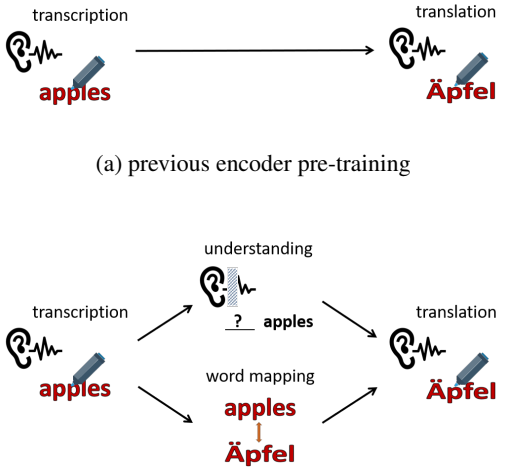
Curriculum Pre-training for End-to-End Speech Translation

Chengyi Wang[*1], **Yu Wu**[2], **Shujie Liu**[2], **Ming Zhou**[2], **Zhenglu Yang**[1]
[1]Nankai University, Tianjin, China
[2]Microsoft Research Asia, Beijing, China
cywang@mail.nankai.edu.cn, Wu.Yu@microsoft.com,
shujliu@microsoft.com, mingzhou@microsoft.com, yangzl@nankai.edu.cn

Abstract

End-to-end speech translation poses a heavy burden on the encoder because it has to transcribe, understand, and learn cross-lingual semantics simultaneously. To obtain a powerful encoder, traditional methods pre-train it on ASR data to capture speech features. However, we argue that pre-training the encoder only through simple speech recognition is not enough, and high-level linguistic knowledge should be considered. Inspired by this, we propose a curriculum pre-training method that includes an elementary course for transcription learning and two advanced courses for understanding the utterance and mapping words in two languages. The difficulty of these courses is gradually increasing. Experiments show that our curriculum pre-training method leads to significant improvements on En-De and En-Fr speech translation benchmarks.



(a) previous encoder pre-training

(b) curriculum encoder pre-training

Figure 1: Comparison between previous encoder pre-training method with our curriculum pre-training method.

1 Introduction

Speech-to-Text translation (ST) is essential to breaking the language barrier for communication. It aims to translate a segment of source language speech to the target language text. To perform this task, prior works either employ a cascaded method, where an automatic speech recognition (ASR) model and a machine translation (MT) model are chained together, or an end-to-end approach, where a single model converts the source language audio sequence to the target language text sequence directly (Berard et al., 2016).

Due to the alleviation of error propagation and lower latency, the end-to-end ST model has been a hot topic in recent years. However, large paired data of source audios and target sentences are required to train such a model, which is not easy to satisfy for most language pairs. To address this issue, previous works resort to pre-training technique (Berard et al., 2018; Bansal et al., 2019), where they leverage the available ASR and MT data to pre-train an ASR model and an MT model respectively, and then initialize the ST model with the ASR encoder and the MT decoder. This strategy can bring faster convergence and better results.

The end-to-end ST encoder has three essential roles: transcribe the speech, extract the syntactic and semantic knowledge of the source sentence and then map it to a semantic space, based on which the decoder can generate the correct target sentence. These pose a heavy burden to the encoder, which can be alleviated by pre-training. However, we argue that the current pre-training method restricts the power of pre-trained representations. The encoder pre-trained on the ASR task mainly

[*]Works are done during internship at Microsoft

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 3728–3738

focuses on transcription, which learns the alignment between the acoustic feature with phonemes or words. It cannot capture linguistic knowledge or understand the semantics, which is essential for translation.

In order to teach the model to understand the sentence and incorporate the required knowledge, extra courses should be taken before learning translation. Motivated by this, we propose a curriculum pre-training method for end-to-end ST. As shown in Figure 1, we first teach the model **transcription** through ASR task. After that, we design two tasks, named frame-based masked language model (FMLM) task and frame-based bilingual lexicon translation (FBLT) task, to enable the encoder to **understand** the meaning of a sentence and **map words** in different languages. Finally, we fine-tune the model on ST data to obtain the **translation** ability.

For the FMLM task, we mask several segments of the input speech feature, each of which corresponds to a complete word. Then we let the encoder predict the masked word. This task aims to force the encoder to recognize the content of the utterance and understand the inner meaning of the sentence. In FBLT, for each speech segment that aligns with a complete word, whether or not it is masked, we ask the encoder to predict the corresponding target word. In this task, we give the model more explicit and strong cross-lingual training signals. Thus, the encoder has the ability to perform simple word translation, and the burden on the ST decoder is largely reduced. Besides, we adopt a hierarchical manner where different layers are guided to perform different tasks (first 8 layers for ASR and FMLM pre-training, and another 4 layers for FBLT pre-training). This is mainly because the three pre-training tasks have different requirements for language understanding and different output spaces. The hierarchical pre-training method can make the division of labor more clear and separate the incorporation of source semantic knowledge and cross-lingual alignments.

We conduct experiments on the LibriSpeech En-Fr and IWSLT18 En-De speech translation tasks, demonstrating the effectiveness of our pre-training method. The contributions of our paper are as follows: (1) We propose a novel curriculum pre-training method with three courses: transcription, understanding and mapping, forcing the encoder to have the ability to generate necessary features for the decoder. (2) We propose two new tasks to learn linguistic features, FMLM and FBLT, which explicitly teach the encoder to do source language understanding and target language meaning mapping. (3) Experiments show that both the proposed courses are helpful for speech translation, and our proposed curriculum pre-training leads to significant improvements.

2 Related Work

2.1 Speech Translation

Early work on speech translation used a cascade of an ASR model and an MT model (Ney, 1999; Matusov et al., 2005; Mathias and Byrne, 2006), which makes the MT model access to ASR errors. Recent successes of end-to-end models in the MT field (Bahdanau et al., 2015; Luong et al., 2015; Vaswani et al., 2017) and the ASR fields (Chan et al., 2016; Chiu et al., 2018) inspired the research on end-to-end speech-to-text translation system, which avoids error propagation and high latency issues.

In this research line, Berard et al. (2016) give the first proof of the potential for an end-to-end ST model. After that, pre-training, multitask learning, attention-passing and knowledge distillation have been applied to improve the ST performance (Anastasopoulos et al., 2016; Duong et al., 2016; Berard et al., 2018; Weiss et al., 2017; Bansal et al., 2018, 2019; Sperber et al., 2019; Liu et al., 2019; Jia et al., 2019). However, none of them attempt to guide the encoder to learn linguistic knowledge explicitly. Recently, Wang et al. (2019b) propose to stack an ASR encoder and an MT encoder as a new ST encoder, which incorporates acoustic and linguistic knowledge respectively. However, the gap between these two encoders is hard to bridge by simply concatenating the encoders. Kano et al. (2017) propose structured-based curriculum learning for English-Japanese speech translation, where they use a new decoder to replace the ASR decoder and to learn the output from the MT decoder (fast track) or encoder (slow track). They formalize learning strategies from easier networks to more difficult network structures. In contrast, we focus on curriculum learning in pre-training and increase the difficulty of pre-training tasks.

2.2 Curriculum Learning

Curriculum learning is a learning paradigm that starts from simple patterns and gradually increases

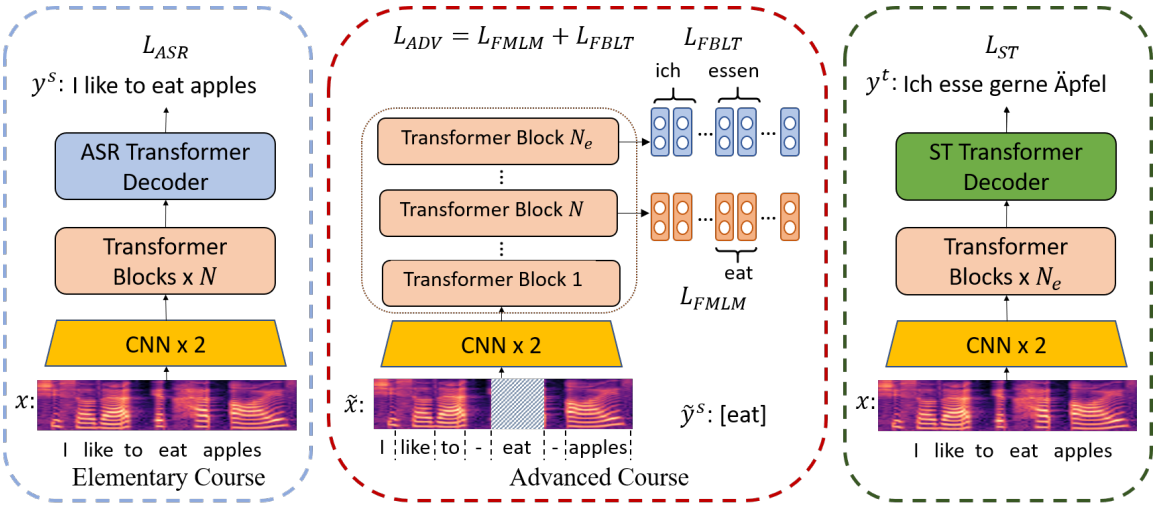


Figure 2: Proposed curriculum pre-training process. $\mathcal{L}_{FMLM}$ only predicts the mask word, while $\mathcal{L}_{FBLT}$ predicts all words in the target language.

to more complex patterns. This idea is inspired by the human learning process and is first applied in the context of machine learning by Bengio et al. (2009). The study shows that this training approach results in better generalization and speeds up the convergence. Its effectiveness has been verified in multiple tasks, including shape recognition (Bengio et al., 2009), object classification (Gong et al., 2016), question answering (Graves et al., 2017), etc. However, most studies focus on how to control the difficulty of the training samples and organize the order of the learning data in the context of single-task learning.

Our method differs from previous works in two ways: (1) We leverage the idea of curriculum learning for pre-training. (2) We do not train the model on the ST task directly with more and more difficult training examples or use more and more complicated structures. Instead, we design a series of tasks with increased difficulty to teach the encoder to incorporate diverse knowledge.

3 Method

3.1 Overview

The overview of our training process is shown in Figure 2. It can be divided into three steps: First, we train the model towards the ASR objective L_{ASR} to learn transcription. We note this as the elementary course. Next, we design two advanced courses (tasks) to teach the model understanding a sentence and mapping words in two languages, named Frame-based Masked Language Model (FMLM) task and Frame-based Bilingual Lexicon Translation (FBLT) task. In the FMLM task, we mask some speech segments and ask the encoder to predict the masked words. In the FBLT task, we ask the encoder to predict the target word for each speech segment which corresponds to a complete source word. In this stage, the encoder is updated by L_{ADV}. We adopt a hierarchical training manner where N encoder blocks are used to perform ASR and FMLM tasks as they both require outputs in source word space, and N_e blocks are used in the FBLT task. After the two-phases pre-training, the encoder is finally combined with a new decoder or a pre-trained MT decoder to perform the ST task towards L_{ST}.

Problem Formulation The speech translation corpus usually contains speech-transcription-translation triples, denoted as $\mathcal{S} = \{(\boldsymbol{x}, \boldsymbol{y^s}, \boldsymbol{y^t})\}$. Specially, $\boldsymbol{x} = (x_1, \cdots, x_{T_x})$ is a sequence of acoustic features which are extracted from the speech signals. $\boldsymbol{y^s} = (y^s_1, \cdots, y^s_{T_s})$ and $\boldsymbol{y^t} = (y^t_1, \cdots, y^t_{T_t})$ represent the corresponding transcription in source language and the translation in target language respectively. To pre-train the encoder, an extra ASR dataset $\mathcal{A} = \{(\boldsymbol{x}, \boldsymbol{y^s})\}$ can be leveraged . Finally, the data for encoder pre-training is denoted as $\{(\boldsymbol{x}, \boldsymbol{y^s}) | (\boldsymbol{x}, \boldsymbol{y^s}) \in \mathcal{A} \vee (\boldsymbol{x}, \boldsymbol{y^s}, \boldsymbol{y^t}) \in \mathcal{S}\}$

After the encoder is pre-trained, we fine-tune the model using only $\mathcal{S}$, to enable it generate $\boldsymbol{y^t}$ from $\boldsymbol{x}$ directly. The model is updated using cross-entropy loss $\mathcal{L}_{ST} = -\log P(\boldsymbol{y^t}|\boldsymbol{x})$.

Model Architecture In this work, we adopt the architecture of Transformer as in (Karita et al.,

2019). The encoder is a stack of two 3×3 2D CNN layers with stride 2 and N_e Transformer encoder blocks. The CNN layers result in downsampling by a factor of 4. The decoder is a stack of N_d Transformer decoder blocks.

3.2 Elementary Course: Transcription

In the elementary course, we train an end-to-end ASR model, which has similar architecture as the ST model. The ASR encoder consists of N blocks, and these blocks are used to initialize the bottom N blocks of the ST encoder. For the ASR task, we follow Karita et al. (2019), to employ a multi-task learning strategy, that is, both the E2E decoder and a CTC module predict the source sentence. Offline experiments indicate that the CTC objective is crucial for attentional encoder-decoder based ASR models. The final objective combines the CTC loss $\mathcal{L}_{ctc}$ and the cross-entropy loss $\mathcal{L}_{CE}$:

$$\begin{aligned}\mathcal{L}_{ASR} &= \alpha\mathcal{L}_{CTC} + (1-\alpha)\mathcal{L}_{CE} \\ &= -\alpha \log P_{ctc}(\boldsymbol{y^s}|\boldsymbol{x}) - (1-\alpha)\log P_{s2s}(\boldsymbol{y^s}|\boldsymbol{x})\end{aligned} \tag{1}$$

In this work, we set α to 0.3. The CTC loss works on the encoder output and it pushes the encoder to learn frame-wise alignment between speech with words.

3.3 Advanced Courses: Understanding and Word Mapping

With the ability of transcription, we further propose two new tasks for the advanced courses.

3.3.1 Frame-based Masked Language Model

The design of the Frame-based Masked Language Model task is inspired by the Masked Language Model (MLM) objective of BERT (Devlin et al., 2019) and semantic mask for ASR task (Wang et al., 2019a). This task enables the encoder to understand the inner meaning of a segment of speech.

As shown in Figure 2, we first perform force-alignment between the speech and the transcript sentence to determine where in time particular words occur in the speech segment. For each word y_i^s, we obtain its corresponding start position s_i and the end position e_i in the sequence $\boldsymbol{x}$ according to force alignment results. At each training iteration, we randomly sample some percentage of the words in the $\boldsymbol{y^s}$ and denote the selected word set as $\tilde{\boldsymbol{y}}^{\boldsymbol{s}}$. Next, for each selected token y_j^s in $\tilde{\boldsymbol{y}}^s$, we mask the corresponding speech piece $[x_{s_j} : x_{e_j}]$. The masked utterance is denoted as $\tilde{\boldsymbol{x}}$ and used as input to the encoder:

$$\boldsymbol{h} = \text{Enc}(\tilde{\boldsymbol{x}}) \tag{2}$$

After that, for a masked piece $[x_{s_j} : x_{e_j}]$, we average the corresponding output hidden states $[h_{\lfloor \frac{s_j}{4} \rfloor} : h_{\lceil \frac{e_j}{4} \rceil}]$[1], and compute the distribution probability over source words as shown in follows:

$$\tilde{h}_j = \text{mean}([h_{\lfloor \frac{s_j}{4} \rfloor} : h_{\lceil \frac{e_j}{4} \rceil}]) \tag{3}$$

$$p(y_j^s|\tilde{\boldsymbol{x}}) = \text{softmax}(\tilde{h}_j \cdot W) \tag{4}$$

In practice, the sentence is represented in BPE tokens and $W \in \mathcal{R}^{d_{model} \times |V_s|}$, where $|V_s|$ is the size of source vocabulary. In this way, a speech piece can be aligned with one or more tokens. We compute KL-Divergence loss as:

$$\mathcal{L}_{FMLM} = -\sum_{y_j^s \in \tilde{\boldsymbol{y}}^s} \sum q(y_j^s) \log \frac{p(y_j^s|\tilde{\boldsymbol{x}})}{q(y_j^s)} \tag{5}$$

$q(y_i^s) \in \mathcal{R}^{|V_s|}$ is a distribution over all BPE tokens in source vocabulary V_s and defined as:

$$q(y_j^s)_{(pos)} = \begin{cases} 1/n_j, & V_s[pos] \in y_j^s \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

where pos represents the dimension index and n_j is the total number of BPE tokens contained in word y_j^s.

In this work, we use a mask ratio of 15% following BERT and the masked speech piece is filled with the mean value of the whole utterance following Park et al. (2019). Because FMLM focuses on the understanding of source language, we computes its loss at the N-th layer of encoder (same with ASR loss), in the hope that the bottom N layers are only concerned with source language.

3.3.2 Frame-based Bilingual Lexicon Translation

Aside from predicting masked source words, we go further to leverage cross-lingual information. Specifically, for each segment of speech features $[x_{s_i} : x_{e_i}]$ which aligned with a source word y_i^s, we assume we can obtain its target counterpart $\tilde{y}_i^t$. Similar to FMLM, we average the output hidden states from position $\lfloor \frac{s_i}{4} \rfloor$ to $\lceil \frac{e_i}{4} \rceil$, and then compute the distribution probability over target vocabulary. The alignment between speech segments and target

[1]The position indexs are divided by 4 due to downsampling.

words is a many-to-many correspondence, so there are cases where $\tilde{y}_i^t$ contains nothing or contains multiple foreign words. For the former case, we set the loss to zero, and for the latter case, we also compute KL-Divergence loss as:

$$\mathcal{L}_{FBLT} = -\sum_{\tilde{y}_i^t}\sum q(\tilde{y}_i^t)\log\frac{p(\tilde{y}_i^t|\tilde{\boldsymbol{x}})}{q(\tilde{y}_i^t)} \quad (7)$$

The definition of $q(\tilde{y}_i^t)$ is the length normalized distribution over all tokens appear in $\tilde{y}_i^t$. Note that the loss is computed on every speech segments, whether or not it is masked.

The only question remaining is how to obtain $\tilde{y}_i^t$ for each speech segment. Since there are two types of data for pre-training, $(\boldsymbol{x}, \boldsymbol{y^s}, \boldsymbol{y^t}) \in \mathcal{S}$ and $(\boldsymbol{x}, \boldsymbol{y^s}) \in \mathcal{A}$, we use two methods to get the alignment:

For training examples $(\boldsymbol{x}, \boldsymbol{y^s}, \boldsymbol{y^t}) \in \mathcal{S}$, we use reference-supervised method. In particular, we simply run Moses[2] scripts to establish word alignments. It begins from running of GIZA++[3] to get source-to-target and target-to-source alignments, and then runs a heuristic grow-diag-final algorithm to get the final results, which means $\forall y_i^s \in \boldsymbol{y^s}$, we choose one word from its translation sentence as the corresponding word $\exists \tilde{y}_i^t \in \boldsymbol{y^t}$ s.t. $\tilde{y}_i^t \sim \boldsymbol{y^s}$.

For training examples $(\boldsymbol{x}, \boldsymbol{y^s}) \in \mathcal{A}$, we apply dictionary-supervised method. Through the above alignment process, we can calculate a bilingual lexical translation table $\mathcal{T}$ with $\{(\boldsymbol{y^s}, \boldsymbol{y^t})|(\boldsymbol{x}, \boldsymbol{y^s}, \boldsymbol{y^t}) \in \mathcal{S}\}$, which estimates the translation probability between a source word w_i^s and a target word w_j^t, denoted as $\mathcal{T} = (w_i^s, w_j^t, p(w_i^s, w_j^t))$. After that, we compute a $\tilde{y}_i^t$ for each y_i^s in $\boldsymbol{y^s}$ according to $\tilde{y}_i^t = \text{argmax}_{w_j^s} p(y_i^s, w_j^s)$.

We compute the $\mathcal{L}_{FBLT}$ at the top layer of the encoder, indicating that the top $N_e - N$ layers are duty on bilingual word mapping. The final training objective in the advanced course combines FMLM and FBLT losses

$$\mathcal{L}_{ADV} = \mathcal{L}_{FMLM} + \mathcal{L}_{FBLT} \quad (8)$$

4 Experiments

4.1 Data and Preprocess

We conduct experiments on two publicly available speech translation datasets: the LibriSpeech En-Fr Corpus (Kocabiyikoglu et al., 2018) and the IWSLT En-De Corpus (Niehues et al., 2018).

LibriSpeech En-Fr: This corpus is a subset of the LibriSpeech ASR corpus (Panayotov et al., 2015) and aligned with French e-books, which contains 236 hours of speech in total. Following previous works, we use the 100 hours clean training set and double the ST size by concatenating the aligned references with the provided Google Translate references, resulting in 90k training instances. We validate on the *dev* set and report results on the *test* set (2048 utterances).

IWSLT En-De: The corpus contains 271 hours of data, with English wave, English transcription, and German translation in each example. We follow Inaguma et al. (2019) to remove utterances of low alignment quality, resulting in 137k utterances. We sample 2k segments from the ST-TED corpus as dev set and *tst2013* is used as the test set (993 utterances).

Data Preprocessing: We run ESPnet[4] (Watanabe et al., 2018) recipes to perform data preprocessing. For both tasks, our acoustic features are 80-dimensional log-Mel filterbanks stacked with 3-dimensional pitch features extracted with a step size of 10ms and window size of 25ms. The features are normalized by the mean and the standard deviation for each training set. Utterances of more than 3000 frames are discarded. We perform speed perturbation with factors 0.9 and 1.1. The alignment results between speech and transcriptions are obtained by Montreal Forced Aligner (McAuliffe et al., 2017).

For references pre-processing, we tokenize and lowercase all the text with the Moses scripts. For pre-training tasks, the vocabulary is generated using sentencepiece (Kudo and Richardson, 2018) with a fixed size of 5k tokens for all languages, and the punctuation is removed. For ST task, we normalize the punctuation using Moses and use the character-level vocabulary due to its better performance (Berard et al., 2018). Since there is no human-annotated segmentation provided in the IWSLT *tst2013*, we use two methods to segment the audios: 1) Following ESPnet, we segment each audio with the LIUM SpkDiarization tool (Meignier and Merlin, 2010). For evaluation, the hypotheses and references are aligned using the MWER method with RWTH toolkit (Bender et al., 2004).

[2] http://www.statmt.org/moses
[3] https://github.com/moses-smt/giza-pp
[4] https://github.com/espnet/espnet

2) We perform sentence-level force-alignment between audio and transcription using aeneas[5] tool and segment the audio according to alignment results.

4.2 Baselines

Experiments are conducted in two settings: **base setting** and **expanded setting**. In base setting, only the corpus described in Section 4.1 is used for each task. In the expanded setting, additional ASR and/or MT data can be used. All results are reported on case-insensitive BLEU with the multi-bleu.perl script unless noted.

4.2.1 End-to-End ST Baselines

We mainly compare our method with the conventional encoder pre-training method which uses only the ASR task to pre-train the encoder. Besides, we also compare with the results of the other works in the literature by copying their numbers.

LibriSpeech: In the context of **base setting**, Berard et al. (2018) and ESPnet have reported results on a LSTM-based ST model with pre-training and/or multi-task learning strategy. Liu et al. (2019) use a Transformer ST model and knowledge distillation method. Wang et al. (2019b) stack an ASR encoder and an MT encoder for final ST task, named as TCEN. Regarding the **expanded setting**, Bahar et al. (2019) apply the SpecAugment on ST task. They use the total 236h of speech for ASR pre-training. Inaguma et al. (2019) combine three ST datasets of 472h training data [6] to train a multilingual ST model. In our work, we use the LibriSpeech ASR corpus as additional pre-training data, including 960h of speech. As the dev and $test$ set of LibriSpeech ST task are extracted from the 960h corpus, we exclude all training utterances with the same speaker that appear in dev or $test$ sets .

IWSLT: Since previous works use different segmentation methods and BLEU-score scripts, it is unfair to copy their numbers. In our work, we choose the ESPnet results as base setting baseline, the multilingual model and TCEN-LSTM model as expanded baselines. Inaguma et al. (2019) use the same multilingual model as described in LibriSpeech baselines. And Wang et al. (2019b) use an additional 272h TEDLIUM2(Rousseau et al., 2014) ASR corpus and 41M parallel data from WMT18 and WIT3[7]. All of them use ESPnet code, LIUM segmentaion method and multi-bleu.perl script. We follow Wang et al. (2019b) to use another 272h ASR data for encoder pre-training and a subset of WMT18[8] for decoder pre-training. We use the same processing method for MT data, resulting in 4M parallel sentences in total. We also re-implement the CL-fast track of Kano et al. (2017) using our model architecture and data as another baseline.

4.2.2 Cacased Baselines

For LibriSpeech ST task, we use results of Berard et al. (2018), Inaguma et al. (2019) and Liu et al. (2019) as base cascaded baselines. The first two use LSTM models for ASR and MT. While the last work trains Transformer ASR and MT models. We build an expanded cascaded system with the pre-trained Transformer ASR model and a LSTM MT model with the default setting in ESPnet recipe. For IWSLT ST task, we use Inaguma et al. (2019) as base cascaded baseline, which is based on LSTM architecture. And we implement a Transformer-based baseline using our pre-trained ASR and MT models in the expanded setting.

4.3 Implementation Details

All our models are implemented based on ESPnet. We set the model dimension d_{model} to 256, the head number H to 4, the feed forward layer size d_{ff} to 2048. For LibriSpeech expanded setting, $d_{model} = 512$ and $H = 8$. For all the ST models, we set the number of encoder blocks $N_e = 12$ and the number of decoder blocks $N_d = 6$. Unless noted, we use $N = 8$ encoder blocks to perform the ASR and the FMLM pre-training tasks. For MT model used in IWSLT expanded setting, we use the Transformer architecture in Vaswani et al. (2017) with $N_e = 6, N_d = 6, H = 4, d_{model} = 256$.

We train the model with 4 Tesla P40 GPUs and batch size is set to 64 per GPU. The pre-training takes 50 and 20 epochs for each phase and the final ST task takes another 50 epochs (a total of 120 epochs). We use the Adam optimizer with warmup steps 25000 in each phase. The learning rate decays proportionally to the inverse square root of the step number after 25000 steps. We

[5] https://www.readbeyond.it/aeneas

[6] LibriSpeech En-Fr, IWSLT En-De and Fisher-CallHome Es-En

[7] https://wit3.fbk.eu/mt.php?release=2017-01-trnted

[8] Europarl v7, Common Crawl, News Comentary v13 and Rapid corpus of EU press releases.

Method	Enc pre-train	Dec pre-train	BLEU
MT(Berard et al., 2018)*	-	-	19.3
MT(Inaguma et al., 2019)	-	-	18.3
base setting			
LSTM ST (Berard et al., 2018)*			12.9
+pre-train+multitask (Berard et al., 2018)*	✓	✓	13.4
LSTM ST+pre-train (ESPnet)	✓	✓	16.68
Transformer+pre-train (Liu et al., 2019)	✓	✓	14.30
+knowledge distillation(Liu et al., 2019)			17.02
TCEN-LSTM (Wang et al., 2019b)	✓	✓	17.05
Transformer+ASR pre-train	✓		15.97
Transformer+curriculum pre-train	✓		**17.66**
expanded setting			
LSTM+pre-train+SpecAugment(Bahar et al., 2019)	✓(236h)	✓	17.0
Multilingual ST+pre-train (Inaguma et al., 2019)	✓(472h)		17.6
Transformer+ASR pre-train	✓(960h)		16.90
Transformer+curriculum pre-train	✓(960h)		**18.01**

Table 1: Comparison on LibriSpeech En-Fr test set. The size of ASR data for base setting is 100h unless labeled. Since inputs of the MT models are ground-truth text, the results of MT models can be seen as the upper-bound of ST models. *: Unknown BLEU score script.

save checkpoints every epoch and average the last 5 checkpoints as the final model. To avoid overfitting, SpecAugment strategy (Park et al., 2019) is used in ASR pre-training with frequency masking (F = 30, mF = 2) and time masking (T = 40, mT=2). The decoding process uses a beam size of 10 and a length penalty of 0.2.

4.4 Experimental Results

4.4.1 Comparison with End-to-End Baselines

LibriSpeech En-Fr: The results on LibriSpeech En-Fr test set are listed in Table 1. In base setting, our method improves the "Transformer+ASR pre-train" baseline by 1.7 BLEU and beats all the previous works, even though we do not pre-train the decoder. It indicates that through a well-designed learning process, the encoder has a strong potential to incorporate large amount of knowledge. Our method beats a knowledge distillation baseline, where an MT model is utilized to teach the ST model. The reason, we believe, is that our method gives the model more training signals and makes it easier to learn. We also outperform a TCEN baseline which includes two encoders. Compared to them, our method is more flexible and incorporates all information into a single encoder, which avoids the representation gap between the two encoders.

As the ASR data size increases, the model performs better. In the expanded setting, we find the FBLT task performs poorly compared with the base setting. This is because the target word prediction task is dictionary-supervised in expanded setting rather than reference-supervised as in base setting. However, our method still outperforms the simple pre-training method by a large margin. Besides, it is surprising to find that the end-to-end ST model is approaching the performance of an MT model, which is the upper bound of the ST model since it accepts golden source sentence without any ASR errors. This further verifies the effectiveness of our method.

IWSLT En-De: The results on IWSLT tst2013 are listed in Table 2, showing a similar trend as in LibriSpeech dataset. We find that the segmentation methods have a big influence on the final results. In the base setting, our method can improve the ASR pre-training baseline by 0.9 to 2.2 BLEU scores, depending on the segmentation methods. In the expanded setting, we find when combined with decoder pre-train, the performance is further improved and beats other expanded baselines.

4.4.2 Comparison with Cascaded Baselines

Table 3 shows comparison with cascaded ST systems. For the base setting of two tasks, our end-to-end model can achieve comparable or better results with cascaded methods. This shows the end-to-end model has powerful learning capabilities and combines the functions of two models. In the LibriSpeech expanded setting, when more ASR data is available, we also obtain a competitive performance. This indicates our method can make a good use of ASR corpus and learn valuable linguistic knowledge other than simple acoustic information. However, when additional MT data is used, there is still a gap between the end-to-end method and the cascaded method. How to utilize bilingual parallel sentences to improve the E2E ST model is worth

Method	Enc pre-train (speech data)	Dec pre-train (text data)	segment method LIUM	segment method aeneas
base setting				
ESPnet			12.50	-
+enc pre-train	✓		13.12	-
+enc dec pre-train	✓	✓	13.54	-
Transformer+ASR pre-train	✓		15.35	17.10
Transformer+curriculum pre-train	✓		**16.27**	**19.29**
expanded setting				
Multilingual ST+pre-train(Inaguma et al., 2019)	✓(472h)		14.6	-
TCEN-LSTM (Wang et al., 2019b)	✓(479h)	✓(40M)	17.65	-
CL-fast(Kano et al., 2017)(re-implemented)	✓(479h)		14.33	16.23
Transformer+curriculum pre-train+dec pre-train	✓(479h)	✓(4M)	**18.15**	**20.35**

Table 2: ST results on IWSLT En-De tst2013 set.

Method	BLEU
LibriSpeech base setting	
LSTM ASR+ MT(Berard et al., 2018)	14.6
LSTM ASR+ MT(Inaguma et al., 2019)	15.8
Transformer ASR + MT(Liu et al., 2019)	17.85
Ours E2E Transformer ST	17.66
LibriSpeech expanded setting	
Transformer ASR+LSTM MT*	18.05
Ours E2E Transformer ST	18.01
IWSLT base setting	
LSTM ASR+ MT(Inaguma et al., 2019)	14.0
Ours E2E Transformer ST	16.27
IWSLT expanded setting	
Transformer ASR+Transformer ST	22.16
Ours E2E Transformer ST	18.15

Table 3: Comparison with cascaded ST. *:we find the LSTM model outperforms Transformer model in our setting since the training data is scarce.

further studying.

4.5 Analysis and Discussion

Ablation Study To better understand the contribution of each component, we perform an ablation study on LibriSpeech expanded setting. The results are shown in Table 4. On the one hand, we show that both of our proposed pre-training tasks are beneficial: In "-FMLM task" and "-FBLT task"[9], we perform single-task pre-training for advanced course. The performance drops when we remove either one of them. On the other hand, we show the two-phases pre-training paradigm is necessary: The "- phase 2" experiment degenerates to the simple ASR pre-training baseline. In "-phase 1" setting, we find that without the ASR pre-training, the training accuracy on FMLM task and FBLT task drops a lot, which further affects the ST performance. This means the ASR task is necessary for both the advanced courses and ST. In "Multi3" setting, we pre-train the model on ASR, FMLM and FBLT tasks in one phase. In this setting, we observe multi-task learning also decrease individual task performances (ASR, FMLM and FBLT) compared to curriculum learning. One reasonable expanation is that it is hard to train on the FMLM and FBLT tasks which takes masked input from randomly initialized parameters, which also leads to performance degradation on the ST task.

[9]we use 12-layer encoder for ASR and FMLM pre-training for a fair comparison.

Method	BLEU
Our method	**18.01**
-FMLM task	17.62
-FBLT task	17.65
-phase 2	16.90
-phase 1	14.26
Multi3	14.82

Table 4: Ablation study on LibriSpeech expanded setting. '-' indicates removing the task or phase from our method.

Hyper-parameter N During pre-training, which layer conducts ASR pre-training and FMLM loss is an important hyper-parameter. We conduct experiments on LibriSpeech base setting to explore the influence of different choices. We keep $N_e = 12$ unchanged and always use the top layer to perform the FBLT task. Then we alter the hyperparameter N. We find if $N = 6$, the model finds it difficult to converge during ST training. That may be because the distance between the decoder and the bottom 6 encoder layers is too far so that the valuable source linguistic knowledge can not be well utilized. Moreover, the model performs undesirable if the choice is 10 or 12, which results in 16.47 and 16.14 BLEU score respectively, since the number of blocks for FBLT task is not enough. The model achieves the best performance when we choose $N = 8$. Thus, we use this strategy in our main experiments.

Unlabeled Speech Data In this work, we also ex-

plore how to utilize the unlabeled speech data in pre-training, but only get negative results. We conduct exploratory experiments on the LibriSpeech ST task. Assume that the $(\boldsymbol{x}, \boldsymbol{y^s})$ from 100h ST corpus as labeled pre-training data and $(\boldsymbol{x})$ from 960h LibriSpeech ASR corpus as unlabeled data. Following Jiang et al. (2019), we design an unsupervised pre-training task for elementary course, in which we randomly mask 15% of fbank features and let the bottom 4 encoder layers predict the masked part. We compute the L1 loss between the prediction and groundtruth filterbanks. However, we find that this method is not helpful for the final ST task, which results in 16.85 BLEU score, lower than our base setting model (without extra data pre-training). It is still an open question about how to use unlabeled speech data.

5 Conclusion and Future Work

This paper investigates the end-to-end method for ST. We propose a curriculum pre-training method, consisting of an elementary course with an ASR loss, and two advanced courses with a frame-based masked language model loss and a bilingual lexicon translation loss, in order to teach the model syntactic and semantic knowledge in the pre-training stage. Empirical studies have demonstrated that our model significantly outperforms baselines. In the future, we will explore how to leverage unlabeled speech data and large bilingual text data to further improve the performance. Besides, we expect the idea of curriculum pre-training can be adopted on other NLP tasks.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grant No.U1636116 and the Ministry of education of Humanities and Social Science project under grant 16YJC790123.

References

Antonios Anastasopoulos, David Chiang, and Long Duong. 2016. An unsupervised probability model for speech-to-translation alignment of low-resource languages. In *EMNLP 2016*, pages 1255–1263. The Association for Computational Linguistics.

Parnia Bahar, Albert Zeyer, Ralf Schlüter, and Hermann Ney. 2019. On using specaugment for end-to-end speech translation. *CoRR*, abs/1911.08876.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *ICLR 2015*.

Sameer Bansal, Herman Kamper, Karen Livescu, Adam Lopez, and Sharon Goldwater. 2018. Low-resource speech-to-text translation. In *Interspeech 2018, 19th Annual Conference of the International Speech Communication Association, Hyderabad, India, 2-6 September 2018*, pages 1298–1302. ISCA.

Sameer Bansal, Herman Kamper, Karen Livescu, Adam Lopez, and Sharon Goldwater. 2019. Pre-training on high-resource speech recognition improves low-resource speech-to-text translation. In *NAACL-HLT 2019*, pages 58–68. Association for Computational Linguistics.

Oliver Bender, Richard Zens, Evgeny Matusov, and Hermann Ney. 2004. Alignment templates: the RWTH SMT system. In *IWSLT 2004*, pages 79–84.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *ICML 2009*, pages 41–48.

Alexandre Berard, Laurent Besacier, Ali Can Kocabiyikoglu, and Olivier Pietquin. 2018. End-to-end automatic speech translation of audiobooks. In *ICASSP 2018*, pages 6224–6228. IEEE.

Alexandre Berard, Olivier Pietquin, Christophe Servan, and Laurent Besacier. 2016. Listen and translate: A proof of concept for end-to-end speech-to-text translation. *CoRR*, abs/1612.01744.

William Chan, Navdeep Jaitly, Quoc V. Le, and Oriol Vinyals. 2016. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In *ICASSP 2016*, pages 4960–4964.

Chung-Cheng Chiu, Tara N. Sainath, Yonghui Wu, Rohit Prabhavalkar, Patrick Nguyen, Zhifeng Chen, Anjuli Kannan, Ron J. Weiss, Kanishka Rao, Ekaterina Gonina, Navdeep Jaitly, Bo Li, Jan Chorowski, and Michiel Bacchiani. 2018. State-of-the-art speech recognition with sequence-to-sequence models. In *ICASSP 2018*, pages 4774–4778.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT 2019*, pages 4171–4186.

Long Duong, Antonios Anastasopoulos, David Chiang, Steven Bird, and Trevor Cohn. 2016. An attentional model for speech translation without transcription. In *NAACL HLT 2016*, pages 949–959. The Association for Computational Linguistics.

Chen Gong, Dacheng Tao, Stephen J. Maybank, Wei Liu, Guoliang Kang, and Jie Yang. 2016. Multi-modal curriculum learning for semi-supervised image classification. *IEEE Trans. Image Processing*, 25(7):3249–3260.

Alex Graves, Marc G. Bellemare, Jacob Menick, Rémi Munos, and Koray Kavukcuoglu. 2017. Automated curriculum learning for neural networks. In *ICML 2017*, pages 1311–1320.

Hirofumi Inaguma, Kevin Duh, Tatsuya Kawahara, and Shinji Watanabe. 2019. Multilingual end-to-end speech translation. In *ASRU 2019*, pages 570–577. IEEE.

Ye Jia, Melvin Johnson, Wolfgang Macherey, Ron J. Weiss, Yuan Cao, Chung-Cheng Chiu, Naveen Ari, Stella Laurenzo, and Yonghui Wu. 2019. Leveraging weakly supervised data to improve end-to-end speech-to-text translation. In *ICASSP 2019*, pages 7180–7184. IEEE.

Dongwei Jiang, Xiaoning Lei, Wubo Li, Ne Luo, Yuxuan Hu, Wei Zou, and Xiangang Li. 2019. Improving transformer-based speech recognition using unsupervised pre-training. *CoRR*, abs/1910.09932.

Takatomo Kano, Sakriani Sakti, and Satoshi Nakamura. 2017. Structured-based curriculum learning for end-to-end english-japanese speech translation. In *Interspeech 2017*, pages 2630–2634.

Shigeki Karita, Xiaofei Wang, Shinji Watanabe, Takenori Yoshimura, Wangyou Zhang, Nanxin Chen, Tomoki Hayashi, Takaaki Hori, Hirofumi Inaguma, Ziyan Jiang, Masao Someki, Nelson Enrique Yalta Soplin, and Ryuichi Yamamoto. 2019. A comparative study on transformer vs RNN in speech applications. In *IEEE Automatic Speech Recognition and Understanding Workshop, ASRU 2019, Singapore, December 14-18, 2019*, pages 449–456. IEEE.

Ali Can Kocabiyikoglu, Laurent Besacier, and Olivier Kraif. 2018. Augmenting librispeech with french translations: A multimodal corpus for direct speech translation evaluation. In *LREC 2018*.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *EMNLP 2018: System Demonstrations*, pages 66–71.

Yuchen Liu, Hao Xiong, Zhongjun He, Jiajun Zhang, Hua Wu, Haifeng Wang, and Chengqing Zong. 2019. End-to-end speech translation with knowledge distillation. In *InterSpeech 2019*, volume abs/1904.08075.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *EMNLP 2015*, pages 1412–1421.

Lambert Mathias and William Byrne. 2006. Statistical phrase-based speech translation. In *ICASSP 2006*, pages 561–564.

Evgeny Matusov, Stephan Kanthak, and Hermann Ney. 2005. On the integration of speech recognition and statistical machine translation. In *INTERSPEECH 2005*, pages 3177–3180.

Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. 2017. Montreal forced aligner: Trainable text-speech alignment using kaldi. In *Interspeech 2017*, pages 498–502.

Sylvain Meignier and Teva Merlin. 2010. Lium spkdiarization: an open source toolkit for diarization. In *CMU SPUD Workshot*.

Hermann Ney. 1999. Speech translation: coupling of recognition and translation. In *ICASSP '99*, pages 517–520.

Jan Niehues, Ronaldo Cattoni, Sebastian Stüker, Mauro Cettolo, Marco Turchi, and Marcello Federico. 2018. The iwslt 2018 evaluation campaign. In *Proceedings of IWSLT*.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An ASR corpus based on public domain audio books. In *ICASSP 2015*, pages 5206–5210.

Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le. 2019. Specaugment: A simple data augmentation method for automatic speech recognition. *CoRR*, abs/1904.08779.

Anthony Rousseau, Paul Deléglise, and Yannick Estève. 2014. Enhancing the TED-LIUM corpus with selected data for language modeling and more TED talks. In *LREC 2014*, pages 3935–3939. European Language Resources Association (ELRA).

Matthias Sperber, Graham Neubig, Jan Niehues, and Alex Waibel. 2019. Attention-passing models for robust and data-efficient end-to-end speech translation. *TACL*, 7:313–325.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NeurIPS 2017*, pages 5998–6008.

Chengyi Wang, Yu Wu, Yujiao Du, Jinyu Li, Shujie Liu, Liang Lu, Shuo Ren, Guoli Ye, Sheng Zhao, and Ming Zhou. 2019a. Semantic mask for transformer based end-to-end speech recognition. *CoRR*, abs/1912.03010.

Chengyi Wang, Yu Wu, Shujie Liu, Zhenglu Yang, and Ming Zhou. 2019b. Bridging the gap between pretraining and fine-tuning for end-to-end speech translation. *CoRR*, abs/1909.07575.

Shinji Watanabe, Takaaki Hori, Shigeki Karita, Tomoki Hayashi, Jiro Nishitoba, Yuya Unno, Nelson Enrique Yalta Soplin, Jahn Heymann, Matthew Wiesner, Nanxin Chen, Adithya Renduchintala, and Tsubasa Ochiai. 2018. Espnet: End-to-end speech processing toolkit. In *Interspeech 2018*, pages 2207–2211. ISCA.

Ron J. Weiss, Jan Chorowski, Navdeep Jaitly, Yonghui Wu, and Zhifeng Chen. 2017. Sequence-to-sequence models can directly translate foreign speech. In *Interspeech 2017*, pages 2625–2629. ISCA.