CoVoST 2: A Massively Multilingual Speech-to-Text Translation Corpus

Changhan Wang*, Anne Wu*, Juan Pino*
Facebook AI
{changhan, annewu, juancarabina}@fb.com

Abstract
Speech-to-text translation (ST) has recently become an increasingly popular topic of research, partly due to the development of benchmark datasets. Nevertheless, current datasets cover a limited number of languages. With the aim to foster research in massive multilingual ST and ST for low resource language pairs, we release CoVoST 2, a large-scale multilingual ST corpus covering translations from 21 languages into English and from English into 15 languages. This represents the largest open dataset available to date from total volume and language coverage perspective. Data sanity checks provide evidence about the quality of the data, which is released under CC0 license. We also provide extensive speech recognition, bilingual and multilingual machine translation and ST baselines.

1 Introduction
The development of benchmark datasets, such as MuST-C (Di Gangi et al., 2019), Europarl-ST (Iranzo-Snchez et al., 2020) or CoVoST (Wang et al., 2020), has greatly contributed to the increasing popularity of speech-to-text translation (ST) as a research topic. MuST-C provides TED talks translations from English into 8 European languages, with data amounts ranging from 385 hours to 504 hours, thereby encouraging research into end-to-end ST (Berard et al., 2016) as well as one-to-many multilingual ST (Di Gangi et al., 2019). Europarl-ST offers translations between 6 European languages, with a total of 30 translation directions, enabling research into many-to-many multilingual ST (Inaguma et al., 2019). The two corpora described so far involve European languages that are in general high resource from the perspective of machine translation (MT) and speech. CoVoST is a multilingual and diversified ST corpus from 11 languages into English, based on the Common Voice project (Ardila et al., 2020). Unlike previous corpora, it involves low resource languages such as Mongolian and it also enables many-to-one ST research. Nevertheless, for all corpora described so far, the number of languages involved is limited.

In this paper, we describe CoVoST 2, an extension of CoVoST (Wang et al., 2020) that provides translations from English (En) into 15 languages—Arabic (Ar), Catalan (Ca), Welsh (Cy), German (De), Estonian (Et), Persian (Fa), Indonesian (Id), Japanese (Ja), Latvian (Lv), Mongolian (Mn), Slovenian (Sl), Swedish (Sv), Tamil (Ta), Turkish (Tr), Chinese (Zh)—and from 21 languages into English, including the 15 target languages as well as Spanish (Es), French (Fr), Italian (It), Dutch (Nl), Portuguese (Pt), Russian (Ru). The overall speech duration is extended from 700 hours to 2880 hours. The total number of speakers is increased from 11K to 78K. We make data available at https://github.com/facebookresearch/covost under CC0 license.

2 Dataset Creation
2.1 Data Collection and Quality Control
Translations are collected from professional translators the same way as for CoVoST. We then conduct sanity checks based on language model perplexity, LASER (Artetxe and Schwenk, 2019) scores and a length ratio heuristic in order to ensure the quality of the translations. Length ratio and LASER score checks are conducted as in the original version of CoVoST. For language model perplexity checks, 20M lines are sampled from the OSCAR corpus (Ortiz Suárez et al.,
### Table 1: Basic statistics of CoVoST 2 using original CV splits and extended CoVoST splits (only for the speech part). Token counts on Chinese (Zh) and Japanese (Ja) are based on characters (there is no word segmentation).

| Language | Hours (CoVoST ext.) | Speakers (CoVoST ext.) | Src./Tgt. Tokens |
|----------|----------------------|-------------------------|-----------------|
|          | Train | Dev | Test | Train | Dev | Test | Train | Dev | Test | Train | Dev | Test | Train | Dev | Test |
| Fr       | 180 (264) | 22 (23) | 23 (24) | 2K (2K) | 2K (2K) | 4K (4K) | 2M/2M | 145K/143K | 143K/141K |
| De       | 119 (184) | 21 (23) | 22 (120) | 1K (1K) | 1K (1K) | 4K (5K) | 1M/1M | 137K/152K | 777K/844K |
| Es       | 97 (113) | 22 (23) | 23 (23) | 1K (1K) | 2K (2K) | 4K (4K) | 2K/2K | 74K/751K | 131K/132K |
| Ca       | 81 (136) | 19 (21) | 20 (25) | 557 (557) | 722 (722) | 2K (2K) | 2K/2K | 939K/972K | 142K/149K |
| It       | 28 (44) | 14 (15) | 15 (15) | 236 (236) | 640 (640) | 2K (2K) | 2K/2K | 307K/329K | 81K/93K |
| Ru       | 16 (18) | 10 (15) | 11 (14) | 8 (8) | 30 (30) | 417 (417) | 2K/2K | 118K/144K | 81K/100K |
| Zh       | 10 (10) | 8 (8) | 8 (8) | 22 (22) | 83 (83) | 784 (784) | 2K/2K | 131K/85K | 88K/57K |
| Pt       | 7 (10) | 4 (5) | 5 (6) | 2 (2) | 16 (16) | 301 (301) | 2K/2K | 76K/68K | 34K/34K |
| Fa       | 5 (49) | 5 (11) | 5 (40) | 532 (545) | 854 (908) | 1K (1K) | 2K/2K | 307K/313K | 244K/271K |
| Et       | 3 (3) | 3 (3) | 3 (3) | 20 (20) | 74 (74) | 135 (135) | 2K/2K | 23K/22K | 10K/20K |
| Mn       | 3 (3) | 3 (3) | 3 (3) | 4 (4) | 24 (24) | 209 (209) | 2K/2K | 20K/23K | 18K/20K |
| Nl       | 2 (7) | 2 (3) | 2 (3) | 74 (74) | 144 (144) | 379 (383) | 2K/2K | 58K/59K | 20K/20K |
| Tr       | 2 (4) | 2 (2) | 2 (2) | 34 (34) | 76 (76) | 324 (324) | 2K/2K | 24K/33K | 11K/15K |
| Ar       | 2 (2) | 2 (2) | 2 (2) | 6 (6) | 13 (13) | 113 (113) | 2K/2K | 10K/13K | 8K/10K |
| Sv       | 2 (2) | 1 (1) | 2 (2) | 4 (4) | 7 (7) | 83 (83) | 2K/2K | 12K/12K | 9K/10K |
| Lv       | 2 (2) | 1 (1) | 2 (2) | 2 (2) | 3 (3) | 54 (54) | 2K/2K | 11K/14K | 8K/10K |
| Sl       | 2 (2) | 1 (1) | 1 (1) | 2 (2) | 1 (1) | 28 (28) | 2K/2K | 11K/3K | 3K/4K |
| Ta       | 2 (2) | 1 (1) | 1 (1) | 3 (3) | 2 (2) | 48 (48) | 2K/2K | 6K/10K | 3K/5K |
| Ja       | 1 (1) | 1 (1) | 1 (1) | 2 (2) | 3 (3) | 37 (37) | 2K/2K | 20K/9K | 12K/6K |
| Id       | 1 (1) | 1 (1) | 1 (1) | 2 (2) | 5 (5) | 44 (44) | 2K/2K | 7K/8K | 5K/6K |
| Cy       | 1 (2) | 1 (12) | 1 (16) | 135 (135) | 234 (371) | 275 (597) | 2K/2K | 11K/10K | 7K/6K |

2020) for each CoVoST 2 language, except for English, Russian for which pre-trained language models (Ng et al., 2019) are utilized. 5K lines are reserved for validation and the rest for training. BPE vocabularies of size 20K are then built on the training data, with character coverage 0.9995 for Japanese and Chinese and 1.0 for other languages. A Transformer base model (Vaswani et al., 2017) is then trained for up to 800K updates. Professional translations are ranked by perplexity and the ones with the lowest perplexity are manually examined and sent for re-translation as appropriate. In the data release, we mark out the sentences that cannot be translated properly.

### 2.2 Dataset Splitting

Original Common Voice (CV) dataset splits utilize only one sample per sentence, while there are potentially multiple samples (speakers) available in the raw dataset. To allow higher data utilization and speaker diversity, we add part of the discarded samples back while keeping the speaker set disjoint and the same sentence assignment across different splits. We refer to this extension as CoVoST splits. As a result, data utilization is increased from 44.2% (1273 hours) to 78.8% (2270 hours). We by default use CoVoST train split for model training and CV dev (test) split for evaluation. The

---

1[https://github.com/pytorch/fairseq/tree/master/examples/language_model](https://github.com/pytorch/fairseq/tree/master/examples/language_model) They are mostly extracted from articles without context, which lack clarity for appropriate translations.
Table 2: Test WER (CER) for monolingual ASR; test BLEU for bilingual MT, cascaded ST (C-ST) and end-to-end ST. All non-English ASR and all ST model encoders are pre-trained on English ASR. * We use CER and character-level BLEU on Chinese and Japanese (no word segmentation).

### 2.3 Statistics

Basic statistics of CoVoST 2 are listed in Table 1, including speech duration, speaker counts as well as token counts for both transcripts and translations. As we can see, CoVoST 2 is diversified with large sets of speakers even on some of the low-resource languages (e.g. Persian, Welsh and Dutch). Moreover, they are distributed widely across 66 accent groups, 8 age groups and 3 gender groups.

### 3 Models

Our speech recognition (ASR) and ST models share the same BLSTM-based encoder-decoder architecture (Berard et al., 2018), which is similar to the Listen, Attend and Spell (LAS) architecture (Chan et al., 2016; Chiu et al., 2017; Park et al., 2019). Specifically, on the encoder side, audio features \( x \in \mathbb{R}^{T \times d_0} \) are first fed into a two-layer DNN with \( \tanh \) activations and hidden sizes \( d_1 \) and \( d_2 \). Then two 2D convolutional layers with kernel size 3x3 and stride 2x2 are applied to reduce the sequence length to \( \frac{T}{4} \). Both convolutional layers have 16 output channels and project the features to \( 4d_2 \) dimensions after flattening. Finally, the features are passed to a stack of \( l_e \) bidirectional LSTM layers of hidden size \( d_3 \) to form encoder output states \( h \in \mathbb{R}^{T \times 2d_3} \). For the decoder side, a stack of \( l_d \) LSTM layers with hidden size \( 2d_3 \) and additive attention (Bahdanau et al., 2014) is applied, followed by a linear projection to size \( d_o \). In the multilingual setting (En→All and All→All), we follow Inaguma et al. (2019) to force decoding into a given language by using a target language ID as the first token.

For MT, we use a Transformer base architecture (Vaswani et al., 2017) with \( l_e \) encoder layers, \( l_d \) decoder layers, 0.3 dropout, and shared embeddings for encoder/decoder inputs and decoder outputs. For multilingual models, encoders and decoders are shared as preliminary experimentation showed that this approach was competitive.

### 4 Experiments

We provide MT, cascaded ST and end-to-end ST baselines under bilingual settings as well as multilingual settings: All→En (A2E), En→All (E2A) and All→All (A2A). Similarly for ASR, we provide both monolingual and multilingual baselines.
Table 3: CV test WER for multilingual ASR and CV test BLEU for multilingual MT/ST on high-resource (Fr, De, Es and Ca) and low-resource X→En. † A multilingual model trained on all 22 languages.

| Bi. ST | Fr | De | Es | Ca | Ni | Tr | Ar | Sv | Lv | Sl | Ta | Ja | Id | Cy |
|--------|----|----|----|----|----|----|----|----|----|----|----|----|----|----|
| ASR†  | 19.4 | 21.1 | 16.3 | 12.6 | 42.4 | 52.3 | 80.1 | 76.5 | 75.6 | 68.6 | 92.6 | 56.8 | 72.4 | 66.7 |
| A2E MT† | 38.0 | 27.0 | 38.2 | 29.8 | 13.5 | 9.2 | 17.3 | 22.0 | 10.2 | 9.3 | 1.1 | 6.3 | 18.7 | 10.0 |
| A2A MT | 40.4 | 31.2 | 40.8 | 32.4 | 18.2 | 12.1 | 18.7 | 26.8 | 11.4 | 10.7 | 1.3 | 6.5 | 22.2 | 14.9 |
| † + 1 | 11.9 | 9.5 | 13.3 | 11.8 | 2.8 | 2.5 | 1.1 | 1.1 | 0.9 | 0.2 | 0.2 | 0.8 | 0.8 | 0.9 |
| † + 2 | 21.5 | 16.3 | 22.1 | 19.9 | 5.9 | 4.3 | 2.8 | 2.3 | 2.3 | 1.6 | 0.1 | 0.9 | 2.9 | 2.7 |
| A2E ST | 26.6 | 19.5 | 26.3 | 23.5 | 8.6 | 2.1 | 0.3 | 1.3 | 0.6 | 1.4 | 0.1 | 0.6 | 0.3 | 0.9 |
| A2A ST | 16.1 | 9.6 | 14.5 | 13.8 | 1.5 | 0.3 | 0.4 | 0.4 | 0.4 | 0.5 | 0.1 | 0.2 | 0.3 | 0.4 |

Table 4: CV Test WER for multilingual ASR and CV test BLEU for multilingual MT/ST on En→X (all directions have equal resource). † A multilingual model trained on all 22 languages.

| Bi. ST | De | Ca | Zh | Fa | Et | Mn | Tr | Ar | Sv | Lv | Sl | Ta | Ja | Id | Cy |
|--------|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|
| ASR†  | 12.5 | 17.1 | 20.0 | 9.1 | 9.3 | 6.4 | 6.7 | 9.1 | 18.1 | 8.7 | 11.6 | 7.4 | 25.6 | 15.2 | 18.9 |
| E2A MT† | 31.9 | 41.6 | 40.9 | 22.2 | 27 | 19.1 | 21.3 | 23.5 | 41.2 | 26.1 | 32.2 | 24.5 | 45.6 | 40.9 | 43.1 |
| A2A MT | 30.2 | 39.7 | 39.3 | 21.4 | 25 | 18.1 | 20 | 21.5 | 39.3 | 23.9 | 29.6 | 22.9 | 44.3 | 39.5 | 41.3 |
| † + 1 | 12.9 | 16.1 | 20.9 | 10.4 | 9.5 | 9.1 | 8.3 | 9.1 | 17 | 9.3 | 11.6 | 10 | 25.6 | 16.6 | 16.9 |
| † + 2 | 11.9 | 16.3 | 15.3 | 10.4 | 8.6 | 7.7 | 7.1 | 8.3 | 15.6 | 8.9 | 11.4 | 8.4 | 22 | 15.1 | 16.3 |
| E2A ST | 12.6 | 17.7 | 22.2 | 9.1 | 9.5 | 6.3 | 7.3 | 8 | 18.3 | 8.9 | 11.4 | 7.3 | 28.2 | 16.0 | 19.3 |
| A2A ST | 12.4 | 18.1 | 22.2 | 9.5 | 9.5 | 6.4 | 7.2 | 8.4 | 18.6 | 8.8 | 11.4 | 7.2 | 28.3 | 15.9 | 19.1 |

4.1 Experimental Settings

For all texts, we normalize the punctuation and build vocabularies with SentencePiece (Kudo and Richardson, 2018) without pre-tokenization. For ASR and ST, character vocabularies with 100% coverage are used. For bilingual MT models, BPE (Sennrich et al., 2016) vocabularies of size 5k are learned jointly on both transcripts and translations. For multilingual MT models, BPE vocabularies of size 40k are created jointly on all available source and target text. For MT and language pair a-b, we also contrast using only a-b training data and both a-b and b-a training data. The latter setting is referred to as +Rev subsequently.

We extract 80-channel log-mel filterbank features (windows with 25ms size and 10ms shift) using Kaldi (Povey et al., 2011), with per-utterance cepstral mean and variance normalization applied. We remove training samples having more than 3,000 frames or more than 512 characters for GPU memory efficiency.

For ASR and ST, we set $d_1 = 256$, $d_2 = 128$, $d_3 = 512$ and $d_o = 128$. We use $l_e = 3$ and $l_e = 2$ for bilingual models and $l_e = 5$ and $l_e = 3$ for multilingual models. We adopt SpecAugment (Park et al., 2019) (LB policy without time warping) to alleviate overfitting. To accelerate model training, we pre-train all non-English ASR and all ST models with English ASR model encoder. For MT, we set $l_d = l_d = 3$ for bilingual models and $l_d = l_d = 6$ for multilingual models. All models are implemented in Fairseq (Ott et al., 2019).

We use a beam size of 5 for all models and length penalty 1. We use the best checkpoint by validation loss for MT, and average the last 5 checkpoints for ASR and ST. For MT and ST, we report case-sensitive detokenized BLEU (Papineni et al., 2002) using sacreBLEU (Post, 2018) with default options, except for English-Chinese and English-Japanese where we report character-level BLEU. For ASR, we report character error rate (CER) on Japanese and Chinese (no word segmentation) and word error rate (WER) on the other languages using VizSeq (Wang et al., 2019). Before calculating WER (CER), sentences are tokenized by sacreBLEU tokenizers, lowercased and with punctuation removed (except for apostrophes and hyphens).
4.2 Monolingual and Bilingual Baselines

Table 2 reports monolingual baselines for ASR and bilingual MT, cascaded ST (C-ST) and end-to-end ST baselines. As expected, the quality of transcriptions and translations is very dependent on the amount of training data per language pair. The poor results obtained on low resource pairs can be improved by leveraging training data from the opposite direction for MT and C-ST. These results serve as baseline for the research community to improve upon, including methods such as multilingual training, self-supervised pre-training and semi-supervised learning.

4.3 Multilingual Baselines

A2E, E2E and A2A baselines are reported in Table 3 for language pairs into English and in Table 4 for language pairs out of English. Multilingual modeling is shown to be a promising direction for low-resource ST.

5 Conclusion

We introduced CoVoST 2, the largest speech-to-text translation corpus to date for language coverage and total volume, with 21 languages into English and English into 15 languages. We also provided extensive monolingual, bilingual and multilingual baselines for ASR, MT and ST. CoVoST 2 is free to use under CC0 license and enables the research community to develop methods including, but not limited to, massive multilingual modeling, ST modeling for low resource languages, self-supervision for multilingual ST, semi-supervised modeling for multilingual ST.

References

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henriety, 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. Marseille, France. European Language Resources Association.

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. Florence, Italy. Association for Computational Linguistics.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473.

Alexandre Bérard, Laurent Besacier, Ali Can Koçabiyıkolu, and Olivier Piotquin. 2018. End-to-end automatic speech translation of audiobooks. In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 6224–6228. IEEE.

Alexandre Berard, Olivier Piotquin, Christophe Servan, and Laurent Besacier. 2016. Listen and translate: A proof of concept for end-to-end speech-to-text translation. In Proceedings of the 2016 NeurIPS Workshop on End-to-end Learning for Speech and Audio Processing.

William Chan, Navdeep Jaitly, Quoc Le, and Oriol Vinyals. 2016. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 4960–4964. IEEE.

Chung-Cheng Chiu, Tara N. Sainath, Yonghui Wu, Rohit Prabhavalkar, Patrick Nguyen, Zhifeng Chen, Anjuli Kannan, Ron J. Weiss, Kanishka Rao, Ekaterina Gonina, Navdeep Jaitly, Bo Li, Jan Chorowski, and Michiel Bacchiani. 2017. State-of-the-art speech recognition with sequence-to-sequence models.

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

Mattia A. Di Gangi, Roldano Cattoni, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2019. MuST-C: a Multilingual Speech Translation Corpus. 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 2012–2017, Minneapolis, Minnesota. Association for Computational Linguistics.

H. Inaguma, K. Duh, T. Kawahara, and S. Watanabe. 2019. Multilingual end-to-end speech translation. In 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), pages 570–577.

J. Iranzo-Sanchez, J. A. Silvestre-Cerd, J. Jorge, N. Rosell, A. Gimnez, 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.

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.

Nathan Ng, Kyra Yee, Alexei Baevski, Myle Ott, Michael Auli, and Sergey Edunov. 2019. Facebook FAIR’s WMT19 news translation task submission. In Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1), pages 314–319, Florence, Italy. Association for Computational Linguistics.

Pedro Javier Ortiz Suárez, Laurent Romary, and Benoît Sagot. 2020. A monolingual approach to contextualized word embeddings for mid-resource languages. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 1703–1714, Online. Association for Computational Linguistics.

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.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In Proceedings of the 40th annual meeting on association for computational linguistics, pages 311–318. Association for Computational Linguistics.

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. arXiv preprint arXiv:1904.08779.

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, Brussels, Belgium. Association for Computational Linguistics.

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 IEEE 2011 workshop on automatic speech recognition and understanding. IEEE Signal Processing Society.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. 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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In Advances in neural information processing systems, pages 5998–6008.

Changhan Wang, Anirudh Jain, Danlu Chen, and Jiatao Gu. 2019. Vizseq: A visual analysis toolkit for text generation tasks. EMNLP-IJCNLP 2019, page 253.

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, Marseille, France. European Language Resources Association.