Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 4197–4203
Marseille, 11–16 May 2020

CoVoST: A Diverse Multilingual Speech-To-Text Translation Corpus

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

Abstract

Spoken language translation has recently witnessed a resurgence in popularity, thanks to the development of end-to-end models and the creation of new corpora, such as Augmented LibriSpeech (Kocabiyikoglu et al., 2018) and MuST-C (Di Gangi et al., 2019). Existing datasets involve language pairs with English as a source language, involve very specific domains or are low resource. We introduce CoVoST, a multilingual speech-to-text translation corpus from 11 languages into English, diversified with over 11,000 speakers and over 60 accents. We describe the dataset creation methodology and provide empirical evidence of the quality of the data. We also provide initial benchmarks, including, to our knowledge, the first end-to-end many-to-one multilingual models for spoken language translation. CoVoST is released under CC0 license and free to use. We also provide additional evaluation data derived from Tatoeba under CC licenses.

Keywords: corpus, multilingual speech-to-text translation, spoken language translation, end-to-end model, CC licensed

1. Introduction

End-to-end speech-to-text translation (ST) has attracted much attention recently (Berard et al., 2016; Duong et al., 2016; Weiss et al., 2017; Bansal et al., 2017; Bérard et al., 2018) given its simplicity against cascading automatic speech recognition (ASR) and machine translation (MT) systems. The lack of labeled data, however, has become a major blocker for bridging the performance gaps between end-to-end models and cascading systems. Several corpora have been developed in recent years. Post et al. (2013) introduced a 180-hour Spanish-English ST corpus by augmenting the transcripts of the Fisher and Callhome corpora with English translations. Di Gangi et al. (2019) created the largest ST corpus to date from TED talks but the language pairs involved are out of English only. Beilharz et al. (2019) created a 110-hour German-English ST corpus from LibriVox audiobooks. Godard et al. (2018) created a Moboshi-French ST corpus as part of a rare language documentation effort. Woldeyohannis et al. (2018) provided an Amharic-English ST corpus in the tourism domain. Boito et al. (2019) created a multilingual ST corpus involving 8 languages from a multilingual speech corpus based on Bible readings (Black, 2019). Previous work either involves language pairs out of English, very specific domains, very low resource languages or a limited set of language pairs. This limits the scope of study, including the latest explorations on end-to-end multilingual ST (Inaguma et al., 2019; Gangi et al., 2019). Our work is mostly similar and concurrent to Iranzo-Sanchez et al. (2019) who created a multilingual ST corpus from the European Parliament proceedings. The corpus we introduce has larger speech durations and more translation tokens. It is diversified with multiple speakers per transcript/translation. Finally, we provide additional out-of-domain test sets.

In this paper, we introduce CoVoST, a multilingual ST corpus based on Common Voice (Ardila et al., 2019) for 11 languages into English, diversified with over 11,000 speakers and over 60 accents. It includes a total 708 hours of French (Fr), German (De), Dutch (Nl), Russian (Ru), Spanish (Es), Italian (It), Turkish (Tr), Persian (Fa), Swedish (Sv), Mongolian (Mn) and Chinese (Zh) speeches, with French and German ones having the largest durations among existing public corpora. We also collect an additional evaluation corpus from Tatoeba[1] for French, German, Dutch, Russian and Spanish, resulting in a total of 9.3 hours of speech. Both corpora are created at the sentence level and do not require additional alignments or segmentation. Using the official Common Voice train-development-test split, we also provide baseline models, including, to our knowledge, the first end-to-end many-to-one multilingual ST models. CoVoST is released under CC0 license and free to use. The Tatoeba evaluation samples are also available under friendly CC licenses. All the data can be obtained at `https://github.com/facebookresearch/covost`.

2. Data Collection and Processing

2.1. Common Voice (CoVo)

Common Voice (Ardila et al., 2019, CoVo) is a crowdsourcing speech recognition corpus with an open CC0 license. Contributors record voice clips by reading from a bank of donated sentences. Each voice clip was validated by at least two other users. Most of the sentences are covered by multiple speakers, with potentially different genders, age groups or accents.

Raw CoVo data contains samples that passed validation as well as those that did not. To build CoVoST, we only use the former one and reuse the official train-development-test partition of the validated data. As of January 2020, the latest CoVo 2019-06-12 release includes 29 languages. CoVoST is currently built on that release and covers the following 11 languages: French, German, Dutch, Russian, Spanish, Italian, Turkish, Persian, Swedish, Mongolian and Chinese.

Validated transcripts were sent to professional translators. Note that the translators had access to the transcripts but not

* Work done as part of the Facebook AI Residency.

[1] https://tatoeba.org/eng/downloads

			Sentences		Speaker		Tokens		Average Length		Word Vocab	
		Hours	All	Unique	Count	Accents	Source	Target	Source	Target	Source	Target
	Train	87.1	78.9K	27.5K	436	9	787.7K	800.8K	10.0	10.1	29.7K	25.3K
Fr	Dev	38.3	34.1K	10.4K	1,001	17	336.0K	339.0K	9.8	9.9	14.6K	12.8K
	Test	46.3	39.2K	10.4K	2,884	24	391.6K	392.0K	10.0	10.0	14.9K	13.2K
	TT	1.6	4.5K	4.5K	3	N/A	25.6K	24.4K	5.7	5.4	3.4K	2.2K
	Train	71.0	60.3K	8.5K	1,109	7	549.5K	605.5K	9.1	10.0	16.3K	11.8K
De	Dev	88.1	77.3K	5.6K	2,337	11	690.8K	759.2K	8.9	9.8	12.0K	9.3K
	Test	168.3	145.8K	5.6K	4,781	13	1.31M	1.43M	9.0	9.8	12.3K	9.5K
	TT	4.0	9.1K	9.1K	5	N/A	45.8K	47.0K	5.0	5.1	4.9K	3.2K
	Train	4.4	4.3K	1.9K	35	2	39.9K	41.5K	9.4	9.8	9.2K	7.7K
Nl	Dev	5.3	5.0K	1.7K	126	2	48.0K	50.0K	9.4	9.8	4.3K	4.0K
	Test	8.2	7.7K	1.7K	461	3	73.6K	76.5K	9.5	9.9	4.3K	3.9K
	TT	0.3	0.6K	0.6K	1	N/A	2.9K	3.2K	5.1	5.5	0.7K	0.7K
	Train	10.2	7.1K	2.1K	6	N/A	75.2K	91.2K	10.6	12.8	7.4K	4.8K
Ru	Dev	9.0	6.4K	1.7K	9	N/A	66.3K	80.5K	10.4	12.7	6.5K	4.3K
	Test	8.2	5.8K	1.7K	61	N/A	59.6K	72.3K	10.3	12.5	6.2K	4.1K
	TT	1.5	2.7K	2.7K	5	N/A	15.2K	18.4K	5.7	6.9	4.2K	2.7K
	Train	20.9	18.3K	6.9K	319	11	162.8K	177.3K	8.9	9.7	5.6K	4.5K
Es	Dev	3.2	2.7K	2.6K	89	10	24.5K	26.6K	9.0	9.8	5.2K	4.2K
	Test	3.5	2.7K	2.6K	457	10	24.2K	26.4K	8.8	9.6	5.2K	4.1K
	TT	1.9	2.8K	2.8K	2	N/A	22.2K	23.6K	7.8	8.3	4.2K	3.3K
	Train	13.4	10.0K	6.4K	28	1	116.7K	127.8K	11.8	12.9	12.8K	9.9K
It	Dev	10.6	8.3K	4.6K	93	1	92.8K	103.1K	11.2	12.4	10.6K	8.1K
	Test	12.8	8.9K	4.6K	577	1	100.8K	110.3K	11.4	12.5	10.4K	8.1K
	Train	2.6	2.5K	1.8K	14	1	18.5K	24.6K	7.3	9.7	4.7K	3.4K
Tr	Dev	3.0	2.9K	1.6K	58	1	21.0K	28.1K	7.2	9.6	4.3K	3.1K
	Test	3.8	3.4K	1.6K	323	1	24.7K	33.2K	7.2	9.7	4.2K	3.1K
	Train	19.9	16.2K	2.4K	352	N/A	133.8K	164.9K	8.3	10.2	5.5K	3.9K
Fa	Dev	22.8	18.4K	2.1K	677	N/A	150.8K	185.0K	8.2	10.1	5.1K	3.7K
	Test	23.9	19.1K	2.1K	1,210	N/A	157.9K	193.5K	8.3	10.2	5.1K	3.7K
	Train	1.2	1.6K	1.6K	2	N/A	10.9K	12.2K	6.8	7.6	2.3K	2.0K
Sv	Dev	1.1	1.2K	1.2K	4	N/A	8.0K	8.9K	6.4	7.2	1.7K	1.6K
	Test	1.0	1.1K	1.1K	41	N/A	7.8K	8.6K	6.8	7.5	1.7K	1.6K
	Train	3.0	2.1K	2.1K	4	N/A	23.0K	27.2K	11.0	13.0	8.2K	4.4K
Mn	Dev	2.5	1.6K	1.4K	22	N/A	17.9K	21.6K	11.1	13.3	6.2K	3.5K
	Test	2.9	1.8K	1.6K	204	N/A	20.2K	24.1K	11.0	13.1	6.8K	3.8K
	Train	4.0	2.3K	2.3K	9	6	50.8K	37.9K	22.1	16.5	2.6K	8.2K
Zh	Dev	3.5	2.0K	2.0K	24	13	44.0K	33.6K	22.5	17.2	2.6K	7.6K
	Test	3.7	2.0K	2.0K	244	22	43.6K	33.0K	22.1	16.7	2.6K	7.5K

Table 1: Basic statistics of CoVoST and TT evaluation set. Token statistics are based on Moses-tokenized sentences. Speaker demographics is partially available.

the corresponding voice clips since clips would not carry additional information. Since transcripts were duplicated due to multiple speakers, we deduplicated the transcripts before sending them to translators. As a result, different voice clips of the same content (transcript) will have identical translations in CoVoST for train, development and test splits.

In order to control the quality of the professional translations, we applied various sanity checks to the translations (Guzmán et al., 2019). 1) For German-English, French-English and Russian-English translations, we computed sentence-level BLEU (Chen and Cherry, 2014) with the NLTK (Bird et al., 2009) implementation between the human translations and the automatic translations produced by a state-of-the-art system (Ng et al., 2019) (the French-English system was a Transformer *big* (Vaswani et al., 2017) separately trained on WMT14). We applied this method to these three language pairs only as we are confident about the quality of the corresponding systems. Translations with a score that was too low were manually inspected and sent back to the translators when needed. 2) We manually inspected examples where the source transcript was identical to the translation. 3) We measured the perplexity of the translations using a language model trained on a large amount of clean monolingual data (Ng et al., 2019). We manually inspected examples where the translation had a high perplexity and sent them back to translators accordingly. 4) We computed the ratio of English characters in the translations. We manually inspected examples with a low ratio and sent them back to translators accordingly. 5) Finally, we used VizSeq (Wang et al., 2019) to calculate similarity scores between transcripts and translations based on LASER cross-lingual sentence embeddings (Artetxe and Schwenk, 2019). Samples with low scores were manually inspected and sent back for translation when needed.

We also checked the overlap between train, development and test sets in terms of transcripts and voice clips (via MD5 file hashing), and confirmed they are disjoint.

2.2. Tatoeba (TT)

Tatoeba (TT) is a community built language learning corpus having sentences aligned across multiple languages with the corresponding speech partially available. Its sentences are on average shorter than those in CoVoST (see also Table 1) given the original purpose of language learning. Sentences in TT are licensed under CC BY 2.0 FR and part of the audio is available under various CC licenses.

We construct an evaluation set from TT (for French, German, Dutch, Russian and Spanish) as a complement to CoVoST development and test sets. We collect (speech, transcript, English translation) triplets for the 5 languages and do not include those whose speech has a broken URL or is not CC licensed. We further filter these samples by sentence lengths (minimum 4 words including punctuations) to reduce the portion of short sentences. This makes the resulting evaluation set closer to real-world scenarios and more challenging.

We run the same quality checks for TT as for CoVoST but we do not find poor quality translations according to our criteria. Finally, we report the overlap between CoVo transcripts and TT sentences in Table 2. We found a minimal overlap, which makes the TT evaluation set a suitable additional test set when training on CoVoST.

CoVo split	Fr	De	Nl	Ru	Es
Train	1.7%	0.2%	0.2%	0.1%	0.1%
Dev	1.0%	0.1%	0.3%	0.0%	0.1%
Test	0.9%	0.3%	0.3%	0.0%	0.4%

Table 2: TT-CoVo transcript overlapping rate.

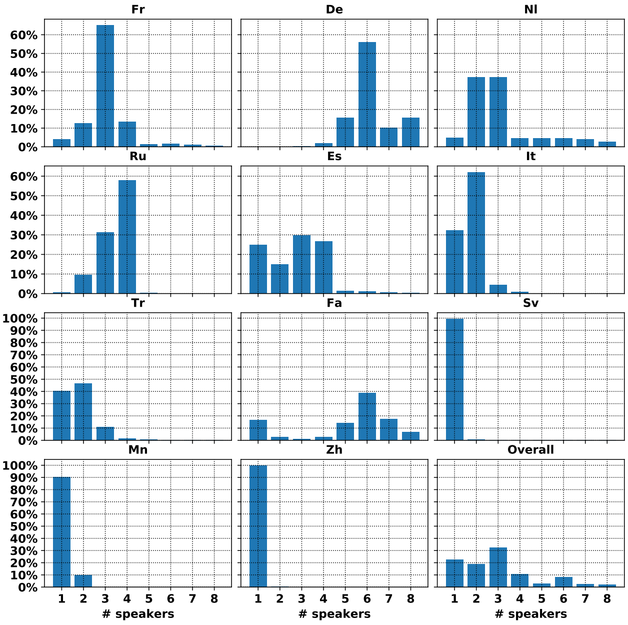

Figure 1: CoVoST transcript distribution by number of speakers.

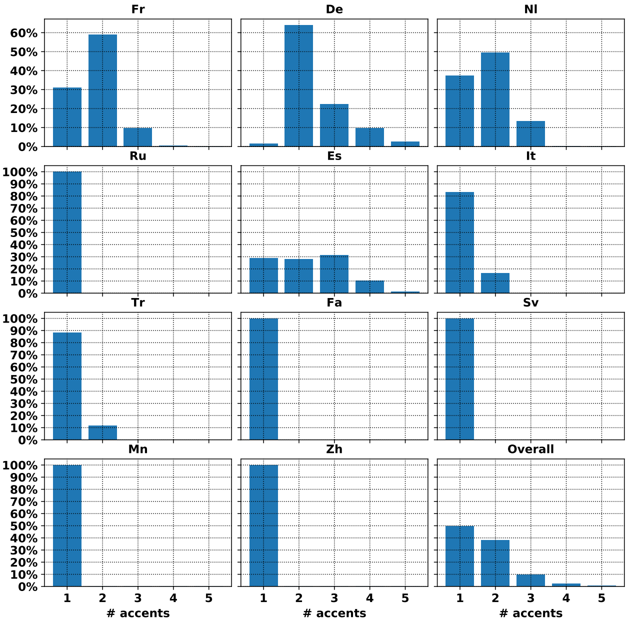

Figure 2: CoVoST transcript distribution by number of speaker accents.

3. Data Analysis

Basic Statistics Basic statistics for CoVoST and TT are listed in Table 1 including (unique) sentence counts, speech durations, speaker demographics (partially available) as well as vocabulary and token statistics (based on Moses-tokenized sentences by sacreMoses[2]) on both transcripts and translations. We see that CoVoST has over 327 hours of German speeches and over 171 hours of French speeches, which, to our knowledge, corresponds to the largest corpus among existing public ST corpora (the second largest is 110 hours (Beilharz et al., 2019) for German and 38 hours (Iranzo-Sanchez et al., 2019) for French). Moreover, CoVoST has a total of 18 hours of Dutch speeches, to our knowledge, contributing the first public Dutch ST resource. CoVoST also has around 27-hour Russian speeches, 37-hour Italian speeches and 67-hour Persian speeches, which is 1.8 times, 2.5 times and 13.3 times of the previous largest public one (Black, 2019). Most of the sentences (transcripts) in CoVoST are covered by multiple speakers with potentially different accents, resulting in a rich diversity in the speeches. For example, there are over 1,000 speakers and over 10 accents in the French and German development / test sets. This enables good coverage of speech variations in both model training and evaluation.

[2] https://github.com/alvations/sacremoses

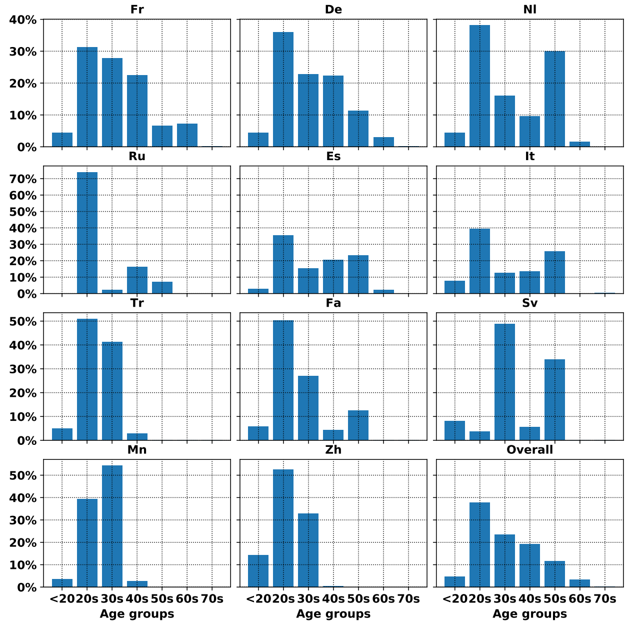

Figure 3: CoVoST transcript distribution by speaker age groups.

Speaker Diversity As we can see from Table 1, CoVoST is diversified with a rich set of speakers and accents. We further inspect the speaker demographics in terms of sample distributions with respect to speaker counts, accent counts and age groups, which is shown in Figure 1, 2 and 3. We observe that for 8 of the 11 languages, at least 60% of the sentences (transcripts) are covered by multiple speakers. Over 80% of the French sentences have at least 3 speakers. And for German sentences, even over 90% of them have at least 5 speakers. Similarly, we see that a large portion of sentences are spoken in multiple accents for French, German, Dutch and Spanish. Speakers of each language also spread widely across different age groups (below 20, 20s, 30s, 40s, 50s, 60s and 70s).

4. Baseline Results

We provide baselines using the official train-development-test split on the following tasks: automatic speech recognition (ASR), machine translation (MT) and speech translation (ST).

4.1. Experimental Settings

Data Preprocessing We convert raw MP3 audio files from CoVo and TT into mono-channel waveforms, and downsample them to 16,000 Hz. For transcripts and translations, we normalize the punctuation, we tokenize the text with sacreMoses and lowercase it. For transcripts, we further remove all punctuation markers except for apostrophes. We use character vocabularies on all the tasks, with 100% coverage of all the characters. Preliminary experimentation showed that character vocabularies provided more stable training than BPE. For MT, the vocabulary is created jointly on both transcripts and translations. We extract 80-channel log-mel filterbank features, computed with a 25ms window size and 10ms window shift using torchaudio[3]. The features are normalized to 0 mean and 1.0 standard deviation. We remove samples having more than 3,000 frames or more than 256 characters for GPU memory efficiency (less than 25 samples are removed for all languages).

	CoVoST Test		TT	
	WER	CER	WER	CER
En	36.1	20.3		
Fr	24.6	10.7	43.9	23.7
De	40.9	16.7	32.5	15.1
Nl	56.9	27.2	53.8	27.0
Ru	54.6	20.6	66.9	32.2
Es	50.6	22.0	61.7	26.3
It	38.2	14.1		
Tr	56.0	22.1		
Fa	65.4	32.3		
Sv	82.1	46.5		
Mn	76.7	38.5		
Zh	59.2	33.2		

Table 3: WER and CER scores for ASR models. Non-English models are pretrained using English model's encoder.

Model Training Our ASR and ST models follow the architecture in Bérard et al. (2018), but have 3 decoder layers like that in Pino et al. (2019). We pretrain their encoders on 120-hour English ASR data from Common Voice (2019-06-12 release). For MT, we use a Transformer *base* architecture (Vaswani et al., 2017), but with 3 encoder layers, 3 decoder layers and 0.3 dropout. We use a batch size of 10,000 frames for ASR and ST, and a batch size of 4,000 tokens for MT. We train all models using Fairseq (Ott et al., 2019) for up to 200,000 updates. We use SpecAugment (Park et al., 2019) for ASR and ST to alleviate overfitting.

Inference and Evaluation We use a beam size of 5 for all models. 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-insensitive tokenized BLEU (Papineni et al., 2002) using sacreBLEU (Post, 2018). For ASR, we report word error rate (WER) and character error rate (CER) using VizSeq where both the hypothesis and reference are tokenized, lowercased and with punctuation removed.

4.2. Automatic Speech Recognition (ASR)

For simplicity, we use the same model architecture for ASR and ST. Table 3 shows the word error rate (WER) and character error rate (CER) for ASR models. We see that French and German perform the best given they are the two highest resource languages in CoVoST. Italian is among the best as well, which is mid-resource and has limited accents. Persian is also mid-resource but is challenging because of rich speaker diversity. Most of the other languages are low resource (especially Swedish and Mongolian) and the ASR models are having difficulties to learn from this data even with pre-trained encoders.

[3] https://github.com/pytorch/audio

	CoVoST Test	TT
Fr	29.8	25.4
De	8.0	8.1
Nl	3.2	5.3
Ru	3.0	0.7
Es	11.0	2.3
It	8.7	
Tr	0.9	
Fa	0.5	
Sv	5.0	
Mn	0.2	
Zh	5.5	

Table 4: BLEU scores for MT models.

	CoVoST Test / TT				
	Fr	De	Nl	Ru	Es
Fr	21.4/10.9				
De		7.6/7.5			
Nl			3.4/5.0		
Ru				4.8/1.1	
Es					6.1/1.9
De+Fr	22.1/**11.9**	**9.3/10.5**			
Nl+Fr	**22.7/13.3**		2.9/3.5		
Ru+Fr	**22.7/13.1**			**7.7**/1.0	
Es+Fr	**22.8/13.2**				5.1/**3.1**
First 5 ⋆	21.8/11.4	**9.8/12.1**	3.4/5.7	**7.0**/1.2	3.9/2.8
All 11	21.5/10.7	**9.8/11.1**	2.8/**6.7**	**6.2**/1.2	4.1/**3.4**

Table 5: BLEU scores for end-to-end ST models. ST model encoders are pre-trained on English ASR. The rows indicate the languages used for training, the columns indicate the CoVoST test / TT BLEU scores on corresponding languages. Multilingual model scores that are better/worse than the bilingual baseline (by at least 1.0) are in bold/underlined. French (Fr) is highest resource among all 11 languages. ⋆ Fr, De, Nl, Ru and Es.

4.3. Machine Translation (MT)

MT models take transcripts (without punctuation) as inputs and outputs translations (with punctuation). For simplicity, we do not change the text preprocessing methods for MT to correct this mismatch. Moreover, this mismatch also exists in cascading ST systems, where MT model inputs are the outputs of an ASR model. Table 4 shows the BLEU scores of MT models. We notice that the results are consistent with what we see from ASR models. For example thanks to abundant training data, French has a decent BLEU score of 29.8/25.4. German doesn't perform well, because of less richness of content (transcripts). The other languages are relatively low resource in CoVoST and it is difficult to train decent models without additional data or pre-training techniques.

4.4. Speech Translation (ST)

CoVoST is a many-to-one multilingual ST corpus. While end-to-end one-to-many and many-to-many multilingual ST models have been explored very recently (Inaguma et al., 2019; Gangi et al., 2019), many-to-one multilingual models, to our knowledge, have not. We hence use CoVoST to examine this setting. Table 5 and 6 show the BLEU scores for both bilingual and multilingual end-to-end ST models trained on CoVoST. We observe that combining speeches from multiple languages brings gains to high-resource languages (Fr and De) consistently. Some mid-resource/low-resource languages (Ru, It and Zh) are improved as well. This includes combinations of distant languages, such as Ru+Fr and Zh+Fr. We simply provide the most basic many-to-one multilingual baselines here, and leave the full exploration of the best configurations to future work. Finally, we note that for some language pairs, absolute BLEU numbers are relatively low as we restrict model training to the supervised data. We encourage the community to improve upon those baselines, for example by leveraging semi-supervised training.

	CoVoST Test / TT						
	Fr	It	Tr	Fa	Sv	Mn	Zh
Fr	21.4/10.9						
It		6.5					
Tr			3.1				
Fa				2.8			
Sv					1.9		
Mn						0.3	
Zh							5.6
It+Fr	**23.1/13.3**	**8.6**					
Tr+Fr	**22.6/12.7**		2.4				
Fa+Fr	**22.5/12.9**			2.3			
Sv+Fr	**22.8/12.7**				0.7		
Mn+Fr	**22.8/13.8**					0.3	
Zh+Fr	**22.4/13.2**						**6.7**
All 11	21.5/10.7	5.9	1.8	1.8	0.9	0.2	5.1

Table 6: BLEU scores for end-to-end ST models (continuation of Table 5).

4.5. Multi-Speaker Evaluation

In CoVoST, large portion of transcripts are covered by multiple speakers with different genders, accents and age groups. Besides the standard corpus-level BLEU scores, we also want to evaluate model output variance on the same content (transcript) but different speakers. We hence propose to group samples (and their sentence BLEU scores) by transcript, and then calculate average per-group mean and average coefficient of variation defined as follows:

$$\text{BLEU}_{MS} = \frac{1}{|G|} \sum_{g \in G} \text{Mean}(g)$$

and

$$\text{CoefVar}_{MS} = \frac{1}{|G'|} \sum_{g \in G'} \frac{\text{StandardDeviation}(g)}{\text{Mean}(g)}$$

where G is the set of sentence BLEU scores grouped by transcript and $G' = \{g | g \in G, |g| > 1, \text{Mean}(g) > 0\}$. BLEU_{MS} provides a normalized quality score as oppose to corpus-level BLEU or unnormalized average of sentence BLEU. And CoefVar_{MS} is a standardized measure of model stability against different speakers (the lower the better). Table 7 shows the BLEU_{MS} and CoefVar_{MS} of

	BLEU$_{MS}$		CoefVar$_{MS}$	
	Breakdown	All	Breakdown	All
Fr		13.38		0.77
De		3.3		2.12
Nl		0.81		1.28
Ru		2.22		0.67
Es		3.36		1.02
It		2.46		0.84
Tr		0.79		0.80
Fa		1.38		1.43
Sv		0.39		-
Mn		0.03		-
Zh		3.24		1.0
De+Fr	4.15/13.31	10.06	2.14/0.79	1.12
Nl+Fr	0.70/14.01	12.14	1.54/0.80	0.82
Ru+Fr	3.93/14.05	12.64	0.76/0.80	0.79
Es+Fr	2.10/14.22	11.69	1.18/0.77	0.80
It+Fr	3.37/14.37	10.99	0.82/0.80	0.80
Tr+Fr	0.60/14.06	12.23	0.80/0.79	0.79
Fa+Fr	0.82/14.07	11.84	1.50/0.78	0.79
Sv+Fr	0.13/14.09	12.71	-/0.80	0.80
Mn+Fr	0.02/14.34	12.41	1.0/0.78	0.78
Zh+Fr	4.71/14.17	12.65	0.13/0.79	0.79
First 5		7.78		1.19
All 11		5.28		1.14

Table 7: Average per-group mean and average coefficient of variation for ST sentence BLEU scores on CoVoST test set (groups correspond to one transcript and multiple speakers). The latter is unavailable for Swedish and Mongolian because models are unable to acheive non-zero scores on multi-speaker samples.

our ST models on CoVoST test set. We see that German and Persian have the worst CoefVar$_{MS}$ (least stable) given their rich speaker diversity in the test set and relatively small train set (see also Figure 1 and Table 1). Dutch also has poor CoefVar$_{MS}$ because of the lack of training data. Multilingual models may improve BLEU$_{MS}$ but have comparable CoefVar$_{MS}$.

5. Conclusion

We introduce a multilingual speech-to-text translation corpus, CoVoST, for 11 languages into English, diversified with over 11,000 speakers and over 60 accents. We also provide baseline results, including, to our knowledge, the first end-to-end many-to-one multilingual model for spoken language translation. CoVoST is free to use with a CC0 license, and the additional Tatoeba evaluation samples are also CC-licensed.

6. Bibliographical References

Ardila, R., Branson, M., Davis, K., Henretty, M., Kohler, M., Meyer, J., Morais, R., Saunders, L., Tyers, F. M., and Weber, G. (2019). Common voice: A massively-multilingual speech corpus.

Artetxe, M. and Schwenk, H. (2019). Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Bansal, S., Kamper, H., Lopez, A., and Goldwater, S. (2017). Towards speech-to-text translation without speech recognition. *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers.*

Beilharz, B., Sun, X., Karimova, S., and Riezler, S. (2019). Librivoxdeen: A corpus for german-to-english speech translation and speech recognition.

Berard, A., Pietquin, O., Servan, C., and Besacier, L. (2016). Listen and translate: A proof of concept for end-to-end speech-to-text translation.

Bérard, A., Besacier, L., Kocabiyikoglu, A. C., and Pietquin, O. (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.

Bird, S., Klein, E., and Loper, E. (2009). *Natural language processing with Python: analyzing text with the natural language toolkit.* " O'Reilly Media, Inc.".

Black, A. W. (2019). Cmu wilderness multilingual speech dataset. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5971–5975, May.

Boito, M. Z., Havard, W. N., Garnerin, M., Ferrand, E. L., and Besacier, L. (2019). Mass: A large and clean multilingual corpus of sentence-aligned spoken utterances extracted from the bible.

Chen, B. and Cherry, C. (2014). A systematic comparison of smoothing techniques for sentence-level BLEU. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 362–367, Baltimore, Maryland, USA, June. Association for Computational Linguistics.

Di Gangi, M. A., Cattoni, R., Bentivogli, L., Negri, M., and Turchi, M. (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, June. Association for Computational Linguistics.

Duong, L., Anastasopoulos, A., Chiang, D., Bird, S., and Cohn, T. (2016). An attentional model for speech translation without transcription. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 949–959, San Diego, California, June. Association for Computational Linguistics.

Gangi, M. A. D., Negri, M., and Turchi, M. (2019). One-to-many multilingual end-to-end speech translation.

Godard, P., Adda, G., Adda-Decker, M., Benjumea, J., Besacier, L., Cooper-Leavitt, J., Kouarata, G.-N., Lamel, L., Maynard, H., Mueller, M., Rialland, A., Stueker, S., Yvon, F., and Zanon-Boito, M. (2018). A very low resource language speech corpus for computational language documentation experiments. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

Guzmán, F., Chen, P.-J., Ott, M., Pino, J., Lample, G., Koehn, P., Chaudhary, V., and Ranzato, M. (2019). The FLORES evaluation datasets for low-resource machine translation: Nepali–English and Sinhala–English. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6097–6110, Hong Kong, China, November. Association for Computational Linguistics.

Inaguma, H., Duh, K., Kawahara, T., and Watanabe, S. (2019). Multilingual end-to-end speech translation.

Iranzo-Sanchez, J., Silvestre-Cerda, J. A., Jorge, J., Rosello, N., Gimenez, A., Sanchis, A., Civera, J., and Juan, A. (2019). Europarl-st: A multilingual corpus for speech translation of parliamentary debates.

Kocabiyikoglu, A. C., Besacier, L., and Kraif, O. (2018). Augmenting librispeech with french translations: A multimodal corpus for direct speech translation evaluation.

Ng, N., Yee, K., Baevski, A., Ott, M., Auli, M., and Edunov, S. (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, August. Association for Computational Linguistics.

Ott, M., Edunov, S., Baevski, A., Fan, A., Gross, S., Ng, N., Grangier, D., and Auli, M. (2019). fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (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.

Park, D. S., Chan, W., Zhang, Y., Chiu, C.-C., Zoph, B., Cubuk, E. D., and Le, Q. V. (2019). Specaugment: A simple data augmentation method for automatic speech recognition. *arXiv preprint arXiv:1904.08779*.

Pino, J., Puzon, L., Gu, J., Ma, X., McCarthy, A. D., and Gopinath, D. (2019). Harnessing indirect training data for end-to-end automatic speech translation: Tricks of the trade. Zenodo.

Post, M., Kumar, G., Lopez, A., Karakos, D., Callison-Burch, C., and Khudanpur, S. (2013). Improved speech-to-text translation with the fisher and callhome spanish–english speech translation corpus. In *Proc. IWSLT*.

Post, M. (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, October. Association for Computational Linguistics.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Wang, C., Jain, A., Chen, D., and Gu, J. (2019). Vizseq: A visual analysis toolkit for text generation tasks. *EMNLP-IJCNLP 2019*, page 253.

Weiss, R. J., Chorowski, J., Jaitly, N., Wu, Y., and Chen, Z. (2017). Sequence-to-sequence models can directly translate foreign speech. *Interspeech 2017*, Aug.

Woldeyohannis, M., Besacier, L., and Meshesha, M., (2018). *A Corpus for Amharic-English Speech Translation: The Case of Tourism Domain*, pages 129–139. 07.