Fused Acoustic and Text Encoding for
Multimodal Bilingual Pretraining and Speech Translation

Renjie Zheng [* 1] **Junkun Chen** [* 2] **Mingbo Ma** [1] **Liang Huang** [1 2]

Abstract

Recently, representation learning for text and speech has successfully improved many language related tasks. However, all existing methods suffer from two limitations: (a) they only learn from one input modality, while a unified representation for both speech and text is needed by tasks such as end-to-end speech translation, and as a result, (b) they can not exploit various large-scale text and speech data and their performance is limited by the scarcity of parallel speech translation data. To address these problems, we propose a Fused Acoustic and Text Masked Language Model (FAT-MLM) which jointly learns a unified representation for both acoustic and text input from various types of corpora including parallel data for speech recognition and machine translation, and even pure speech and text data. Within this cross-modal representation learning framework, we further present an end-to-end model for Fused Acoustic and Text Speech Translation (FAT-ST). Experiments on three translation directions show that by fine-tuning from FAT-MLM, our proposed speech translation models substantially improve translation quality by up to +5.9 BLEU.

1. Introduction

In recent years, task-agnostic text representation learning (Peters et al., 2018; Devlin et al., 2019; Sun et al., 2019) has attracted much attention in the NLP community due to its strong performance to many downstream tasks. More recently, unsupervised speech representation learning (Baevski et al., 2020; Chen et al., 2020; Liu et al., 2020a) also successfully improved many speech related tasks, such
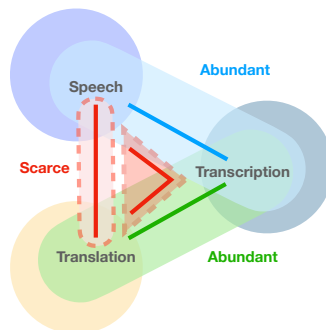

Figure 1. The quality of end-to-end speech translation models has been limited by the scarcity of speech translation datasets. However, there is an abundance of datasets for speech, text, speech recognition, and machine translation data that can be leveraged.

as speech recognition and speech translation.

However all these existing methods can only handle one modality, either text or speech, while joint acoustic and text representation is desired for many end-to-end spoken language processing tasks, such as spoken question answering (Chuang et al., 2019) and end-to-end speech-to-text translation (Liu et al., 2020b). For example, end-to-end speech translation (ST) is desired due to its advantages over the pipeline paradigm, such as low latency, alleviation of error propagation, and fewer parameters (Weiss et al., 2017; Bérard et al., 2018; Jia et al., 2019; Sperber et al., 2017; Zheng et al., 2020; Chen et al., 2021). However, its translation quality is limited by the scarcity of large-scale parallel speech translation data while there exists sufficient data for speech recognition and text machine translation (Fig. 1). It would be helpful if source speech and bilingual text can be encoded into a unified representation via abundant speech recognition and text machine translation data. Liu et al. (2020b) show that jointly training a multi-modal ST encoder can largely improve the translation quality. However, their proposed representation learning method is constrained to the sequence-to-sequence framework and there is no experiment showing whether their proposed method can benefit from extra speech recognition and machine translation data.

Inspired by recent cross-lingual language model pre-training work (Lample & Conneau, 2019) which shows the potential
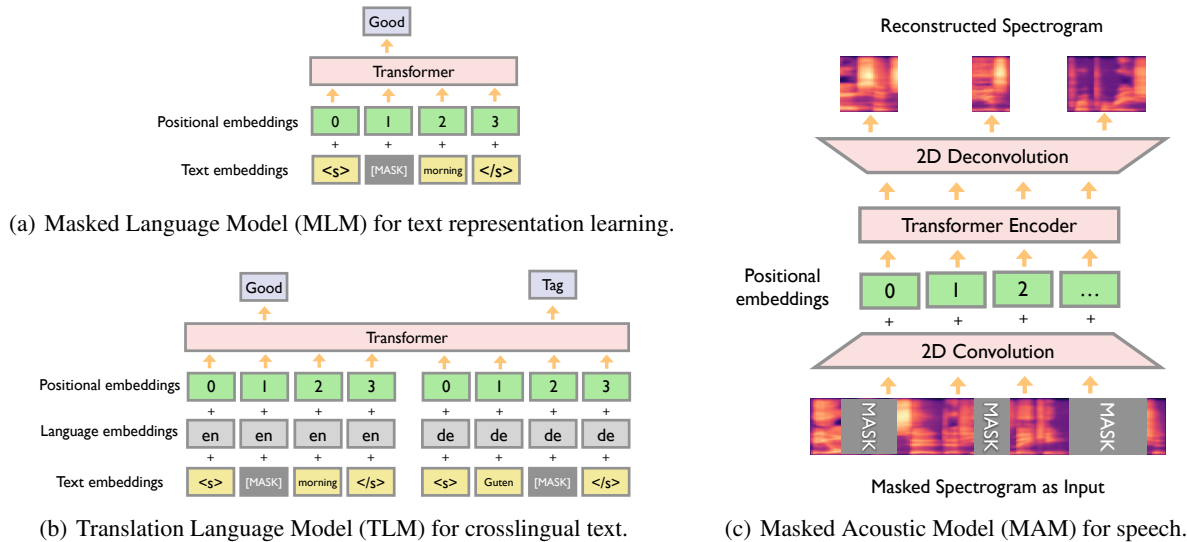
(a) Masked Language Model (MLM) for text representation learning.

(b) Translation Language Model (TLM) for crosslingual text.

(c) Masked Acoustic Model (MAM) for speech.

Figure 2. Previous work for speech or text monomodal representation learning.

to unify the representations of different languages into one encoder, we propose a Fused Acoustic and Text Masked Language Model (FAT-MLM). This model jointly learns a unified representation for both acoustic and text input. In this way, we extend the masked language model's input from only acoustic or text data to multimodal corpora containing both acoustic and text data, such as speech recognition and speech translation for the first time (Fig. 1).

We further extend this Fused Acoustic and Text encoder to a sequence-to-sequence framework and present an end-to-end Speech Translation model (FAT-ST). This enables the model to be trained from both speech and text machine translation data into one single encoder-decoder model. Meanwhile, this model can also learn from speech recognition data using an extra FAT-MLM loss. This resolves the limitation of existing single encoder and decoder speech translation models, which can only learn from scarce parallel speech translation data, but neglects much larger scale speech recognition and text machine translation data (Fig. 1).

We make the following contributions:

- We propose the Fused Acoustic and Text Masked Language Model (FAT-MLM), which can learn a unified acoustic and text representation.

- Based on FAT-MLM, we propose the Fused Acoustic and Text Speech Translation model (FAT-ST), which can do speech recognition and machine translation in a single encoder-decoder framework.

- Spontaneous speech translation experiments on three language pairs show that by finetuning FAT-MLM, the accuracy of FAT-ST improves end-to-end speech translation model by +4.65 BLEU in average and achieves

state-of-the-art. This is the first time that an end-to-end speech translation model achieves similar performance with the strong cascaded system in these three translation directions of this dataset, while still maintaining a smaller model size and faster decoding time.

- We show that FAT-MLM trained with additional speech recognition, machine translation, and monolingual text data can improve FAT-ST by +1.25 BLEU. FAT-ST can be further improved by using additional speech recognition and machine translation data.

2. Previous Work

2.1. Masked Language Modeling

Radford et al. (2018), Howard & Ruder (2018) and Devlin et al. (2019) investigate language modeling for pretraining Transformer encoders. Unlike Radford et al. (2018) using unidirectional language models for pretraining, Devlin et al. (2019) proposes BERT which enables deep bidirectional representation pretraining by a masked language modeling (MLM) objective inspired by the Cloze task (Taylor, 1953) which randomly masks some of the tokens from the input, with an objective to recover the masked word based only on its context. Their approaches lead to drastic improvements on several natural language understanding tasks including text classification (Wang et al., 2018),and question answering (Rajpurkar et al., 2016).

2.2. Translation Language Modeling

Lample & Conneau (2019) extend MLM to cross-lingual pretraining by proposing two methods: one unsupervised that only relies on monolingual data, and one supervised that leverages parallel data with a new cross-lingual lan-
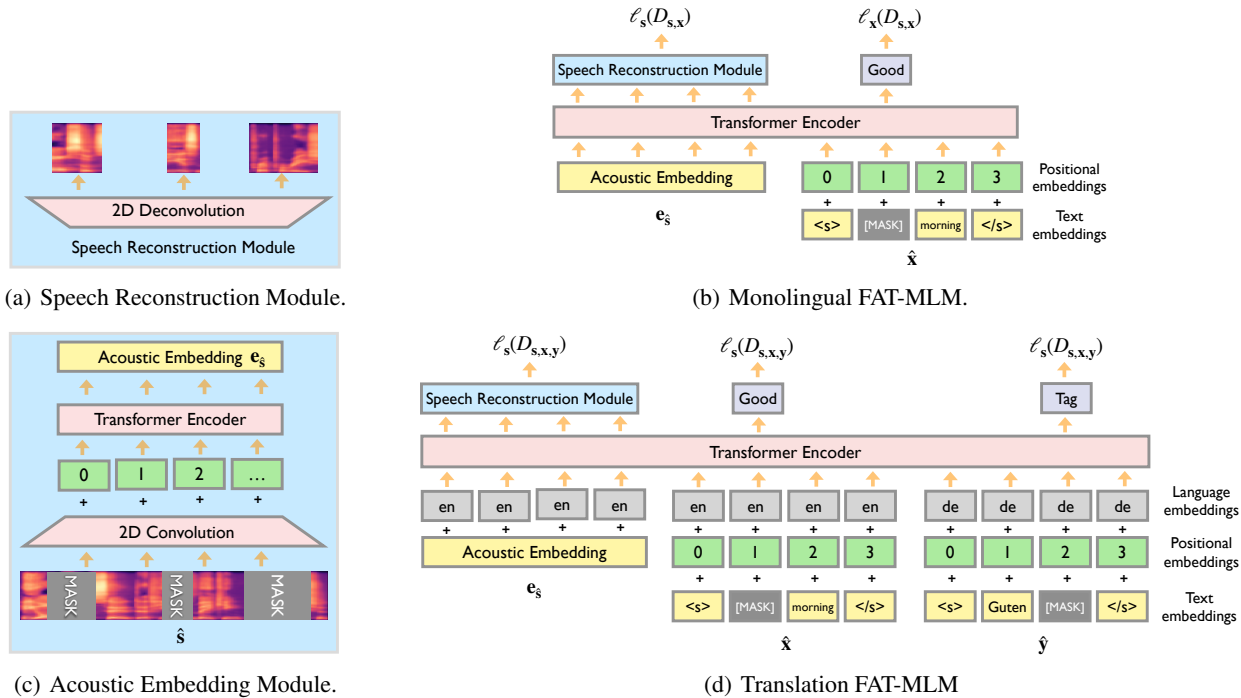
(a) Speech Reconstruction Module.

(b) Monolingual FAT-MLM.

(c) Acoustic Embedding Module.

(d) Translation FAT-MLM

Figure 3. Fused Acoustic and Text-Masked Language Model (FAT-MLM).

guage model objective which is called Translation Language Model (TLM). As shown in Fig. 2(b), TLM encodes both source and target sentences from a parallel data after masking several tokens with [MASK], and then learn to recover the masked tokens. Experiments show that TLM achieves state-of-the-art results on cross-lingual classification, unsupervised and supervised machine translation.

2.3. Masked Acoustic Model

Recently, Chen et al. (2020) propose to learn a speech encoder in a self-supervised fashion on the speech side, which can utilize speech data without transcription. This technique termed Masked Acoustic Modeling (MAM), can also perform pretraining on any acoustic signals (including non-speech ones) without annotation. Fig. 2(c) demonstrate the architecture of MAM. Similar with MLM, MAM replaces a span of speech spectrogram with mask tokens [MASK]. After a 2D Convolution layer and a Transformer Encoder, MAM learns to recover the masked spectrogram via a 2D De-convolution layer during training. Chen et al. (2020) shows that MAM can improve end-to-end speech translation as either an additional loss or a pretraining model. Parallel to MAM, Baevski et al. (2020) proposes the wav2vec 2.0 pretraining model, which masks the speech input in the latent space and pretrains the model via a contrastive task defined over a quantization of the latent representations.

3. Fused Acoustic and Text Masked Language Model (FAT-MLM)

Although existing pretraining models show a strong representation learning ability and significantly improve upon many down-streaming tasks, they all can only learn the representation for either text or speech. However, a unified speech and text multi-modal representation is useful for many end-to-end spoken language processing tasks.

To address this problem, we propose the Fused Acoustic and Text Masked Language Model (FAT-MLM), a multimodal pretraining model which encodes acoustic, text into a unified representation. The idea is similar with Lample & Conneau (2019) who propose to learn a unified representation of different languages. They first propose a method relying on the shared sub-word vocabulary to align different languages' representation. However this is unapplicable in our case because of the modality difference. Thus we propose a method similar to their second approach TLM which uses parallel speech recognition data. In the following sections, we first introduce the monolingual FAT-MLM and then show how to extend it to translation scenario.

3.1. Monolingual FAT-MLM

The monolingual FAT-MLM takes speech and transcription tuples as input, denotes as $D_{\mathbf{s},\mathbf{x}} = \{(\mathbf{s}, \mathbf{x})\}$, where

$\mathbf{s} = (s_1, ..., s_{|s|})$ is a sequence of acoustic features $s_i \in \mathbb{R}^{d_s}$ which can be the spectrogram or mel-spectrogram of the speech audio, and each s_i represents the frame-level speech feature, and $\mathbf{x} = (x_1, ..., x_{|\mathbf{x}|})$ is the sequence of corresponding transcription.

As shown in Fig. 3(c), we first randomly mask several spans of $\mathbf{s}$ by a random masking function over the input $\mathbf{s}$:

$$\hat{\mathbf{s}} \sim \text{Mask}_{\text{span}}(\mathbf{s}, \lambda) \tag{1}$$

where $\text{Mask}_{\text{span}}(\cdot)$ replaces several random spans of $\mathbf{s}$ by probability of λ (30% in our work) with a random initialized vector $\epsilon_{\mathbf{s}} \in \mathbb{R}^{d_{\mathbf{s}}}$. Then we encode $\hat{\mathbf{s}}$ with Convolutions and a Transformer encoder for acoustic embeddings $e_{\hat{\mathbf{s}}}$. Similarly, we randomly mask tokens in $\mathbf{x}$ by a random masking function over the input $\mathbf{s}, \mathbf{x}$:

$$\hat{\mathbf{x}} \sim \text{Mask}_{\text{token}}(\mathbf{x}, \lambda) \tag{2}$$

where $\text{Mask}_{\text{token}}(\cdot)$ replaces several tokens of $\mathbf{x}$ by probability of λ with a random initialized vector $\epsilon_{\text{token}} \in \mathbb{R}^{d_{\mathbf{x}}}$. Then we concatenate acoustic embeddings and source text embeddings $[e_{\hat{\mathbf{s}}}; \hat{\mathbf{x}}]$, and obtain the latent representation $f([e_{\hat{\mathbf{s}}}; \hat{\mathbf{x}}])$ using another Transformer encoder, denoted as f. Same with Lample & Conneau (2019), we reset the positional embeddings for different types of input.

The training objective of monolingual FAT-MLM includes a speech reconstruction loss $\ell_{\mathbf{s}}(D_{\mathbf{s},\mathbf{x}})$ and a text reconstruction loss $\ell_{\mathbf{x}}(D_{\mathbf{s},\mathbf{x}})$. For speech input $\mathbf{s}$, we have the following training objective to reconstruct the original speech signal with the surrounding context information[1]:

$$\ell_{\mathbf{s}}(D_{\mathbf{s},\mathbf{x}}) = \sum_{(\mathbf{s},\mathbf{x}) \in D_{\mathbf{s},\mathbf{x}}} ||\mathbf{s} - g(f([e_{\hat{\mathbf{s}}}; \hat{\mathbf{x}}]))||_2^2 \tag{3}$$

where g is a reconstruction function (we use 2D deconvolution in this work) which tries to recover the original signal from encoded representation $f([e_{\hat{\mathbf{s}}}; \hat{\mathbf{x}}])$. We use mean squared error for measuring the difference between s and the reconstructed spectrogram. For transcription input $\mathbf{x}$, following Devlin et al. (2019) we use cross entropy loss, denoted as

$$\ell_{\mathbf{x}}(D_{\mathbf{s},\mathbf{x}}) = -\sum_{(\mathbf{s},\mathbf{x}) \in D_{\mathbf{s},\mathbf{x}}} \log p(\mathbf{x} \mid [e_{\hat{\mathbf{s}}}; \hat{\mathbf{x}}]) \tag{4}$$

to reconstruct the masked token. The final loss for monolingual FAT-MLM is:

$$\ell_{\text{FAT-MLM}}(D_{\mathbf{s},\mathbf{x}}) = \ell_{\mathbf{s}}(D_{\mathbf{s},\mathbf{x}}) + \ell_{\mathbf{x}}(D_{\mathbf{s},\mathbf{x}}) \tag{5}$$

3.2. Translation FAT-MLM

To support multimodal crosslingual tasks such as speech translation, We propose Translation FAT-MLM which extends Monolingual FAT-MLM by using additional target

language translation of the source language transcription as input. Formally Translation FAT-MLM takes $D_{\mathbf{s},\mathbf{x},\mathbf{y}} = \{(\mathbf{s}, \mathbf{x}, \mathbf{y})\}$ as input, where $\mathbf{y} = [y_1, ..., y_{|y|}]$ denotes the sequence of target language translation. This kind of triplet input is very common in speech translation corpus.

As shown in Fig. 3(d), we incorporate source language embedding e_{src} and target language embedding e_{tgt} for different languages to show the language difference. Similar to Monolingual FAT-MLM, Translation FAT-MLM randomly masks the translation input $\hat{\mathbf{y}} \sim \text{Mask}_{\text{token}}(\mathbf{y}, \lambda)$ and concatenate it with another two embeddings:

$$\mathbf{h}_{\mathbf{s},\mathbf{x},\mathbf{y}} = [e_{\hat{\mathbf{s}}} + e_{\text{src}}; \hat{\mathbf{x}} + e_{\text{src}}; \hat{\mathbf{y}} + e_{\text{tgt}}]$$

Then we reconstruct masked input from concatenated embeddings $\mathbf{h}_{\mathbf{s},\mathbf{x},\mathbf{y}}$ via a Transformer encoder. The reconstruction loss for different masked input is:

$$\ell_{\mathbf{s}}(D_{\mathbf{s},\mathbf{x},\mathbf{y}}) = \sum_{(\mathbf{s},\mathbf{x},\mathbf{y}) \in D_{\mathbf{s},\mathbf{x},\mathbf{y}}} ||\mathbf{s} - g(f(\mathbf{h}_{\mathbf{s},\mathbf{x},\mathbf{y}}))||_2^2$$
$$\ell_{\mathbf{x}}(D_{\mathbf{s},\mathbf{x},\mathbf{y}}) = -\sum_{(\mathbf{s},\mathbf{x},\mathbf{y}) \in D_{\mathbf{s},\mathbf{x},\mathbf{y}}} \log p(\mathbf{x} \mid \mathbf{h}_{\mathbf{s},\mathbf{x},\mathbf{y}})$$
$$\ell_{\mathbf{y}}(D_{\mathbf{s},\mathbf{x},\mathbf{y}}) = -\sum_{(\mathbf{s},\mathbf{x},\mathbf{y}) \in D_{\mathbf{s},\mathbf{x},\mathbf{y}}} \log p(\mathbf{y} \mid \mathbf{h}_{\mathbf{s},\mathbf{x},\mathbf{y}})$$

We sum these loss functions for the final loss function of Translation FAT-MLM:

$$\ell_{\text{FAT-MLM}}(D_{\mathbf{s},\mathbf{x},\mathbf{y}}) = \ell_{\mathbf{s}}(D_{\mathbf{s},\mathbf{x},\mathbf{y}}) + \ell_{\mathbf{x}}(D_{\mathbf{s},\mathbf{x},\mathbf{y}}) + \ell_{\mathbf{y}}(D_{\mathbf{s},\mathbf{x},\mathbf{y}})$$

To fully utilize the corpora for different tasks, FAT-MLM can take any combination of speech, transcription, translation triplets $D_{2\{\mathbf{s},\mathbf{x},\mathbf{y}\}}$ as input.[2] Specifically, these combinations include speech only data $\{\mathbf{s}\}$, monolingual text data, $\{\mathbf{x}\}$ or $\{\mathbf{y}\}$, speech and transcription tuple $\{(\mathbf{s}, \mathbf{x})\}$ for speech recognition, transcription and translation tuple $\{(\mathbf{x}, \mathbf{y})\}$ for machine translation, speech and translation tuple $\{(\mathbf{s}, \mathbf{y})\}$ for direct speech translation and speech transcription translation triplets $\{(\mathbf{s}, \mathbf{x}, \mathbf{y})\}$. For different combinations of input, FAT-MLM encodes the full concatenation of their embeddings and recover the masked portion. The loss function is:

$$\ell_{\text{FAT-MLM}}(D_{2\{\mathbf{s},\mathbf{x},\mathbf{y}\}}) = \ell_{\mathbf{s}}(D_{\mathbf{s}\star}) + \ell_{\mathbf{x}}(D_{\mathbf{x}\star}) + \ell_{\mathbf{y}}(D_{\mathbf{y}\star}) \tag{6}$$

where $D_{\mathbf{s}\star}$, $D_{\mathbf{x}\star}$, $D_{\mathbf{y}\star}$ means any input including speech, source language text and target language text respectively. Note that in this framework, we can denote MLM as $\ell_{\mathbf{x}}(D_{\mathbf{x}})$, TLM as $\ell_{\mathbf{x},\mathbf{y}}(D_{\mathbf{x},\mathbf{y}})$, MAM as $\ell_{\mathbf{s}}(\mathbf{s})$.

3.3. Attention Visualization

To demonstrate FAT-MLM's ability to unify the representation of different modality and language, we show the self-attention layers of a translation FAT-MLM in Fig. 4 and 5. The clear monotonic attention in Fig. 4 shows that our

---

[1]Similar with previous work using masked language model objective, this loss only takes the masked input into consideration.

[2]$2^{\{\mathbf{s},\mathbf{x},\mathbf{y}\}}$ is the power set of $\{\mathbf{s}, \mathbf{x}, \mathbf{y}\}$ triplets.
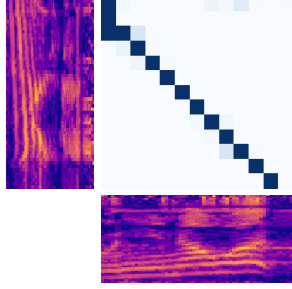
Figure 4. One speech self-attention head's output at the first transformer layer in acoustic embedding module and its corresponding spectrogram. This is a Translation FAT-MLM model trained with Must-C En→De dataset.

proposed method can learn good representation for speech (Chen et al., 2020). Fig. 5(a) shows that FAT-MLM can learn a good crosslingual alignment between two languages, such as and to Und and you to Sie. Fig. 5(b) shows that FAT-MLM is able to learn a clear monotonic speech-to-text crossmodal attention like many speech recognition models.

4. Fused Acoustic and Text Speech Translation (FAT-ST)

In this section, we present how to adapt FAT-MLM to speech translation and enable speech translation models to learn from speech recognition and text machine translation.

## 4.1. From Text Translation to Speech Translation

Regardless of the particular design of different seq-to-seq models, the text machine translation encoder always takes the input sequence $\mathbf{x} = (x_1, ..., x_n)$ where each $x_i \in \mathbb{R}^{d_x}$ is a word embedding of d_x dimensions, and produces a new sequence of hidden states $\mathbf{h} = f(\mathbf{x}) = (h_1, ..., h_n)$. On the other hand, a decoder predicts the next output word y_t given the source sequence (actually its representation $\mathbf{h}$) and previously generated words, denoted $\mathbf{y}_{<t} = (y_1, ..., y_{t-1})$. The decoder stops when it emits <eos>, and the final hypothesis $\mathbf{y} = (y_1, ..., \text{<eos>})$ has probability

$$p(\mathbf{y} \mid \mathbf{x})_{\text{MT}} = \prod_{t=1}^{|\mathbf{y}|} p(y_t \mid \mathbf{x}, \mathbf{y}_{<t}) \qquad (7)$$

At training time, we maximize the conditional probability of each ground-truth target sentence $\mathbf{y}^\star$ given input $\mathbf{x}$ over the whole training data $D_{\mathbf{x},\mathbf{y}}$, or equivalently minimizing the following loss:

$$\ell_{\text{MT}}(D_{\mathbf{x},\mathbf{y}}) = -\sum_{(\mathbf{x},\mathbf{y}) \in D_{\mathbf{x},\mathbf{y}}} \log p(\mathbf{y} \mid \mathbf{x}) \qquad (8)$$

Different from text machine translation, speech translation takes speech features $\mathbf{s} = (s_1, ..., s_{|\mathbf{s}|})$ as input. Same as the speech input portion of FAT-MLM, these speech features are converted from the speech signals (e.g. spectrogram).
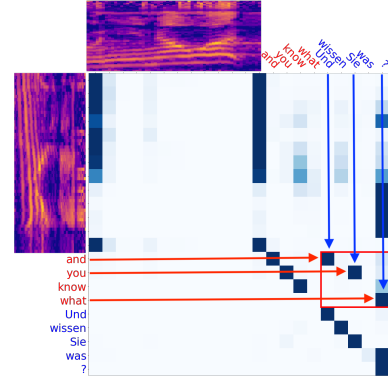
(a) This self-attention head shows bilingual alignment between "and" and "Und", "you" and "Sie", "what" and "?" in transcription and translation respectively.
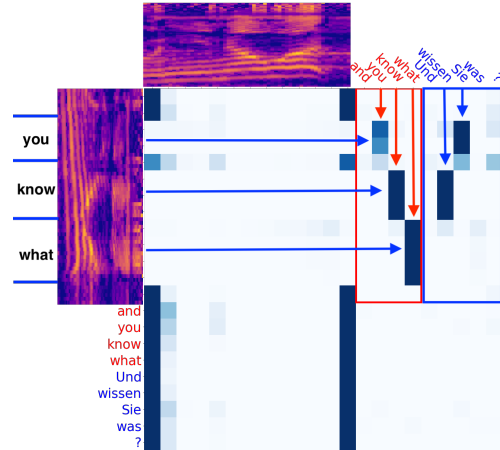
(b) Left side spectrogram shows gold speech-transcription alignment. This self-attention head shows monotonic crossmodal attention in red box. Meanwhile, the speech-to-translation attention (in blue box) clearly show the alignment between "you" and "Sie", "know" and "wissen" in speech and translation respectively. Note that in this speech, the pronounciation of "and" is very weak.

Figure 5. Two self-attention heads' output at the first layer of acoustic and text shared transformer from a Translation FAT-MLM model trained with Must-C En→De dataset, annotated with corresponding spectrogram, transcription (red) and translation (blue).

Formally, the decoding and training of speech translation models can be defined as follows:

$$p(\mathbf{y} \mid \mathbf{s})_{\text{ST}} = \prod_{t=1}^{|\mathbf{y}|} p(y_t \mid \mathbf{s}, \mathbf{y}_{<t}) \qquad (9)$$

$$\ell_{\text{ST}}(D_{\mathbf{s},\mathbf{y}}) = -\sum_{(\mathbf{s},\mathbf{y}) \in D_{\mathbf{s},\mathbf{y}}} \log p(\mathbf{y} \mid \mathbf{s}) \qquad (10)$$
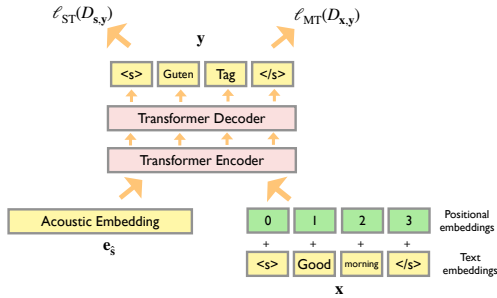
Figure 6. Fused Acoustic and Text-Speech Translation (FAT-ST).

4.2. FAT-ST

To boost the performance of end-to-end speech translation, we propose to enable speech translation to encode both acoustic and text features as input by simply adapting the architecture of monolingual FAT-MLM to a Fused Acoustic and Text Speech Translation model (FAT-ST).

FAT-ST's encoder shares identical architecture with monolingual FAT-MLM. In this way, we can simply encode either acoustic or text features by this encoder and the FAT-ST model can be optimized by speech translation loss ℓ_{ST}, machine translation loss ℓ_{MT} and FAT-MLM loss $\ell_{FAT-MLM}$. For a speech translation dataset $D_{s,x,y}$, we decouple the triplets into three part $D_{s,y}$ for ℓ_{ST}, $D_{s,x}$ for $\ell_{FAT-MLM}$ and $D_{x,y}$ for ℓ_{MT}. The loss function of FAT-ST is:

$$\ell_{FAT-ST}(D_{s,y} \cup D_{s,x} \cup D_{x,y}) = \ell_{ST}(D_{s,y}) + \ell_{MT}(D_{x,y})$$
$$+ \ell_{FAT-MLM}(D_{s,x})$$

Please note that the speech recognition and machine translation data can either be included in speech translation data or additional datasets. Meanwhile, in practice, we find that CTC loss (Graves et al., 2006) is useful to improve the translation quality so that we include it in all the experiments.

4.3. Finetuning FAT-ST from Translation FAT-MLM

Similar to Lample & Conneau (2019) we can further improve FAT-ST by finetuning from FAT-MLM. Since the FAT-ST decoder predicts text only, we initialize it from the acoustic and text shared Transformer encoder. Although Transformer decoder is unidirectional which is different from bidirectional FAT-MLM, it can still benefit from FAT-MLM in our experiments, This is also observed by Lample & Conneau (2019) and Devlin et al. (2019).

5. Experiments

We conducted speech translation experiments in 3 directions: English to German (En→De), English to Spanish (En→Es), and English to Dutch (En→Nl) to show the translation quality of baselines and our proposed methods.

5.1. Dataset

(a) Bilingual Dataset

Type	Name	En → De		En → Es		En → Nl	
		Hours	#Sent	Hours	#Sent	Hours	#Sent
$D_{s,x,y}$	Must-C ST	408	226K	504	262K	442	245K
$D_{x,y}$	Europarl MT	-	1.9M	-	2.0M	-	2.0M

(b) Monolingual Dataset

Type	Name	En		De	Es	Nl
		Hours	#Sent	#Sent	#Sent	#Sent
$D_{s,x}$	Librispeech ASR	960	281K	-	-	-
D_s	Libri-light Speech	3,748	579K	-	-	-
D_x/D_y	Europarl / Wiki Text	-	2.3M	2.1M	2.0M	2.3M

Table 1. Statistics of all datasets used in our experiments. Note that we use Europarl for En, De, Es monolingual text and Wiki Text for Nl because there is no monolingual Nl portion in Europarl. #Sent means the number of sentences.

We use 5 corpora with different modalities and languages: speech translation data $D_{s,x,y}$ Must-C (Di Gangi et al., 2019), speech recognition data $D_{s,x}$ Librispeech (Panayotov et al., 2015), machine translation and monolingual text data $D_{x,y}, D_x, D_y$ Europarl V7 (Koehn, 2005), speech only data D_s Libri-Light (medium version) (Kahn et al., 2020) and monolingual text data Wiki Text (only for Nl). The statistical results of the dataset are shown in Table. 1. We evaluate our models on Must-C dev and test set. Note that Must-C is collected based on spontaneous speeches (TED) which are very different from other audiobook speech dataset used in our experiments. Spontaneous speeches are much harder for speech translation than audiobook dataset such as Libri-trans (Kocabiyikoglu et al., 2018). That is one of the reasons why the translation accuracy of end-to-end speech translation is much worse than cascaded systems on Must-C than other speech translation corpus.

5.2. Training Detail

Raw audio files are processed by Kaldi (Povey et al., 2011) to extract 80-dimensional log-Mel filterbanks stacked with 3-dimensional pitch features using a window size of 25 ms and step size of 10 ms. We train sentencepiece (Kudo & Richardson, 2018) models with a joint vocabulary size of 8K for text in each dataset. Training samples that have more than 3000 frames have been ignored for GPU efficiency. Our basic Transformer-based E2E-ST framework has similar settings with ESPnet-ST(Inaguma et al., 2020). the speech input is first down-sampled the speech input with 2 layers of 2D convolution of size 3 with stride size of 2. Then there is a standard 12-layers Transformer with feed-forward layer of 2048 hidden size to bridge the source and target side. We only use 4 attention heads on each side of the transformer and each of them has a dimensionality of 256. We also show the results of FAT-ST big model with 4096 hidden

Pretrain Method	Models	En→De	En→Es	En→Nl	Avg.	Model Size
No Pretraining	ST	19.64	23.68	23.01	22.11	31.25M
	ST + ASR	21.70	26.83	25.44	24.66 (+2.55)	44.82M
	ST + ASR & MT	21.58	26.37	26.17	24.71 (+2.60)	56.81M
	ST + MAM	20.78	25.34	24.46	23.53 (+1.42)	33.15M
	ST + MAM + ASR	22.41	26.89	26.49	25.26 (+3.15)	46.72M
	Liu et al. (2020b)	22.55	-	-	-	-
	Le et al. (2020)	23.63	28.12	27.55	26.43 (+4.32)	51.20M
	Cascade§	23.65	28.68	27.91	26.75 (+4.64)	83.79M
	FAT-ST (base).	22.70	27.86	27.03	25.86 (+3.75)	39.34M
ASR & MT	ST	21.95	26.83	26.03	24.94 (+2.83)	31.25M
	ST + ASR & MT	22.05	26.95	26.15	25.05 (+2.94)	56.81M
MAM	FAT-ST (base)	22.29	27.21	26.26	25.25 (+3.14)	39.34M
FAT-MLM	FAT-ST (base)	**23.68**	28.61	**27.84**	26.71 (+4.60)	39.34M
	FAT-ST (big)	23.64	**29.00**	27.64	**26.76** (+4.65)	58.25M

Table 2. BLEU comparisons on Must-C test set between our proposed methods and other baselines over 3 translation directions using MuST-C ($D_{\mathbf{s,x,y}}$) only (including pretraining methods). § are reported in Inaguma et al. (2020).

Pretrain Data	Pretrain Method	Train Data	Models	En→De	En→Es	En→Nl	Avg.
∅			ST	19.64	23.68	23.01	22.11
			Cascade§	23.65	28.68	27.91	26.75 (+4.64)
$D_{\mathbf{s,x,y}} \cup D_{\mathbf{s,x}} \cup D_{\mathbf{x,y}}$	ASR & MT	$D_{\mathbf{s,x,y}}$	ST	22.20	27.16	26.15	25.17 (+3.06)
			ST + ASR & MT	22.73	27.99	27.12	25.95 (+3.84)
	FAT-MLM		FAT-ST (base)	23.98	28.95	28.08	27.00 (+4.89)
			FAT-ST (big)	24.34	29.41	28.86	27.54 (+5.43)
$D_{\mathbf{s,x,y}} \cup D_{\mathbf{s,x}} \cup D_{\mathbf{x,y}}$ $\cup D_{\mathbf{s}} \cup D_{\mathbf{x}} \cup D_{\mathbf{y}}$	FAT-MLM		FAT-ST (base)	24.02	29.25	28.28	27.18 (+5.07)
			FAT-ST (big)	24.58	30.10	29.36	28.01 (+5.90)
$D_{\mathbf{s,x,y}} \cup D_{\mathbf{s,x}} \cup D_{\mathbf{x,y}}$ $\cup D_{\mathbf{s}} \cup D_{\mathbf{x}} \cup D_{\mathbf{y}}$	FAT-MLM	$D_{\mathbf{s,x,y}}$ $D_{\mathbf{s,x}} \cup D_{\mathbf{x,y}}$	FAT-ST (base)	23.91	29.01	28.18	27.03 (+4.92)
			FAT-ST (big)	**25.47**	**30.75**	**30.08**	**28.77 (+6.66)**
∅		$D_{\mathbf{s,x,y}} + D'_{\mathbf{s,y}}$	Pino et al. (2020)	25.2	-	-	-

Table 3. BLEU comparisons on Must-C test set between our proposed methods using additional data. $D_{\mathbf{s,x}}$: Librispeech, $D_{\mathbf{x,y}}$: Europarl MT, $D_{\mathbf{s}}$: libri-light, $D_{\mathbf{x}}, D_{\mathbf{y}}$: monolingual data from Europarl or Wiki Text. § are reported in Inaguma et al. (2020). Pino et al. (2020) use extra $D'_{\mathbf{s,y}}$ which includes Librispeech ($D_{\mathbf{s,x}}$) and 35,217 hour version of Libri-light speech data (almost $10\times$ of our $D_{\mathbf{s}}$) paired with their corresponding pseudo-translations generated by ASR and MT models. Their model size is 435.0M.

size for feed-forward layers of all transformer layer. For speech reconstruction module, we simply linearly project the outputs of the Transformer encoder to another latent space, then upsample the latent representation with 2-layers deconvolution to match the size of the original input signal. We choose 30% for the random masking ratio λ across all the experiments including pre-training. During inference, we do not perform any masking over the speech input. We average the last 5 checkpoints for testing. For decoding, we use a beam search with beam-size 5 and length penalty 0.6 for German, 0.0 for Spanish and 0.3 for Dutch.

5.3. Translation Quality Comparisons

We showcase the translation accuracy of FAT-ST comparing against to the baselines in Table 2 and Table 3:

- **ST**: this is the vanilla speech translation system which does not use transcriptions.

- **ST + ASR MTL**: ST model with an additional ASR decoder and is trained with ASR multi-task learning using the transcriptions.

- **ST + ASR & MT MTL**: ST model with an additional ASR decoder and a MT encoder. It is trained with ASR and MT multi-task learning.

- **ST + MAM**: ST trained with additional MAM loss (Chen et al., 2020) which can be formalized as $\ell_{\mathbf{s}}(D_{\mathbf{s}})$ (See Fig. 2(c)).

- **ST + MAM + ASR MTL**: ST trained with MAM loss and ASR multi-task learning.

- **Liu et al. (2020b)**: An end-to-end ST system with a multimodal encoder.

Model	# Parameters
MAM	23.69 M
FAT-MLM (base)	25.76 M
FAT-MLM (big)	38.36 M

Table 4. Models sizes of different models.

- **Le et al. (2020)**: The state-of-the-art end-to-end ST model with an extra ASR decoder.

- **Cascade**: cascaded model which first transcribes the speech into transcription then passes the results to a machines translation system.

- **ST + ASR & MT pretraining**: the encoder of ST is initialized by a pretrained ASR encoder and decoder initialized by a pretrained MT decoder

- **Pino et al. (2020)**: They propose to leverage additional speech data by generating pseudo-translations using a cascaded or an end-to-end speech translation model.

5.3.1. MODEL SIZE OF PRETRAINING MODELS

Table 4 shows the number of parameters of different pretraining models. We can see that our FAT-MLM base model is a little bit larger than the MAM pretraining model, and the FAT-MLM big model is much larger than the base model.

5.3.2. TRAINING WITH $D_{s,x,y}$

In Table 2, with no pretraining, we can see that our proposed FAT-ST base model achieves the best results except Le et al. (2020) and the cascaded model. However, our base model has much less parameters than both of them. Models with ASR or MT MTL and Liu et al. (2020b) all use the transcription data in Must-C dataset but show worse performance, thus our model can use transcription data more efficiently. Similar to other open source ST implementation results on Must-C [3], our implementation of ST + ASR & MT MTL is worse than ST + ASR.

We also compare the performance of models pretrained from different pretraining models. With pretrained on Must-C, FAT-ST (base) is improved by 0.85 BLEU by being finetuned from FAT-MLM, while it's performance drops by finetuning from MAM. Meanwhile, our proposed methods achieve much better performance compared with ASR & MT pretraining baselines. We also note that our FAT-ST base model for the first time achieves similar performances compared with Cascade baselines in these three translation directions of Must-C, while comparing with the cascaded model, our our base model is much smaller in size and faster in inference (see Fig. 7).

[3]ESPnet: https://github.com/espnet/espnet

Train Data	Pretrain Data	Models	→De	→Es	→Nl
$D_{s,x,y}$	No pretraining	MT[§]	27.63	32.61	32.08
		FAT-ST (base)	24.41	30.81	29.18
	$D_{s,x,y}$	FAT-ST (base)	27.24	31.98	31.27
		FAT-ST (big)	26.92	32.29	31.48
	$D_{s,x,y}$ $\cup D_{s,x} \cup D_{x,y}$	FAT-ST (base)	27.43	32.38	32.44
		FAT-ST (big)	27.60	32.95	32.37
	$D_{s,x,y} \cup D_{s,x} \cup D_{x,y}$ $\cup D_s \cup D_x \cup D_y$	FAT-ST (base)	27.63	32.75	32.52
		FAT-ST (big)	28.13	33.39	32.72
$D_{s,x,y}$ $\cup D_{s,x} \cup D_{x,y}$	$D_{s,x,y} \cup D_{s,x} \cup D_{x,y}$ $\cup D_s \cup D_x \cup D_y$	FAT-ST (base)	27.89	32.96	32.43
		FAT-ST (big)	28.80	34.28	34.22

Table 5. Comparisons of the auxiliary MT task between MT baselines and our proposed methods. [§] are reported in Inaguma et al. (2020).

5.3.3. PRETRAINING WITH ADDITIONAL DATA

Table 3 shows that FAT-MLM can further improve FAT-ST by simply adding speech recognition data $D_{s,x}$ (Librispeech) text machine translation data $D_{x,y}$ (Europarl) and even speech only data D_s (Libri-light) and monolingual text data $D_x \cup D_y$. This shows good representation learning ability of our proposed FAT-MLM models. We can see that using larger data, the performance of our big model is increased much faster than the base model. That's because the number of parameters of the base model is too limited to learn from such big data.

5.3.4. FINETUNING WITH ADDITIONAL DATA

The last part of Table 2 show that FAT-ST can be improved by learning from extra speech recognition and machine translation data. This is promising because speech translation data is very limited compared with much more abundant speech recognition and machine translation data. Different from Pino et al. (2020) who propose to leverage additional speech data by generating pseudo-translations, our method doesn't use any pseudo-labels. Our best model outperforms their result on En→De by using much $7\times$ smaller model size and almost $10\times$ smaller speech data.

Model	En→De
FAT-ST with FAT-MLM (base)	23.68
- FAT-MLM decoder init.	23.20
- FAT-MLM encoder init.	22.70
- CTC loss	22.30
- Hierarchical Transformer	22.07
- FAT-MLM loss	20.64
- MT loss	19.64

Table 6. Ablation study. Here, hierarchical transformer means the model only shares the 6 layers of the transformer encoder for acoustic feature input and text feature input.

Speech transcription	those are their expectations of who you are not yours
Target reference	那 是 他们 所期望的 你的 样子 而不是 你自己的 期望
	that is they expected your appearance not yourself expectation
Cascade-ASR	those are **there** expectations **to do** you are not yours
Cascade-Translation	那些 都是 希望 做到的， 你 不是 你的 。
	those are expect achievement you not yours
FAT-ST	这些 是 他们 对 你的 期望， 而不是 你的 期望 。
	these are they to your expectation not your expectation

Table 7. English-to-Chinese speech translation example. The cascaded system is our implementation using the TED training data. The errors of cascaded model is highlighted in red.

Models	En→Zh
KD (Liu et al., 2019)	19.55
LUT (Dong et al., 2020)	20.84
COSTT (Dong et al., 2021)	21.12
Cascade (Dong et al., 2020)	21.36
ST*	22.07
FAT-ST	23.73
FAT-MLM + FAT-ST	25.49

Table 8. BLEU comparisons on English-to-Chinese speech translation. * is our implementation. Cascaded model is implemented by Dong et al. (2020).

5.3.5. PERFORMANCE OF AUXILIARY MT TASK

Table 5 shows the translation quality of auxiliary MT task of FAT-ST. Although our models trained with Must-C are worse than the MT baseline, by using FAT-MLM trained with more data, our proposed methods can easily outperform the MT baseline. Note that these models' parameters are tuned to optimize speech translation task and MT is just an auxiliary task.

5.3.6. ABLATION STUDY

Table 6 shows an ablation study of our proposed method. we can see that all the components contribute to the final performance.

5.3.7. ENGLISH→CHINESE SPEECH TRANSLATION

We also compare several models in TED English→Chinese speech translation task (Liu et al., 2019) with 524 hours speech in training set, 1.5 hours validation set (dev2010) and 2.5 hours test set (tst2015). We follow our previous experiments to preprocess the data. Same with previous work, we evaluate the performance with character-level BLEU. Table 8 shows that our proposed model can largely outperform other baselines. Table 7 shows one example in this dataset. The translation of the cascaded model is wrong because of the errors in the its ASR (their→their, of who→ to do), while our FAT-ST produces the right translation.
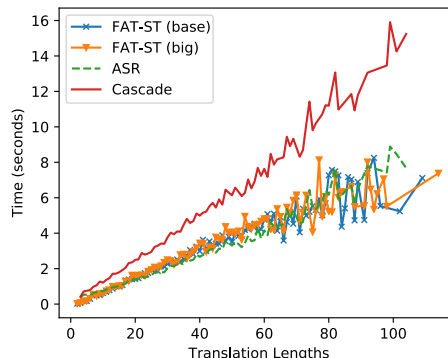

Figure 7. Decoding time comparison between Cascaded model (including its ASR) and FAT-ST.

5.3.8. DECODING SPEED

Fig. 7 shows the decoding speed comparison between the Cascade model and our proposed FAT-ST. Our proposed FAT-ST model is almost 2× faster than the Cascade system which needs to wait for the speech recognition module to finish before starting to translate. The decoding time of FAT-ST (big) is almost the same as FAT-ST (base) because we only increase the feedforward network in Transformers.

Conclusion

In this paper, we propose Fused Acoustic and Text Masked Language Model (FAT-MLM) which learns a unified representation for text and speech from any data that combines speech and text. We further extend this framework to a sequence-to-sequence speech translation model which enables learning from speech recognition and text-based machine translation data at the first time. Our results show significant improvement on three translation directions of the Must-C dataset and outperform the cascaded baseline.

Acknowledgements

References

Baevski, A., Zhou, H., Mohamed, A., and Auli, M. wav2vec 2.0: A framework for self-supervised learning of speech representations. *NeurIPS 2020*, 2020.

Bérard, A., Besacier, L., Kocabiyikoglu, A. C., and Pietquin, O. End-to-end automatic speech translation of audiobooks. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6224–6228. IEEE, 2018.

Chen, J., Ma, M., Zheng, R., and Huang, L. Mam: Masked acoustic modeling for end-to-end speech-to-text translation. *arXiv preprint arXiv:2010.11445*, 2020.

Chen, J., Ma, M., Zheng, R., and Huang, L. Direct simultaneous speech-to-text translation assisted by synchronized streaming asr. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics: Findings*, 2021.

Chuang, Y.-S., Liu, C.-L., Lee, H.-y., and Lee, L.-s. Speechbert: An audio-and-text jointly learned language model for end-to-end spoken question answering. *arXiv preprint arXiv:1910.11559*, 2019.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, 2019.

Di Gangi, M. A., Cattoni, R., Bentivogli, L., Negri, M., and Turchi, M. MuST-C: a Multilingual Speech Translation Corpus. In *NAACL*, 2019.

Dong, Q., Ye, R., Wang, M., Zhou, H., Xu, S., Xu, B., and Li, L. ” listen, understand and translate”: Triple supervision decouples end-to-end speech-to-text translation. *arXiv preprint arXiv:2009.09704*, 2020.

Dong, Q., Wang, M., Zhou, H., Xu, S., Xu, B., and Li, L. Consecutive decoding for speech-to-text translation. In *The Thirty-fifth AAAI Conference on Artificial Intelligence, AAAI*, 2021.

Graves, A., Fernández, S., Gomez, F., and Schmidhuber, J. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*, pp. 369–376, 2006.

Howard, J. and Ruder, S. Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*, 2018.

Inaguma, H., Kiyono, S., Duh, K., Karita, S., Soplin, N. E. Y., Hayashi, T., and Watanabe, S. Espnet-st: All-in-one speech translation toolkit. *arXiv preprint arXiv:2004.10234*, 2020.

Jia, Y., Johnson, M., Macherey, W., Weiss, R. J., Cao, Y., Chiu, C.-C., Ari, N., Laurenzo, S., and Wu, Y. Leveraging weakly supervised data to improve end-to-end speech-to-text translation. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7180–7184. IEEE, 2019.

Kahn, J., Rivière, M., Zheng, W., Kharitonov, E., Xu, Q., Mazaré, P. E., Karadayi, J., Liptchinsky, V., Collobert, R., Fuegen, C., Likhomanenko, T., Synnaeve, G., Joulin, A., Mohamed, A., and Dupoux, E. Librilight: A benchmark for asr with limited or no supervision. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7669–7673, 2020. https://github.com/facebookresearch/libri-light.

Kocabiyikoglu, A. C., Besacier, L., and Kraif, O. Augmenting librispeech with french translations: A multimodal corpus for direct speech translation evaluation. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

Koehn, P. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pp. 79–86. Citeseer, 2005.

Kudo, T. and Richardson, J. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 66–71, Brussels, Belgium, November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-2012. URL https://www.aclweb.org/anthology/D18-2012.

Lample, G. and Conneau, A. Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*, 2019.

Le, H., Pino, J., Wang, C., Gu, J., Schwab, D., and Besacier, L. Dual-decoder transformer for joint automatic speech recognition and multilingual speech translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 3520–3533, 2020.

Liu, A. T., Li, S.-W., and Lee, H.-y. Tera: Self-supervised learning of transformer encoder representation for speech. *arXiv preprint arXiv:2007.06028*, 2020a.

Liu, Y., Xiong, H., He, Z., Zhang, J., Wu, H., Wang, H., and Zong, C. End-to-end speech translation with knowledge distillation. *arXiv preprint arXiv:1904.08075*, 2019.

Liu, Y., Zhu, J., Zhang, J., and Zong, C. Bridging the modality gap for speech-to-text translation. *arXiv preprint arXiv:2010.14920*, 2020b.

Panayotov, V., Chen, G., Povey, D., and Khudanpur, S. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5206–5210. IEEE, 2015.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*, 2018.

Pino, J., Xu, Q., Ma, X., Dousti, M. J., and Tang, Y. Self-training for end-to-end speech translation. *Proc. Interspeech 2020*, pp. 1476–1480, 2020.

Povey, D., Ghoshal, A., Boulianne, G., Goel, N., Hannemann, M., Qian, Y., Schwarz, P., and Stemmer, G. The kaldi speech recognition toolkit. In *In IEEE 2011 workshop*, 2011.

Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I. Improving language understanding by generative pre-training. 2018.

Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016.

Sperber, M., Neubig, G., Niehues, J., and Waibel, A. Neural lattice-to-sequence models for uncertain inputs. *arXiv preprint arXiv:1704.00559*, 2017.

Sun, Y., Wang, S., Li, Y., Feng, S., Chen, X., Zhang, H., Tian, X., Zhu, D., Tian, H., and Wu, H. Ernie: Enhanced representation through knowledge integration. *arXiv preprint arXiv:1904.09223*, 2019.

Taylor, W. L. "cloze procedure": A new tool for measuring readability. *Journalism quarterly*, 30(4):415–433, 1953.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. R. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*, 2018.

Weiss, R. J., Chorowski, J., Jaitly, N., Wu, Y., and Chen, Z. Sequence-to-sequence models can directly translate foreign speech. *Proc. Interspeech 2017*, pp. 2625–2629, 2017.

Zheng, R., Ma, M., Zheng, B., Liu, K., Yuan, J., Church, K., and Huang, L. Fluent and low-latency simultaneous speech-to-speech translation with self-adaptive training. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pp. 3928–3937, 2020.