Multilingual Speech Translation from Efficient Finetuning of Pretrained Models

Xian Li*, Changhan Wang*, Yun Tang, Chau Tran, Yuqing Tang, Juan Pino, Alexei Baevski, Alexis Conneau, Michael Auli
Facebook AI
{xianl,changhan,yuntang,chau,yuqtang,juancarabina,
abaevski,aconneau,michaelauli}@fb.com

Abstract

We present a simple yet effective approach to build multilingual speech-to-text (ST) translation through efficient transfer learning from a pretrained speech encoder and text decoder. Our key finding is that a minimalistic LNA (**L**ayer**N**orm and **A**ttention) finetuning can achieve zero-shot crosslingual and cross-modality transfer ability by only finetuning $10 \sim 50\%$ of the pretrained parameters. This effectively leverages large pretrained models at low training cost such as wav2vec 2.0 for acoustic modeling, and mBART for multilingual text generation. This sets a new state-of-the-art for 36 translation directions (and surpassing cascaded ST for 30 of them) on the large-scale multilingual ST benchmark CoVoST 2 (Wang et al., 2020b) (+6.4 BLEU on average for En-X directions and +6.7 BLEU for X-En directions). Our approach demonstrates strong zero-shot performance in a many-to-many multilingual model (+5.6 BLEU on average across 28 directions), making it an appealing approach for attaining high-quality speech translation with improved parameter and data efficiency.

1 Introduction

Recent advances in pretraining over unlabeled data and then finetuning on labeled data leads to significant performance improvement in text understanding and generation tasks (Devlin et al., 2019; Liu et al., 2020; Conneau et al., 2019; Radford, 2018). Lately, such text pretraining and finetuning paradigms have been extended to other modalities: audio (Schneider et al., 2019; Baevski et al., 2020), images (Su et al., 2019; Lu et al., 2019), and video (Sun et al., 2019). At the same time, pretraining and finetuning techniques have improved multi-tasking applications significantly, such as multilingual translation, cross-lingual representations, question-answering and so on (Raffel et al., 2020; Yang et al., 2019; Tang et al., 2020). In this paper, we advance the one-model-for-all paradigm further by adapting audio and multilingual text pretraining and finetuning to improve multilingual speech-to-text translation.

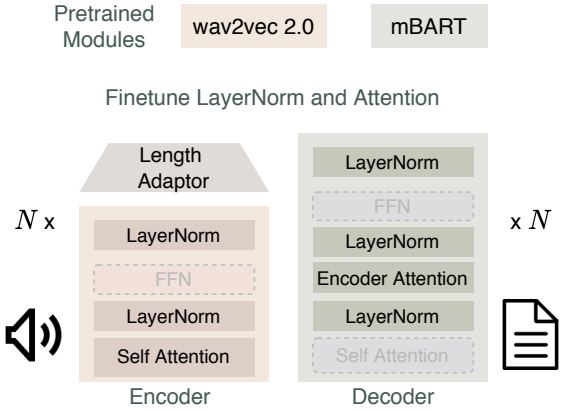

Figure 1: An overview of the proposed speech-to-text translation via transfer learning from efficient finetuning of single-modality pretrained models. The proposed LNA finetuning is applied to each layer.

Our contributions are as follows:

- We propose a simple and effective approach to combine pretrained single-modality modules to perform speech-to-text translation. With minimal architecture change, we add a cross-modal adaptor to bridge the length discrepancy between audio encoder output and text decoder input. Our approach can also perform multi-task finetuning with both speech-to-text translation and text-to-text translation tasks where we find joint training with the latter brings further gains.

- We present an efficient transfer learning strategy by only finetuning the **L**ayer**N**orm and **A**ttention (LNA) parameters of pretrained models. This approach is not only parameter- and data-efficient but also effective for zero-

Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, pages 827–838
August 1–6, 2021.

shot crosslingual transfer to unseen languages (train on A → B, test on A → C and C → B).

- Our approach is also effective for zero-shot multilingual translation (train on A → B and B → C, test on A → C), which provides an efficient approach for many-to-many speech-to-text translation without dependency for parallel data for every direction.

- Using a pretrained audio encoder (wav2vec (Baevski et al., 2020)) and multilingual text decoder (mBART (Liu et al., 2020)), this approach sets a new state-of-the-art (SOTA) on two large-scale speech translation benchmarks. On CoVoST 2 (Wang et al., 2020b), we pushed the SOTA for end-to-end approach for all 21 X-En directions(+6.7 BLEU on average) and 15 En-X directions (+6.4 BLEU on average) by finetuning only $10 \sim 50\%$ of parameters. Similarly on Europarl (Iranzo-Sánchez et al., 2020), our zero-shot multilingual many-to-many model is not only data efficient, but also brings +5.7 BLEU (on average) when translating 18 non-English directions compared to a many-to-many model training on 1.6× training data with all pairwise (both to/from English and non-English) directions.

We describe our approach in Section 2, namely pretrained models, length adaptor, LNA finetuning and joint speech-text finetuning as is illustrated in Figure 1. Experiments setup and results are elaborated in Section 3 and Section 4. Section 5 provides ablation studies of the proposed finetuning strategy.

2 Methods

2.1 Pretrained Modules

Our model leverages a pretrained wav2vec 2.0 (Baevski et al., 2020) as encoder for acoustic modeling, a pretrained multilingual BART (mBART) (Liu et al., 2020) as decoder for language modeling. Both models are pretrained on unlabelled data via self-supervised learning. We provide an overview of the pretraining procedure in A.1.

2.2 Length Adaptor

We add a lightweight adaptor module in between encoder and decoder to better align the two modules pretrained with different modalities. The adaptor module performs projection and downsampling to alleviate length inconsistency between the audio and text sequences. Specifically, we use a stack of n 1-dimensional convolutional layers with stride m to shrink the speech sequence (encoder output) by a factor of m^n.

2.3 LNA Finetuning

Instead of finetuning all parameters in pretrained models, we propose parameter efficient finetuning strategy (**LNA**) of only finetuning the layer normalization (**L**ayer**N**orm) and multi-head attention (MH**A**) parameters. LNA is motivated to bridge the discrepancy between pretraining and downstream (ST) task, which we hypothesize are accounted by the following parameters:

LayerNorm parameters from pretrained models were trained based on the statistics of the data used in pretraining and thus need to be adapted to downstream tasks during finetuning. The importance of finetuning LayerNorm has been observed in multilingual (text-only) translation (Stickland et al., 2020).

Attention Encoder attention (EA, attention to encoder outputs) parameters from pretrained MT decoder were trained on the text-to-text MT task, so we hypothesize that they are crucial to be adapted to the speech encoder output. Combined with LayerNorm parameter is the proposed LNA-Minimalist finetuning. In addition, we also investigate the role of self attention (SA) parameters in facilitating crosslingual transfer ability.

2.4 Joint Speech-text Finetuning

Multi-task learning has been shown as an effective approach to improve the performance of the speech translation task using other related tasks, such as MT and ASR (Weiss et al., 2017; Anastasopoulos and Chiang, 2018; Bahar et al., 2019; Tang et al., 2021a,b). We jointly train MT and ST tasks in the finetuning with pretrained models. The speech transcripts are used as input for the MT task and the corresponding speech data is used as input for the ST task. As a result, we can leverage abundant parallel text data to further improve the performance.

3 Experimental Setup

3.1 Datasets

We evaluate our proposed models on two large-scale multilingual speech translation benchmarks.

Statistics of the datasets and implementation details are reported in the A.2 and A.3.

CoVoST 2 (Wang et al., 2020b) is a multilingual speech-to-text translation corpus with English into 15 languages (En-X) and 21 languages into English (X-En). It provides a comprehensive test bed for low-resource scenarios, with 4 X-En directions between 10 hours and 20 hours training data, and 11 X-En directions less than 4 hours training data.

Europarl ST (Iranzo-Sánchez et al., 2020) has both English-centric as well as non-English directions, which allow us to evaluate the proposed method's effectiveness of multilingual translation between any pair, especially zero-shot performance. We experiment on all 6 languages (de, en, es, fr, it, pt). We compare to a multilingual baseline trained with all pair-wise parallel data.

3.2 Training

We evaluate the following instantiation of the proposed method which is referred to as XMEF (Cross-Modal Efficient Finetuning).

Encoder. We initialize the encoder using the open-sourced[1] wav2vec 2.0 large architecture pretrained on unlabelled English-only (XMEF-En) audio from LibriVox (Baevski et al., 2020). For many-to-one experiments, we also experiment with a multilingual wav2vec 2.0 (XMEF-X), which was pretrained on raw audio from 53 languages (Conneau et al., 2020). Encoder output is followed by 3 1-D convolution layers with stride 2 to achieve 8x down-sampling of audio encoder outputs.

Decoder. We initialize the decoder with open-sourced[2] mBART50 models and the same vocabulary (Tang et al., 2020). We use mBART50N1 (49 languages to English) for X-En ST directions and mBART501N (English to 49 languages) for translating En-X ST directions.

LNA Finetuning. We study the parameter efficiency and crosslingual transfer ability of LNA finetuning in the bilingual setting without the additional effect from multilingual training. Drawing learnings on that, we then evaluate applying LNA finetuning to encoder only (LNA-E), decoder only (LNA-D), and both (LNA-E,D) respectively. For multilingual finetuning on CoVoST 2, we use all X-En training data (except zero-shot crosslingual transfer experiments) for evaluating X-En performance, and En-X data from all directions for evaluating En-X performance. For evaluating multilingual zero-shot performance on Europarl, we only use X-En and En-X for finetuning and evaluate on all (X-X) pairs.

[1] https://github.com/pytorch/fairseq/tree/master/examples/wav2vec.

[2] https://github.com/pytorch/fairseq/tree/master/examples/multilingual

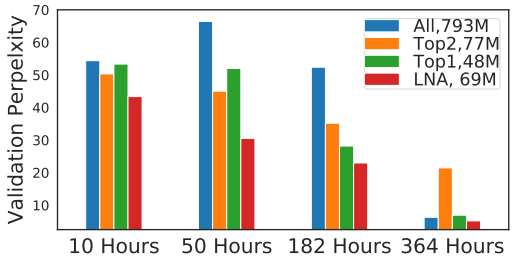

Figure 2: Comparison of LNA finetuning with alternative finetuning strategies: finetuning all parameters (All), finetuning top-k layers (Top1, Top2). We evaluate generalization (perplexity on dev set) performance with different amounts of training data. LNA achieves the best generalization with substantially less parameters. Experiments are done using CoVoST En-De.

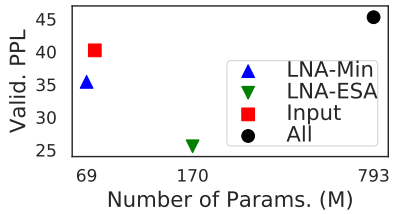

Figure 3: Comparison of LNA finetuning with alternatives: finetuning all parameters (All) and finetuning feature extractor (Input), on adapting wav2vec English encoder to translate non-English speech input. Experiments are done using CoVoST De-En.

Joint Training. Two encoders are initialized with the pretrained mBART encoder and wav2vec 2.0 encoder mentioned above, and are used for text and speech input respectively. The last 12 transformer layers in the wav2vec encoder are replaced with 12 mBART encoder layers. Parameters in those 12 layers are shared between the two encoders during joint training (Tang et al., 2021b). The decoder is also shared between two tasks and is initialized with the pretrained mBART decoder model. We also experimented with adding additional bitext used in ML50 (Tang et al., 2020) as training data for the MT task. Only the language pairs present in the CoVoST 2 dataset are chosen and they cover all language pairs except English to and from "Ca" and "Cy". We fine-tune all parameters in this experiments due to the large mismatch of the pretrained model (mBART encoder as part of the speech encoder) and more available training data.

3.3 Baselines

From scratch: The first baseline trains a sequence-to-sequence model with Transformer architecture without any pretraining.For CoVoST 2 experiments, we use the same model configuration as is provided by (Wang et al., 2020b).

ASRPT+Multi: Pretraining encoder on ASR task was shown to be an effective method to improve speech translation and accelerates convergence (Bansal et al., 2019). We compare our results to a strong baseline provided by (Wang et al., 2020b), consisting of a multilingual Transformer model trained on CoVoST 2 with multilingual ASR pretraining (ST). For the Europarl ST many-to-many baseline, we use Transformer architecture with 12-layer encoder, 6-layer decoder, and trained on all 30 directions. To provide the strongest baseline, encoder was pre-trained on LibriSpeech English ASR).

XMEF-BL: Multilingual models for En-X (one-to-many) usually face more challenges from interference as they were found to underperform the bilingual counterparts (Arivazhagan et al., 2019). Therefore, we compare to applying our method (XMEF, LNA) to bilingual (BL) finetuning, i.e. finetuning on parallel data from a single language pair.

Previous SOTAs: We compare to the best end-to-end (E2E) model from previous literature (Wang et al., 2020b; Iranzo-Sánchez et al., 2020) on each translation direction, which is usually the best-performing multilingual model trained with parallel data from all directions (both X-En and En-X) and also pretrained with ASR. Even though the focus of the proposed method is E2E model, we also compare to the best performing cascade approach (Cascade SOTA) which is composed of Transformer-large encoder from ASR pretraining and a multilingual MT model trained on all X-En and En-X data.

4 Results

4.1 Parameter Efficiency

First, we evaluate the transfer learning performance of finetuning the entire pretrained model as well as the proposed efficient finetuning (LNA). To separate the additional crosslingual transfer learning from multilingual finetuning, we evalute on bilingual ST (En-De and De-En in CoVoST) task. We first evaluate LNA-Minimalist (69M params), comparing to finetuning all parameters and only top layers which were found effective in transfer learning in NLP tasks with pretrained BERT (Wu and Dredze, 2019; Kovaleva et al., 2019). Figure 2 show that in both low data and high data regimes, the proposed LNA-Minimalist both generalizes better (lower perplexity on dev set) and substantially improves training efficiency (only 10% of parameters to train leading to lower memory cost and faster training).

4.2 Transfer from Pretraining

To assess transfer ability from encoder pretrained on English to other (speech) input languages, we evaluate the performance of XMEF-En on CoVoST 2 De-En ST task. We investigate the role of finetuning encoder self-attention (LNA-ESA) in facilitating crosslingual transfer. We compare to baselines of finetuning the entire encoder (All), and finetuning feature extractor which are commonly used in adaptation in ASR (Rivière et al., 2020). Results are summarized in Figure 3. LNA still demonstrates improved generalization than alternative finetuning approaches, with finetuning encoder self attention (LNA-ESA) being crucial for adapting pretrained English encoder to other languages.

4.3 Zero-shot Crosslingual Transfer

Next, we evaluate XMEF's crosslingual transfer performance from multilingual finetuning. To precisely measure the transfer capability, we evaluate the zero-shot setting, i.e. finetune XMEF-En with parallel ST data from multiple languages, and evaluate on an unseen language. We study the transfer performance in source (speech) and target (text) separately.

Source-side (speech) transfer. We evaluate whether the proposed approach enables positive crosslingual transfer to translate speech *from* unseen languages in Table 1. We finetune on labelled data for 5 to-English language pairs, and evaluate the finetuned model's zero-shot performance when translating speech input from unseen languages (Pt). First, we found that comparing to finetuning more parameters (LNA-D, and All), LNA finetuning (LNA-E,D) not only trains more than 2 $\times$ faster but also achieves better generalization both for seen and unseen languages. Especially, it attains remarkable performance as unsupervised speech translation for Portuguese-English, achieving 8.2 BLEU (compared to the supervised bilingual baseline 0.5 BLEU as is provided in Table 3, and even beats

				Train					Zero-shot
	Enc	Dec	Params.	Fr	De	Es	Ca	It	Pt
LNA-E,D	LN+SA	LN+EA	170.7M	**32.4**	**24.9**	**31.6**	**28.6**	**24.0**	**8.2**
LNA-D	All	LN+EA	384.8M	31.6	23.7	31.0	27.8	23.2	7.6
Finetune All	All	All	793.0M	27.1	17.7	27.8	21.7	18.9	5.1
			ASRPT+Multi	23.1	15.3	21.2	19.9	14.9	4.4
Supervised (Multi) SOTA			(Wang et al., 2020b)	26.5	17.6	27.0	23.1	18.5	6.3

Table 1: Performance on zero-shot transfer on the source-side (speech). Each model is finetuned on 5 directions from {*Fr, De, Es, Ca, It*} → *En*, and evaluated on **unsupervised translation** of a new language (*Pt*). We report BLEU scores on test set, and compare to the zero-shot transfer performance of a supervised multilingual baseline (ASRPT+Multi), as well as previous state-of-the-art which is also the supervised and multilingually trained.

				Train				Zero-shot
	Enc	Dec	Params.	De	Fa	Tr	Zh	Ja
LNA-E,D	LN	LN+EA	69.4M	22.1	17.7	13.4	29.2	22.9
LNA-E,D	LN+SA	LN+EA	170.7M	23.8	19.2	14.2	30.6	29.2
LNA-D	All	LN+EA	384.8M	**24.9**	**19.8**	15.2	**32.7**	**30.6**
LNA-E	LN+SA	All	477.6M	22.0	18.1	14.2	29.5	0.8
Finetune All	All	All	793.0M	24.1	19.6	**15.6**	32.4	0.4
			ASRPT+Multi	9.5	10.9	6.8	23.5	0.0
Supervised (Multi) SOTA			(Wang et al., 2020b)	17.3	14.5	10.7	28.2	31.9

Table 2: Performance on zero-shot transfer on the target-side (text). Each model is finetuned on 4 directions *En* → {*De, Fa, Tr, Zh*}, and evaluated on **unsupervised translation** to a new language (*Ja*). We report BLEU scores on test set, and compare to the zero-shot transfer performance of a supervised multilingual baselines (ASRPT+Multi), as well as previous state-of-the-art which is also the supervised and multilingually trained.

(+1.9 BLEU) the previous state-of-the-art for this direction which is a supervised multilingual model.

Target-side (text) transfer. Table 2 shows the proposed approach also achieves zero-shot transfer capability for translating *to* new languages, with unsupervised translation for English-Japanese only 1.3 BLEU behind the best supervised result. Furthermore, an interesting finding is that applying LNA finetuning to decoder is crucial for zero-shot transfer to unseen languages (Ja), as finetuning the entire decoder tends to optimize the model on target languages seen during training.

4.4 Multilingual Speech Translation

We evaluate the performance of XMEF with multilingual finetuning on all 36 translation directions in CoVoST 2, respectively all 21 languages into English (many-to-one) and from English into 15 languages (one-to-many).

Many to one. Consistent with the observation of source-side crosslingual transfer in Sec 4.1, XMEF-En perform very well on Romance, Germanic and Slavic language families in both high-resource ($\geq$ 100 hours training data) and low-resource directions (7 $\sim$ 44 hours training data) as is summarized in Table 3, and even surpassing the best cascade results on 8 languages. Our multilingual model also improves distant (from English) and extremely low resource (mostly $\leq$ 5 hours training data) languages as is shown in second panel of Table 3. For crosslingual adaptation from XMEF-En to speech input of other languages, LNA-E,D (only finetune 21.5% of pretrained parameters) outperforms finetuning the entire model (Finetune All) by 0.7 BLEU (averaged across 21 directions), while finetuning the entire encoder (LNA-D) brings +1.2 BLEU. Finetuning XMEF-X achieves the best average BLEU score, however, major improvement is from finetuning encoder (LNA-D).

One to many. Table 4 summarizes performance on translating (from English) to 15 languages where multilingual models from XMEF-En have improved previous state-of-the-art (both E2E and cascade) on all directions (+6.4 BLEU on average). The performance of applying LNA finetuning to encoder only (LNA-E) is very close to (24.2 vs. 24.5 averaged BLEU) that of finetuning the entire model (Finetune All) while has 40% less parameters to train. Applying LNA to both encoder and decoder (LNA-Min, LNA-E,D) further reduces the amount of parameters to train to only 8 $\sim$ 20% of all parameters in the pretrained models yet still maintain strong performance compared to strong baselines such as ASR PT with multilingual finetuning (ASR PT+Multi) as well as the best cascade models. The only two languages (Ca, Cy) it did not

		High Resource				Low Resource					
	→ **En**	Fr	De	Es	Ca	It	Ru	Pt	Nl	Sl	Sv
	Train Hours	264	184	113	136	44	18	10	7	2	2
	Scratch-BL	24.3	8.4	12.0	14.4	0.2	1.2	0.5	0.3	0.3	0.2
	+ ASR PT	26.3	17.1	23.0	18.8	11.3	14.8	6.1	3.0	3.0	2.7
	+ Multi.	26.5	17.5	27.0	23.1	18.5	4.7	6.3	5.0	0.7	0.5
	+mBART	28.1	19.7	28.1	24.0	19.9	2.7	6.2	8.1	0.5	1.4
XMEF-En	LNA-E,D (170.7M)	**33.8***	**26.7***	**34.0***	**29.5***	**26.1***	**21.1**	**19.2**	**14.1***	**4.6**	**5.9**
XMEF-En	LNA-D (384.8M)	**35.0***	**28.2***	**35.2***	**31.1***	**27.6***	**22.8**	**24.1***	**14.2***	**5.0**	**5.0**
XMEF-En	Finetune All (793.0M)	**33.0***	**24.5***	**33.6***	**28.0***	**25.2***	**20.2**	**19.5**	**9.4**	**4.6**	**4.8**
XMEF-En	Joint Training (1.05B)	**33.5***	**28.6***	**33.5***	**30.6***	**26.6***	**17.6**	**12.0**	**15.0***	**3.9**	2.6
XMEF-En	+ Extra MT Data	**34.4***	**29.6***	**34.4***	**30.6***	**27.7***	**27.7***	**14.6**	**14.5***	**5.2**	**3.4**
XMEF-X	LNA-E,D (170.7M)	**32.8***	**28.6***	**34.0***	**29.7***	**27.9***	**25.1**	**19.5**	**24.1**	**3.0**	**4.0**
XMEF-X	LNA-D (384.8M)	**34.2***	**30.8***	**35.8***	**31.7***	**29.4***	**26.5**	**19.6**	**25.7**	**4.3**	**3.2**
XMEF-X	Finetune All (793.0M)	**36.1***	**30.6***	**38.1***	**31.8***	**31.9***	**30.9**	**20.7**	**24.0**	**5.6**	**4.0**
	Prev. E2E SOTA	27.0	18.9	28.0	24.0	11.3	14.8	6.1	8.4	3.0	2.7
	Cascade SOTA	29.1	23.2	31.1	27.2	22.9	25.0	22.7	10.4	7.0	11.9

		Fa	Zh	Tr	Et	Mn	Ar	Lv	Cy	Ta	Ja	Id	Avg.
	→ **En**	Fa	Zh	Tr	Et	Mn	Ar	Lv	Cy	Ta	Ja	Id	Avg.
	Train Hours	49	10	4	3	3	2	2	2	2	1	1	
	ASR (WER)	62.4	45.0	51.2	65.7	65.2	63.3	51.8	72.8	80.8	77.1	63.2	
	Baseline	1.9	1.4	0.7	0.1	0.1	0.3	0.1	0.3	0.3	0.3	0.4	
	+ ASR PT	3.7	5.8	3.6	0.1	0.2	4.3	2.5	2.7	0.3	1.5	2.5	
	+ Multi.	2.4	5.9	2.3	0.6	0.1	0.4	0.6	1.9	0.1	0.1	0.3	7.0
	+ mBART	3.3	5.4	2.4	0.7	0.2	0.5	0.6	1.4	0.1	0.2	0.2	7.3
XMEF-En	LNA-E,D (170.7M)	**4.0**	**6.2**	**5.5**	**1.3**	**1.0**	3.7	**4.6**	2.8	**0.7**	**1.7**	**2.9**	11.9
XMEF-En	LNA-D (384.8M)	3.6	**6.0**	**4.8**	**1.5**	**0.9**	2.8	**4.9**	2.3	**0.8**	**1.7**	**3.7**	12.4
XMEF-En	Finetune All (793.0M)	3.7	**6.5**	**4.0**	**1.4**	**1.0**	3.3	**4.9**	2.1	**0.5**	**2.1**	**3.4**	11.2
XMEF-En	Joint Training (1.05B)	**6.1***	5.4	3.3	0.7	0.2	0.8	**2.7**	1.0	0.1	0.3	0.5	10.7
XMEF-En	+ Extra MT Data	**5.0**	**6.2**	**4.0**	0.8	0.3	1.0	**3.6**	1.1	0.2	0.5	0.5	11.7
XMEF-X	LNA-E,D (170.7M)	**6.6***	**8.0**	**8.6**	**2.0**	**1.0**	**6.2**	**3.3**	**4.9**	**0.6**	0.8	2.3	13.0
XMEF-X	LNA-D (384.8M)	**11.0***	**9.1**	**11.2***	**2.9**	**1.1**	**6.2**	**3.8**	**9.0***	**0.7**	0.8	2.3	14.3
XMEF-X	Finetune All (793.0M)	**8.5***	**8.9**	**9.4***	**2.5**	**1.2**	**6.4**	**5.0**	**8.1***	**0.9**	1.0	2.8	14.7
	Prev. SOTA	3.7	5.9	3.7	0.9	0.2	4.3	2.5	3.3	0.3	1.5	2.5	
	Cascade	5.8	11.4	9.3	3.8	1.0	12.3	7.2	7.4	0.4	3.8	11.8	

Table 3: Performance of $X \rightarrow En$ multilingual model. We report BLEU scores on test set. For each XMEF method, we report the number of parameters trained in brackets. Previous E2E SOTA is the best-performing end-to-end multilingual (with ASR pretraining) model from (Wang et al., 2020b). Results in **bold** are where the proposed approach improves previous E2E SOTA, and sets new SOTA as underlined. * means our new E2E SOTA also beats the previous cascade SOTA.

improve with LNA finetuning of the decoder were never seen during mBART pretraining.

Joint Training In the many to one case (Table 3), language pairs with reasonable amount speech training data (+ 18 hours) and large amount of parallel text data (+1 million sentences) ("Fr-En", "De-En", "Es-En", "It-En", "Ru-En" and "Fa-En"), outperform the corresponding single task trained models and achieve state-of-art results . However, if the amount of speech data is too small (10 hours or less), joint training is ineffective and may even make the performance worse. In one to many case ("En-X"), where there are 364 hours English audio data for training, joint training improves the results further by another 0.6 BLEU (Table 4).

4.5 Zero-shot Many-to-Many Speech to Text Translation

Finally, we evaluate how the proposed approach performs in zero-shot multilingual translation (translating X → Y after training on X → En and En → Y. We apply LNA-D multilingual finetuning using En-X and X-En training data only from the Europarl corpus. Table 5 reports both the supervised performance on to- and from-English directions and zero-shot performance translating between non-Engligh languages without training on their parallel data. We compare to the strong baseline of a many-to-many multilingual model trained from scratch *using all parallel data* from non-English directions as well as English-centric directions. Our approach improves both to- and from-English directions (+6.8 BLEU and +8.2 BLEU on

En →	Ar	Ca	Cy	De	Et	Fa	Id	Ja
Scratch-BL	8.7	20.2	22.2	13.6	11.1	11.5	18.9	26.9
+ ASR PT	12.1	21.8	23.9	16.5	13.4	13.5	20.8	29.6
+ Multi.	13.0	22.3	23.7	17.3	13.9	14.5	20.3	31.9
LNA-Min-BL (69.4M)	12.0	18.8	12.9	**20.3***	15.0	**15.9***	**24.4***	31.4
LNA-Min (69.4M)	**15.3***	20.3	13.2	**23.2***	**18.6***	**19.6***	**26.5***	**36.9***
LNA-E,D (170.7M)	**17.4***	22.2	14.8	**25.3***	**21.0***	**20.1***	**27.6***	**38.4***
LNA-E (477.6M)	**17.2***	**29.5***	**30.3***	**25.2***	**20.7***	**19.8***	**28.5***	**37.8***
Finetune All (793.0M)	**17.7***	**30.1***	**30.0***	**25.2***	**21.1***	**20.3***	**28.9***	**38.1***
Joint Training (1.05B)	**18.0***	**30.9***	**30.6***	**25.8***	**22.1***	**21.5***	**29.9***	**39.3***
+ Extra MT Data	**18.6***	**30.4***	**29.2***	**26.6***	**21.4***	**20.6***	**28.8***	**39.1***
Prev. E2E SOTA	13.9	23.6	25.1	18.4	15.1	15.5	22.0	33.0
Cascade SOTA	14.3	25.0	25.6	19.4	15.4	14.1	23.1	33.8

En →	Lv	Mn	Sl	Sv	Ta	Tr	Zh	Avg.
Scratch-BL	11.5	6.6	11.5	20.1	9.9	8.9	20.6	
+ ASR PT	13.1	9.2	16.1	22.3	11.2	10.2	25.7	
+ Multi.	14.1	10.2	17.1	22.3	11.7	10.7	28.2	18.1
LNA-Min-BL (69.4M)	14.3	6.9	17.9	**26.1***	12.6	10.8	21.8	
LNA-Min (69.4M)	**17.9***	**12.0***	**21.1***	**27.5***	**14.6***	**14.1***	**32.1***	20.9
LNA-E,D (170.7M)	**20.1***	**13.3***	**23.0***	**29.6***	**16.4***	**15.5***	**33.0***	22.5
LNA-E (477.6M)	**20.2***	**14.1***	**23.5***	**30.0***	**16.8***	**16.2***	**32.8***	24.2
Finetune All (793.0M)	**20.8***	**14.1***	**23.6***	**30.4***	**17.1***	**16.3***	**33.7***	24.5
Joint Training (1.05B)	**21.5***	**14.8***	**25.1***	**29.6***	**17.8***	**17.0**	**33.3**	25.1
+ Extra MT Data	**21.3***	**14.7***	**24.8***	**30.0***	**17.4***	**16.3***	**34.4***	24.5
Prev. E2E SOTA	15.2	11.0	18.3	24.1	12.8	11.7	31.3	
Cascade SOTA	15.6	11.7	18.9	24.8	13.7	11.7	26.9	

Table 4: Performance on *En* → *X* multilingual ST. We report BLEU scores on test set. For each XMEF method, we report the number of parameters trained in brackets. 'BL' refers to using the same XMEF and LNA-E,D finetuning but only on bilingual corpus. Results in **bold** are where the proposed approach improves previous E2E SOTA, and sets new SOTA as is underlined. * means our new E2E SOTA also beats the previous cascade SOTA. For multilingual models (i.e. the same model evaluated on multiple directions), we also report the average (Avg.) BLEU scores across *all 15* directions.

averge respectively) and our zero-shot results also beats (+5.6 BLEU) the supervised many-to-many model on 28 pair-wise (except for It-Pt and Pt-Es) translation directions.

5 Ablation Studies

Ablation on LNA Finetuning. In Table 6 we analyze how individual components of LNA contribute to the generalization performance and training efficiency. Specifically, we examine the key components of LNA-Minimalist (LNA-Min) finetuning. We find finetuning LayerNorm parameter (far less compared to the amount of multi-head attention parameters) is important for training stability when finetuning pretrained models without which (-LN) training diverges. Finetuning the encoder attention (EA) parameters is important for adapting the pretrained text decoder for ST task. For adapting to a single language pair downstream ST task (English-German), we find finetuning self attention (+SA) parameters in the decoder did not bring further improvement while significantly increasing the amount of parameters to train.

Ablation on Length Adaptor. We study whether the performance is sensitive to downsampling ratio in the adaptor module. We conduct the experiments on CoVoST 2 many-to-one experiments, and report perplexity on dev set of three directions with diverse input languages: German-English (De-En), Chinese-English (Zh-En) and Estonian-English (Et-En). Table 7 shows our approach is not sensitive to common downsampling ratios (4 or 8) while extreme downsampling (27) hurts performance.

6 Related Work

Speech Translation. Sequence-to-sequence based speech translation has shown very good potential over the traditional cascaded system (Berard et al., 2016; Goldwater et al., 2017; Weiss et al., 2017) with end-to-end approaches surpassing cascaded system for the first time at IWSLT (Ansari et al., 2020) in a shared task setting. However, previous work also indicates that its success heavily relies on large amounts of labelled training data, which is difficult to acquire. In order

		Target					
		De	En	Es	Fr	It	Pt
Source	De		12.8/**20.6**	10.2/**13.8**	11.6/**14.9**	6.6/**8.6**	10.4/**13.0**
	En	13.1/**22.5***		23.1/**32.3***	22.1/**30.0***	14.9/**21.5**	20.7/**28.4**
	Es	9.2/**12.1**	18.9/**26.0**		19.0/**21.8**	13.3/**15.4**	20.0/**21.9**
	Fr	9.8/**13.6**	19.8/**27.9***	18.6/**21.7**		13.8/**15.2**	19.7/**21.4**
	It	10.1/**11.9**	19.8/**25.6**	18.8/**20.8**	19.1/**20.0***		19.8/19.2
	Pt	9.0/**11.4**	19.0/**24.1**	19.8/19.6	18.1/**18.6**	15.6/**16.1**	

Table 5: Zero-shot performance (baseline/XMEF) on Europarl. Baseline is a many-to-many multilingual model trained on parallel data from all 30 directions. For our approach (XMEF), only to- and from- English directions (shaded) were used in multilingual finetuning while the rest are results of zero-shot translation. **Bold** are where our model (En-only and zero-shot for the rest) outperforms a supervised many-to-many model. * means that our zero-shot model also beats the supervised cascade model in (Iranzo-Sánchez et al., 2020).

Enc	Dec	PPL ↓	Params (%)
LN	LN + EA	5.17	69.4M (8.8%)
- LN	- LN	37.66	69.3M (8.7%)
	- EA	5.97	19.0M (2.4%)
	+ SA	5.26	119.8M (15.1%)
+ SA		5.53	170.2M (21.5%)

Table 6: Ablation on LNA-Minimalist finetuning, where we evaluate the effect of finetuning LayerNorm (**LN**) and **A**ttention parameters. The experiment was conducted on the CoVoST English-German dataset and we report perplexity on the dev set. **%** indicates what percentage of total parameters of pretrained modules are trained during finetuning.

# Layers	Stride	De	Zh	Et
3	2	5.88	26.72	35.46
2	2	5.79	25.92	34.01
3	3	11.92	32.07	42.33

Table 7: Ablation on length adaptor with different downsampling ratios of speech input. The experiment was conducted on the CoVoST X-English multilingual finetuning and we report perplexity on the dev set (PPL ↓) for three distinct languages.

to mitigate the data scarcity issue, recent research work focuses on multi-task learning (Weiss et al., 2017; Anastasopoulos and Chiang, 2018; Bahar et al., 2019; Wang et al., 2020c,d; Indurthi et al., 2020; Di Gangi et al., 2019), pretraining different components of the model (Bérard et al., 2018; Bansal et al., 2019), transfer learning (Gaido et al., 2020; Liu et al., 2019) and generating synthetic data (Jia et al., 2018; Pino et al., 2020).

Pretraining and Finetuning. Our work is motivated by the recent success of self-supervised learning for NLP and speech processing applications (Radford, 2018; Devlin et al., 2019; Clark et al., 2019; Lewis et al., 2019; Lample and Conneau, 2019; Dong et al., 2019; Liu et al., 2020; Tang et al., 2020; Rivière et al., 2020; Kawakami et al., 2020; Chung and Glass, 2020; Baevski et al., 2020), which has achieved state-of-the-art results when finetuning on downstream tasks in NLP (Liu et al., 2020; Devlin et al., 2019; Raffel et al., 2020; Tang et al., 2020). Our work attempts to leverage pretrained components from different modalities (text and speech) to perform the ST task. How to efficiently adapt large pretrained models has gained growing interest. (Houlsby et al., 2019) and (Pfeiffer et al., 2020) represent the stream of work which adds additional "adaptor modules" to achieve fast adaptation to downstream tasks. Another category of solutions focus selective finetuning (only subset of parameters) suitable for downstream tasks. Our work belongs to the second category of efficient finetuning without adding extra parameters (e.g. adaptor modules). Empirical studies shows that finetuning the final layers of BERT account for most of the quality gains on downstream tasks (Kovaleva et al., 2019; Lee et al., 2019). Finetuning LayerNorm parameters was also found effective for adapting pretrained BART or mBART for machine translation (Stickland et al., 2020). A general approach is to automatically learn which layers/parameters from a large-pretrained model to finetune and freeze (Guo et al., 2019), which we found is an exciting direction for future work.

7 Conclusion

We proposed a simple and effective approach to leverage pretrained single-modality models (such as wav2vec 2.0, mBART) to perform speech-to-text translation. On two large-scale multilingual speech translation benchmarks, our approach advances the state-of-the-art (+6.6 BLEU on average for 36 translation directions in CoVoST 2, and +5.6 BLEU for 28 translation directions in Europarl).

We provide an efficient finetuning strategy which is not only data- and parameter-efficient, but also demonstrates crosslingual transfer ability by only finetuning $10 \sim 50\%$ of the parameters of large pretrained models.

References

Antonios Anastasopoulos and David Chiang. 2018. Tied multitask learning for neural speech translation. In *NAACL-HLT*.

Ebrahim Ansari, Amittai Axelrod, Nguyen Bach, Ondřej Bojar, Roldano Cattoni, Fahim Dalvi, Nadir Durrani, Marcello Federico, Christian Federmann, Jiatao Gu, Fei Huang, Kevin Knight, Xutai Ma, Ajay Nagesh, Matteo Negri, Jan Niehues, Juan Pino, Elizabeth Salesky, Xing Shi, Sebastian Stüker, Marco Turchi, Alexander Waibel, and Changhan Wang. 2020. FINDINGS OF THE IWSLT 2020 EVALUATION CAMPAIGN. In *Proceedings of the 17th International Conference on Spoken Language Translation*, pages 1–34, Online. Association for Computational Linguistics.

Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Dmitry Lepikhin, Melvin Johnson, Maxim Krikun, Mia Xu Chen, Yuan Cao, George Foster, Colin Cherry, et al. 2019. Massively multilingual neural machine translation in the wild: Findings and challenges. *arXiv preprint arXiv:1907.05019*.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations.

Parnia Bahar, Tobias Bieschke, and Hermann Ney. 2019. A comparative study on end-to-end speech to text translation. In *ASRU*.

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 *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 58–68, Minneapolis, Minnesota. Association for Computational Linguistics.

Alexandre Bérard, Laurent Besacier, Ali Can Kocabiyikoglu, and Olivier Pietquin. 2018. End-to-end automatic speech translation of audiobooks. In *ICASSP*.

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. In *NIPS*.

Yu-An Chung and James Glass. 2020. Improved speech representations with multi-target autoregressive predictive coding. In *ACL*.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2019. Electra: Pre-training text encoders as discriminators rather than generators.

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. 2020. Unsupervised cross-lingual representation learning for speech recognition. *arXiv preprint arXiv:2006.13979*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*.

Mattia A Di Gangi, Matteo Negri, and Marco Turchi. 2019. One-to-many multilingual end-to-end speech translation. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 585–592. IEEE.

Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2019. Unified language model pre-training for natural language understanding and generation. In *NeurIPS*.

Marco Gaido, Mattia Antonino Di Gangi, Matteo Negri, and Marco Turchi. 2020. End-to-end speech-translation with knowledge distillation: Fbk@iwslt2020.

S. Goldwater, Adam Lopez, Sameer Bansal, and H. Kamper. 2017. Towards speech-to-text translation without speech recognition. In *EACL*.

Yunhui Guo, Honghui Shi, Abhishek Kumar, Kristen Grauman, Tajana Rosing, and Rogerio Feris. 2019. Spottune: transfer learning through adaptive fine-tuning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4805–4814.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. *arXiv preprint arXiv:1902.00751*.

Sathish Reddy Indurthi, HouJeung Han, Nikhil Kumar Lakumarapu, Beom seok Lee, Insoo Chung, Sang-Ha Kim, and Chanwoo Kim. 2020. End-end speech-to-text translation with modality agnostic meta-learning. In *ICASSP*.

J. Iranzo-Sánchez, J. A. Silvestre-Cerdà, J. Jorge, N. Roselló, A. Giménez, A. Sanchis, J. Civera, and A. Juan. 2020. Europarl-st: A multilingual corpus for speech translation of parliamentary debates. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8229–8233.

Ye Jia, Melvin Johnson, Wolfgang Macherey, Ron J. Weiss, Yuan Cao, Chung-Cheng Chiu, Naveen Ari, Stella Laurenzo, and Yonghui Wu. 2018. Leveraging weakly supervised data to improve end-to-end speech-to-text translation. *ICASSP*, pages 7180–7184.

Kazuya Kawakami, Luyu Wang, Chris Dyer, Phil Blunsom, and Aaron van den Oord. 2020. Learning robust and multilingual speech representations. *ArXiv*.

Olga Kovaleva, Alexey Romanov, Anna Rogers, and Anna Rumshisky. 2019. Revealing the dark secrets of bert. *arXiv preprint arXiv:1908.08593*.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. In *NeurIPS*.

Jaejun Lee, Raphael Tang, and Jimmy Lin. 2019. What would elsa do? freezing layers during transformer fine-tuning. *arXiv preprint arXiv:1911.03090*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension.

Yinhan Liu, Jiatao Gu, Naman Goyal, X. Li, Sergey Edunov, Marjan Ghazvininejad, M. Lewis, and L. Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *ArXiv*, abs/2001.08210.

Yuchen Liu, Hao Xiong, Zhongjun He, Jiajun Zhang, Hua Wu, Haifeng Wang, and Chengqing Zong. 2019. End-to-end speech translation with knowledge distillation.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*, pages 13–23.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020. Mad-x: An adapter-based framework for multi-task cross-lingual transfer. *arXiv preprint arXiv:2005.00052*.

Juan Pino, Qiantong Xu, Xutai Ma, Mohammad Javad Dousti, and Yun Tang. 2020. Self-Training for End-to-End Speech Translation. In *INTERSPEECH*.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

A. Radford. 2018. Improving language understanding by generative pre-training.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Morgane Rivière, Armand Joulin, Pierre-Emmanuel Mazaré, and Emmanuel Dupoux. 2020. Unsupervised pretraining transfers well across languages. In *ICASSP*.

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

Asa Cooper Stickland, Xian Li, and Marjan Ghazvininejad. 2020. Recipes for adapting pre-trained monolingual and multilingual models to machine translation.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2019. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*.

Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. 2019. Videobert: A joint model for video and language representation learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7464–7473.

Y. Tang, C. Tran, X. Li, P. Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and A. Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning. *ArXiv*, abs/2008.00401.

Yun Tang, J. Pino, Changhan Wang, Xutai Ma, and Dmitriy Genzel. 2021a. A general multi-task learning framework to leverage text data for speech to text tasks. In *ICASSP*.

Yun Tang, Juan Pino, Xian Li, Changhan Wang, and Dmitriy Genzel. 2021b. Improving speech translation by understanding and learning from the auxiliary text translation task. In *ACL*.

Changhan Wang, Yun Tang, Xutai Ma, Anne Wu, Dmytro Okhonko, and Juan Pino. 2020a. fairseq s2t: Fast speech-to-text modeling with fairseq. In *Proceedings of the 2020 Conference of the Asian Chapter of the Association for Computational Linguistics (AACL): System Demonstrations*.

Changhan Wang, Anne Wu, and Juan Pino. 2020b. Covost 2 and massively multilingual speech-to-text translation. *arXiv e-prints*, pages arXiv–2007.

Chengyi Wang, Yu Wu, Shujie Liu, Zhenglu Yang, and Ming Zhou. 2020c. Bridging the gap between pre-training and fine-tuning for end-to-end speech translation.

Chengyi Wang, Yunzhao Wu, Shujie Liu, Ming Zhou, and Zhenglu Yang. 2020d. Curriculum pre-training for end-to-end speech translation. In *ACL*.

Ron J. Weiss, Jan Chorowski, Navdeep Jaitly, Yonghui Wu, and Zhifeng Chen. 2017. Sequence-to-sequence models can directly translate foreign speech. In *INTERSPEECH*.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of bert. *arXiv preprint arXiv:1904.09077*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in neural information processing systems*, pages 5753–5763.

A Appendix

A.1 Description of Pretrained Models

wav2vec 2.0 is a simple and powerful framework to learn high quality speech representation from unlabelled audio data. It mainly consists of two components: feature encoder and context encoder. The feature encoder, which is built from temporal convolution layers, takes raw audio signal O as input and generates latent speech representation $Z = [z_1, \cdots, z_T]$. They are fed to the transformer based context encoder to generate context representations $C = [c_1, \cdots, c_T]$ with sequence level information. During pre-training, the model is optimized with a contrastive task to distinguish true latent from distractors. The input to the context encoder is with span masked. The latent speech representation Z is discretized to $Q = [q_1, \cdots, q_T]$ and used as targets for the frames in the masked span.

mBART is a sequence-to-sequence generative pre-training scheme, specifically a denoising autoencoder (DAE) to predict the original text x given $g(x)$ where g is a noising function that corrupts text such as random span masking and order permutation (Liu et al., 2020). The model is trained with monolingual data of N languages: $\mathcal{D} = \{\mathcal{D}_1, ..., \mathcal{D}_N\}$ where each $\mathcal{D}_i$ is a collection of documents in language i. The pretraining objective optimizes $\mathcal{L}_\theta$:

$$\mathcal{L}_\theta = \sum_{\mathcal{D}_i \in \mathcal{D}} \sum_{x \in \mathcal{D}_i} \log P(x|g(x); \theta) \,, \qquad (1)$$

where x is an instance in language i and the distribution P is parameterized by the sequence-to-sequence model.

A.2 Data

The CoVoST 2 dataset (Wang et al., 2020b) is a large-scale multilingual ST corpus which covers translations from English into 15 languages—Arabic, Catalan, Welsh, German, Estonian, Persian, Indonesian, Japanese, Latvian, Mongolian, Slovenian, Swedish, Tamil, Turkish, Chinese, and translations from 21 languages into English, including Spanish, French, Italian, Dutch, Portuguese, Russian in addition to the 15 target languages. It has total 2,880 hours of speech from 78K speakers. The data could be downloaded from `https://github.com/facebookresearch/covost`.

We provide the list of languages used in our experiments and their ISO codes.

Code	Language
Ar	Arabic
Ca	Catalan
Cy	Welsh
De	German
En	English
Et	Estonian
Es	Spanish
Fa	Persian
Fr	French
Ja	Japanese
Id	Indonesian
It	Italian
Lv	Latvian
Mn	Mongolian
Nl	Dutch
Pt	Portuguese
Ru	Russian
Sv	Swedish
Ta	Tamil
Tr	Turkish
Zh	Chinese (Sim)

Table 8: A list of 21 languages and their ISO codes with experiment results reported in this paper.

A.3 Implementation Details

Preprocessing. When using wav2vec 2.0 encoder, we use 16-bit 16kHz mono-channel audios as inputs. When using a traditional speech recognition (ASR) encoder, we extract 80-channel log mel-filter bank features (25ms window size and 10ms shift) with utterance-level cepstral mean and variance normalization applied. We remove training samples with more than 3,000 frames for GPU memory efficiency. For preprocessing the target (text) data, we use the same vocabulary as is used in the pretrained mBART model.

Pretrained models. We use the open-sourced models from wav2vec 2.0 and mBART50 pretrained with multilingual parallel text data. These models can be downloaded from https://github.com/pytorch/fairseq/tree/master/examples/wav2vec and https://github.com/pytorch/fairseq/tree/master/examples/multilingual. For XMEF-En, we use the 960-hour Wav2Vec 2.0 Large (LV-60) model. For XMEF-X, we use the 56K-hour XLSR-53 Large model. For decoder, we use the pretrained "mMBART 50 finetuned many-to-one" model for many-to-one experiments and "mMBART 50 finetuned one-to-many" for one-to-many experiments.

Training. We implement all our experiments using fairseq S2T (Ott et al., 2019; Wang et al., 2020a). Our experiments are run with 32 Nvidia V100 GPUs (32GB) with batch size of 256k tokens. We use FP16 training implemented in fairseq (Ott et al., 2019). We apply the same regularization as the baseline models such as label smoothing 0.3, attention dropout probablity 0.3. We choose learning rate among $[1e-5, 5e-5, 1e-4]$ based on validation accuracy (measured on dev set). For multilingual wav2vec 2.0, we enable normalization flag to be consistent with pretraining. We did not apply any temperature adjustment in sampling language pairs in training, but simply train on the empirical distribution of training data volume.

Evaluation. We use the best checkpoint (without checkpoint averaging) according to validation loss and a beam size of 5 for decoding. We report case-sensitive detokenized BLEU using sacreBLEU (Post, 2018), except for Japanese and Chinese translations (no word segmentation) where we report character-level BLEU.