Swietojanski, Joao Monteiro, Jan Trmal, and Yoshua Bengio. Multi-task self-supervised learning for robust speech recognition. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6989–6993. IEEE, 2020.

Yi Ren, Yangjun Ruan, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. Fastspeech: Fast, robust and controllable text to speech. *Advances in Neural Information Processing Systems*, 32, 2019.

Morgane Rivière, Armand Joulin, Pierre-Emmanuel Mazaré, and Emmanuel Dupoux. Unsupervised pretraining transfers well across languages. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7414–7418. IEEE, 2020.

Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. wav2vec: Unsupervised pre-training for speech recognition. *arXiv preprint arXiv:1904.05862*, 2019.

Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. Self-attention with relative position representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pp. 464–468, June 2018. doi: 10.18653/v1/N18-2074.

Yun Tang, Hongyu Gong, Ning Dong, Changhan Wang, Wei-Ning Hsu, Jiatao Gu, Alexei Baevski, Xian Li, Abdelrahman Mohamed, Michael Auli, and Juan Pino. 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)*, pp. 1488–1499, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.105. URL https://aclanthology.org/2022.acl-long.105.

Pascale Tremblay and Anthony Steven Dick. Broca and wernicke are dead, or moving past the classic model of language neurobiology. *Brain and language*, 162:60–71, 2016.

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st Conference on Neural Information Processing Systems*, volume 30, pp. 6000–6010, 2017. URL https://dl.acm.org/doi/pdf/10.5555/3295222.3295349.

Changhan Wang, Anne Wu, and Juan Pino. Covost 2 and massively multilingual speech-to-text translation. *arXiv preprint arXiv:2007.10310*, 2020.

Changhan Wang, Anne Wu, Juan Miguel Pino, Alexei Baevski, Michael Auli, and Alexis Conneau. Large-scale self- and semi-supervised learning for speech translation. *CoRR*, abs/2104.06678, 2021a. URL https://arxiv.org/abs/2104.06678.

Chengyi Wang, Yu Wu, Yao Qian, Kenichi Kumatani, Shujie Liu, Furu Wei, Michael Zeng, and Xuedong Huang. Unispeech: Unified speech representation learning with labeled and unlabeled data. In *International Conference on Machine Learning*, pp. 10937–10947. PMLR, 2021b.

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

Chengyi Wang, Yu Wu, Sanyuan Chen, Shujie Liu, Jinyu Li, Yao Qian, and Zhenglu Yang. Improving self-supervised learning for speech recognition with intermediate layer supervision. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2022, Virtual and Singapore, 23-27 May 2022*, pp. 7092–7096. IEEE, 2022b. doi: 10.1109/ICASSP43922.2022.9747022. URL `https://doi.org/10.1109/ICASSP43922.2022.9747022`.

Shu-wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Yist Y Lin, Andy T Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, et al. Superb: Speech processing universal performance benchmark. *arXiv preprint arXiv:2105.01051*, 2021.

Ziqiang Zhang, Long Zhou, Junyi Ao, Shujie Liu, Lirong Dai, Jinyu Li, and Furu Wei. Speechut: Bridging speech and text with hidden-unit for encoder-decoder based speech-text pre-training. *arXiv preprint arXiv:2210.03730*, 2022.

A Appendix

A.1 Tokenizer Details

Phone-unit tokenizer for speech In the Base setting, we train a hybrid GMM-HMM ASR model on 100 hours labeled LibriSpeech data following Kaldi recipe (Povey et al., 2011). In the Large setting, we use the neural network in place of GMM with 960 hours labeled LibriSpeech data, which can boost the performance and alignment accuracy. Once the hybrid model is trained, unlabeled speech data is decoded and transduced to the best phoneme-level alignment paths. The frame shift is 10ms for the Base model setup, and 30ms for the Large model setup, respectively. We then re-sample the phonemes to a frame rate of 20ms by linear interpolation.

Phone-unit tokenizer for text We use the 200K word-to-phone lexicon provided by LibriSpeech to convert words to phonemes, the OOV words are replaced by `<unk>` symbol. Following Baevski et al. (2021), we randomly insert `<SIL>` phoneme between words with a probability of 25%. Then we upsample the phoneme sequence by repeating the phonemes. The length of phonemes follows Gaussian distribution estimated from the `train` set of LibriSpeech, specifically, the mean is 5 and the variance is 25 except for the `<SIL>` phoneme which has a mean of 14.

Hidden-unit tokenizer for speech We use the released HuBERT (Hsu et al., 2021) model following a K-Means model as the tokenizer for speech. The K-Means model has 500 classes with a frame rate of 50.

Hidden-unit tokenizer for text To build a text-to-hidden-unit tokenizer, we modify FastSpeech (Ren et al., 2019) by replacing the prediction head from predicting the spectrum to predicting the probability of hidden units. Specifically, the tokenizer has 4 layers of encoders and 4 layers of decoders, with the model dimension of 256. The input to the model is a phoneme sequence converted from raw text. The upsampling is performed by a duration model between the encoder and the decoder, which predicts the length of each phoneme and repeats the phonemes before feeding them into the decoder. We train the model on LibriSpeech `train-clean-100` subset for 10K steps, with a learning rate of 5e-4 and a batch size of 10K phonemes. The final model achieves 41.3 and 34.6 BLEU scores on `dev-clean` and `dev-other` subsets.

A.2 Effect of Speech/Text Pre-Training Ratio

Table 5 shows the fine-tuning performance of the different pre-trained models with respect to the pre-training loss ratio λ. It is noticed that a lower weight (0.1) of the text pre-training task achieves the best performance in the dev set. Hence, we use $\lambda = 0.1$ for all other experiments.

Ratio (λ)	WER (↓) w/o LM		WER (↓) With LM		
	dev-clean	dev-other	LM	dev-clean	dev-other
0.01	3.65	8.72	4-gram	2.10	5.94
0.1	**3.24**	**7.33**	4-gram	2.02	**5.34**
0.5	2.96	7.68	4-gram	**1.95**	5.56
1.0	2.93	7.82	4-gram	1.97	5.77
10.0	3.45	9.76	4-gram	2.18	7.24

Table 5: ASR performance on 100-hour LibriSpeech benchmark. Different ratio of text pre-training loss in SpeechLM-P model.

A.3 Experimental details

Pre-training configuration The Base model has 12 Transformer layers with the attention dimension of 768 and attention heads of 12, the Large model has 24 Transformer layers with the attention dimension of 1024 and attention heads of 16. The convolutional layers have 512 channels and kernel sizes of [10,3,3,3,3,2,2], resulting in a downsampling rate of 320. The CTC layer is a single 1-D convolutional layer with a kernel size of 2, whose channel matches the Transformer dimension. It is then followed by a linear projection to the text characters. All models are pre-trained on 32 GPUs for 400K steps including 32K warming-up steps. We use Adam (Kingma & Ba, 2014) with β_1=0.9, β_2=0.98 for optimization. The maximum learning rate is set to $5e-4$ and decays linearly to zero after the warming-up steps.

Fine-tuning configuration For Base models fine-tuned on 100-hour LibriSpeech, the total steps are 30K with a batch size of 800 seconds. For Large models fine-tuned on the full LibriSpeech, the total steps are 200K with a batch size of 1800 seconds. All LibriSpeech models are tuned with a maximum learning rate of 1e-5 and a tri-stage learning rate schedule with the warming-up, holding, and decay periods of $[0.1, 0.4, 0.5]$. And for CoVoST-2, both the Base and the Large models are fine-tuned for 50K steps with a batch size of 1600 seconds. The learning rate warms up to 1e-4 in 5K steps and then decays linearly to zero. After fine-tuning, we select the model with the best accuracy on the valid set in the Base setting and average the top 5 models with the best accuracy on the valid set in the Large setting. The decoding beam size is 5 without external language model fusion.