SpeechMatrix: A Large-Scale Mined Corpus of Multilingual Speech-to-Speech Translations

Paul-Ambroise Duquenne[1,2,*] and **Hongyu Gong**[1,*]
Ning Dong[1] and **Jingfei Du**[1]
Ann Lee[1] and **Vedanuj Goswani**[1] and **Changhan Wang**[1]
Juan Pino[1] and **Benoît Sagot**[2] and **Holger Schwenk**[1]
[1] Meta AI Research, [2] Inria
padqn,hygong,dnn,jingfeidu,schwenk@meta.com

Abstract

We present SpeechMatrix, a large-scale multilingual corpus of speech-to-speech translations mined from real speech of European Parliament recordings. It contains speech alignments in 136 language pairs with a total of 418 thousand hours of speech. To evaluate the quality of this parallel speech, we train bilingual speech-to-speech translation models on mined data only and establish extensive baseline results on EuroParl-ST, VoxPopuli and FLEURS test sets. Enabled by the multilinguality of SpeechMatrix, we also explore multilingual speech-to-speech translation, a topic which was addressed by few other works. We also demonstrate that model pre-training and sparse scaling using Mixture-of-Experts bring large gains to translation performance. The mined data and models are freely available.

1 Introduction

Research has progressed in the area of speech-to-speech translation (S2ST) with the goal of seamless communication among people who speak different languages. Direct S2ST models attract increasing research interest, e.g. (Jia et al., 2019). Compared to conventional cascaded models, direct models do not rely on intermediate text representations which make them applicable to the translation of languages without a well-defined writing script. Moreover, direct S2ST have the advantage of higher training and inference efficiency (Lee et al., 2022a).

Despite the benefits of direct approaches, model training is faced with the major issue of data scarcity. Human labeled speech data is expensive to create, there are very few data resources providing parallel speech, and the data amount is quite limited. To mitigate the data scarcity, some works have leveraged multitask learning (Jia et al., 2019; Lee et al., 2022a), data augmentation with speech variation (Jia et al., 2019), or with synthesized speech (Jia et al., 2022a; Popuri et al., 2022), and knowledge from pre-trained models (Lee et al., 2022b; Popuri et al., 2022) such as HuBERT (Hsu et al., 2021), wav2vec 2.0 (Baevski et al., 2020) and mBART (Liu et al., 2020).

Recently, the multilingual speech/text sentence embedding space from Duquenne et al. (2021) enabled the first speech mining results, aligning speech and text in different languages. Using this mined data to train direct speech-to-text and speech-to-speech translation systems can improve the performance of such models (Duquenne et al., 2021; Lee et al., 2022b). Finally, Duquenne et al. (2022) showed that such multilingual and multimodal sentence embeddings could be decoded into different languages and/or modalities in a zero-shot way, which suggests that multilingual speech content is well encoded in these fixed-size representations.

In this work, we trained speech encoders for 17 languages[1] and mined speech-to-speech alignments for all possible language pairs. To the best of our knowledge, SpeechMatrix is by far the largest freely available speech-to-speech translation corpus, with 136 language directions and an average of 1,537 hours of source speech in each direction for a total of 418 thousand hours. We demonstrate that strong S2ST models can be trained with these mined data and validate the good quality of the speech alignments across languages. We are open-sourcing the mined data, the speech encoders used for mining, multilingual HuBERT models in four language families for target unit generation, language-specific vocoders for speech synthesis from discrete units, and S2S models trained and presented in this work.[2]

[*] Equal contributions

[1] Czech (cs), German (de), English (en), Spanish (es), Estonian (et), Finnish (fi), French (fr), Croatian (hr), Hungarian (hu), Italian (it), Lithuanian (lt), Dutch (nl), Polish (pl), Portuguese (pt), Romanian (ro), Slovak (sk) and Slovenian (sl).

[2] https://github.com/facebookresearch/fairseq/tree/ust/examples/speech_matrix

2 Related Works

From bitext mining to speech mining. State-of-the-art machine translation models are trained on labeled translation data, commonly called bitexts. In addition to human-labeled data, automatic methods have been proposed to find parallel sentences in a source and target language in monolingual resources. This task is known as bitext mining. In order to decide if sentences should be aligned, some works used document meta-information (Resnik, 1999), cross-lingual document retrieval (Munteanu and Marcu, 2005) or machine translation and information retrieval (Abdul-Rauf and Schwenk, 2009; Bouamor and Sajjad, 2018). More recent work use multilingual sentence embeddings to perform bitext mining, calculating cosine similarity (Schwenk, 2018) or other margin-based similarity (Artetxe and Schwenk, 2018; Yang et al., 2019a) in the embedding space to decide which sentences should be aligned. A large variety of methods has been explored to learn multilingual sentence embedding spaces (España-Bonet et al., 2017; Schwenk and Douze, 2017; Artetxe and Schwenk, 2019; Yang et al., 2019a; Reimers and Gurevych, 2019; Yang et al., 2019b; Feng et al., 2020). Such methods enabled the creation of massively multilingual mined corpora of bitexts like the CCMatrix project (Schwenk et al., 2021a). These bitexts were successfully used to train state-of-the-art machine translation models, e.g. (Fan et al., 2020; Wang et al., 2022). Finally, existing sentence embedding spaces can be extended to new languages (Reimers and Gurevych, 2020; Heffernan et al., 2022) or the speech modality (Duquenne et al., 2021; Khurana et al., 2022) with knowledge distillation, also called teacher-student approach. These multilingual and multimodal sentence embeddings enabled to perform large-scale speech-text mining, or speech-speech mining for a small set of languages.

Speech-to-speech translation. Early works on speech-to-speech translation are cascaded systems typically consisting of automatic speech recognition (ASR), machine translation (MT) and text-to-speech synthesis (TTS) (Nakamura et al., 2006; Do et al., 2015). The reliance on intermediate text outputs makes cascaded models unable to support unwritten languages. Moreover multiple separate components make the training and inference inefficient. Given these limitations, there has been a recent surge of research interest in direct approaches to speech translation without the need of texts. Translatotron (Jia et al., 2019) is the first end-to-end S2ST model built upon a sequence-to-sequence architecture, which is trained to generate target spectrograms from source speech using multitask learning. As an improved version, Translatotron2 (Jia et al., 2022b) has the ability to preserve voice in translated speech. It adopts two decoders: a linguistic decoder to predict phoneme and an acoustic decoder to predict target spectrograms.

Another line of research replaces the target spectrograms in S2ST modeling with discrete units which are learned from a large amount of unlabeled speech (Lee et al., 2022a,b). Discrete units are shown to better capture linguistic content than spectrograms, making the translations more robust to speech variations from different speakers or prosody (Hsu et al., 2021).

Despite these progress on direct S2ST, direct approaches are faced with the challenge of limited parallel speech. We will describe existing speech-to-speech datasets in the next part. Data augmentation is a straightforward way of increasing the data such as speech transformation (Jia et al., 2019) and speech synthesis via TTS (Jia et al., 2022a). Multitask learning is a commonly used technique to better train the model on small amount of labeled data (Jia et al., 2022b; Lee et al., 2022a). Another effective method is to leverage pre-trained components for model initialization.

Speech translation corpora. The Fisher dataset, a collection of approximately 170 hours of telephone conversations in Spanish (Post et al., 2014), is commonly used as training data for Spanish-English S2ST, However it does not have parallel English speech. Previous works generate synthesized English speech from English text translations provided by Fisher. Another S2S dataset containing synthesized speech is CVSS which covers parallel S2ST translations from 21 languages into English. It is derived from Common Voice (Ardila et al., 2020) and CoVoST 2 (Wang et al., 2021b), and synthesizes speech from translated texts via TTS models. The release of VoxPopuli dataset provided the largest S2S translations in real speech so far (Wang et al., 2021a). It covers pairwise speech-to-speech translations among 15 languages, and each direction has less than 500 hours of speech. In another initiative named FLEURS, the text-to-text evaluation data of the FLoRes-101 benchmark (Goyal et al., 2022) was extended to the speech modality. Supporting 102 languages, FLEURS has a larger

Dataset	# of Languages	Avg. duration (h)	Source speech	Target speech
Fisher (Post et al., 2014)	2	127	Telephone conversation	Synthetic
MaSS (Boito et al., 2020)	8	20	Bible reading	Bible reading
VoxPopuli (Wang et al., 2021a)	15	82	European Parliament speech	Simultaneous interpretation
CVSS (C+T) (Jia et al., 2022c)	21	181	Read	Synthetic
FLEURS (Conneau et al., 2022)	102	12	Read	Read
SpeechMatrix (ours)	17	1537	European Parliament speech	European Parliament speech

Table 1: A comparison of existing speech-to-speech datasets.

language coverage than VoxPopuli, but it contains speech of only around 12 hours per language and it is intended to be used as N-way parallel test set.

In this work, we present SpeechMatrix, a large-scale multilingual speech-to-speech corpus mined from VoxPopuli (Wang et al., 2021a). It contains speech alignments in 136 language pairs with an average of $1,537$-hour source speech per direction. The main characteristics of these speech corpora are summarized in Table 1.

3 Speech-to-Speech Mining

The mining approach used in this work is built upon the core idea of encoding multilingual speech utterances into a shared embedding space. Speech encoders project utterances with similar semantic content to fixed-size representations, the resulting embedding vectors are close in the embedding space regardless of their languages. The closeness of speech embeddings thus reflects the similarity of speech content, which serves as the speech alignment score in the mining process. In this section, we discuss speech encoders and speech alignment mining.

3.1 Speech Encoders

We followed the teacher-student approach introduced in (Duquenne et al., 2021) and trained speech encoders with the supervision of the multilingual LASER text encoder (Schwenk et al., 2021b). The LASER text space has proven to have interesting semantic properties for mining purposes (Schwenk et al., 2021a; NLLB Team et al., 2022). In order to have similar semantic properties in the multilingual speech embedding space, the LASER text encoder is used as the teacher to train speech encoders. Transcriptions or written translation of the audio utterances are encoded with LASER text encoder as target vectors for speech encoders. During training, we minimize the cosine loss between fixed-size representations output by speech encoders, and the outputs of LASER text encoder (which weights are fixed during training). Speech encoders are initialized with the 2B parameter XLS-R model (Babu et al., 2021), which was pre-trained on nearly half a million hours of publicly available speech audio in 128 languages. Following (Duquenne et al., 2022), the fixed-size representation for speech are obtained with max-pooling of the encoder outputs which appeared to work better compared to other pooling methods. We summarize the architecture of the speech encoder training in Figure 1.

Duquenne et al. (2021) presented an ablation study on different text teacher choices: using transcription embeddings as target, written translation embeddings as target, or both. Their conclusion was that using both transcriptions and written translations embeddings as target yield the best similarity search results. We used various publicly available ASR data sets which cover our languages to train the speech encoders, like CoVoST 2 (Wang et al., 2020, 2021b), Common Voice (citation), EuroParl (Ardila et al., 2020), mTedx (Salesky et al., 2021), Must-C (Di Gangi et al., 2019) or VoxPopuli (Wang et al., 2021a), as well as speech translation data from the foreign languages into English and from English into German. We removed training samples where the transcription or the written translation consisted of multiple sentences, as LASER has been trained on single sentences only. For bet-

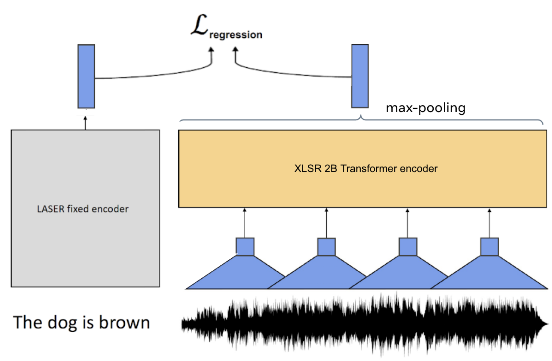

Figure 1: Architecture of speech encoders training.

Sim Search	cs	de	en	es	et	fi	fr	hr	hu	it	lt	nl	pl	pt	ro	sk	sl
# test sentences	1k	1.7k	1.5k	1.4k	47	0.4k	1.5k	0.3k	1k	1k	39	1k	1.6k	—	1.3k	0.6k	0.3k
Audio vs. transcriptions	0.6	1.0	0.2	0.7	0.0	0.7	0.5	0.3	1.1	4.9	0.0	0.8	0.9	—	0.9	0.7	3.1

Table 2: Similarity search error rates (in %) on VoxPopuli ASR test set.

ter training efficiency, we trained speech encoders for each language family instead of each language. The language grouping is provided in Table 6. To better handle imbalanced training data, we sample the training data from different languages with the same method and hyper-parameter as (Duquenne et al., 2021). For English (en), Slovenian (sl), Lithuanian (lt) and Dutch (nl), we also trained separate monolingual speech encoders that appeared to have lower valid cosine loss compared to multilingual speech encoders, and these ones were used for mining.

3.2 Evaluation of speech encoders

Similarity search is frequently used to evaluate multilingual text encoders, e.g. (Artetxe and Schwenk, 2018; Feng et al., 2020; Heffernan et al., 2022). The underlying approach is to first calculate the embeddings of all source and reference sentences in a test set. Then, for each source sentence, the closest embedding among all reference sentences is searched, if it is not the expected one, an error is counted. Following Duquenne et al. (2021), we extend similarity search to the multimodal setting and use the following score to measure similarity between the source audio, and the target transcriptions or translations:

$$sim(x,y) = cos(x,y) - \left(\sum_{z \in NN_k(x)} \frac{cos(x,z)}{2k} + \sum_{z \in NN_k(y)} \frac{cos(y,z)}{2k} \right) \quad (1)$$

where x and y are the source and target embeddings, and $NN_k(x)$ denotes the k nearest neighbors of x. We used $k = 4$. We evaluated similarity search of audio against transcriptions on VoxPopuli ASR test set in Table 2, which is our target domain as we plan to mine unlabeled speech from VoxPopuli (see subsection 3.3). We also evaluated similarity search of audio against written translations or transcriptions on CoVoST 2 test set in Table 3, in order to compare with previous work for German, English, Spanish and French (details can be found in the Appendix). Finally, we report text-to-text similarity search using the LASER text encoder as lower bound for the speech translation similarity search error rate since we use gold transcriptions to search against written translations. We report error rates (in %) that correspond to the percentage of audio utterances wrongly matched with text transcripts from the same test set. We notice that error rates are very low for all languages (below 5% and around 1 or 2% for most languages), which is an initial validation of the quality if the speech encoders before large-scale mining.

	de	en	es	et	fr	it	nl	pt	sl
# test sentences	14k	16k	13k	2k	15k	9k	2k	4k	0.4k
Audio									
vs. transcriptions	1.4	2.9	0.4	0.1	0.5	0.5	1.0	1.1	1.7
vs. en translations	3.3	—	1.3	1.0	1.5	1.7	4.4	1.9	4.4
Text transcription									
vs. en translations	2.0	—	1.0	0.1	1.0	1.3	2.4	0.7	0.8

Table 3: Similarity search error rates (in %) on CoVoST 2 test set.

3.3 Large-scale speech mining

We used VoxPopuli as our source of unlabeled speech for the 17 languages of focus. We downloaded the full unsegmented parliament session recordings from VoxPopuli github repository.[3] We present in Table 4 the number of hours of unlabeled speech for each language.

In principle, performing speech-to-speech or speech-to-text mining can be done with exactly the same pipeline than text-to-text mining. The calculation of the embeddings uses of course different encoders, but the rest of the pipeline is exactly the same. We follow the global mining approach as described in Schwenk et al. (2021a) and compare all segments in the source language with all segments in the target language. Similarity scores are calculated in both directions using the margin as described in Equation 1 considering $k = 16$ neighbors. Segments are considered to be parallel if the margin score exceeds a threshold, we use 1.06 if not specified otherwise. The reader is referred to Schwenk et al. (2021a) for a detailed description of the generic mining pipeline.

[3] https://github.com/facebookresearch/voxpopuli

Lang	Hours	Lang	Hours	Lang	Hours
cs	18.7k	fr	22.8k	pl	21.2k
de	23.2k	hr	8.1k	pt	17.5k
en	24.1k	hu	17.7k	ro	17.9k
es	21.4k	it	21.9k	sk	12.1k
et	10.6k	lt	14.4k	sl	11.3k
fi	14.2k	nl	19.0k		

Table 4: Number of hours of raw unlabeled speech from VoxPopuli by language.

There is however one important difference when processing speech: it is not straight-forward to segment the audio signal into parts which have the optimal granularity for mining. When aligning texts, the monolingual data is usually segmented into sentences, since a sentence is a well defined semantic concept and there are reliable algorithms to perform this sentence segmentation. The VoxPopuli recordings have a rather long duration, e.g. one hour and a half in average for English. We apply Voice Activity Detection (VAD) using Silero-VAD[4] which supports over 100 languages. The resulting segments do not necessarily correspond to sentences. On one hand, there may be a silence in the middle of a sentence, e.g. a hesitation. And on the other hand, two sentences may follow each other without a long silence separating them. We follow the "over segmentation approach" outlined in Duquenne et al. (2021): several possible segments are created and we let the mining algorithm decide which ones match best. Initial experiments suggest that sentences shorter than 1 sec or longer than 20 sec are unlikely to be aligned and were excluded.

After mining, the resulting speech alignments may have overlap as we over-segmented the unlabeled speech. Duquenne et al. (2021) introduced a post-processing method to remove overlaps between mined speech segments on the source speech side. We relax a little bit the post-processing of the mined data, allowing for some overlap between mined speech segments: for two audio segments that overlap on the source side, if the overlap represent more than 20% of the first segment and of the second segment, we discard the alignment with the lowest mining score. We did an ablation study on different percentage thresholds for one low resource, one mid-resource and one high-resource pair and found that 20% was the best threshold for all settings.

We report the statistics of the mined speech-to-speech translation pairs in Table 5, with a mining score threshold of 1.06. The mined data totals 418k hours of parallel speech with an average of 1,537 hours of source speech on all translation directions. While some high resource languages like English (en), Spanish (es) or French (fr) can reach up to 5k hours of aligned speech with other spoken languages; lower resource languages such as Estonian (et), Croatian (hr), Slovenian (sl) and Lithuanian (lt) obtain much less aligned speech, with only a few hours of aligned speech for Lithuanian.

We also performed mining of the source speech in sixteen languages against more than twenty billion English sentences from Common Crawl. This yielded between 827 and 3966 hours of speech-text alignments (see the last column of Table 5). Training and evaluation of speech-to-text translation is left for future research. An important advantage of our teacher-student approach is that all speech and text encoders are derived from the same LASER teacher and are mutually compatible. This enables us to perform speech-to-text mining for many more languages (NLLB Team et al., 2022).

3.4 Evaluation Data

Besides the speech-to-speech data mined as the train set, we also leverage labeled public speech datasets as the evaluation sets.

Test set. In our experiments, we derive test sets in speech translation from three public corpora, evaluating translation models trained on mined data across different domains.

- EuroParl-ST (EPST) (Iranzo-Sánchez et al., 2020). It is a multilingual speech-to-text translation corpus built from recordings of debates from the European Parliament, containing 72 translation directions in 9 languages.[5]

- VoxPopuli (Wang et al., 2021a). S2S data, as part of VoxPopuli release, provides aligned source and target speech together with source transcription. We prepare the speech-to-text data with target speech and source transcription as our test set. To ensure that there is no overlap between the mined data and VoxPopuli test sets, we need to remove speech from mined alignments which are from the same session as test samples. In order to keep as much mined data as possible, we use

[4] https://github.com/snakers4/silero-vad

[5] en, fr, de, it, es, pt, pl, ro and nl

	Speech targets																	Text
Src/Tgt	**cs**	**de**	**en**	**es**	**et**	**fi**	**fr**	**hr**	**hu**	**it**	**lt**	**nl**	**pl**	**pt**	**ro**	**sk**	**sl**	**en**
cs	-	2381	3208	2290	952	1312	2476	726	1396	2410	84	2377	2516	1867	1190	2146	452	2528
de	2386	-	4734	3113	901	1477	3536	498	1871	3476	41	3384	2632	2250	1281	1646	361	3073
en	3172	4676	-	4715	1585	2169	5178	824	2266	4897	82	4422	3583	3572	2258	2306	586	-
es	2240	3041	4708	-	862	1373	4446	528	1599	4418	47	3067	2646	3484	1857	1603	308	3966
et	943	892	1593	877	-	1201	934	265	1119	1019	39	1055	949	721	419	780	196	1578
fi	1296	1463	2180	1393	1197	-	1449	306	1473	1599	47	1654	1350	1128	621	977	260	1969
fr	2424	3457	5171	4455	923	1435	-	560	1711	4618	50	3273	2822	3384	1991	1657	326	3966
hr	736	507	854	553	273	317	588	-	328	615	24	546	660	433	277	586	136	1311
hu	1417	1897	2346	1672	1140	1507	1787	328	-	1855	68	1839	1566	1315	808	1064	311	2301
it	2404	3460	4948	4500	1028	1614	4700	607	1823	-	103	3414	2848	3421	1995	1656	474	2891
lt	78	38	79	46	37	44	48	21	61	95	-	77	80	35	18	64	6	827
nl	2322	3305	4396	3066	1040	1633	3269	521	1768	3355	80	-	2459	2399	1352	1646	458	2708
pl	2530	2646	3662	2735	967	1378	2913	656	1554	2883	88	2540	-	2121	1301	1892	431	2871
pt	1849	2224	3606	3525	722	1131	3421	421	1279	3403	37	2436	2087	-	1579	1358	247	3540
ro	1187	1275	2290	1894	423	627	2024	271	789	1996	19	1384	1288	1592	-	870	125	2784
sk	2127	1628	2329	1631	781	982	1685	574	1038	1650	69	1676	1869	1361	867	-	370	2090
sl	436	350	579	307	192	254	324	128	295	461	6	454	413	241	121	359	-	1267

Table 5: Duration statistics (hours of source speech) of aligned speech-to-speech data between each pair of 17 languages (for mining threshold of 1.06). The last column provides statistics for alignments of source speech against 21.5 billion sentences of English texts.

VoxPopuli test set only when a language direction is not covered by EPST considering their domain similarity. Moreover, similarity scores are provided to indicate the quality of VoxPopuli samples. To choose high-quality data, we sort all sessions in the VoxPopuli S2S data in a decreasing order of the average similarity score of their samples. We keep adding samples from highly ranked sessions to the test set until the test size hits 1000.

- FLEURS (Conneau et al., 2022). Built upon N-way text translations from FLoRes (Goyal et al., 2022), FLEURS provides speech for these texts and creates speech-to-speech data covering all mined directions. We take its source speech and target texts as the test data. In the case where multiple utterances correspond to one piece of source text, we generate one test pair for each source utterance respectively. FLEURS texts are from English Wikipedia, which is a different domain from VoxPopuli and EPST.

Valid set. Valid sets are prepared using VoxPopuli and FLEURS data in a similar way as test sets. For VoxPopuli, we extract a valid set of about 1000 samples by adding data from highly scored sessions which are not in the test set. FLEURS valid set is derived from its valid samples. We prepare speech-to-unit data from these selected valid samples by transforming the target speech into target units for speech-to-unit modeling which will be discussed in section 4.

4 Experiments & Results

To evaluate the quality of the mined data, we trained S2ST models on SpeechMatrix data and report the translation performance. We hope that these results will serve as baselines for future studies in speech translation.

4.1 Experimental Setup

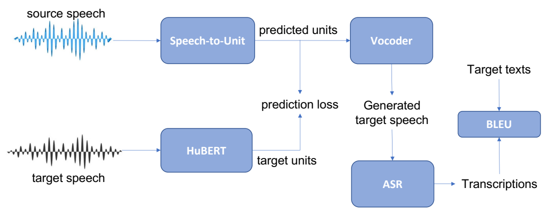

Figure 2: A Pipeline of Speech-to-Speech Translation and Evaluation.

The training and evaluation pipeline of speech-to-speech translation is shown in Figure 2. Recent progress in speech-to-speech translation modeling suggests to discretize the target speech waveform into a unit sequence, relieving models from the complexity of predicting continuous waveform values. We borrow the idea of training speech-to-unit (S2U) model where units are pre-generated from target speech with a pre-trained HuBERT model (Lee et al., 2022a). During S2U training, models are periodically evaluated on the valid set of speech-to-unit samples, and the best checkpoint with the lowest valid loss is saved for model inference.

When it comes to inference, speech could be synthesized from the predicted units with a vocoder, as the output of the S2S pipeline. It is then transcribed into texts by an off-the-shelf ASR model. The BLEU score is calculated by comparing the transcriptions against the ground truth target texts, which serves as the quantitative metric of mined data quality. We note that the ASR BLEU score is not a perfect metric for data quality, as it is unavoidably affected by the quality of ASR models. Next we discuss each module of the pipeline.

Speech-to-Unit. The S2U model takes the source speech and predicts a sequence of target units. It typically has an encoder-decoder architecture, where the encoder consists of convolutional and Transformer encoder layers, and the decoder is a Transformer decoder. We have experimented with different model variants, and discuss bilingual and multilingual training in section 5 and section 6 respectively.

HuBERT. We reuse the same HuBERT model and k-means clusters for English, Spanish and French as in (Lee et al., 2022b) for a fair comparison with existing results. We also train HuBERT models for each language family as shown in Table 6 to cover other languages in SpeechMatrix. We collect unlabeled VoxPopuli speech for all languages of the same family as the training data. Each HuBERT model is trained for three iterations, and more details of HuBERT training can be found in Appendix A.2. We select the best layer for speech feature extraction and the best label size (i.e., the number of k-means clusters) using speech resynthesis together with the vocoder, which is discussed in Appendix A.4. With the optimal layer and label size decided, we could generate the target unit sequence for the given source speech.

Family	Languages
Romance	es, fr, it, pt, ro
Slavic	cs, pl, sk, sl, hr, lt
Germanic	en, de, nl
Uralic	fi, et, hu

Table 6: Language families in VoxPopuli data.

Vocoder. Unit-based HiFi-GAN vocoders are trained to synthesize speech from unit sequence (Polyak et al., 2021). In our experiments, vocoders are separately trained from S2U model. We train vocoders on three datasets:

- CSS10 (Park and Mulc, 2019). CSS10 is a single-speaker corpus which we use to train vocoders in German, Finnish, Hungarian and Dutch.
- VoxPopuli (Wang et al., 2021a). Given the ASR data with speaker information, we sort speakers based on their speech duration, and keep adding the top speakers until the speech is more than 20 hours.
- Common Voice (Ardila et al., 2020). Two languages, Portuguese and Estonian, are not covered by the two corpora above, and thus we turn to Common Voice. Again, we select top speakers and prepare 12-hour and 10-hour speech for the vocoder training in Portuguese and Estonian respectively.

We applied a denoiser[6] (Defossez et al., 2020) to the speech of VoxPopuli and Common Voice as the speech preprocessing to increase signal-to-noise ratio (SNR) given that they are noisier than CSS10 audios. Then we prepare vocoder labels with HuBERT models generating k-means cluster labels for each utterance. Single-speaker vocoders are trained in CSS10, and languages from VoxPopuli and Common Voice have multi-speaker vocoders where speaker embeddings are learned. During inference, we select the speaker with the longest speech duration to synthesize speech from predicted unit sequences, who has the most data for the vocoder to learn good speaker embeddings.

ASR. We use off-the-shelf ASR models to transcribe the speech generated by vocoders, and details about the ASR models and their benchmark results are provided in the subsection A.3.

5 Bilingual Speech-to-Speech Baselines

In this part, we discuss the bilingual S2S models trained in each of 272 language directions in SpeechMatrix. The architecture of Textless model is used for bilingual translation in our experiments (Lee et al., 2022a). A Textless model consists of a speech encoder with 2 convolution layers and 12 Transformer encoder layers. Transformer layer has the embedding dimension of 512 and the forward dimension of 2048. It has two unit decoders with 6 and 2 Transformer decoder layers for target and source unit prediction respectively. The target unit decoder has the embedding dimension of 512 and the forward dimension of 2048, and the source unit decoder's dimensions are 256 and 2048.

[6] https://github.com/facebookresearch/denoiser

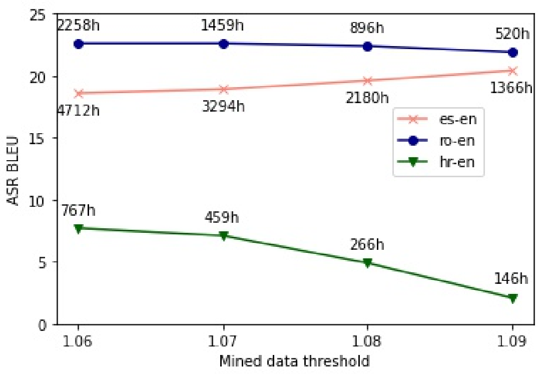

Figure 3: Bilingual S2S BLEU by mined data at different thresholds.

Training. For a given direction, we extract units for source and target speech with their corresponding HuBERT models (Hsu et al., 2021). Taking source speech, the model is trained to predict target unit sequence with cross-entropy loss as well as to reconstruct source units as an auxiliary task.

For the sake of good translation performance and training efficiency, it is important to set a reasonable threshold to mined data and select samples above the threshold as the train set. We performed an analysis of translation performance varying with thresholds from 1.06 to 1.09 on three language pairs: es-en, ro-en and hr-en. Figure 3 shows the threshold, the corresponding speech data size and BLEU score.

For low-resource directions such as hr-en, it is best to include all the mined data. For high-resource directions, es-en and ro-en, the optimal amount of mined data is around 1k hours and it does not bring further gains to increase the data size. Given these observations, we choose the highest threshold that keeps the source speech duration in mined data more than 1k hour for each direction. For example, we use a threshold of 1.09 for es-en and of 1.06 for hr-en.

Comparison with existing results. Since we adopt the same model as the previous work (Lee et al., 2022a) and the difference only lies in the train set, it is straightforward to compare with its reported S2S translation results. Table 7 shows the results of S2ST models which are trained on our SpeechMatrix mined data compared to VoxPopuli S2S data in each of four language directions: es-en, fr-en, en-es and en-fr. The threshold of mined data is set as 1.09 to these four directions, yielding an average of 1,436-hour train set. Compared with 480-hour labeled speech from VoxPopuli,

Train set		Es-En	Fr-En	En-Es	En-Fr
VoxPopuli S2S	Hours	532	523	415	451
	BLEU	13.1	15.4	16.4	15.8
SpeechMatrix ($t = 1.09$)	Hours	1,353	1,507	1,366	1,518
	BLEU	20.4	20.7	21.9	19.3

Table 7: BLEU scores on EPST test sets by S2ST models with different training data.

SpeechMatrix achieves an an average improvement of 5.4 BLEU, indicating the good quality and usefulness of the mined data.

5.1 Large-Scale Bilingual Evaluation

A large-scale evaluation is launched covering 272 mined languages directions, and bilingual models are trained for each direction to establish baseline results in speech-to-speech translation.

Table 8 summarizes performance of bilingual S2ST models on three test sets. In each direction, the first BLEU score is for European Parliament domain, either EPST or VoxPopuli set. EPST BLEU is underlined to be distinguished from VoxPopuli BLEU. The second score is for Wikipedia domain, i.e., FLEURS test data.

Bilingual results. Empirically we find that translations into high-resource languages such as en, es and fr outperform those into low-resource languages such as lt and sl based on the language resource sizes in Table 5. Another observation is the performance difference across test domains, i.e., BLEU on FLEURS is lower than that on EPST and VoxPopuli data likely because of the domain mismatch between train ad test data.

It is also found that translation results are not symmetric for some language pairs, for example, ro-en has a BLEU of 22.6 while en-ro BLEU is only 7.6 on EPST. Besides different complexity levels of target languages and test sets, such asymmetry also results from the dependency of BLEU score on the speech synthesis quality of the vocoder and transcription quality of the ASR model. For languages whose vocoder and ASR models are not well trained, they are likely to receive low BLEU scores. In this case, Romanian vocoder and ASR are not as strong as English models as reflected by its higher word error rate in speech resynthesis as reported in Appendix A.4.

	cs	de	en	es	et	fi	fr	hr	hu	it	lt	nl	pl	pt	ro	sk	sl
cs	- / -	12.9 / 2.0	22.7 / 4.2	16.7 / 4.6	- / 0.1	0.6 / 0.2	21.1 / 7.5	4.4 / 2.1	0.5 / 0.2	10.2 / 2.5	0.1 / 0.1	6.1 / 1.0	8.5 / 2.3	- / 2.8	4.3 / 1.4	16.9 / 3.5	3.0 / 1.7
de	7.3 / 2.3	- / -	16.3 / 8.3	11.7 / 3.8	- / 0.1	1.2 / 0.2	10.7 / 6.5	4.5 / 2.2	0.6 / 0.2	3.8 / 1.8	0.1 / 0.0	10.4 / 1.2	3.5 / 0.9	7.1 / 3.1	5.2 / 2.1	3.0 / 0.8	4.1 / 1.0
en	8.2 / 2.7	10.1 / 2.7	- / -	21.9 / 6.0	- / 0.7	1.9 / 0.6	19.2 / 10.4	8.4 / 2.4	1.1 / 0.3	11.5 / 3.6	0.3 / 0.1	15.1 / 3.8	8.2 / 1.3	11.8 / 5.1	7.6 / 2.0	5.7 / 1.2	5.5 / 1.2
es	5.2 / 1.9	6.1 / 1.8	20.4 / 7.5	- / -	- / 0.1	1.3 / 0.2	16.3 / 9.2	3.6 / 1.0	0.7 / 0.2	11.1 / 4.2	0.1 / 0.1	8.0 / 1.5	3.9 / 1.4	13.3 / 5.9	5.2 / 2.3	2.2 / 0.9	2.2 / 0.8
et	- / 2.1	- / 0.7	- / 8.2	- / 3.0	- / -	- / 0.7	- / 6.3	- / 1.0	- / 0.7	- / 2.3	- / 0.1	- / 1.5	- / 1.2	- / 1.7	- / 1.4	- / 0.4	- / 0.8
fi	3.0 / 1.5	9.0 / 0.9	19.7 / 5.5	11.4 / 3.8	- / 0.5	- / -	14.1 / 6.2	1.5 / 0.5	0.0 / 0.0	5.8 / 1.2	0.1 / 0.0	6.6 / 0.8	4.5 / 1.2	- / 2.0	4.4 / 1.1	1.7 / 0.7	1.6 / 0.7
fr	5.4 / 1.5	6.3 / 2.1	20.7 / 9.8	18.4 / 7.6	- / 0.1	0.8 / 0.2	- / -	5.4 / 1.7	0.7 / 0.2	10.2 / 3.1	0.1 / 0.1	8.4 / 1.3	4.8 / 1.5	13.4 / 5.8	5.6 / 2.4	1.6 / 0.6	1.5 / 0.6
hr	- / 2.5	- / 0.9	- / 7.7	- / 3.1	- / 0.2	- / 0.1	- / 5.8	- / -	- / 0.2	- / 1.1	- / 0.0	- / 0.9	- / 1.1	- / 2.0	- / 0.6	- / 0.9	- / 0.8
hu	2.6 / 1.3	7.3 / 1.0	15.3 / 4.6	9.5 / 3.0	- / 0.1	0.7 / 0.2	13.8 / 5.7	1.9 / 0.7	- / -	6.3 / 1.2	0.1 / 0.0	3.0 / 0.1	1.6 / 0.4	- / 2.3	2.4 / 0.9	0.9 / 0.2	1.2 / 0.3
it	6.4 / 1.3	4.9 / 1.0	18.9 / 6.3	19.6 / 8.3	- / 0.1	0.4 / 0.1	15.3 / 11.3	5.2 / 1.3	0.7 / 0.2	- / -	0.1 / 0.0	6.5 / 0.9	3.6 / 1.1	12.4 / 5.6	3.7 / 1.9	2.1 / 0.4	2.8 / 0.6
lt	0.2 / 0.1	0.0 / 0.0	3.1 / 0.9	0.8 / 0.2	- / 0.0	0.0 / 0.0	0.7 / 0.2	0.1 / 0.0	0.0 / 0.0	0.6 / 0.4	- / -	0.7 / 0.1	0.1 / 0.0	- / 0.0	0.0 / 0.0	0.0 / 0.0	0.1 / 0.0
nl	3.5 / 1.4	8.1 / 3.1	18.0 / 5.7	13.2 / 4.9	- / 0.2	0.5 / 0.2	13.0 / 7.5	3.3 / 1.8	0.4 / 0.2	5.2 / 1.7	0.1 / 0.0	- / -	3.4 / 0.9	6.7 / 3.3	4.1 / 1.4	1.7 / 0.4	2.1 / 1.0
pl	7.2 / 1.6	2.8 / 1.6	4.9 / 4.9	6.3 / 4.4	- / 0.1	1.0 / 0.2	5.5 / 5.4	4.5 / 1.2	0.5 / 0.1	5.8 / 1.5	0.2 / 0.0	1.6 / 0.3	- / -	6.1 / 2.5	3.2 / 1.2	4.7 / 1.1	2.4 / 0.7
pt	- / 1.2	4.7 / 1.0	21.2 / 6.1	23.2 / 8.7	- / 0.1	- / 0.3	18.1 / 11.1	- / 1.1	- / 0.1	4.4 / 1.1	- / 0.1	5.0 / 0.6	3.6 / 0.8	- / -	4.4 / 1.5	- / 0.6	- / 0.6
ro	4.6 / 1.9	6.5 / 2.2	22.6 / 7.8	20.1 / 7.0	- / 0.4	0.8 / 0.3	18.6 / 11.3	2.4 / 0.9	0.4 / 0.2	8.7 / 3.8	0.1 / 0.1	3.5 / 0.9	4.6 / 1.1	10.3 / 6.0	- / -	2.3 / 0.7	0.7 / 0.2
sk	28.2 / 9.1	10.7 / 2.1	21.4 / 5.5	15.5 / 5.1	- / 0.3	1.0 / 0.2	19.2 / 7.8	5.0 / 3.0	0.5 / 0.4	4.7 / 2.1	0.1 / 0.0	4.2 / 0.7	5.3 / 1.9	- / 2.3	4.4 / 1.9	- / -	3.6 / 1.5
sl	4.0 / 2.2	11.1 / 2.0	19.5 / 7.3	8.6 / 3.4	- / 0.2	0.8 / 0.3	13.2 / 4.5	4.8 / 1.1	0.4 / 0.1	6.0 / 1.2	0.1 / 0.0	4.5 / 1.0	6.7 / 1.2	- / 1.5	1.1 / 0.1	1.7 / 0.3	- / -

Table 8: BLEU scores of bilingual S2S models on three test sets. The first score is either on EPST or VoxPopuli data, and EPST score is underscored. The second score is on FLEURS data.

6 Multilingual Speech-to-Speech Translation

Multilingual modeling has been explored for the tasks of language models and machine translation, demonstrating knowledge transfer among languages. However, to our best knowledge, there are very few studies of multilingual speech-to-speech translation, partially due to the lack of multilingual speech-to-speech resources. With the massively multilingual data we have mined, we are able to explore multilingual S2ST training.

In this work, we focus on many-to-English translation, studying the translation from 6 Slavic languages to English in subsection 6.1 and the translation from all 16 languages in SpeechMatrix to English in subsection 6.2. English-to-many or many-to-many translation are left to future work.

Multilingual models in our experiments include

- **Textless model**. The same model that we use for bilingual evaluation is reused in the multilingual experiments. The bilingual Textless model has 70M parameters. Given the increased amount and diversity of multilingual data, we increase the model size for larger model capacity, trying multilingual models with 70M, 260M and 424M parameters respectively.

- **XM Transformer**. Inspired by the recent finding that crossmodal pre-training is beneficial for speech translation (Popuri et al., 2022), we introduce XM Transformer to multilingual training, whose encoder is initialized from pre-trained XLS-R model with 1B parameters (Babu et al., 2021) and decoder is initialized from a unit decoder pre-trained in an mBART style (Popuri et al., 2022). With multilingual speech-to-unit data, the model is further fine-tuned to minimize the cross-entropy loss in unit prediction.

- **XM Transformer with Sparsity**. Sparse modeling, in particular Mixture-of-Experts (MoE), has been widely studied in multilingual machine translation as an efficient approach to encourage knowledge transfer and mitigate interference among languages. MoE increases the number of parameters of the model in magnitude without sacrificing computation efficiency. In this work, we experiment with two variants of sparse modeling, GShard (Lepikhin et al., 2020) and Base Layer (Lewis et al., 2021).

GShard. GShard is a sparse scaling technique for transformer proposed in (Lepikhin et al., 2020). We replace every other transformer layer with an MoE layer. FFN layers in an MoE transformer layer are shared across experts. One GPU could host one or more experts and a learned gating function is used to determine which expert a token is routed to. Similar approaches have also been adopted recently in the multilingual text translation area to address language interference issues (NLLB Team et al., 2022). We apply GShard architecture on the decoder of XM Transformer, to expand the capacity without the need to retrain larger unit mBART models. All expert weights are initialized with unit mBART.

Base Layer. Base Layer is a variant of the sparse model proposed by (Lewis et al., 2021), which applies a balanced assignment to expert layers. It formulates token-to-expert allocation as a linear assignment problem, encour-

	Bilingual				Multilingual					
	70M		260M		70M		260M		424M	
	EP / VP	FLEURS	EP / VP	FLEURS	EP / VP	FLEURS	EP / VP	FLEURS	EP / VP	FLEURS
cs	22.7	4.2	24.7	11.2	19.7	2.3	27.5	13.7	25.3	10.2
hr	-	7.7	-	4.6	-	3.1	-	12.8	-	9.2
lt	3.1	0.9	0.2	0.0	2.8	0.3	14.7	4.8	10.7	3.3
pl	4.9	4.9	17.6	7.7	14.4	1.9	19.9	9.5	16.4	6.9
sk	21.4	5.5	24.4	11.0	18.9	4.1	27.2	15.4	24.9	11.1
sl	19.5	7.3	16.9	4.7	14.6	3.1	22.9	10.7	21.0	7.6
avg	14.3	5.1	16.8	6.5	14.1	2.5	22.4	11.2	19.7	8.1

Table 9: BLEU of Slavic-to-English multilingual Textless model across domains (for EP / VP column, underlined scores are on EPST data, and others on VoxPopuli data).

	Bilingual (1.2B)		Multilingual Dense(1.2B)		Multilingual MoE-GShard64 (4.3B)	
	EP / VP	FLEURS	EP / VP	FLEURS	EP / VP	FLEURS
cs	28.3	17.8	29.7	18.2	30.6	19.3
hr	-	12.1	-	17.1	-	17.6
lt	0.0	0.0	20.9	9.6	22.2	10.2
pl	17.4	7.4	21.1	12.9	21.4	12.6
sk	24.7	14.5	30.8	19.3	31.8	20.0
sl	20.1	8.5	27.4	14.0	29.1	13.0
avg	18.1	10.1	26.0	15.2	27.0	15.5

Table 10: BLEU of Slavic-to-English multilingual XM Transformer models in different domains (for EP / VP column, underlined scores are on EPST data, and others on VoxPopuli data).

aging each expert to receive an equal number of tokens. The assignment scheme improves efficiency by balancing compute loads, and simplifies training without any new hyperparameters or auxiliary losses. In our experiments, we add one additional Base Layer in the middle (the 7th layer) of decoder. The weights of Base Layer experts are randomly initialized.

Appendix B provides more details about these multilingual model configurations.

6.1 Slavic-to-English Translation

The six Slavic languages include Czech (cs), Croatian (hr), Lituanian (lt), Polish (pl), Slovak (sk), and Slovenian (sl). In the multilingual setting, mined data into English at the threshold of 1.06 are combined from each Slavic language as the train set. As for the evaluation of multilingual models, we report ASR BLEU in each direction, respectively.

We first extend Textless model from the bilingual to multilingual setting. For Textless models with different parameter sizes, their multilingual as well as bilingual translation results are presented in Table 9.

Results. With the Textless model size fixed as 70M, multilingual training hurts the performance of most languages compared with bilingual training. This is due to the insufficient model capacity, and the language interference is reflected by an average of -2.6 BLEU in FLEURS. We increase model parameters to 260M in both bilingual and multilingual settings. With a larger model capacity, bilingual models achieve gains in high-resource languages including cs, pl and sk, while suffering from performance loss in low-resource directions such as hr, lt and sl.

Given model sizes of 260M, we observe consistent gains of multilingual models over the bilingual models across different language directions and test domains. An average gain of 5.6 BLEU is achieved in EP/VP and the gain of 4.7 BLEU in FLEURS. It demonstrates the positive transfer enabled by multilingual training. As the multilingual model size continues to increase to 424M, we don't observe further gains likely due to the bottleneck of training data amount.

Pre-training has been shown to be beneficial for speech-to-speech translation (Popuri et al., 2022). XM Transformer leveraging pre-trained modules is also trained on Slavic-to-English data, and results are reported in Table 10. Comparing against bilingual Textless model (70M), bilingual XM Transformer outperforms it in all directions

	Dense (1.2B)		MoE-GShard64 (4.3B)		Base Layer (1.7B)	
	EP / VP	FLEURS	EP / VP	FLEURS	EP / VP	FLEURS
cs	29.9	18.7	30.9	18.2	29.9	17.3
de	<u>18.8</u>	19.0	<u>19.3</u>	20.3	<u>19.4</u>	19.5
es	<u>22.8</u>	15.2	<u>23.3</u>	15.9	<u>22.9</u>	14.9
et	-	16.7	-	16.7	-	16.4
fi	26.8	14.1	28.2	14.0	28.5	13.9
fr	<u>23.5</u>	18.3	<u>24.1</u>	18.9	<u>23.4</u>	18.2
hr	-	16.6	-	16.8	-	16.3
hu	20.2	12.0	21.3	12.5	20.5	12.1
it	36.3	16.2	37.8	14.9	37.4	14.0
lt	21.9	9.8	23.8	10.3	23.4	10
nl	<u>21.4</u>	16.4	<u>22.1</u>	17.3	<u>21.5</u>	16.6
pl	<u>21.2</u>	12.4	<u>21.3</u>	13.4	<u>20.9</u>	12.5
pt	<u>23.8</u>	21.8	<u>24.2</u>	22.3	<u>23.8</u>	21.1
ro	<u>25.1</u>	19.7	<u>25.0</u>	19.8	<u>25.3</u>	19.0
sk	30.8	19.6	32.2	18.2	31.5	18.4
sl	28.3	13.7	29.9	13.7	28.8	13.5
avg	25.1	16.3	26.0	16.5	25.5	15.9

Table 11: BLEU of All-to-English multilingual models in different domains (for EP / VP column, underlined scores are on EPST data, and others on VoxPopuli data).

except lt-en. The gain in EP/VP is 3.8 BLEU on average, and a larger gain of 5.0 BLEU is achieved in FLEURS. Multilingual training brings further gains to XM Transformer with +7.9 and +5.1 BLEU over bilingual training in EP/VP and FLEURS test set respectively.

Comparing against dense XM Transformer, GShard with 64 experts has +1.0 BLEU gains on average over 5 directions on EP/VP, and +0.3 BLEU gains for FLEURS. We believe it's due to a phenomena mentioned in (Zoph et al., 2022) that MoE specializes in multilingual settings but not by language. GShard in our setting brings larger improvements to in-domain test set.

Overall the best Slavic-to-English translation is achieved by XM Transformer with GShard trained in multilingual setting. This demonstrates that pretraining, model sparsity and multilinguality are of help to speech-to-speech translation.

6.2 All-to-English Translation

We move forward to a larger-scale multilinguality by extending from Slavic language family to all languages in SpeechMatrix. We adopt the best models in Slavic-to-English translation, i.e., multilingual XM Transformer with both dense and sparse architectures. In terms of sparse modeling, we try both GShard and Base Layer in all-to-English translation.

Results. Compared with XM Transformer (1.2B) dense model, MoE-GShard64 (4.3B) with the same forward computation time brings gains of +0.9 and +0.2 BLEU to EP/VP and FLEURS respectively. Similar to our findings in Slavic-to-English setting, increasing the capacity with sparse modeling benefits more on in-domain (EP/VP) than out-of-domain FLEURS test set.

With the sparse architecture of XM Transformer with GShard, all-to-English model shows +0.6 and -0.4 BLEU difference compared with Slavic-to-English model on EP/VP and FLEURS respectively, averaged over Slavic languages. Multilingual sparse model benefits from the additional in-domain data in other languages when evaluated in EP/VP domain, while sees performance degradation in out-of-domain data.

The other sparse variant, Base Layer (1.7B) performs comparably to the dense XM Transformer, with an average of +0.4 BLEU in EP/VP test sets and -0.4 BLEU in FLEURS. The sparsity in Base Layer does not bring obvious gains to all-to-English translation. This is likely because we only add one Base Layer to the decoder with a small expert size. The number of increased model parameters is only 0.5B in Base Layer, while it is 3.1B in GShard. As suggested by (Lewis et al., 2021), the Base Layer performance might improve with more GPUs and a larger expert size.

7 Limitations

The HuBERT model is critical to the speech-to-speech translation performance as its extracted units are used by both speech-to-unit model and vocoder. We have not explored the optimal strategy of multilingual HuBERT training. One research question is how to choose a group of languages that a multilingual HuBERT model is trained on. For example, it is arguable whether Lithuanian (lt) should be included in Slavic or Uralic family. Other questions could be whether a larger HuBERT with more model capacity should be used and how we should deal with language imbalance in multilingual training.

We provide benchmark results of bilingual speech translation with mined data selected by heuristics. One of our future directions is to come up with a better strategy of mined data selection to improve translation performance and training efficiency.

As mentioned in our results analysis, the reported BLEU scores are heavily dependent on the ASR quality, which may not reflect the speech translation performance accurately. Future directions could be improving ASR quality or exploring other evaluation metrics without reliance on ASR models.

8 Conclusion

In this paper, we introduce a large-scale multilingual speech-to-speech corpus mined from VoxPopuli. It is the largest resource of speech alignments with a coverage of 17 languages. We perform an extensive evaluation of the mined parallel speech, showing good quality of the speech alignments. Multilingual speech-to-speech models can be efficiently trained on this corpus and we suggest different methods, like sparse scaling using Mixture-of-Experts, to further boost translation performance in the multilingual setting.

References

Sadaf Abdul-Rauf and Holger Schwenk. 2009. On the Use of Comparable Corpora to Improve SMT performance. In *EACL*, pages 16–23.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis M. Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*, pages 4218–4222. European Language Resources Association.

Mikel Artetxe and Holger Schwenk. 2018. Margin-based Parallel Corpus Mining with Multilingual Sentence Embeddings. https://arxiv.org/abs/1811.01136.

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *TACL*, pages 597–610.

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. 2021. XLS-R: self-supervised cross-lingual speech representation learning at scale. *CoRR*, abs/2111.09296.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Marcely Zanon Boito, William Havard, Mahault Garnerin, Éric Le Ferrand, and Laurent Besacier. 2020. Mass: A large and clean multilingual corpus of sentence-aligned spoken utterances extracted from the bible. In *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*, pages 6486–6493. European Language Resources Association.

Houda Bouamor and Hassan Sajjad. 2018. H2@BUCC18: Parallel Sentence Extraction from Comparable Corpora Using Multilingual Sentence Embeddings. In *BUCC*.

Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. 2022. FLEURS: few-shot learning evaluation of universal representations of speech. *CoRR*, abs/2205.12446.

Alexandre Defossez, Gabriel Synnaeve, and Yossi Adi. 2020. Real time speech enhancement in the waveform domain. In *Interspeech*.

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.

Quoc Truong Do, Sakriani Sakti, Graham Neubig, Tomoki Toda, and Satoshi Nakamura. 2015. Improving translation of emphasis with pause prediction in speech-to-speech translation systems. In *Proceedings of the 12th International Workshop on Spoken Language Translation: Papers, IWSLT 2015, Da Nang, Vietnam, December 3-4, 2015*.

Paul-Ambroise Duquenne, Hongyu Gong, Benoît Sagot, and Holger Schwenk. 2022. T-modules: Translation modules for zero-shot cross-modal machine translation. In *EMNLP*.

Paul-Ambroise Duquenne, Hongyu Gong, and Holger Schwenk. 2021. Multimodal and multilingual embeddings for large-scale speech mining. *Advances in Neural Information Processing Systems*, 34:15748–15761.

Cristina España-Bonet, Ádám Csaba Varga, Alberto Barrón-Cedeño, and Josef van Genabith. 2017. An Empirical Analysis of NMT-Derived Interlingual Embeddings and their Use in Parallel Sentence Identification. *IEEE Journal of Selected Topics in Signal Processing*, pages 1340–1348.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2020. Beyond english-centric multilingual machine translation. *JMLR*.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020. Language-agnostic bert sentence embedding. *arXiv preprint arXiv:2007.01852*.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Trans. Assoc. Comput. Linguistics*, 10:522–538.

Kevin Heffernan, Onur Çelebi, and Holger Schwenk. 2022. Bitext mining using distilled sentence representations for low-resource languages. In *EMNLP*.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE ACM Trans. Audio Speech Lang. Process.*, 29:3451–3460.

Javier Iranzo-Sánchez, Joan Albert Silvestre-Cerdà, Javier Jorge, Nahuel Roselló, Adrià Giménez, Albert Sanchís, Jorge Civera, and Alfons Juan. 2020. Europarl-st: A multilingual corpus for speech translation of parliamentary debates. In *2020 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2020, Barcelona, Spain, May 4-8, 2020*, pages 8229–8233. IEEE.

Ye Jia, Yifan Ding, Ankur Bapna, Colin Cherry, Yu Zhang, Alexis Conneau, and Nobuyuki Morioka. 2022a. Leveraging unsupervised and weakly-supervised data to improve direct speech-to-speech translation. *CoRR*, abs/2203.13339.

Ye Jia, Michelle Tadmor Ramanovich, Tal Remez, and Roi Pomerantz. 2022b. Translatotron 2: High-quality direct speech-to-speech translation with voice preservation. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 10120–10134. PMLR.

Ye Jia, Michelle Tadmor Ramanovich, Quan Wang, and Heiga Zen. 2022c. CVSS corpus and massively multilingual speech-to-speech translation. *CoRR*, abs/2201.03713.

Ye Jia, Ron J. Weiss, Fadi Biadsy, Wolfgang Macherey, Melvin Johnson, Zhifeng Chen, and Yonghui Wu. 2019. Direct speech-to-speech translation with a sequence-to-sequence model. In *Interspeech 2019, 20th Annual Conference of the International Speech Communication Association, Graz, Austria, 15-19 September 2019*, pages 1123–1127. ISCA.

Sameer Khurana, Antoine Laurent, and James Glass. 2022. Samu-xlsr: Semantically-aligned multimodal utterance-level cross-lingual speech representation. *arXiv preprint arXiv:2205.08180*.

Ann Lee, Peng-Jen Chen, Changhan Wang, Jiatao Gu, Sravya Popuri, Xutai Ma, Adam Polyak, Yossi Adi, Qing He, Yun Tang, Juan Pino, and Wei-Ning Hsu. 2022a. Direct speech-to-speech translation with discrete units. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 3327–3339. Association for Computational Linguistics.

Ann Lee, Hongyu Gong, Paul-Ambroise Duquenne, Holger Schwenk, Peng-Jen Chen, Changhan Wang, Sravya Popuri, Yossi Adi, Juan Pino, Jiatao Gu, and Wei-Ning Hsu. 2022b. Textless speech-to-speech translation on real data. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 860–872. Association for Computational Linguistics.

Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. 2020. Gshard: Scaling giant models with conditional computation and automatic sharding. *arXiv preprint arXiv:2006.16668*.

Mike Lewis, Shruti Bhosale, Tim Dettmers, Naman Goyal, and Luke Zettlemoyer. 2021. Base layers: Simplifying training of large, sparse models. In *International Conference on Machine Learning*, pages 6265–6274. PMLR.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and

Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Dragos Stefan Munteanu and Daniel Marcu. 2005. Improving Machine Translation Performance by Exploiting Non-Parallel Corpora. *Computational Linguistics*, 31(4):477–504.

Satoshi Nakamura, Konstantin Markov, Hiromi Nakaiwa, Gen-ichiro Kikui, Hisashi Kawai, Takatoshi Jitsuhiro, Jinsong Zhang, Hirofumi Yamamoto, Eiichiro Sumita, and Seiichi Yamamoto. 2006. The ATR multilingual speech-to-speech translation system. *IEEE Trans. Speech Audio Process.*, 14(2):365–376.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

Kyubyong Park and Thomas Mulc. 2019. CSS10: A collection of single speaker speech datasets for 10 languages. In *Interspeech 2019, 20th Annual Conference of the International Speech Communication Association, Graz, Austria, 15-19 September 2019*, pages 1566–1570. ISCA.

Adam Polyak, Yossi Adi, Jade Copet, Eugene Kharitonov, Kushal Lakhotia, Wei-Ning Hsu, Abdelrahman Mohamed, and Emmanuel Dupoux. 2021. Speech Resynthesis from Discrete Disentangled Self-Supervised Representations. In *Proc. Interspeech 2021*.

Sravya Popuri, Peng-Jen Chen, Changhan Wang, Juan Pino, Yossi Adi, Jiatao Gu, Wei-Ning Hsu, and Ann Lee. 2022. Enhanced direct speech-to-speech translation using self-supervised pre-training and data augmentation. *CoRR*, abs/2204.02967.

Matt Post, Gaurav Kumar, Adam Lopez, Damianos Karakos, Chris Callison-Burch, and Sanjeev Khudanpur. 2014. Fisher and callhome spanish–english speech translation. *LDC2014T23. Web Download. Philadelphia: Linguistic Data Consortium*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using siamese bert-networks. In *https://arxiv.org/abs/1908.10084*.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *EMNLP*, pages 4512–4525.

Philip Resnik. 1999. Mining the Web for Bilingual Text. In *ACL*.

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. In *Interspeech 2021, 22nd Annual Conference of the International Speech Communication Association, Brno, Czechia, 30 August - 3 September 2021*, pages 3655–3659. ISCA.

Holger Schwenk. 2018. Filtering and mining parallel data in a joint multilingual space. In *ACL*, pages 228–234.

Holger Schwenk and Matthijs Douze. 2017. Learning joint multilingual sentence representations with neural machine translation. In *ACL workshop on Representation Learning for NLP*.

Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, Armand Joulin, and Angela Fan. 2021a. CCMatrix: Mining billions of high-quality parallel sentences on the web. In *ACL*, pages 6490–6500.

Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, Armand Joulin, and Angela Fan. 2021b. Ccmatrix: Mining billions of high-quality parallel sentences on the web. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 6490–6500. Association for Computational Linguistics.

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, Morgane Rivière, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Miguel Pino, and Emmanuel Dupoux. 2021a. Voxpopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 993–1003. Association for Computational Linguistics.

Changhan Wang, Anne Wu, Jiatao Gu, and Juan Pino. 2021b. Covost 2 and massively multilingual speech translation. In *Interspeech*, pages 2247–2251.

Hongyu Wang, Shuming Ma, Li Dong, Shaohan Huang, Dongdong Zhang, and Furu Wei. 2022. Deepnet: Scaling transformers to 1,000 layers.

Yinfei Yang, Gustavo Hernandez Abrego, Steve Yuan, Qinlan Shen Mandy Guo, Daniel Cer, Brian Strope Yun-hsuan Sun and, and Ray Kurzweil. 2019a. Improving multilingual sentence embedding using bidirectional dual encoder with additive margin softmax. In *IJCAI*, pages 5370–5378.

Yinfei Yang, Daniel Cer, Amin Ahmad, Mandy Guo, Jax Law, Noah Constant, Gustavo Hernandez Abrego, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2019b. Multilingual universal sentence encoder for semantic retrieval. In https://arxiv.org/abs/1907.04307.

Barret Zoph, Irwan Bello, Sameer Kumar, Nan Du, Yanping Huang, Jeff Dean, Noam Shazeer, and William Fedus. 2022. Designing effective sparse expert models. *arXiv preprint arXiv:2202.08906*.

A Appendix

A.1 Similarity search on CoVoST

We compared our similarity search results with previous work (Duquenne et al., 2021) in Table 12. We notice that our new speech encoders have lower error rates compared to previous work.

Audio vs. en translations	de	es	fr
Previous work	3.36	1.66	2.05
This work	**3.27**	**1.26**	**1.55**

Table 12: Similarity search error rates (in %) on CoVoST 2 test set.

A.2 HuBERT

We train a multilingual HuBERT model for each family on the collection of speech in each component language as shown in Table 6. The HuBERT model consists of 7 convolutional layers and 12 Transformer encoder layers. Each encoder layer has 12 attention heads, the embedding dimension of 768 and the forward dimension of 3072. Models are trained for 3 iterations, and in each iteration pseudo-labels are prepared as the training target for utterances. In the first iteration, the target labels are MFCC features. In the second iteration, we extract speech features from the 6-th layer of the trained HuBERT model and apply k-means clustering to derive a set of 500 labels. In the third iteration, speech features from the 9-th layer are clustered into 500 labels. Lastly after these three iterations, we try feature extraction from different layers including layer 10, 11 and 12 of trained HuBERT. As for feature clustering, we also try different numbers of clusters, 800, 1000 and 1200, to derive multiple sets of target units.

To choose the optimal setup, we launch a resynthesis evaluation to select the HuBERT layer to extract speech features and the number of k-means clusters. We train a vocoder on each set of target units, i.e., vocoder takes the units and synthesizes target speech. The synthesized speech is sent to off-the-shelf ASR models, and Word Error Rate (WER) is reported to measure the speech quality. The set of target units is selected if the corresponding vocoder achieves lowest WER.

A.3 ASR models

We use ASR models publicly released on HuggingFace to transcribe the generated speech in order to calculate WER or BLEU scores in comparison with ground truth texts. ASR models used in our evaluation are listed in Table 13.

A.4 Vocoder

Vocoders are trained to synthesize speech from a given sequence of units. The train sets are speech data from CSS10, VoxPopuli and Common Voice. As mentioned before, units are derived from HuBERT models for these speech. Table 14 summarizes WER of ASR models, which reflects the transcription quality in each language. Besides, we report the training dataset, vocoder WER of synthesized speech from vocoders, and here we include the vocoder results obtained from the optimal HuBERT layer and k-means cluster size. Layer 11 is the best HuBERT layer for feature extraction in all languages, and most languages have the best k-means size of 1000 except Italian (it) whose best label size is 800.

As shown in Table 14, ASR models are of good quality for high-resource languages such as de, fi and pt, while suffering from high error rates in languages such as ro, lt and sl. It is expected to have higher vocoder WER than ASR WER since the former is for synthesized speech. By measuring the gap between the two error rates, we can tell how good a vocoder is and also infer the quality of HuBERT units. For et, pt and lt, the gaps are obviously larger than other languages. It not surprising since we do not have much good-quality speech data for these languages, for example, there is only around 10-hour noisy speech from Common Voice for et and pt vocoder training.

B Multilingual Speech-to-Speech Translation

We provide details of models and experiment setups in multilingual speech-to-speech translation.

B.1 Slavic-to-English Translation

Textless model. Textless model (260M) has a speech encoder with 4 convolution layers and 12 Transformer encoder layers with the embedding dimension of 1024 and the forward dimension of 4096. It has two unit decoders with 6 and 2 Transformer decoder layers for target and source unit prediction respectively. The target unit decoder has the embedding dimension of 1024 and the forward dimension of 4096, and the source unit decoder's dimensions are 256 and 2048.

Lang	cs	de
ASR	comodoro/wav2vec2-xls-r-300m-cs-250	jonatasgrosman/wav2vec2-xls-r-1b-german
Lang	**et**	**fi**
ASR	RASMUS/wav2vec2-xlsr-1b-et	jonatasgrosman/wav2vec2-large-xlsr-53-finnish
Lang	**hr**	**hu**
ASR	classla/wav2vec2-xls-r-parlaspeech-hr	jonatasgrosman/wav2vec2-large-xlsr-53-hungarian
Lang	**it**	**lt**
ASR	jonatasgrosman/wav2vec2-large-xlsr-53-italian	sammy786/wav2vec2-xlsr-lithuanian
Lang	**nl**	**pl**
ASR	jonatasgrosman/wav2vec2-xls-r-1b-dutch	jonatasgrosman/wav2vec2-xls-r-1b-polish
Lang	**pt**	**ro**
ASR	jonatasgrosman/wav2vec2-xls-r-1b-portuguese	gigant/romanian-wav2vec2
Lang	**sk**	**sl**
ASR	anuragshas/wav2vec2-xls-r-300m-sk-cv8-with-lm	anuragshas/wav2vec2-xls-r-300m-sl-cv8-with-lm

Table 13: HuggingFace ASR models for each language.

Lang	Data	ASR WER	HuBERT	Vocoder WER	Lang	Data	ASR WER	HuBERT	Vocoder WER
de	CSS10	0.10	Germanic HuBERT layer 11, km 1000	0.16	nl	CSS10	0.19	Germanic HuBERT layer 11, km 1000	0.27
fi	CSS10	0.02	Uralic HuBERT layer 11, km 1000	0.15	hu	CSS10	0.21	Uralic HuBERT layer 11, km 1000	0.21
et	Common Voice	0.14	Uralic HuBERT layer 11, km 1000	0.44	it	VoxPopuli	0.23	Uralic HuBERT layer 11, km 800	0.27
pt	Common Voice	0.06	Uralic HuBERT layer 11, km 1000	0.31	ro	VoxPopuli	0.42	Uralic HuBERT layer 11, km 1000	0.50
cs	VoxPopuli	0.15	Slavic HuBERT layer 11, km 1000	0.23	pl	VoxPopuli	0.14	Slavic HuBERT layer 11, km 1000	0.23
hr	VoxPopuli	0.21	Slavic HuBERT layer 11, km 1000	0.29	lt	VoxPopuli	0.38	Slavic HuBERT layer 11, km 1000	0.57
sk	VoxPopuli	0.28	Slavic HuBERT layer 11, km 1000	0.41	sl	VoxPopuli	0.37	Slavic HuBERT layer 11, km 1000	0.46

Table 14: Benchmark results of ASR models and vocoder resynthesis.

For the Textless model (424M), its speech encoder contains 6 convolution layers and 16 Transformer encoder layers with the embedding dimension of 1024 and the forward dimension of 4096. It has two unit decoders with 12 and 2 Transformer decoder layers for target and source unit prediction respectively. The target unit decoder has the embedding dimension of 1024 and the forward dimension of 4096, and the source unit decoder's dimensions are 256 and 2048.

XM Transformer. XM Transformer (1.2B) is initialized from XLS-R encoder with 7 convolution layers and 48 Transformer encoder layers with the embedding dimension of 1280 and the forward dimension of 5120. Its unit decoder is initialized from a pre-trained mbart-style decoder with 12 layers, embedding dimension of 1024 and forward dimension of 4096.

B.2 All-to-English Translation

XM Transformer-GShard. XM Transfomer (1.2B) is initialized with the same XLS-R encoder and unit decoder used in Slavic-English. On the decoder side of XM Transformer-GShard, each expert is initialized with the same unit decoder. We set MoE frequency as 2, i.e., every other transformer layer is an MoE layer.

XM Transformer-Base Layer. For our XM Transformer with Base Layer sparsity (1.7B), the encoder is initialized with the same XLS-R encoder, and the dense layers of the decoder is initialized with the same unit decoder as GShard. We add an additional Base Layer which is randomly initialized as the 7th layer of decoder. There is one expert in each GPU and we used 64 GPUs in our experiments, which means we have 64 Base Layer experts in total.