Maastricht University's Multilingual Speech Translation System for IWSLT 2021

Danni Liu, Jan Niehues
Department of Data Science and Knowledge Engineering, Maastricht University
{danni.liu,jan.niehues}@maastrichtuniversity.nl

Abstract

This paper describes Maastricht University's participation in the IWSLT 2021 multilingual speech translation track. The task of this track is to build multilingual speech translation systems in supervised and zero-shot directions. Our primary system is an end-to-end model that performs both speech transcription and translation. We observe that the joint training for the two tasks is complementary especially when the speech translation data is scarce. On the source and target side, we use data augmentation and pseudo-labels respectively to improve the performance of our systems. We also introduce an ensembling technique that consistently improves the quality of transcriptions and translations. The experiments show that the end-to-end system is competitive with its cascaded counterpart especially in zero-shot conditions.

1 Introduction

In this paper, we describe our systems for the multilingual speech translation track of IWSLT 2021. Speech translation (Bérard et al., 2016; Weiss et al., 2017) is the task of converting speech utterances to their translation in other languages. While "end-to-end" modeling (Di Gangi et al., 2019; Sperber et al., 2019) of the speech translation pipeline has become the dominant approach, an open challenge remains in terms of data scarcity. As the amount of speech directly paired with translation is lower compared to speech transcription or text-to-text translation, it is especially crucial for models to be data-efficient. In this context, multilingual speech translation (Inaguma et al., 2019; Li et al., 2021) presents itself as a promising direction to alleviate data scarcity by leveraging commonalities across languages.

In this multilingual translation track, we submit: 1) an end-to-end system (§5.2) that directly translates from speech and 2) a cascaded system (§5.1) that consists of a multilingual speech transcription module (§3) followed by a multilingual text translation module (§4).

Our efforts to improve the speech translation system can be categorized as follows. When **training**, on the source side, we augment the speech data by speed perturbation. On the target side, we apply pseudo-labeling[1] by translating the ASR transcriptions. Furthermore, we train multilingual systems for both speech transcription and translation to alleviate the scarcity of training data. When **testing**, we use different ensembling techniques to increase the diversity of output distribution and improve output quality.

The main findings from our experiments are:

- Multilingual training and jointly training speech transcription and translation are beneficial when data scarcity limits the performance of mono- or bilingual systems.
- The gain in the overall speech-to-text systems also propagates to cascaded systems as a result of stronger ASR performance.
- Pseudo-labeling strongly improves speech translation quality, especially in directions that are originally zero-shot.

2 Setup

2.1 Corpus Statistics

Our systems are trained on the multilingual TEDx (mTEDx) speech recognition and translation corpus (Salesky et al., 2021). We do not use any data outside this corpus. Table 1 outlines some statistics about the training set of the mTEDx corpus.

2.2 Preprocessing

For the audio data, we downsample the original audio files from 48kHz to 16kHz and mix the two channels into one. We then extract 40-dimensional

[1] or forward-translation in analogy to back-translation

Proceedings of the 18th International Conference on Spoken Language Translation, pages 138–143
Bangkok, Thailand (Online), August 5–6, 2021.

Source	transcription (hour, # utts.)	Target (# utts.) en	es	fr	pt	it
es	178, 102k	36k		4k	21k	6k
fr	176, 116k	30k	20k		13k	–
pt	153, 90k	31k	–	–		–
it	101, 50k	–	–	–	–	

Table 1: Data amount of speech transcription and translation in the training set of mTEDx.

Mel Frequency Cepstral Coefficients (MFCC) with 3-dimensional pitch using Kaldi (Povey et al., 2011). We concatenate adjacent 4 audio frames, resulting in an input dimension of 172.

For the text data, we combine all transcriptions and translations from the training set and learn a joint byte pair encoding (BPE) (Sennrich et al., 2016b) of size 16k using SentencePiece (Kudo and Richardson, 2018). With this joint BPE, we can translate from tokenized ASR transcriptions in our cascaded system.

2.3 Training Details

We use the dev partition of mTEDx as validation set and average the model weights from last 5 best checkpoints. When decoding, we use a beam size of 8. The specific models for different tasks will be described in the corresponding sections.

3 Automatic Speech Recognition (ASR)

The ASR performance is summarized in Table 2. We report case-insensitive word error rates (WER) after removing all punctuation marks.

3.1 Model Description

Multilingual Baseline We start from a Transformer (Vaswani et al., 2017) with stochastic layer dropout (Pham et al., 2019a) rate of 0.5. We use 36 encoder layers and 12 decoders layers, following the original work (Pham et al., 2019a). The hidden dimension is 512 and the inner dimension 2048. We use dropout rate of 0.2 and label smoothing rate of 0.1.

The model is jointly trained on all four languages. As the data volume for each individual language is relatively low, after initially seeing poor performance of monolingual ASR models, we proceed with a multilingual system for all four languages, with the intention of better utilizing common acoustic features.

Language Embedding While the multilingual ASR system does not need to explicitly know the target language, we find it beneficial to provide the decoder more guidance by feeding in target language embeddings. Specifically, we achieve this by language embeddings concatenated with decoder input embeddings (Pham et al., 2019b). Meanwhile, the decoder begin token is replaced by the target language embedding. With this approach, we reduce the WER on average by 0.6% absolute (2.4% relative; model A2 in Table 2). More importantly, this approach allows us to easily extend the model to speech translation, where the number of target languages can be more than one.

Speed Perturbation We augment the training data by speed perturbation with factor 0.9 and 1.1 (Ko et al., 2015) using the corresponding Kaldi script[2]. After speed perturbation, we further observe a reduction of 2.4% absolute WER (9.3% relative; model A3 in Table 2). Here we did not use SpecAugument (Park et al., 2019), but would expect further gains from this approach.

Ensembling By ensembling two independently trained models on the output distributions, we further reduce WER by 1% absolute (4.4% relative; model A4 in Table 2).

Joint Training with Speech Translation We can directly apply the same ASR model to speech translation, as we control the output language by the target language embedding. As described later in §5.2, we train end-to-end systems using both ASR and ST data. The strongest system from ASR and ST training (model E5) achieves a large reduction of WER from 21.9% to 18.7% (14.7% relative) on average.

ID	Model	es	fr	it	pt
A1	Multilingual baseline	24.3	24.5	25.9	28.7
A2	A1 + language emb.	23.8	23.9	25.5	27.7
A3	A2 + speed perturb	21.0	22.1	23.1	25.3
A4	A3 + ensembling	20.4	21.0	22.0	24.1
E5	A3 + ST joint training	17.6	18.4	18.6	20.0

Table 2: ASR performance in WER↓ (%) (lowercased, no punctuation) of the multilingual ASR system on mTEDx test set.

3.2 Main Findings

As summarized in Table 2, we reduce the WER of our baseline multilingual Transformer from 25.8%

[2] https://github.com/kaldi-asr/kaldi/blob/master/egs/wsj/s5/utils/data/perturb_data_dir_speed_3way.sh

ID	Model	es-en	fr-en	fr-es	pt-en	pt-es*	it-en**	it-es*
M1	Transformer (6-6 encoder-decocder layers)	32.3	38.0	41.3	37.1	42.3	23.0	32.5
M2	M1 + residual drop	32.9	38.1	40.7	37.0	42.5	24.1	32.8
M3	Ensemble M1 and M2	33.4	39.4	41.8	37.9	43.3	24.8	34.0
M4	Ensemble M1×2 and M2×2	33.7	39.3	42.1	38.3	44.0	24.9	34.8

Table 3: Machine translation performance in BLEU↑[3]of the multilingual MT system on mTEDx test set by directly translating from ground-truth transcriptions. *: zero-shot directions for speech translation. **: zero-shot direction for text translation.

	en	es	fr	it	pt
en	-	36	30	0	30
es		-	24	6	21
fr			-	0	13
it				-	0
pt					-

Table 4: Overview of MT parallel training data amount (in 1k sentences) after including all directions with text-to-text translation data.

to 18.7% by a combination of techniques. Among these, the largest gain comes from joint training for speech translation. This highlights the benefit of multilingual training, especially when data scarcity limits the performance of monolingual end-to-end systems.

4 Machine Translation (MT)

When translating from speech, the MT module ingests ASR outputs. To assess the quality of the MT component alone, we first report the performance of directly translating from the ground truth transcriptions in Table 3. The results of cascading the ASR and MT systems are reported later in Table 5.

4.1 Data

For the MT component, we train our models on all translation directions from {en, es, fr, it, pt} with all text translation data in the training set, including both directions of transcription ↔ translation. In doing so, we cover more directions than tested in the evaluation campaign. A main advantage of this is additional training data on the target side. For instance, although the evaluation task does not involve translating from English, incorporating en→X directions provides around 30k sentences with each of {es, fr, pt} on the target side. Including these data largely expands the data amount when translating into the three target languages. The data amount for our MT training is outlined in Table 4. Note that while {pt→es, it→en, it→es} are zero-shot directions for speech translation, only it→en is zero-shot for MT.

[3]sacreBLEU: BLEU+case.mixed+numrefs.1+smooth .exp+tok.13a+version.1.4.12

4.2 Model Description

Multilingual Baseline We start with a Transformer-base with 6 encoder and decoder layers respectively (model M1 from Table 3). We use dropout rate of 0.2 and label smoothing rate of 0.1. The source and target embeddings are shared. The output language is controlled by the language embedding described in §3.1. As we observe no performance gain by increasing the number of encoder and decoder layers, we keep the Transformer-base setup.

Residual Drop We additionally use the Transformer with residual connections removed from the middle encoder layer (Liu et al., 2020) that was shown to improve zero-shot performance under English-centric scenarios. We see that the model (M2 from Table 3) outperforms the vanilla Transformer in the zero-shot direction (it-en) by 1.1 BLEU, while being on-par on other directions.

Ensembling We ensemble the two models above by averaging the output distributions (model M3 in Table 3). This brings a gain of 0.9 BLEU on average. By further incorporating another two independently trained vanilla model and residual-drop model (hence ensembling four models), we see a further gain of 0.4 BLEU (model M4 in Table 3). This MT system and will be used in the later cascaded speech translation system.

4.3 Main Findings

We build a multilingual translation model with results summarized in Table 3. We first confirm that the residual drop approach (Liu et al., 2020) improves zero-shot translation performance. Furthermore, ensembling different models brings gains up to 1.5 BLEU.

Type	ID	Model	es-en	fr-en	fr-es	pt-en	pt-es*	it-en*	it-es*	ASR (avg.)
Cascaded	C1	A4 + M2	25.6	30.1	32.2	28.1	31.4	19.1	26.0	-
	C2	A4 + M4	26.1	30.6	33.3	29.0	32.0	19.5	26.8	-
	C3	C2 + ASR multi-view ensemble	26.5	30.6	33.6	28.9	32.2	19.7	27.0	-
	C4	E5 + M4	27.3	31.6	34.2	31.0	34.6	20.5	27.8	-
End-to-end	E1	Transformer	17.0	20.1	21.2	17.5	11.7	5.8	6.6	-
	E2**	E1 + ASR joint training	18.0	20.8	24.7	20.1	19.0	8.2	10.2	25.3
	E3	E2 + pseudo labels (zero-shot dir.)	21.9	25.3	29.1	24.9	33.3	19.2	28.2	20.4
	E4	E2 + pseudo labels (all dir.)	25.0	30.0	33.3	28.5	34.4	20.4	28.8	19.5
	E5	E4 + multi-view ensemble (3 speeds)	25.2	30.1	33.3	28.7	34.5	20.5	29.1	18.7

Table 5: Speech translation performance in BLEU↑ on mTEDx test set. We mark the cascaded systems with "ASR-ID + MT-ID". For e2e systems trained to jointly perform ST and ASR, we additionally report average WER↓ over the 4 source languages {es, fr, it, pt}. *: zero-shot directions. **: Due to computation constraints, we terminated the training of model E2 early to combine with the other approaches.

5 Speech Translation (ST)

In Table 5, we report the performance of our cascaded (§5.1) and end-to-end (§5.2) speech translation systems.

5.1 Cascaded System

The performance of the cascaded systems is summarized in the upper section of the Table 5. We combine the stronger ASR system and MT system and derive cascaded models C1 and C2. Compared to the MT results in Table 3 that utilizes ground-truth transcriptions, we observe a clear drop in BLEU. This highlights the importance of high-quality transcriptions for the cascaded system.

Multi-View Ensemble (Transcription) Since at test time the ASR transcriptions are likely noisy, we propose an ensembling approach that incorporates multiple variants (or views) of ASR transcriptions. At test time, given an utterance, we transcribe it with different ASR models. The MT module then translates from these slightly different transcriptions and ensembles by averaging the output distribution. The results from this technique are shown in C3 in Table 3. With this ensembling technique, on average we see an improvement of 0.2 BLEU, with the all other modules unchanged from the previous model C2.

5.2 End-to-End System

For the end-to-end ST system, we use the provided ST training data augmented with three-way speed perturbation (Ko et al., 2015). We initialize the models with pre-trained encoder weights from our trained ASR system.

ASR Joint Training Since our decoder utilizes target language embeddings, we can conveniently incorporate ASR data for jointly training the ST system (Model E2 in Table 5). Upon seeing improvements over the setup without ASR data, we terminated the training of E2 and continued by combining with other approaches described next. Therefore if trained till convergence, the final performance of E2 would be better than reported here.

Pseudo-Labels Since the provided corpus contains no Italian ST data, the BLEU scores when translating from Italian are poor (8.2 and 10.2 for it-en and it-es from model E2 in Table 5). To have more training signals, we create pseudo-labels by translating the ASR transcription using our MT system. The model trained with the additional pseudo-labeled data (pt-es, it-en, it-es) is E3 in Table 5. As expected, incorporating pseudo-labels largely improves the performance on the three zero-shot directions (pt-es, it-en, it-es). It is worth noting that on these zero-shot directions the end-to-end system already surpassed the strongest cascaded system so far (C3), achieving 33.3, 19.2, 28.2 compared to 32.2, 19.7, 27.0 BLEU points.

Observing the strength of the pseudo-labeling, we take a step further and create pseudo-labels also for the supervised directions (model E4 in Table 5). This further improves the overall ST and ASR performance by $+2.6$ BLEU and -4.4% WER (relative) on average.

Multi-View Ensemble (Speech Speed) Similar to the motivation for the ensembling approach in §5.1, we utilize multiple views of the same input to create an ensemble. Since the input here is audio, we take the speed-perturbed variants with factors 0.9, 1.0, 1.1 (Ko et al., 2015) of the test utterances and ensemble the output distributions (model E5 of Table 5). This simple technique slightly yet

Type	ID	es-en	fr-en	fr-es	pt-en	pt-es*	it-en*	it-es*	avg.
Cascaded	C3	34.5	21.9	24.3	24.3	29.3	21.7	26.8	26.1
End-to-end	E5	33.9	25.4	27.6	25.7	33.7	22.8	29.4	28.4

Table 6: Speech translation performance in BLEU↑ on the blind test set. We mark the cascaded systems with "ASR-ID + MT-ID". *: zero-shot directions for speech translation.

consistently improves ST and ASR quality, gaining +0.2 BLEU and −4.1% WER (relative) on average. It is worth noting that the model has already been trained on speech data perturbed with the same speed factors. This suggests that we can further improve our model's prediction consistency for perturbed versions of the the same utterance, e.g. by consistency regularization (Sohn et al., 2020). Furthermore, although this ensembling approach leads to improvements in the current offline setting, we note that it could be difficult to apply under real-time constraints due to the computation load of generating 3 variants of speech utterances and applying ensembling on top of that.

Feeding Back to Cascaded System Till now, the series of improvements of the speech-to-text model also lead to better ASR performance. We therefore use the improved ASR transcriptions from model E5 as the MT input for the cascaded system. The resulting model is C4 in Table 5, which brings a gain of 1.2 BLEU for the cascaded system.

5.3 Main Findings

The results for cascaded and end-to-end ST systems are summarized in Table 5. First, using a unified end-to-end speech-to-text system for both ASR and ST improves the output quality for both tasks. This gain further propagates to the cascaded systems as a result of higher ASR quality. Second, confirming findings from the literature (Kahn et al., 2020; Pino et al., 2020), training with pseudo-labels is a strong method to improve end-to-end systems. Last but not least, by ensembling from different views of the same data, we can achieve further gains at inference time.

6 Results on Blind Test Set

We submitted systems C3 and E5 for evaluation on the blind test set. The results are summarized in Table 6. In line with the results on the public test set in Table 5, the end-to-end system outperforms the cascaded system on zero-shot directions. Different from on the public test set, the end-to-end system also shows large gains when translating from French speech. A potential reason is errors propagated from the French ASR transcriptions that led to weaker performance of the MT module in the cascaded system.

7 Conclusion

This paper summarizes our participation in the IWSLT 2021 multilingual speech translation track. We improved our end-to-end speech-to-text systems from different angles. On the source side, we augmented the input utterance. On the target side, we created pseudo-labels from ASR transcriptions. Furthermore, at test time we used different ensembling approaches to improve the performance of trained models. By experimenting under different data scenarios, we showed the benefit of multilingual training and the joint training speech transcription and translation.

We note a few directions to further improve our systems: First, we expect that utterances augmented by SpecAugment (Park et al., 2019) could improve the quality of the ASR and ST systems. Second, our MT module can be improved by synthetic data from back-translation (Sennrich et al., 2016a), especially for the zero-shot directions. Regarding upcoming work, since the source languages all belong to the same family, an interesting next step is to investigate how to better utilize the relatedness between these languages.

Acknowledgement

We thank the anonymous reviewers for their helpful feedback. This work is supported by a Facebook Sponsored Research Agreement.

References

Alexandre Bérard, 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 Workshop on end-to-end learning for speech and audio processing*.

Mattia A. Di Gangi, Matteo Negri, and Marco Turchi. 2019. Adapting Transformer to End-to-End Spoken

Language Translation. In *Proc. Interspeech 2019*, pages 1133–1137.

Hirofumi Inaguma, Kevin Duh, Tatsuya Kawahara, and Shinji Watanabe. 2019. Multilingual end-to-end speech translation. In *Proc. ASRU 2019*.

Jacob Kahn, Ann Lee, and Awni Hannun. 2020. Self-training for end-to-end speech recognition. In *Proc. ICASSP 2020*, pages 7084–7088.

Tom Ko, Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur. 2015. Audio augmentation for speech recognition. In *Proc. Interspeech 2015*, pages 3586–3589.

Taku Kudo and John Richardson. 2018. 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*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Xian Li, Changhan Wang, Yun Tang, Chau Tran, Yuqing Tang, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. 2021. Multilingual speech translation with efficient finetuning of pretrained models.

Danni Liu, Jan Niehues, James Cross, Francisco Guzmán, and Xian Li. 2020. Improving zero-shot translation by disentangling positional information. *CoRR*, abs/2012.15127.

Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le. 2019. SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition. In *Proc. Interspeech 2019*, pages 2613–2617.

Ngoc-Quan Pham, Thai-Son Nguyen, Jan Niehues, Markus Müller, and Alex Waibel. 2019a. Very Deep Self-Attention Networks for End-to-End Speech Recognition. In *Proc. Interspeech 2019*, pages 66–70.

Ngoc-Quan Pham, Jan Niehues, Thanh-Le Ha, and Alexander Waibel. 2019b. Improving zero-shot translation with language-independent constraints. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 13–23, Florence, Italy. Association for Computational Linguistics.

Juan Pino, Qiantong Xu, Xutai Ma, Mohammad Javad Dousti, and Yun Tang. 2020. Self-Training for End-to-End Speech Translation. In *Proc. Interspeech 2020*, pages 1476–1480.

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al. 2011. The kaldi speech recognition toolkit. In *Proc. ASRU 2011*.

Elizabeth Salesky, Matthew Wiesner, Jacob Bremerman, Roldano Cattoni, Matteo Negri, Marco Turchi, Douglas W. Oard, and Matt Post. 2021. The multilingual tedx corpus for speech recognition and translation. *CoRR*, abs/2102.01757.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. 2020. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In *Advances in Neural Information Processing Systems*, volume 33, pages 596–608. Curran Associates, Inc.

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

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