CMU's IWSLT 2024 Offline Speech Translation System: A Cascaded Approach For Long-Form Robustness

Brian Yan*[1] Patrick Fernandes*[1] Jinchuan Tian[1] Siqi Ouyang[1]
William Chen[1] Karen Livescu[1,2] Lei Li[1] Graham Neubig[1] Shinji Watanabe[1,3]
[1]Language Technologies Institute, Carnegie Mellon University, USA
[2]Toyota Technological Institute at Chicago, University of Chicago, USA
[3]Human Language Technology Center of Excellence, Johns Hopkins University, USA
{byan, pfernand}@cs.cmu.edu

Abstract

This work describes CMU's submission to the IWSLT 2024 Offline Speech Translation (ST) Shared Task for translating English speech to German, Chinese, and Japanese text. We are the first participants to employ a *long-form* strategy which directly processes unsegmented recordings without the need for a separate voice-activity detection stage (VAD). We show that the Whisper automatic speech recognition (ASR) model has a hallucination problem when applied out-of-the-box to recordings containing non-speech noises, but a simple noisy fine-tuning approach can greatly enhance Whisper's long-form robustness across multiple domains. Then, we feed English ASR outputs into fine-tuned NLLB machine translation (MT) models which are decoded using COMET-based Minimum Bayes Risk. Our VAD-free ASR+MT cascade is tested on TED talks, TV series, and workout videos and shown to outperform prior winning IWSLT submissions and large open-source models.

1 Introduction

CMU's submission to the IWSLT 2024 Offline Speech Translation shared task is a cascaded automatic speech recognition (ASR) and machine translation (MT) system designed to effectively translate English speech from long unsegmented recordings, such as TED talks, TV series, and workout videos, into German, Chinese, and Japanese text.

Typically systems are *short-form*, meaning they are dependent on some voice-activity detection to first convert long recordings which contain speech and non-speech noises into short segments of speech. This makes it relatively easy to train a short-form model and test it on similar clean speech segments. However, these systems exhibit alarming brittleness in the wild; results from recent iterations of the Offline ST track have shown large fluctuations in performance between different segmentations of the same test set (Anastasopoulos et al., 2021, 2022; Agarwal et al., 2023).

Why are these short-form systems brittle in-the-wild (or in IWSLT by proxy)? Our view is that these systems are plagued by train/test mismatch. Common training sets, e.g. MuST-C (Di Gangi et al., 2019), are produced using sentence-level forced alignment. In other words, this training segmentation can only be obtained given a reference. For a blind test set however, forced alignment is not possible. Instead, practitioners have resorted to using VAD with additional tricks to reduce the train/test mismatch, such as heuristically replicating segment characteristics (Inaguma et al., 2021) or modeling the segmentation pattern of training data (Tsiamas et al., 2022). These methods of approximating the training data segmentation may work within a single domain but are complex to configure for multi-domain scenarios.

In this work, we explore *long-form* processing of unsegmented recordings via a 30 second sliding window as an alternative to segment-dependent speech processing. Our system consists of:

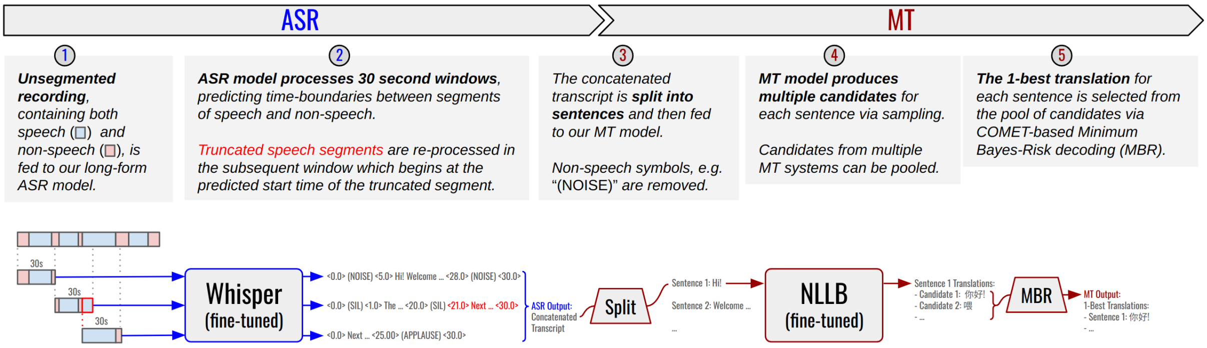
1. Whisper-based ASR (Radford et al., 2023) applied in long-form inference §3.1.1, after a simple noisy fine-tuning procedure which greatly enhances robustness to non-speech noises §3.1.2
2. NLLB-based MT (Costa-jussà et al., 2022), fine-tuned and decoded via Minimum Bayes-Risk §3.2

Our experiments first show that Whisper out-of-the-box has a hallucination problem caused by non-speech noises during long-form inference. We then show that our noisy fine-tuning broadly addresses these hallucinations. Finally, we show the ultimate cascaded ST performance across multiple domains: TED talks, TV series, and workout videos.

2 Task Description

The IWSLT 2024 Offline Speech Translation shared task consists of three language

Proceedings of the 21st International Conference on Spoken Language Translation (IWSLT 2024), pages 212–217
August 15-16, 2024 ©2024 Association for Computational Linguistics



Unsegmented recording, containing both speech (□) and non-speech (□), is fed to our long-form ASR model.

ASR model processes 30 second windows, predicting time-boundaries between segments of speech and non-speech.

Truncated speech segments are re-processed in the subsequent window which begins at the predicted start time of the truncated segment.

The concatenated transcript is split into sentences and then fed to our MT model.

Non-speech symbols, e.g. "(NOISE)" are removed.

MT model produces multiple candidates for each sentence via sampling.

Candidates from multiple MT systems can be pooled.

The 1-best translation for each sentence is selected from the pool of candidates via COMET-based Minimum Bayes-Risk decoding (MBR).

Figure 1: Summary of our cascaded system. ASR: long-form processing of unsegmented recordings. MT: sentence-based translation with Minimum Bayes-Risk decoding.

pairs: English-to-German, English-to-Chinese, and English-to-Japanese. For all three language pairs, unsegmented TED talks (5-20 min) are given as shared task evaluation data. As a dev set, we use the provided tst2020 (TED'20), tst2021 (TED'21), and tst2022 (TED'22) for English-to-German and tst2022 (TED'22) for English-to-Chinese and English-to-Japanese.

For English-to-German, systems are also tested on additional domains: TV series (45-60 min), workout videos (10-20 min), and accented speech (5-20 min). We therefore use two additional dev sets obtained from the IWSLT 2024 Subtitling shared task: ITV and Peloton.

We evaluate ASR using case-sensitive punctuated word error-rate (WER ↓) against recording-level references. We evaluate MT systems using COMET ↑ (Rei et al., 2020) against sentence-level references. We evaluate ST systems using COMET after first performing minimum WER alignment of our hypothesis to sentence-level references. Note that for Chinese and Japanese, this alignment is done at the character level.

We use MuST-C v3 (Di Gangi et al., 2019) for fine-tuning ASR models on the TED domain. We use TED2020 for fine-tuning English-to-German MT and MuST-C for English-to-Chinese and English-to-Japanese. For multi-domain fine-tuning we also add Bazinga TV series ASR data (Lerner et al., 2022) and a 500k subset of OpenSubtitles MT data (Creutz, 2018). Note that our use of Bazinga (as well as the use of Whisper) puts our system under the "Unconstrained" designation.

3 System Description

Figure 1 summarizes the components in our ASR+MT cascade. The following section describes the system in greater detail, referring at times to the summary figure.

3.1 ASR

3.1.1 Long-Form Inference

As illustrated in Steps 1 and 2 of Figure 1, we deploy Whisper in a long-form mode. Under this scheme, the window size is always 30 seconds (or the remainder of the recording). Although the window size is fixed, the hop size is dynamic and based on the predicted time-boundaries of speech segments. As shown in Step 2, the final speech segment in a window is considered to be truncated if the predicted end-time is within 1 second of the end of the window. To avoid transcribing with a truncated utterance, the next window starts from the start-time of the truncated utterance.

For non-speech noises, the expected behavior is that the model produces a special symbol, e.g. (NOISE), along with time-boundaries. **However, we found that Whisper Large-v2 frequently hallucinates on non-speech such as music and applause.**[1] These errors can be categorized as oscillations in which the auto-regressive decoder enters a bad state causing long repeated garbage outputs.

Whisper applies an inference time patch to address these oscillations, somewhat obscuring the lack of long-form robustness in the model out-of-the-box. This patch detects oscillations via a heuristic repetition factor, then if high repetitions are detected then it falls back to sampling. If the sampling output is still high in repetitions, then it falls back to sampling with greater and greater temperature. Eventually, the model either escapes from the oscillations (typically by producing EOS) or exhausts

[1] We also tested Large-v3 and found that hallucinations to be more severe than Large-v2, perhaps due to error compounding from the semi-supervision used in Large-v3.

Model	TED'20	TED'21	TED'22	ITV	Peloton
Whisper	54.9	8.3	5.8	38.9	47.9
+ Fallback on Oscillation	1.6	2.4	2.1	6.5	4.2
Whisper ft on TED + Baz	**0.9**	**1.2**	1.9	6.9	5.9
+ Fallback on Oscillation	**0.9**	**1.2**	**1.1**	**3.5**	**3.4**

Table 1: Insertion error-rates of Whisper out-of-the-box vs Whisper after Noisy Fine-tuning. High insertions indicates frequent oscillation.

the allotted number of fallback decodings.

3.1.2 Noisy Fine-tuning

Motivated by the apparent lack of long-form robustness described in the previous section, we propose a simple fine-tuning strategy to improve the Whisper's ability to predict the special non-speech token: (NOISE). We prepare fine-tuning data by taking consecutive 30 second segments from the unsegmented recordings. Using given sentence-level forced alignments, we obtain references containing transcribed speech and noises. Critically, the 30 second segments also include untranscribed noises; these non-speech portions were originally cut out via forced alignment (they represent the durations between speech segments). If these untranscribed non-speech portions exceed 1 second in duration, we add a new (NOISE) token in the target.

In practice (see §4.1), this noisy fine-tuning encompasses non-speech noises that Whisper out-of-the-box struggles with. After fine-tuning, the model does not produce oscillations and rather produces the (NOISE) token which is cleaned before scoring ASR and feeding into MT.

3.2 MT

ASR outputs, which are cased and punctuated, are concatenated at a recording-level (Step 3). This recording-level ASR output is then split into sentences and subsequently fed into our MT model.

For each language-pair, we fine-tune NLLB 1B and NLLB 3B on TED data. For English-German we also fine-tune a separate NLLB 1B model on TED + OpenSubtitles data.

During inference, we generate a set of candidate translations via epsilon-sampling (Step 4). We then (optionally) pool the candidate translations across multiple MT systems. Finally, the 1-best translation is chosen using COMET-based Minimum Bayes-Risk decoding (Yan et al., 2022).

4 Results

4.1 Noisy Fine-Tuning Improves Whisper's Long-Form Robustness

Table 1 shows ASR insertion error-rates for Whisper out-of-the-box versus Whisper with noisy fine-tuning. As can be seen from the high insertion error-rates in row 1, Whisper without fine-tuning and without relying on the fallback-based inference-time patch (described in §3.1.1) has a major oscillation problem. Noisy fine-tuning greatly reduces this problem, as can be seen from row 3. Our results show that noisy fine-tuning improved performance on all domains, so we have reason to believe that the improved long-form robustness generalizes to some extent. The fallback method still improves the fine-tuned model, indicating that some oscillations still remain, but this inference-time patch is not critical as it was out-of-the-box.

Note that fallback is applied in all subsequent ASR results unless otherwise indicated.

4.2 ST Results

Table 2 shows the ASR, MT, and ST performances of our fine-tuned models versus their out-of-the-box counterparts for English-German. For ASR, fine-tuning on TED + Bazinga versus fine-tuning on TED-only improved the TV series performance (ITV) while maintaining the performance on TED.

For MT, the NLLB 3B fine-tuned model was the best across all sets. The NLLB 1B models fine-tuned on TED versus on TED + OpenSubtitles performed similarly. We use all three MT models in our final ensemble.

Table 3 shows a single-domain version of the same story for English-Chinese and English-Japanese. For these pairs, we use the TED-only fine-tuned ASR model and we do not use any TED + OpenSubtitles fine-tuned MT models.

4.3 COMET-Based Minimum Bayes-Risk

Table 4 shows the impact of COMET-based MBR compared to beam search. We observed improvements up to 50 samples per system. Further, ensembling slightly improves results.

4.4 Benchmarking vs. Prior Works

Finally, Table 5 compares our VAD-free cascaded approach to prior works. Note we're showing BLEU score (Post, 2018) in this table for compatibility with prior studies.

MODEL	TED'20	TED'21	TED'22	ITV	PELOTON	TED AVG	NON-TED AVG
ASR			WER ↓				
Whisper (Large-v2)	10.8	10.1	9.3	30.9	24.2	10.1	27.6
Whisper ft on TED	8.8	**7.5**	**7.8**	27.5	**22.3**	**8.0**	24.9
Whisper ft on TED + Bazinga (Baz)	**8.7**	7.7	**7.8**	**25.0**	22.4	8.1	**23.7**
MT			COMET ↑				
NLLB 1B	0.8093	0.7825	0.7845	0.6322	0.6162	0.7921	0.6242
NLLB 1B ft on TED	0.8229	0.8006	0.7977	0.6638	0.6348	0.8071	0.6493
NLLB 1B ft on TED + OpenSubtitles (OS)	0.8219	0.7991	0.7943	0.6598	0.6396	0.8051	0.6497
NLLB 3B	0.8171	0.7892	0.7892	0.6472	0.6200	0.7985	0.6336
NLLB 3B ft on TED	**0.8242**	**0.8053**	**0.8010**	**0.6697**	**0.6548**	**0.8102**	**0.6623**
ST (ASR→MT)			COMET ↑				
Whisper → NLLB 1B	0.7891	0.7622	0.7691	0.5920	0.6119	0.7735	0.6020
Whisper → NLLB 3B	0.7954	0.7705	0.7779	0.6012	0.6152	0.7813	0.6082
Whisper ft on TED → NLLB 1B ft on TED	0.8050	0.7872	0.7844	0.6311	0.6111	0.7922	0.6211
Whisper ft on TED + Baz → NLLB 1B ft on TED (①)	0.8053	0.7856	0.7855	0.6501	0.6087	0.7921	0.6294
Whisper ft on TED + Baz → NLLB 1B ft on TED + OS (②)	0.8018	0.7872	0.7827	0.6537	0.6086	0.7906	0.6312
Whisper ft on TED + Baz → NLLB 3B ft on TED (③)	**0.8059**	**0.7911**	**0.7875**	**0.6562**	**0.6183**	**0.7948**	**0.6373**
MBR Ensemble (① + ② + ③)	-	-	**0.8104**	**0.6647**	**0.6293**	-	**0.6470**

Table 2: ASR/MT/ST results for English-German across TED and non-TED domains.

LANG	MODEL	MT	ST
En-Zh	NLLB 1B	0.7864	0.7309
En-Zh	NLLB 1B ft on TED (①)	**0.8362**	**0.8082**
En-Zh	NLLB 3B	0.7464	0.7279
En-Zh	NLLB 3B ft on TED (②)	**0.8362**	0.8078
En-Zh	MBR Ensemble (① + ②)	-	**0.8295**
En-Ja	NLLB 1B	0.8300	0.7568
En-Ja	NLLB 1B ft on TED (①)	0.8625	**0.8086**
En-Ja	NLLB 3B	0.7854	0.7715
En-Ja	NLLB 3B ft on TED (②)	**0.8639**	0.8046
En-Ja	MBR Ensemble (① + ②)	-	**0.8363**

Table 3: MT/ST results for English-Chinese and English-Japanese.

MODEL	DECODING	TED'22	ITV	PELOTON
NLLB 1B ft on TED	Beam (5)	0.7855	0.6501	0.6087
NLLB 1B ft on TED (①)	MBR (50)	**0.8038**	**0.6570**	**0.6180**
NLLB 1B ft on TED + OS	Beam (5)	0.7827	0.6537	0.6086
NLLB 1B ft on TED + OS (②)	MBR (50)	**0.8009**	**0.6628**	**0.6207**
NLLB 3B ft on TED	Beam (5)	0.7875	0.6562	0.6183
NLLB 3B ft on TED (③)	MBR (50)	**0.8076**	**0.6632**	**0.6286**
Ensemble (① + ② + ③)	MBR (50 ea.)	**0.8104**	**0.6647**	**0.6293**

Table 4: Beam search vs. MBR decoding.

5 Conclusion

We describe our IWSLT 2024 Offline Speech Translation system which is based on long-form processing of unsegmented recordings. Our system consists of fine-tuned Whisper and NLLB components of a cascade. We evaluate our system on TED talks, TV series, and workout videos.

TYPE	MODEL	USES VAD	TED'22
Cascade	IWSLT 2022 Top (Zhang et al., 2022)	✓	23.9
Cascade	Our Single Best Model	✗	**24.5**
Direct	SeamlessM4T (Barrault et al., 2023)	✓	16.2
Direct	WavLM+mBART (Yan et al., 2023)	✓	19.2
Direct	OWSM 3.1 (Peng et al., 2024b)	✗	18.4
Direct	OWSM-CTC (Peng et al., 2024a)	✗	**19.6**

Table 5: BLEU score comparison with prior works.

Acknowledgements

Brian Yan and Shinji Watanabe are supported by the Human Language Technology Center of Excellence. This work used the Extreme Science and Engineering Discovery Environment (XSEDE) (Towns et al., 2014), which is supported by National Science Foundation grant number ACI-1548562; specifically, the Bridges system (Nystrom et al., 2015), as part of project cis210027p, which is supported by NSF award number ACI-1445606, at the Pittsburgh Supercomputing Center. This work also used GPUs donated by the NVIDIA Corporation.

References

Milind Agarwal, Sweta Agrawal, Antonios Anastasopoulos, Ondřej Bojar, Claudia Borg, Marine Carpuat, Roldano Cattoni, Mauro Cettolo, Mingda Chen, William Chen, Khalid Choukri, Alexandra Chronopoulou, Anna Currey, Thierry Declerck, Qianqian Dong, Yannick Estève, Kevin Duh, Marcello Federico, Souhir Gahbiche, Barry Haddow, Benjamin Hsu, Phu Mon Htut, Hirofumi Inaguma, Dávid Javorský, John Judge, Yasumasa Kano, Tom Ko, Rishu

Kumar, Pengwei Li, Xutail Ma, Prashant Mathur, Evgeny Matusov, Paul McNamee, John P. McCrae, Kenton Murray, Maria Nadejde, Satoshi Nakamura, Matteo Negri, Ha Nguyen, Jan Niehues, Xing Niu, Atul Ojha Kr., John E. Ortega, Proyag Pal, Juan Pino, Lonneke van der Plas, Peter Polák, Elijah Rippeth, Elizabeth Salesky, Jiatong Shi, Matthias Sperber, Sebastian Stüker, Katsuhito Sudoh, Yun Tang, Brian Thompson, Kevin Tran, Marco Turchi, Alex Waibel, Mingxuan Wang, Shinji Watanabe, and Rodolfo Zevallos. 2023. Findings of the IWSLT 2023 Evaluation Campaign. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*. Association for Computational Linguistics.

Antonios Anastasopoulos, Loc Barrault, Luisa Bentivogli, Marcely Zanon Boito, Ondřej Bojar, Roldano Cattoni, Anna Currey, Georgiana Dinu, Kevin Duh, Maha Elbayad, et al. 2022. Findings of the iwslt 2022 evaluation campaign. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, pages 98–157. 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.

Loïc Barrault, Yu-An Chung, Mariano Cora Meglioli, David Dale, Ning Dong, Paul-Ambroise Duquenne, Hady Elsahar, Hongyu Gong, Kevin Heffernan, John Hoffman, et al. 2023. Seamlessm4t-massively multilingual & multimodal machine translation. *arXiv preprint arXiv:2308.11596*.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Mathias Creutz. 2018. Open subtitles paraphrase corpus for six languages. *arXiv preprint arXiv:1809.06142*.

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.

Hirofumi Inaguma, Brian Yan, Siddharth Dalmia, Pengcheng Guo, Jiatong Shi, Kevin Duh, and Shinji Watanabe. 2021. Espnet-st iwslt 2021 offline speech translation system. *IWSLT 2021*, page 100.

Paul Lerner, Juliette Bergoënd, Camille Guinaudeau, Hervé Bredin, Benjamin Maurice, Sharleyne Lefevre, Martin Bouteiller, Aman Berhe, Léo Galmant, Ruiqing Yin, et al. 2022. Bazinga! a dataset for multi-party dialogues structuring. In *13th Conference on Language Resources and Evaluation (LREC 2022)*, pages 3434–3441.

Nicholas A Nystrom, Michael J Levine, Ralph Z Roskies, and J Ray Scott. 2015. Bridges: a uniquely flexible hpc resource for new communities and data analytics. In *Proceedings of the 2015 XSEDE Conference: Scientific Advancements Enabled by Enhanced Cyberinfrastructure*, pages 1–8.

Yifan Peng, Yui Sudo, Muhammad Shakeel, and Shinji Watanabe. 2024a. Owsm-ctc: An open encoder-only speech foundation model for speech recognition, translation, and language identification. *arXiv preprint arXiv:2402.12654*.

Yifan Peng, Jinchuan Tian, William Chen, Siddhant Arora, Brian Yan, Yui Sudo, Muhammad Shakeel, Kwanghee Choi, Jiatong Shi, Xuankai Chang, et al. 2024b. Owsm v3. 1: Better and faster open whisper-style speech models based on e-branchformer. *arXiv preprint arXiv:2401.16658*.

Matt Post. 2018. A call for clarity in reporting bleu scores. *WMT 2018*, page 186.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pages 28492–28518. PMLR.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. Comet: A neural framework for mt evaluation. *arXiv preprint arXiv:2009.09025*.

J. Towns, T. Cockerill, M. Dahan, I. Foster, K. Gaither, A. Grimshaw, V. Hazlewood, S. Lathrop, D. Lifka, G. D. Peterson, R. Roskies, J. R. Scott, and N. Wilkins-Diehr. 2014. Xsede: Accelerating scientific discovery. *Computing in Science & Engineering*, 16(5):62–74.

Ioannis Tsiamas, Gerard I Gállego, José AR Fonollosa, et al. 2022. Shas: Approaching optimal segmentation for end-to-end speech translation.

Brian Yan, Patrick Fernandes, Siddharth Dalmia, Jiatong Shi, Yifan Peng, Dan Berrebbi, Xinyi Wang, Graham Neubig, and Shinji Watanabe. 2022. Cmu's iwslt 2022 dialect speech translation system. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, pages 298–307.

Brian Yan, Jiatong Shi, Soumi Maiti, William Chen, Xinjian Li, Yifan Peng, Siddhant Arora, and Shinji Watanabe. 2023. Cmu's iwslt 2023 simultaneous speech translation system. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pages 235–240.

Weitai Zhang, Zhongyi Ye, Haitao Tang, Xiaoxi Li, Xinyuan Zhou, Jing Yang, Jianwei Cui, Pan Deng, Mohan Shi, Yifan Song, Dan Liu, Junhua Liu, and Lirong Dai. 2022. The USTC-NELSLIP offline speech translation systems for IWSLT 2022. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, pages 198–207, Dublin, Ireland (in-person and online). Association for Computational Linguistics.