FST: the FAIR Speech Translation System for the IWSLT21 Multilingual Shared Task

Yun Tang* **Hongyu Gong*** **Xian Li** **Changhan Wang**
Juan Pino **Holger Schwenk** **Naman Goyal**
Facebook AI Research
{yuntang,hygong,xianl,changhan,juancarabina,schwenk,naman}@fb.com

Abstract

In this paper, we describe our end-to-end multilingual speech translation system submitted to the IWSLT 2021 evaluation campaign on the Multilingual Speech Translation shared task. Our system is built by leveraging transfer learning across modalities, tasks and languages. First, we leverage general-purpose multilingual modules pretrained with large amounts of unlabelled and labelled data. We further enable knowledge transfer from the text task to the speech task by training two tasks jointly. Finally, our multilingual model is finetuned on speech translation task-specific data to achieve the best translation results. Experimental results show our system outperforms the reported systems, including both end-to-end and cascaded based approaches, by a large margin. In some translation directions, our speech translation results evaluated on the public Multilingual TEDx test set are even comparable with the ones from a strong text-to-text translation system, which uses the oracle speech transcripts as input.

1 Introduction

Multilingual speech translation (Inaguma et al., 2019) enables translation from audio to text in multiple language directions with a single model. Similar to multilingual text translation, it is sample efficient as the model supports more languages. Furthermore, multilingual speech models can facilitate positive transfer across languages by learning a common representation space from speech inputs, typically either raw audio or filterbank features.

In this paper, we provide a description of our submission to the first multilingual speech translation task at IWSLT 2021. The task evaluates speech translations from Spanish (es), French (fr), Portuguese (pt) and Italian (it) into English (en) and Spanish (es). Among them, three translation directions (it-en, it-es and pt-es) are considered zero-shot with respect to the constrained track. In addition, participants are encouraged to submit transcriptions for the relevant languages.

Our team, FAIR Speech Translation (FST), participated in the unconstrained track, where we submitted one primary system and four contrastive systems. We are interested in exploring the effectiveness of building a general-purpose multilingual multi-modality model. We leverage large amounts of data, including unlabelled and labelled data from different modalities, to alleviate the data scarcity issue. We build the multilingual model to perform speech translation and speech recognition tasks for all evaluation directions. Our model leverages self-supervised pretraining on both the encoder and the decoder. The model is further improved by knowledge transferring from the text-to-text translation task to the speech-to-text translation task under the multitask learning framework. Finally, we finetune the model on parallel speech translation corpora as well as weakly aligned speech translation data through data mining to achieve the best result.

In section 2, we described data sources and our method for speech translation data mining. Models and training methods are then described in section 3. Finally, we present the results for the primary and contrastive systems in section 4.

2 Data

Provided by the IWSLT 2021 shared task, the multilingual TEDx dataset collected from TED talks provides speech translations in 13 directions (Salesky et al., 2021). We focus on the seven competition directions in the shared task: es-en, fr-en, pt-en, it-en, fr-es, pt-es and it-es.

*Yun Tang and Hongyu Gong have equal contribution to this work.

Proceedings of the 18th International Conference on Spoken Language Translation, pages 131–137
Bangkok, Thailand (Online), August 5–6, 2021. ©2021 Association for Computational Linguistics

Table 1: Audio Length in Hours of TEDx, CoVoST, EuroParl and Mined Data

	es-en	fr-en	it-en	pt-en	fr-es	pt-es	it-es
TEDx	163.7	119.9	-	134.2	85.5	-	-
CoVoST	113.0	264.1	10.3	44.1	-	-	-
EuroParl	20.7	31.0	35.5	14.6	20.0	9.5	20.6
Common Voice (mined data)	52.7	39.6	12.8	9.6	18.7	4.4	6.6
MLS (mined data)	23.9	64.7	2.3	-	42.7	1.3	3.4

2.1 Public data

Besides TEDx dataset provided by the shared task, we also include two other public datasets, CoVoST and EuroParl, which provides parallel audio-text samples in some of the test directions of TEDx.

- CoVoST (Wang et al., 2020). As a large scale dataset for multilingual speech translation, CoVoST contains translations from 11 languages to English. We use its data in 5 language directions [2].

- EuroParl (Iranzo-Sánchez et al., 2020). Collected from debates in European Parliment, EuroParl provides speech-to-text translations in 6 European languages. Its data in 11 language directions [3] is used in model training.

2.2 Mined data

We also mined additional speech-to-text data from unlabeled corpora. The audio corpora used in our experiments include Common Voice and Multilingual LibriSpeech (MLS).

- Common Voice (Ardila et al., 2020). It is a massive collection of multilingual audios and their transcriptions in 29 languages.

- MLS (Pratap et al., 2020). It is a speech corpus collected from audiobooks of LibriVox in 8 languages.

The text corpus used for mining is CCNet, which serves as the source of target translations (Wenzek et al., 2020). Collected from snapshots of CommonCrawl dataset, CCNet provides a large-scale and high-quality monolingual datasets.

Since the audio corpora provide transcripts for audios, we could align source audios with target translations by finding the alignments between source transcripts and target texts. LASER alignment is applied for the crosslingual text alignment (Artetxe and Schwenk, 2019). It generates sentence embeddings with a pre-trained multilingual text encoder (Schwenk and Douze, 2017), and use them to measure the semantic similarity between sentences.

Table 1 summarizes the statistics of the data used in our experiments. It reports the total length of audios in TEDx, CoVost and EuroParl datasets. Moreover, we include the statistics of mined speech from Common Voice and MLS. The mined data has an equivalent size to TEDx dataset in training directions. It also provides a good amount of speech data in zero-shot directions including it-en, pt-es and it-es.

2.3 Text Data

We use additional text data to train mBART model, which later is used to initialize our speech-to-text model. mBART model is first trained with monolingual text data from five languages[4] using self-supervised training. Then they are finetuned with parallel text data from seven evaluation directions as a multilingual text-to-text translation model. The monolingual text data comes from the CC100 dataset (Conneau et al., 2020b) and the parallel text data are downloaded from OPUS (Tiedemann, 2012). [5]

3 Methods

Our evaluation system is based on an encoder decoder model with the state-of-the-art Transformer architecture. The submitted model is developed

[2] {es, fr, it, pt, ru}-en

[3] es-{en, fr, it, pt}, fr-{en, es, pt}, it-{en, es}, pt-{en, es}, ru-en

[4] Five languages include en,es,fr,it and pt.

[5] The following datasets are used: CommonCrawl, OPUS-Books v1, CAPES v1, DGT v2019, ECB v1, ELRA-W0138 v1, ELRA-W0201 v1, ELRC 2682 v1, EMEA v3, EUbookshop v2, EuroPat v1, Europarl v8, GlobalVoices v2018q4, JRC-Acquis v3.0, JW300 v1b, Multi ParaCrawl v7.1, MultiUN v1, News-Commentary v14, QED v2.0a, SciELO v1, TED2013 v1.1, TED2020 v1, Tanzil v1, Tatoeba v2020-11-09, TildeMODEL v2018,UNPC v1.0, and UN v20090831, Wikipedia v1.0.

	→ en				→ es			
	es	fr	pt	it	fr	pt	it	Ave.
MT (M2M-100) (Salesky et al., 2021)	34.0	40.9	38.7	34.6	42.4	45.8	44.2	40.1
Cascaded System (Salesky et al., 2021)	21.5	25.3	22.3	21.9	26.9	26.3	28.4	24.7
Multilingual E2E (Salesky et al., 2021)	12.3	12.0	12.0	10.7	13.6	13.7	13.1	12.5
ST Baseline	27.8	32.4	26.6	20.6	35.0	28.7	28.3	28.5
XLSR-IPA	32.1	36.8	35.1	30.0	38.3	38.5	37.5	35.5
XLSR-SPM	33.2	37.8	35.0	29.3	39.5	36.7	35.3	35.3
VP100K-IPA	31.6	37.1	35.3	29.3	38.2	37.9	37.1	35.2
Ensemble (3 models)	34.0	38.7	37.2	30.9	39.7	40.4	38.6	37.1

Table 2: Main results on the public test set from the Multilingual TEDx Corpus (Salesky et al., 2021).

with a transfer learning approach (Li et al., 2020), including three ingredients: single-modality modules pretrained from self-supervised learning, multitask joint training, and task-specific fine-tuning. The pretrained modules make use of a large amount of unlabeled data, joint training focuses on transferring knowledge from a relatively simple text-to-text task to a speech-to-text task, and the model is fine-tuned on speech-to-text translation task to boost in-task performance.

3.1 Modality Dependent Pretraining

Our model leverages large amounts of unlabelled data from different modalities through two pretrained models: a wav2vec 2.0 (Baevski et al., 2020) and a multilingual BART (mBART) (Liu et al., 2020).

wav2vec 2.0 is a simple and powerful framework to learn high quality speech representation from unlabelled audio data. Given the raw input audio samples, the model learns both latent representations and context representations through a contrastive task to distinguish true latent from distractors. Two multilingual wav2vec 2.0 models are explored during our development. One ("XLSR-53") is trained on 56K-hour speech in 53 languages (Conneau et al., 2020a), and another ("VP-100K") is trained on 100K-hour speech in 23 languages (Wang et al., 2021). The pretrained wav2vec 2.0 models are used to initialize the speech encoder in the jointly trained model of the next stage.

As will be discussed in our experiments, the two encoders are strong in different language directions. We ensemble models with XLSR-53 encoder and VP-100K encoder respectively to achieve the best performance.

mBART is a sequence-to-sequence generative pretraining scheme, specifically a denoising autoencoder (DAE) to predict the original text from its noisy version such as random span masking and order permutation (Liu et al., 2020). The model is pretrained with monolingual data and finetuned with parallel data as described in subsection 2.3. The encoder and decoder in mBART model are used to initialize the encoder and decoder in the joint trained model of the second stage.

Previous study (Tang et al., 2021b) shows that it makes the knowledge transfer from the text-to-text task to speech-to-text task easier by representing the input text as its pronunciation form, i.e., the phoneme sequence. We also investigate representing the input text as its pronunciation forms rather than sentencepiece tokens during our development. We choose International Phonetic Alphabet (IPA) as input text representation since it can be shared across different languages. espeak[6] is used to convert the text word into IPA phonemes.

3.2 Multitask Joint Training

In the second stage, we choose to optimize the speech-to-text translation model along with a text-to-text translation model. Two encoders are used to process text input and speech input respectively. The speech encoder is with the large wav2vec 2.0 configuration. The feature extractor and the bottom 12 transformer layers in the context extractor are initialized with the corresponding parameters from the pretrained wav2vec 2.0 model in subsection 3.1. The top 12 transformer layers in the speech encoder are shared with the text encoder. They are initialized with the pretrained mBART encoder (Tang et al., 2021a). An adaptor (Li et al., 2020), which consists of 3 1-D convolution layers with stride 2 to achieve 8x down-sampling of speech encoder out-

[6] http://espeak.sourceforge.net/index.html

	BLEU							WER			
	→ en				→ es						
	es	fr	pt	it	fr	pt	it	es	fr	it	pt
ST Baseline	34.1	28.4	19.8	20.0	29.3	25.3	25.8	18.6	25.7	33.2	44.5
XLSR-IPA	40.4	36.4	29.0	28.4	34.4	34.4	34.6	13.0	21.8	21.8	29.9
Ensemble (3 models)	42.2	38.7	31.0	29.4	36.5	38.2	37.3	11.2	18.7	19.6	27.4

Table 3: Main results on the blind test set from the Multilingual TEDx Corpus.

	Data			→ en				→ es			Ave.
	ASR	Public	Mined	es	fr	pt	it	fr	pt	it	
M0	✗	✗	✗	22.3	26.7	21.7	5.9	28.2	23.6	8.4	19.5
M1	✓	✗	✗	24.2	29.1	26.3	18.1	31.7	28.9	27.3	26.5
M2	✓	✓	✗	25.2	30.8	26.9	19.2	32.5	29.4	28.1	27.4
M3 (ST Baseline)	✓	✓	✓	27.8	32.4	26.6	20.6	35.0	28.7	28.3	28.5

Table 4: Ablation studies of training data (Public: CoVoST and EuroParl, Mined: mined data from Common Voice, MLS and CCMatrix, ASR: ASR data in mTEDx). The results are BLEU scores on TEDx test set. The models considered here are built upon pretrained XLSR-53 encoder and mbart decoder, and they do not have joint training. The speeach translation data from mTEDx is used by all models.

puts, is placed between the last non-shared speech encoder layer and the first shared speech text encoder layer. The decoder is shared by two tasks and initialized with the pretrained mBART decoder.

Two techniques (Tang et al., 2021a): cross attentive regularization (CAR) and online knowledge distillation (online KD), are employed to enhance the knowledge transferring. Text input data comes from the corresponding transcripts in the speech translation dataset. Due to time limits, we don't use extra parallel text data to enhance the performance.

3.3 Speech only Finetuning

In the last stage, the model is fine-tuned in the speech-to-text translation task with speech input only. The text encoder is dropped and no text input data is used.

4 Experiments

4.1 Experimental Setting

Both wav2vec 2.0 model and mBART model are trained with the large configuration. There are 24 transformer layers in the wav2vec 2.0 model, and 12 transformer layers in both mBART encoder and decoder. We build the mBART model with a target vocabulary of 64,000 SentencePiece (Kudo and Richardson, 2018) tokens, which are shared among all 6 evaluation languages[7]. For the mBART model with IPA phoneme input, the vocabulary size is 516 which includes phoneme variants with "_" attached to denote the word leading phoneme.

A language id symbol "⟨LID⟩" is used as the initial token to predict the sentence. Speech recognition task is treated as the same as the speech translation task but with the source speech language id symbol.

The primary system results submitted are from an ensemble system with three models. All three models are trained with 3-stage optimization discussed in section 3 with different initialization models. The first one is initialized with "XLSR-53" wav2vec model and IPA mBART model ("XLSR-IPA"). Compared with the first model, the second model chooses sentence piece mBART model ("XLSR-SPM") while the third one is initialized with "VP-100K" wav2vec model ("VP100K-IPA") [8].

We use 8 V100 GPUs for each model during the jointly training and fine-tuning stages. It takes approximate five days to jointly train the models for 15 epochs and another two days for speech only fine tuning. The last 5 checkpoints are averaged for inference with beam size 5.

To provide a deep insight into the factors affect-

[7] In our evaluation, the new mBART model achieves comparable results as the public available mBART model, which is with 250k vocabulary, but with much smaller memory footprint.

[8] We will release the training and evaluation recipe under https://github.com/pytorch/fairseq/tree/master/examples/speech_text_joint_to_text

	Train		→ en				→ es			Ave.
	JT	FT	es	fr	pt	it	fr	pt	it	
M3	✗	✗	27.8	32.4	26.6	20.6	35.0	28.7	28.3	28.5
M4	✓	✗	32.3	36.6	33.8	28.4	38.3	35.9	35.7	34.4
M5	✓	✓	33.2	37.8	35.0	29.3	39.5	36.7	35.3	35.3

Table 5: Ablation studies of training approaches (JT: joint training of text and speech translation, FT: finetuning a trained model on TEDx data in speech translation). The results are BLEU scores reported on TEDx test set. The models considered here are built upon pretrained XLSR-SPM encoder and mbart decoder. They are trained with the combination of TEDx including the ASR portion, public data as well as mined data.

	Encoder	→ en				→ es			Ave.
		es	fr	pt	it	fr	pt	it	
M4	XLSR-SPM	32.3	36.6	33.8	28.4	38.3	35.9	35.7	34.4
M6	VP100K-SPM	30.5	35.6	33.7	28.5	36.9	36.9	36.2	34.0

Table 6: Ablation studies of different encoders. BLEU scores are reported on TEDx test set. Models are jointly trained on all data, but they are not further finetuned on speech translation.

ing translation performance, we conduct ablation studies on different model configurations.

4.2 Main Results

We summarize our main results at Table 2. The first row presents results from a large text-to-text translation system (M2M-100) using oracle speech transcripts (Salesky et al., 2021) as input. The second two rows list best results from the cascaded system and multilingual end to end system from literature (Salesky et al., 2021). The fourth row to eighth row are results from our systems. The fourth row presents our multilingual baseline, which is initialized with pretrained wav2vec 2.0 model ("XLSR-53") for encoder and mBART model ("SPM") for decoder. The model is fine-tuned with Multilingual TEDx data, public data and mined data listed in section 2. No joint training is applied. "XLSR-IPA", "XLSR-SPM" and "VP100K-IPA" from row 5 to 8 are results from the 3 best systems we built. Compared with the baseline in the third row, these three systems have an extra step to co-train with the text-to-text translation task.

It is clear that we create a very strong baseline (row 4) with the help from the large amounts of speech/text training data. In comparison to the previous reported cascaded system (row 2) or multilingual end-to-end system (row 3), the results are 3.8 and 16.0 BLEU scores higher on average.

Row 5 to 8 provide evaluation results from our 3 best single models built with single-modality based pre-training, multitask joint training and task-specific fine-tuning. They are built with different pre-training data or input text representations. Compared with the baseline in row 4, another 6.7 $\sim$ 7.0 BLEU improvement are observed. IPA phoneme based text representation gives slight gain compared with text units separated with SentencePiece model ("XLSR-IPA" v.s. "XLSR-SPM"), which is smaller than we expected. We hypothesis that it is due to the imperfect text to phoneme conversion for different languages. The difference due to different pre-training data is also small that there are only 0.3 BLEU in average when the speech pre-training is changed ("XLSR-IPA" v.s. "VP100K-IPA").

The ensemble of three models achieves the best performance with a 1.6 BLEU improvement over the best single model. It indicates those three models are complementary to each other, though they give similar BLEU scores in our test. The results are even close to the ones from the strong text-to-text translation system (M2M-100 in row 1), which takes speech transcript as translation input. Our primary system achieves the same BLEU score as the text-to-text translation system on translation direction "es-en" and the average BLEU score gap from 7 directions is 3.0.

The corresponding blind test results are reported in Table 3. Similar to our observation in Table 2, the model trained with the 3-stage approach significantly improves the translation accuracy compared with the baseline. The ensemble system outperforms other systems in all speech translation directions as well as the speech recognition tasks.

4.3 Analysis

Data. Table 4 compares models trained with different sets of data. Additional data is shown to improve the translation performance. In our multitask training, we combine the text-to-text and speech-to-text translation tasks together. We don't include ASR task as separated task, instead we treat ASR task as a special translation direction. It shows ASR data is helpful for speech translation, especially for translation directions with small amount of speech training data ("it-en" and "it-es"). On average, we observe a significant gain of 7.0 BLEU from the comparison between M0 and M1 .

When we continue adding public datasets including CoVost and EuroParl to the training set, M2 has an average improvement of 0.9 BLEU over M1. The mined data brings another 1.1 BLEU gain as we compare M3 and M2.

Training. Different training approaches are compared in Table 5. We observe significant gains brought by joint training of text and speech translation. Compared against M3, M4 with joint training demonstrates an improvement of 5.9 BLEU over 7 language directions. When the jointly trained model is further finetuned with speech translation data, an extra gain of 0.9 is achieved as we compare M5 against M4.

Encoder. We compare XLSR-53 and VP-100K encoder in Table 6. XLSR-53 is strong at encoding audios in Spanish and French, achieving BLEU gains of 1.8 and 1.0 in es-en and fr-en respectively. VP100k encoder outperforms XLSR-53 in pt-es and it-es directions with gains of 1.0 and 0.5 respectively. This can be explained by the fact that VP100K encoder is trained on more Portuguese and Italian Speech.

5 Conclusion

In this work, we described our multilingual end-to-end speech translation system submitted to IWSLT 2021. We leverage the large amount of training data from different domains and modalities to improve the speech translation performance. We adopt a progressive approach to build the model with three stages. Compared with our strong baseline, the proposed system achieves 8.6 BLEU score improvement, which also outperforms other reported systems, including both end-to-end and cascaded based, by a large margin.

References

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 4218–4222.

Mikel Artetxe and Holger Schwenk. 2019. Margin-based parallel corpus mining with multilingual sentence embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3197–3203.

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

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. 2020a. 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, E. Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020b. Unsupervised cross-lingual representation learning at scale. In *ACL*.

Hirofumi Inaguma, Kevin Duh, Tatsuya Kawahara, and Shinji Watanabe. 2019. Multilingual end-to-end speech translation. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 570–577. IEEE.

Javier Iranzo-Sánchez, Joan Albert Silvestre-Cerdà, Javier Jorge, Nahuel Roselló, Adrià Giménez, Albert Sanchis, Jorge Civera, and Alfons 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. IEEE.

T. Kudo and J. Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *EMNLP*.

Xian Li, Changhan Wang, Yun Tang, C. Tran, Yuqing Tang, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. 2020. Multilingual speech translation with efficient finetuning of pretrained models. *arXiv: Computation and Language*.

Yinhan Liu, Jiatao Gu, Naman Goyal, X. Li, Sergey Edunov, Marjan Ghazvininejad, M. Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Vineel Pratap, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve, and Ronan Collobert. 2020. Mls: A large-scale multilingual dataset for speech research. *Proc. Interspeech 2020*, pages 2757–2761.

Elizabeth Salesky, Matthew Wiesner, Jacob Bremerman, Roldano Cattoni, Matteo Negri, Marco Turchi, Douglas W. Oard, and Matt Post. 2021. Multilingual tedx corpus for speech recognition and translation.

Holger Schwenk and Matthijs Douze. 2017. Learning joint multilingual sentence representations with neural machine translation. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 157–167.

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

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

J. Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *LREC*.

Changhan Wang, Juan Pino, Anne Wu, and Jiatao Gu. 2020. Covost: A diverse multilingual speech-to-text translation corpus. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 4197–4203.

Changhan Wang, M. Rivière, A. Lee, Anne Wu, C. Talnikar, Daniel Haziza, Mary Williamson, J. Pino, and Emmanuel Dupoux. 2021. Voxpopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. *ArXiv*, abs/2101.00390.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Édouard Grave. 2020. Ccnet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 4003–4012.