A Case Study on Filtering for End-to-End Speech Translation

Anonymous ACL submission

Abstract

It is relatively easy to mine a large parallel corpus for any machine learning task, such as speech-to-text or speech-to-speech translation. Although these mined corpora are large in volume, their quality is questionable. This work shows that the simplest filtering technique can trim down these big, noisy datasets to a more manageable, clean dataset. We also show that using this clean dataset can improve the model's performance, as in the case of the multilingual-to-English Speech Translation (ST) model, where, on average, we obtain a 4.65 BLEU score improvement.

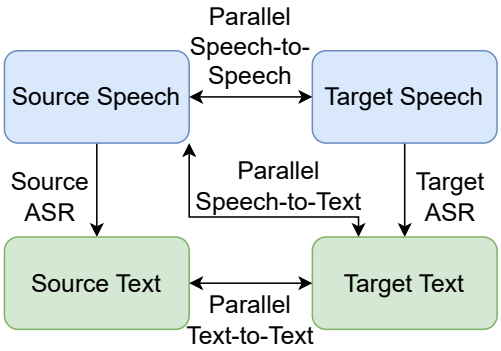

Figure 1: Pipeline of our noisy data creation.

1 Introduction

End-to-end speech translation (E2E ST) represents a paradigm shift from cascaded speech translation by utilizing a direct model for converting source language speech into target-language text. This approach offers advantages such as low latency and avoiding the "error propagation" issue often encountered in traditional cascade methods (Weiss et al., 2017; Sperber and Paulik, 2020). In recent times, end-to-end speech translation (ST) models have demonstrated noteworthy progress, achieving comparable or, in some instances, superior results when compared to traditional cascaded ST models (Bentivogli et al., 2021; Anastasopoulos et al., 2021, 2022).

Despite these benefits, E2E ST models face challenges when compared to the cascaded methods (Di Gangi et al., 2019), as there are robust training data available for automatic speech recognition (ASR) and machine translation (MT) which can be employed in cascade methods. The relatively limited training data for E2E ST models may lead to sub-optimal performance in certain scenarios. Prior research endeavors frequently tap into machine translation (MT) data to enhance training by applying multi-task learning techniques, as observed in works by Ye et al. (2022); Tang et al. (2021). Here the encoder and decoder are shared between speech translation (ST) and machine translation (MT) and the model acquires similar representations from distinct modalities.

Another research direction can be automatically creating large-scale mined corpus (Duquenne et al., 2023). However, the created data is often too noisy mostly because it is too hard to align data in two different languages. On top of that, for Speech Translation we have to also align between two different modalities. Then the research focus falls on filtering the noisy data. This work focuses on filtering Speech Translation data and observing if by doing so the systems' performance can be improved.

Our main contributions in a nutshell:

- We try different filtering methods for Speech Translation.
- We show that the simplest ratio-based filtering can effectively differentiate clean data from noisy data.
- We show that a somewhat good ST model trained by a small clean data can efficiently filter a noisy corpus and thus improve the models' performance.

Language	Model	WER ↓	Dataset
English (EN)	(Di Gangi et al., 2019)	25.81	MuST-C
	Monolingual	14.26	
	Multilingual	-	
French (FR)	(Salesky et al., 2021)	45.6	MTedX
	Monolingual	19.2	
	Multilingual	22.24	
Spanish (ES)	(Salesky et al., 2021)	46.4	MTedX
	Monolingual	17.06	
	Multilingual	18.17	
Portuguese (PT)	(Salesky et al., 2021)	54.8	MTedX
	Monolingual	22.86	
	Multilingual	24.48	
Italian (IT)	(Salesky et al., 2021)	48.0	MTedX
	Monolingual	-	
	Multilingual	21.37	

Table 1: WER score of the ASR systems used to transcribe the SpeechMatrix dataset. The multilingual model uses FR, ES, PT, and IT languages. The monolingual models outperform the multilingual models.

2 Approach

Our approach is of three parts: Firstly, we create noisy speech-text data pairs. Secondly, we score each pair using our sorting techniques. Thirdly, we filter the pairs based on the scores and create subsets of pairs to train the model.

2.1 Noisy Data Creation

For our work, we have used a speech-to-speech translation dataset, SpeechMatrix (SM). To create the speech-to-text dataset for ST training, we have used ASR models to transcribe both the source and target sides. Figure 1 outlines our approach. We created an ASR system, using each language's MuST-C or MTedX dataset. The WER of these ASR systems is shown in Table 1. The speech-to-speech translation dataset itself is noisy, and this approach adds one more layer of noise as the ASR models are imperfect. We explore if filtering approaches can work on this super noisy data.

2.2 Scoring Techniques

For our experiments, we explore ratio-based scoring techniques and one NLL(Negative Log likelihood)-based technique. The ratio-based techniques are faster to compute than NLL-based ones.

Text-Text Ratio To calculate this score, we divide the number of tokens from the source side transcript sentence by the number of tokens from the target side transcript sentence.

Speech-Text Ratio This is the same as the Text-Text Ratio, but now, on the source side, instead of the token count of the transcript text, we use the length in seconds of the speech as the numerator.

Speech-Speech Ratio As our original data is Speech-to-Speech data, we can calculate this score where on both the source and target side, we find the length (audio seconds) to compute the ratio.

Text-Speech Ratio Here, on the source side, we calculate the number of tokens, and on the target side, we calculate seconds of speech.

NLL(Negative Log Likelihood)-based We have used our baseline systems to calculate this score. Given a speech-text data pair $<x, y>$, we use Equation 1 to calculate the NLL loss. Here, f_θ is our baseline model.

$$\mathcal{L}(\theta) = -\left(y \, log \, \hat{y}_\theta + (1-y) \, log \, (1-\hat{y}_\theta)\right) \quad (1)$$

$$\hat{y}_\theta = \sigma(f_\theta(x)) \quad (2)$$

2.3 Filtering

After obtaining the score of each data instance using the techniques explained above, we use two approaches to select a subset of clean data. For ratio-based techniques, we calculate z-scores, and for NLL-based techniques, we use percentiles.

z-score calculation requires first determining the mean, μ and standard deviation, σ of the ratio-based scores of all the instances. Once we have these, we can calculate the z-score using Equation 3. Here, x is the ratio of a data instance. After calculating the z-score from each technique, we create four subsets where the z-score of the instances are less than or equal to 0.25, 0.5, 0.75, and 1.0. z-score here ranges from 0 to ∞.

$$z = |(x - \mu)/\sigma| \quad (3)$$

Percentile calculation requires sorting the data instances based on NLL loss score. After sorting in ascending order, we create four subsets that take the first 20%, 40%, 60%, and 80% data instances.

Union and Intersection We create another subset of data instances by performing a union and intersection operation between the best two subsets of each language pair.

3 Experimental Settings

3.1 Datasets

MuST-C (Di Gangi et al., 2019) are multilingual speech translation corpora whose size and quality facilitate end-to-end ST training from English into German, Spanish, French, Italian, Dutch, Portuguese, Romanian and Russian. MuST-C includes at least 385 hours of audio recordings from TED

Talks in English for each target language. They are automatically aligned with their manual transcriptions and translations on a sentence level.

MTedX (Salesky et al., 2021) supports speech recognition (ASR) and speech translation (ST) research across multiple non-English languages. The corpus is a collection of audio recordings from TEDx talks in 8 source languages. Transcriptions are segmented into sentences and aligned to the audio in the source language and the translation in the target language.

SpeechMatrix (Duquenne et al., 2023) is a groundbreaking initiative that introduces a vast and extensive multilingual corpus focused on speech-to-speech translations (S2ST). This corpus consists of authentic speech recordings of the European Parliament, offering a rich and diverse collection. In total, SpeechMatrix boasts speech alignments across 136 language pairs, providing a comprehensive resource with a staggering cumulative duration of 418 thousand hours of speech data.

3.2 Baseline Models

We have created three baseline models. First, we have created a bilingual baseline model using the MuST-C or MTedX dataset. We consider these datasets as clean datasets. Thus, we expect these models' performance to be quite high in quality. Next, we have created another baseline model using the SpeechMatrix dataset. This dataset is quite noisy; thus, we expect this model not to perform very well. Our last baseline model combines the MTedX or MuST-C datasets and the SpeechMatrix datasets without filtering. These baselines aim to quantify how adding the unfiltered SpeechMatrix affects the models' performance.

3.3 ASR and ST Model

For our experiments in ASR and ST, we use the same neural architecture. This setting allows us to initialize the ST encoder with the ASR encoder for faster convergence (Bahar et al., 2019; Bansal et al., 2019). We also created a multilingual ASR model and used it to initialize the multilingual ST model's encoder. We used FAIRSEQ (Ott et al., 2019) to train end-to-end Transformer models for ST, using 80-dimensional log mel filterbank features with cmvn and SpecAugment (Park et al., 2019), and 2-d convolutions for downsampling of speech in time. We use the *s2t_transformer_s* architecture of FAIRSEQ, which has a network depth of 12 encoder layers and 6 decoder layers. All other hyperparameters are discussed in the Appendix A.1.

3.4 Language Pairs

According to the clean data availability of the language pairs, we divided the language pairs that we experimented on into three categories: high-, mid-, and low-resource. High-resourced language pairs like English-French have a large amount of clean data as opposed to mid- or low-resourced. Thus, we can get a high-quality ST system in the high-resourced language pairs. In contrast, the models of the mid-resourced pairs like French-English and Spanish-Portuguese will have moderate quality. The low-resourced ones like Portuguese-Spanish will be even more degraded in quality.

4 Results

Using the above approach we create different subsets of data to train our systems with. We run experiments in two different settings: Bilingual, and Multilingual. This is to get an understanding of the filtering approaches' generalization ability.

4.1 Bilingual Systems

We have experimented on four language pairs. Due to space constraints, all results of this experiment are in Appendix A.2. Table 2 presents a summarized version of the results where we only show the highest-scoring models. We have two setups in each pair: trained only using the filtered SpeechMatrix data and trained by combining the filtered SpeechMatrix data with either MuST-C or MTedX.

High Resource (EN-FR) This language pair has the most clean data (270,000 instances) and the highest baseline score (31.51 BLEU) among the four pairs. This is the only pair where adding the whole noisy SpeechMatrix data substantially hurts the system's performance. All models created from five filtering techniques lead to higher BLEU scores than the Baseline (MuST-C + unfiltered SM). This shows that even a simple filtering technique is helpful instead of using all the noisy data.

We achieved the best performance (32.07 BLEU) by using only 20% clean data from SpeechMatrix on top of the clean data, a 0.56 BLEU score improvement over the baseline model, Baseline(MuST-C). We find a similar trend for all filtering techniques, where adding more data hurts performance. Note that systems trained using just the SpeechMatrix are awful, yet our best model earns 3.55 BLEU points just by filtering compared to the Baseline(unfiltered SM) model.

Lang Pair	ST System (Only SM)	# of pairs	BLEU	ST System (MuST-C/MTedX + SM)	# of pairs	BLEU
En-Fr	Baseline (unfiltered SM)	1384112	4.87	Baseline (MuST-C)	269256	31.51
	80% NLL	1107289	**8.42**	MuST-C + SM (20% NLL)	546078	**32.07**
Fr-En	Baseline (unfiltered SM)	1403985	5.20	Baseline (MTedX + unfiltered SM)	1433931	26.74
	80% NLL	1107289	**6.31**	MTedX + SM (60% NLL)	872337	**27.66**
Es-Pt	Baseline (unfiltered SM)	1074027	0.57	Baseline (MTedX + unfiltered SM)	1094848	26.28
	0.5 z Text-Text	559285	**5.13**	MTedX + SM (0.5 z T-T $\bigcup$ 0.5 z Speech-Speech)	834039	**27.07**
Pt-Es	Baseline (unfiltered SM)	1107283	3.23	Baseline (MTedX + unfiltered SM)	1118635	22.64
	60% NLL	664369	**4.38**	MTedX + SM (80% NLL)	897178	**23.7**

Table 2: Results of filtering approaches for Bilingual ST systems. The models created using filtering approaches outperform the baseline in each language direction.

Multi ST System	ES-EN (Δ)	FR-EN (Δ)	PT-EN (Δ)	IT-EN (Δ)	Average (Δ)	# of pairs
(Salesky et al., 2021)	12.3	12.0	12.0	10.7	11.75	120538
(i) Baseline (MTedX)	17.34	18.92	20.02	15.74	18.01	120538
(i) + SM (100%)	15.46 (-1.88)	19.93 (1.01)	16.61 (-3.41)	17.43 (1.69)	17.36 (-0.65)	7292751
(i) + SM (80% NLL)	17.04 (-0.30)	21.01 (2.09)	18.53 (-1.49)	18.59 (2.85)	18.79 (0.78)	5858310
(i) + SM (60% NLL)	18.15 (0.81)	22.86 (3.94)	21.1 (1.08)	18.92 (3.24)	20.26 (2.25)	4423867
(i) + SM (40% NLL)	19.86 (2.52)	24.76 (5.84)	22.35 (2.33)	20.51 (4.77)	21.87 (3.86)	2989425
(i) + SM (20% NLL)	**20.4 (3.06)**	**24.96 (6.04)**	**24.45 (4.43)**	**20.81 (5.07)**	22.66 (4.65)	1554982

Table 3: Results of filtering approaches for Multilingual ST system. Δ is the difference from the baseline model.

Mid Resource (FR-EN & ES-PT) With data sizes of 30,000 and 21,000, respectively, FR-EN and ES-PT language pairs are mid-resourced. When trained using only the SpeechMatrix, the BLEU score of the best model for FR-EN increases by 1.11 BLEU, whereas ES-PT increases by 4.56 in comparison to the Baseline (unfiltered SM) system.

Combining the unfiltered SpeechMatrix data with the MTedX data for FR-EN gives us an additional 8-BLEU points boost compared to only using MTedX data. Filtering the SpeechMatrix with the NLL-based technique and using 60% of the data again gives an additional boost of 0.92.

For ES-PT, the baseline score increases when combining the unfiltered SpeechMatrix data with the MTedX by 5.23 BLEU compared to Baseline (MTedX). Our best model is the union of our two best subsets (0.5 z Text-Text $\bigcup$ 0.5 z Speech-Speech), with an increase of 0.79 BLEU points over the Baseline (MTedX + unfiltered SM). Even using the intersection of our two best subsets, the system is better than the baseline one. This exhibits the prowess of the filtering techniques.

Low Resource (PT-ES) This language pair has the lowest amount of clean data (11,000 instances). Thus, the baseline score (12.62 BLEU) is the lowest among the four pairs. Just adding the SpeechMatrix gives an astonishing 10.02 BLEU score improvement. Noisy parallel data apparently helps improve the system to a certain level when the models are low-resourced. Beyond that, showing the model noisy data only hurts the performance. Filtering is crucial in that aspect to trim down the noisy dataset to a cleaner version. The best model uses NLL-based filtering and 80% of the dataset, which increases its performance by 1.06 BLEU points on top of the Baseline (MTedX + unfiltered SM).

4.2 Multilingual Systems

After observing the success of the NLL-based filtering for four language pairs, we have also explored a multilingual setup. Table 3 shows the result of our multilingual experiments. The language directions are to English from Spanish, French, Portuguese, and Italian. Compared to the bilingual FR-EN results trained on the same dataset, we obtained a higher BLEU score (14.74 to 18.92).

We achieve even better scores by combining MTedX data with the first 20% SpeechMatrix data filtered by NLL-loss for all language pairs. On average, this boosts our BLEU score by 4.65. We obtained the highest jump for FR-EN, a 6.04 BLEU score improvement. We obtain a similar trend for all other subsets, where the quality degrades the more noisy data we use.

5 Conclusion

All language pairs benefit from the simplest filtering techniques, from high-resourced to low-resourced pairs. The multilingual models' increase in model robustness is much more consistent and higher than that of bilingual ones.

6 Limitations

Even though we observed a huge increase in BLEU points by incorporating the simplest scoring and filtering techniques for all the language pairs that we have experimented with, one limitation arises when computing the NLL-based score. The limitation is that the computation time is huge. On the other hand ratio-based scoring techniques are faster by a huge margin. So, even if our best models are from NLL-based scoring, ratio-based scoring can be an obvious choice where time is an issue.

References

Antonios Anastasopoulos, Loïc Barrault, Luisa Bentivogli, Marcely Zanon Boito, Ondřej Bojar, Roldano Cattoni, Anna Currey, Georgiana Dinu, Kevin Duh, Maha Elbayad, Clara Emmanuel, Yannick Estève, Marcello Federico, Christian Federmann, Souhir Gahbiche, Hongyu Gong, Roman Grundkiewicz, Barry Haddow, Benjamin Hsu, Dávid Javorský, Věra Kloudová, Surafel Lakew, Xutai Ma, Prashant Mathur, Paul McNamee, Kenton Murray, Maria Nădejde, Satoshi Nakamura, Matteo Negri, Jan Niehues, Xing Niu, John Ortega, Juan Pino, Elizabeth Salesky, Jiatong Shi, Matthias Sperber, Sebastian Stüker, Katsuhito Sudoh, Marco Turchi, Yogesh Virkar, Alexander Waibel, Changhan Wang, and Shinji Watanabe. 2022. Findings of the IWSLT 2022 evaluation campaign. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, pages 98–157, Dublin, Ireland (in-person and online). Association for Computational Linguistics.

Antonios Anastasopoulos, Ondřej Bojar, Jacob Bremerman, Roldano Cattoni, Maha Elbayad, Marcello Federico, Xutai Ma, Satoshi Nakamura, Matteo Negri, Jan Niehues, Juan Pino, Elizabeth Salesky, Sebastian Stüker, Katsuhito Sudoh, Marco Turchi, Alexander Waibel, Changhan Wang, and Matthew Wiesner. 2021. FINDINGS OF THE IWSLT 2021 EVALUATION CAMPAIGN. In *Proceedings of the 18th International Conference on Spoken Language Translation (IWSLT 2021)*, pages 1–29, Bangkok, Thailand (online). Association for Computational Linguistics.

Parnia Bahar, Tobias Bieschke, and Hermann Ney. 2019. A comparative study on end-to-end speech to text translation. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 792–799.

Sameer Bansal, Herman Kamper, Karen Livescu, Adam Lopez, and Sharon Goldwater. 2019. Pre-training on high-resource speech recognition improves low-resource speech-to-text translation. 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 58–68, Minneapolis, Minnesota. Association for Computational Linguistics.

Luisa Bentivogli, Mauro Cettolo, Marco Gaido, Alina Karakanta, Alberto Martinelli, Matteo Negri, and Marco Turchi. 2021. Cascade versus direct speech translation: Do the differences still make a difference? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2873–2887, Online. Association for Computational Linguistics.

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.

Paul-Ambroise Duquenne, Hongyu Gong, Ning Dong, Jingfei Du, Ann Lee, Vedanuj Goswami, Changhan Wang, Juan Pino, Benoît Sagot, and Holger Schwenk. 2023. SpeechMatrix: A large-scale mined corpus of multilingual speech-to-speech translations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16251–16269, Toronto, Canada. Association for Computational Linguistics.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le. 2019. Specaugment: A simple data augmentation method for automatic speech recognition. In *Interspeech 2019*, interspeech_2019. ISCA.

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

Matthias Sperber and Matthias Paulik. 2020. Speech translation and the end-to-end promise: Taking stock of where we are. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7409–7421, Online. Association for Computational Linguistics.

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

translation task. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4252–4261, Online. Association for Computational Linguistics.

Ron J. Weiss, Jan Chorowski, Navdeep Jaitly, Yonghui Wu, and Zhifeng Chen. 2017. Sequence-to-sequence models can directly transcribe foreign speech. *CoRR*, abs/1703.08581.

Rong Ye, Mingxuan Wang, and Lei Li. 2022. Cross-modal contrastive learning for speech translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5099–5113, Seattle, United States. Association for Computational Linguistics.

A Appendix

A.1 Hyper-parameters

Training Parameters	
Criterion	Label Smoothing Cross Entropy
Label Smoothing	0.1
Optimizer	Adam
Learning Rate (LR)	2e-3
LR Scheduler	Inverse Sqrt
Max Tokens	40000
Warm-up Updates	10000
Max Update	100000
Clip Norm	10.0
Update Frequency	8

Table 4: Hyper-parameters during training our ASR and ST models.

Inference Parameters	
Criterion	Label Smoothing Cross Entropy
Beam	5
Max Tokens	10000
Scoring (ST)	sacrebleu
Scoring (ASR)	WER
- WER Tokenizer	13a
- WER Lowercase	True
- WER remove Punct	True

Table 5: Hyper-parameters during inference of our ASR and ST models.

Model Architecture	
Vocab Size (ASR)	5000
Vocab Size (ST)	8000
Vocab Type	Unigram
Conv1d Kernel Size	5,5
Conv1d Channels	1024
Con21d Channels	256
Encoder Embedding Dim	256
Encoder FFN Embedding Dim	256 * 8
Encoder Attention Head	4
Encoder Layer	12
Decoder Embedding Dim	256
Decoder FFN Embedding Dim	256 * 8
Decoder Attention Head	4
Decoder Layer	6
Dropout	0.1
Activation Function	RELU

Table 6: Hyper-parameters of our ASR and ST model architecture.

A.2 Bilingual Systems Full Results

ST System (Only SM)	# of pairs	BLEU	ST System (MuST-C + SM)	# of pairs	BLEU	Overlap
			(Di Gangi et al., 2019)	269256	22.29	
			(i) Baseline (MuST-C)	269256	31.51	
Baseline (unfiltered SM)	1384112	4.87	**(ii) Baseline (MuST-C + unfiltered SM)**	1653367	28.06	
20% NLL	276823	6.01	**(i) + 20% NLL**	546078	**32.07**	100.00
40% NLL	553645	5.41	**(i) + 40% NLL**	822900	31.86	66.36
60% NLL	830467	7.74	**(i) + 60% NLL**	1099722	31	49.66
80% NLL	1107289	**8.42**	**(i) + 80% NLL**	1376544	30.13	39.67
0.25 z Text-Text	251201	4.64	**(i) + 0.25 z Text-Text**	520456	31.82	62.14
0.5 z Text-Text	582453	7.39	**(i) + 0.5 z Text-Text**	851708	31.2	74.57
0.75 z Text-Text	820294	7.79	**(i) + 0.75 z Text-Text**	1089549	30.04	84.66
1 z Text-Text	1017152	5.54	**(i) + 1 z Text-Text**	1286407	29.52	91.43
0.25 z Speech-Speech	409934	2.76	**(i) + 0.25 z Speech-Speech**	679189	31.28	58.38
0.5 z Speech-Speech	723714	4.63	**(i) + 0.5 z Speech-Speech**	992969	29.95	66.60
0.75 z Speech-Speech	945962	5.05	**(i) + 0.75 z Speech-Speech**	1215217	29.6	73.84
1 z Speech-Speech	1095298	4.81	**(i) + 1 z Speech-Speech**	1364553	29.13	79.89
0.25 z Speech-Text	458789	5.77	**(i) + 0.25 z Speech-Text**	728044	31.05	55.30
0.5 z Speech-Text	860457	7.3	**(i) + 0.5 z Speech-Text**	1129712	29.64	67.51
0.75 z Speech-Text	1133782	7.57	**(i) + 0.75 z Speech-Text**	1403037	29.04	75.42
1 z Speech-Text	1266324	6.82	**(i) + 1 z Speech-Text**	1535579	28.89	82.75
0.25 z Text-Speech	311053	2.13	**(i) + 0.25 z Text-Speech**	580308	31.81	57.82
0.5 z Text-Speech	601697	3.88	**(i) + 0.5 z Text-Speech**	870952	30.57	66.56
0.75 z Text-Speech	849956	5.23	**(i) + 0.75 z Text-Speech**	1119211	29.91	74.09
1 z Text-Speech	1042400	6.21	**(i) + 1 z Text-Speech**	1311655	29.69	80.55
			(i) + 20% NLL $\bigcup$ 0.25 z Text-Text	764533	31.58	100.00
			(i) + 20% NLL $\bigcap$ 0.25 z Text-Text	302001	30.94	100.00

Table 7: Results of filtering approaches for En-Fr direction

ST System (Only SM)	# of pairs	BLEU	ST System (MTedX + SM)	# of pairs	BLEU	Overlap
			(Salesky et al., 2021)	29946	8.9	
			(i) Baseline (MTedX)	29946	14.74	
Baseline (unfiltered SM)	1403985	5.2	**(ii) Baseline (MTedX + unfiltered SM)**	1433931	26.74	
20% NLL	280797	5.15	**(i) + 20% NLL**	310743	22.33	100.00
40% NLL	561594	**6.31**	**(i) + 40% NLL**	591540	25.46	100.00
60% NLL	842391	5.36	**(i) + 60% NLL**	872337	**27.66**	100.00
80% NLL	1123188	5.68	**(i) + 80% NLL**	1153134	27.23	75.65
0.25 z Text-Text	475032	3.8	**(i) + 0.25 z Text-Text**	504978	26.52	37.18
0.5 z Text-Text	893780	4.78	**(i) + 0.5 z Text-Text**	923726	26.58	66.73
0.75 z Text-Text	1159658	4.11	**(i) + 0.75 z Text-Text**	1189604	26.81	85.20
1 z Text-Text	1285668	4.61	**(i) + 1 z Text-Text**	1315614	26.33	93.89
0.25 z Speech-Speech	413584	4.24	**(i) + 0.25 z Speech-Speech**	443530	26.85	32.76
0.5 z Speech-Speech	729455	4.28	**(i) + 0.5 z Speech-Speech**	759401	26.68	55.00
0.75 z Speech-Speech	952797	4.51	**(i) + 0.75 z Speech-Speech**	982743	27.06	70.73
1 z Speech-Speech	1102227	4.94	**(i) + 1 z Speech-Speech**	1132173	26.62	81.25
0.25 z Speech-Text	308869	4.92	**(i) + 0.25 z Speech-Text**	338815	25.21	25.28
0.5 z Speech-Text	600976	4.36	**(i) + 0.5 z Speech-Text**	630922	26.35	45.88
0.75 z Speech-Text	849820	4.56	**(i) + 0.75 z Speech-Text**	879766	26.55	63.44
1 z Speech-Text	1041531	4.25	**(i) + 1 z Speech-Text**	1071477	26.7	76.93
0.25 z Text-Speech	458975	4.37	**(i) + 0.25 z Text-Speech**	488921	26.07	35.79
0.5 z Text-Speech	861614	3.4	**(i) + 0.5 z Text-Speech**	891560	26.22	64.12
0.75 z Text-Speech	1134478	4.91	**(i) + 0.75 z Text-Speech**	1164424	26.26	83.33
1 z Text-Speech	1266947	4.31	**(i) + 1 z Text-Speech**	1296893	26.64	92.59
			(i) + 60% NLL $\bigcup$ 0.75 z Speech-Speech	1250087	27.18	100.00
			(i) + 60% NLL $\bigcap$ 0.75 z Speech-Speech	604993	26.22	100.00

Table 8: Results of filtering approaches for Fr-En direction

ST System (Only SM)	# of pairs	BLEU	ST System (MTEdX + SM)	# of pairs	BLEU	Overlap
			(Salesky et al., 2021)	20821	12.2	
			(i) Baseline (MTedX)	20821	21.05	
Baseline (unfiltered SM)	1074027	0.57	**(ii) Baseline (MTedX + unfiltered SM)**	1094848	26.28	
20% NLL	214806	2.54	**(i) + 20% NLL**	235627	24.44	49.43
40% NLL	429611	2.76	**(i) + 40% NLL**	450432	24.57	48.60
60% NLL	644416	4.07	**(i) + 60% NLL**	665237	24.45	49.72
80% NLL	859221	4.24	**(i) + 80% NLL**	880042	26.16	51.21
0.25 z Text-Text	294944	3.9	**(i) + 0.25 z Text-Text**	315765	25.62	54.43
0.5 z Text-Text	559285	**5.13**	**(i) + 0.5 z Text-Text**	580106	26.77	100.00
0.75 z Text-Text	762804	4.55	**(i) + 0.75 z Text-Text**	783625	26.22	100.00
1 z Text-Text	904329	4.2	**(i) + 1 z Text-Text**	925150	26.13	100.00
0.25 z Speech-Speech	370267	3.34	**(i) + 0.25 z Speech-Speech**	391088	26.3	39.95
0.5 z Speech-Speech	584872	4.09	**(i) + 0.5 z Speech-Speech**	605693	26.67	60.64
0.75 z Speech-Speech	739279	4.39	**(i) + 0.75 z Speech-Speech**	760100	26.57	74.69
1 z Speech-Speech	847397	4.51	**(i) + 1 z Speech-Speech**	868218	26.35	83.81
0.25 z Speech-Text	295074	3.68	**(i) + 0.25 z Speech-Text**	315895	24.74	36.25
0.5 z Speech-Text	570658	5.02	**(i) + 0.5 z Speech-Text**	591479	26.13	64.95
0.75 z Speech-Text	788590	4.95	**(i) + 0.75 z Speech-Text**	809411	26.47	84.40
1 z Speech-Text	924162	4.56	**(i) + 1 z Speech-Text**	944983	26.64	93.82
0.25 z Text-Speech	288588	2.32	**(i) + 0.25 z Text-Speech**	309409	24.98	34.47
0.5 z Text-Speech	549748	3.79	**(i) + 0.5 z Text-Speech**	570569	25.55	60.88
0.75 z Text-Speech	755145	3.2	**(i) + 0.75 z Text-Speech**	775966	24.96	79.37
1 z Text-Speech	897009	2.81	**(i) + 1 z Text-Speech**	917830	25.65	89.99
			(i) + 0.5 z Text-Text $\bigcup$ 0.5 z Speech-Speech	834039	**27.07**	100.00
			(i) + 0.5 z Text-Text $\bigcap$ 0.5 z Speech-Speech	351760	26.46	100.00

Table 9: Results of filtering approaches for Es-Pt direction

ST System (Only SM)	# of pairs	BLEU	ST System (MTedX + SM)	# of pairs	BLEU	Overlap
			(Salesky et al., 2021)	11353	8.7	
			(i) Baseline (MTedX)	11353	12.62	
Baseline (SM)	1107283	3.23	**(ii) Baseline (MTedX + unfiltered SM)**	1118635	22.64	
20% NLL	221457	3.24	**(i) + 20% NLL**	232810	20.01	100.00
40% NLL	442913	4.11	**(i) + 40% NLL**	454266	20.07	100.00
60% NLL	664369	**4.38**	**(i) + 60% NLL**	675722	21.02	100.00
80% NLL	885825	4.02	**(i) + 80% NLL**	897178	**23.7**	100.00
0.25 z Text-Text	314739	1.35	**(i) + 0.25 z Text-Text**	326092	21.66	30.41
0.5 z Text-Text	603732	2.98	**(i) + 0.5 z Text-Text**	615085	22.88	57.11
0.75 z Text-Text	818558	2.97	**(i) + 0.75 z Text-Text**	829911	22.91	76.83
1 z Text-Text	950625	3.53	**(i) + 1 z Text-Text**	961978	23.07	88.98
0.25 z Speech-Speech	363865	2.2	**(i) + 0.25 z Speech-Speech**	375218	22.51	35.40
0.5 z Speech-Speech	596387	2.97	**(i) + 0.5 z Speech-Speech**	607740	22.08	56.79
0.75 z Speech-Speech	756379	3.22	**(i) + 0.75 z Speech-Speech**	767732	23.19	71.46
1 z Speech-Speech	865675	3.34	**(i) + 1 z Speech-Speech**	877028	22.46	81.42
0.25 z Speech-Text	290424	2.24	**(i) + 0.25 z Speech-Text**	301777	20.97	28.18
0.5 z Speech-Text	552538	2.51	**(i) + 0.5 z Speech-Text**	563891	22.7	52.39
0.75 z Speech-Text	760190	2.78	**(i) + 0.75 z Speech-Text**	771543	22.67	71.52
1 z Speech-Text	901333	3.88	**(i) + 1 z Speech-Text**	912686	21.88	84.46
0.25 z Text-Speech	295699	1.27	**(i) + 0.25 z Text-Speech**	307052	20.07	28.66
0.5 z Text-Speech	571514	2.77	**(i) + 0.5 z Text-Speech**	582867	22.69	54.09
0.75 z Text-Speech	790076	3.29	**(i) + 0.75 z Text-Speech**	801429	23.32	74.20
1 z Text-Speech	927495	2.4	**(i) + 1 z Text-Speech**	938848	22.96	86.83
			(i) + 80% NLL $\bigcup$ 0.75 z Speech-Speech	1043876	22.6	100.00
			(i) + 80% NLL $\bigcap$ 0.75 z Speech-Speech	621034	22.94	100.00

Table 10: Results of filtering approaches for Pt-Es direction