FINDINGS OF THE IWSLT 2023 EVALUATION CAMPAIGN

Milind Agarwal[1] **Sweta Agrawal**[2] **Antonios Anastasopoulos**[1] **Luisa Bentivogli**[3]
Ondřej Bojar[4] **Claudia Borg**[5] **Marine Carpuat**[2] **Roldano Cattoni**[3]
Mauro Cettolo[3] **Mingda Chen**[6] **William Chen**[7] **Khalid Choukri**[8]
Alexandra Chronopoulou[9] **Anna Currey**[10] **Thierry Declerck**[11] **Qianqian Dong**[12]
Kevin Duh[13] **Yannick Estève**[14] **Marcello Federico**[10] **Souhir Gahbiche**[15]
Barry Haddow[16] **Benjamin Hsu**[10] **Phu Mon Htut**[10] **Hirofumi Inaguma**[6]
Dávid Javorský[4] **John Judge**[17] **Yasumasa Kano**[18] **Tom Ko**[12]
Rishu Kumar[4] **Pengwei Li**[6] **Xutai Ma**[6] **Prashant Mathur**[10]
Evgeny Matusov[19] **Paul McNamee**[13] **John P. McCrae**[20] **Kenton Murray**[13]
Maria Nadejde[10] **Satoshi Nakamura**[18] **Matteo Negri**[3] **Ha Nguyen**[14]
Jan Niehues[21] **Xing Niu**[10] **Atul Kr. Ojha**[20] **John E. Ortega**[22]
Proyag Pal[16] **Juan Pino**[6] **Lonneke van der Plas**[23] **Peter Polák**[4]
Elijah Rippeth[2] **Elizabeth Salesky**[13] **Jiatong Shi**[7] **Matthias Sperber**[24]
Sebastian Stüker[25] **Katsuhito Sudoh**[18] **Yun Tang**[6] **Brian Thompson**[10]
Kevin Tran[6] **Marco Turchi**[25] **Alex Waibel**[7] **Mingxuan Wang**[12]
Shinji Watanabe[7] **Rodolfo Zevallos**[26]
[1]GMU [2]UMD [3]FBK [4]Charles U. [5]U. Malta [6]Meta [7]CMU [8]ELDA
[9]LMU [10]AWS [11]DFKI [12]ByteDance [13]JHU [14]Avignon U. [15]Airbus
[16]U. Edinburgh [18]NAIST [19]AppTek [20]U. Galway [21]KIT [22]Northeastern U.
[23]IDIAP [24]Apple [25]Zoom [26]U. Pompeu Fabra

Abstract

This paper reports on the shared tasks organized by the 20th IWSLT Conference. The shared tasks address 9 scientific challenges in spoken language translation: simultaneous and offline translation, automatic subtitling and dubbing, speech-to-speech translation, multilingual, dialect and low-resource speech translation, and formality control. The shared tasks attracted a total of 38 submissions by 31 teams. The growing interest towards spoken language translation is also witnessed by the constantly increasing number of shared task organizers and contributors to the overview paper, almost evenly distributed across industry and academia.

1 Introduction

The International Conference on Spoken Language Translation (IWSLT) is the premier annual scientific conference for all aspects of spoken language translation (SLT). IWSLT is organized by the Special Interest Group on Spoken Language Translation (SIG-SLT), which is supported by ACL, ISCA and ELRA. Like in all previous editions (Akiba et al., 2004; Eck and Hori, 2005; Paul, 2006; Fordyce, 2007; Paul, 2008, 2009; Paul et al., 2010; Federico et al., 2011, 2012; Cettolo et al., 2013, 2014, 2015, 2016, 2017; Niehues et al., 2018, 2019; Ansari et al., 2020; Anastasopoulos et al., 2021, 2022b),this year's conference was preceded by an evaluation campaign featuring shared tasks addressing scientific challenges in SLT.

This paper reports on the 2023 IWSLT Evaluation Campaign, which offered the following 9 shared tasks:

- **Offline SLT**, with focus on speech-to-text translation of recorded conferences and interviews from English to German, Japanese and Chinese.
- **Simultaneous SLT**, focusing on speech-to-text translation of streamed audio of conferences and interviews from English to German, Japanese and Chinese.
- **Automatic Subtitling**, with focus on speech-to-subtitle translation of audio-visual documents from English to German and Spanish.
- **Multilingual SLT**, with focus on speech-to-text translation of recorded scientific talks from

Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023), pages 1–61

Team	Organization
ALEXA AI	Amazon Alexa AI, USA (Vishnu et al., 2023)
APPTEK	AppTek, Germany (Bahar et al., 2023)
BIGAI	Beijing Institute of General Artificial Intelligence, China (Xie, 2023)
BIT	Beijing Institute of Technology, China (Wang et al., 2023b)
BUT	Brno University of Technology, Czechia (Kesiraju et al., 2023)
CMU	Carnegie Mellon University, USA (Yan et al., 2023)
CUNI-KIT	Charles University, Czechia, and KIT, Germany (Polák et al., 2023)
FBK	Fondazione Bruno Kessler, Italy (Papi et al., 2023)
GMU	George Mason University, USA (Mbuya and Anastasopoulos, 2023)
HW-TSC	Huawei Translation Services Center, China (Li et al., 2023; Wang et al., 2023a) (Guo et al., 2023; Shang et al., 2023; Rao et al., 2023)
I2R	Institute for Infocomm Research, A*STAR, Singapore (Huzaifah et al., 2023)
JHU	Johns Hopkins University, USA (Hussein et al., 2023; Xinyuan et al., 2023)
KIT	Karlsruhe Institute of Technology, Germany (Liu et al., 2023)
KU	Kyoto University, Japan (Yang et al., 2023)
KU X UPSTAGE	Korea University X Upstage, South Korea (Wu et al., 2023; Lee et al., 2023)
MATESUB	Translated Srl, Italy (Perone, 2023)
MINETRANS	U. of Sci. and Techn. of China, Tancient AI Lab, State Key Lab. of Cognitive Intelligence (Du et al., 2023)
NAIST	Nara Institute of Science and Technology, Japan (Fukuda et al., 2023)
NAVER	NAVER Labs Europe, France (Gow-Smith et al., 2023)
NIUTRANS	NiuTrans, China (Han et al., 2023)
NPU-MSXF	Northwestern Polytechnical U., Nanjing U., MaShang Co., China (Song et al., 2023)
NEURODUB	NeuroDub, Armenia
NEMO	NVIDIA NeMo, USA(Hrinchuk et al., 2023)
ON-TRAC	ON-TRAC Consortium, France (Laurent et al., 2023)
QUESPA	Northeastern U, USA, U. de Pompeu Fabra, Spain, CMU, USA(Ortega et al., 2023)
UPC	Universitat Politècnica de Catalunya, Spain (Tsiamas et al., 2023)
SRI-B	Samsung R&D Institute Bangalore, India (Radhakrishnan et al., 2023)
UCSC	U. of California, Santa Cruz, USA (Vakharia et al., 2023)
UM-DFKI	U. of Malta, Malta, and DFKI, Germany (Williams et al., 2023)
USTC	U. of Science and Technology of China (Deng et al., 2023; Zhou et al., 2023)
XIAOMI	Xiaomi AI Lab, China (Huang et al., 2023)

Table 1: List of Participants

English into Arabic, Chinese, Dutch, French, German, Japanese, Farsi, Portuguese, Russian, and Turkish.

- **Speech-to-speech translation**, focusing on natural-speech to synthetic-speech translation of recorded utterances from English to Chinese.

- **Automatic Dubbing**, focusing on dubbing of short video clips from German to English.

- **Dialect SLT**, focusing on speech translation of recorded utterances from Tunisian Arabic to English.

- **Low-resource SLT**, focusing on speech translation of recorded utterances from Irish to English, Marathi to Hindi, Maltese to English, Pashto to French, Tamasheq to French, and Quechua to Spanish.

- **Formality Control for SLT**, focusing on formality/register control for spoken language translation from English to Korean, Vietnamese, EU Portuguese, and Russian.

The shared tasks attracted 38 submissions by 31 teams (see Table 1) representing both academic and industrial organizations. The following sections report on each shared task in detail, in particular: the goal and automatic metrics adopted for the task, the data used for training and testing data, the received submissions and the summary of results. Detailed results for some of the shared tasks are reported in a corresponding appendix.

2 Offline SLT

Offline speech translation is the task of translating audio speech in one language into text in a different target language, without any specific time or structural constraints (as, for instance, in the simultaneous, subtitling, and dubbing tasks). Under this general problem definition, the goal of

the offline ST track (one of the speech tasks with the longest tradition at the IWSLT campaign) is to constantly challenge a technology in rapid evolution by gradually introducing novelty aspects that raise the difficulty bar.

2.1 Challenge

In continuity with last year, participants were given three sub-tasks corresponding to three language directions, namely English→German/Japanese/Chinese. Participation was allowed both with *cascade* architectures combining automatic speech recognition (ASR) and machine translation (MT) systems as core components, or by means of *end-to-end* approaches that directly translate the input speech without intermediate symbolic representations. Also this year, one of the main objectives was indeed to measure the performance difference between the two paradigms, a gap that recent research (Bentivogli et al., 2021) and IWSLT findings (Ansari et al., 2020; Anastasopoulos et al., 2021, 2022b) indicate as gradually decreasing.

The other main objective of this round was to assess the ability of SLT technology to deal with complex scenarios involving different types of input characterized by phenomena like spontaneous speech, noisy audio conditions and overlapping speakers. In light of this, the main novelty of the 2022 offline SLT task lies in a richer variety of speech data to be processed. To this aim, in addition to the classic TED talks test set, two novel test sets were released:

- **ACL presentations**, in which a single speaker is presenting on a stage. Although similar to the TED talks scenario, additional challenges posed by this test set include the presence of non-native speakers, different accents, variable recording quality, terminology, and controlled interactions with a second speaker.
- **Press conferences and interviews**, in which two persons interact on different topics. Inherent challenges, therefore, include the presence of spontaneous speech, non-native speakers, different accents, and controlled interaction with a second speaker.

All the test sets were used for evaluation in the English-German sub-task, while only TED Talks and ACL presentations were used to test the submissions to the English-Japanese and English-Chinese sub-tasks.

2.2 Data and Metrics

Training and development data. Participants were offered the possibility to submit systems built under three training data conditions:

1. **Constrained**: the allowed training data is limited to a medium-sized framework in order to keep the training time and resource requirements manageable. The complete list[1] of allowed training resources (speech, speech-to-text-parallel, text-parallel, text-monolingual) does not include any pre-trained language model.
2. **Constrained with large language models** (constrained^{+LLM}): in addition to all the constrained resources, a restricted selection[1] of large language models is allowed to give participants the possibility to leverage large language models and medium-sized resources.
3. **Unconstrained**: any resource, pre-trained language models included, can be used with the exception of evaluation sets. This setup is proposed to allow the participation of teams equipped with high computational power and effective in-house solutions built on additional resources.

The development data allowed under the constrained condition consist of the dev set from IWSLT 2010, as well as the test sets used for the 2010, 2013-2015 and 2018-2020 IWSLT campaigns. Besides this TED-derived material, additional development data were released to cover the two new scenarios included in this round of evaluation. For the ACL domain, 5 presentations from the ACL 2022 conference with translations and transcriptions were provided. Due to additional constraints, these references were generated by human post-editing of automatic transcriptions and translation. For the press conferences and interviews domain, 12 videos (total duration: 1h:3m) were selected from publicly available interviews from the Multimedia Centre of the European Parliament (EPTV)[2].

[1] See the IWSLT 2023 offline track web page: https://iwslt.org/2023/offline

[2] https://multimedia.europarl.europa.eu

Test data. Three new test sets were created for the three language directions. The new test sets include heterogeneous material drawn from each scenario. For the traditional TED scenario, a new set of 42 talks not included in the current public release of MuST-C was selected to build the en-de test set.[3] Starting from this material, the talks for which Japanese and Chinese translations are available were selected to build the en-zh and en-ja test sets (respectively, 38 and 37 talks). Similar to the 2021 and 2022 editions, we consider two different types of target-language references, namely:

- The original TED translations. Since these references come in the form of subtitles, they are subject to compression and omissions to adhere to the TED subtitling guidelines.[4] This makes them less literal compared to standard, unconstrained translations;
- Unconstrained translations. These references were created from scratch[5] by adhering to the usual translation guidelines. They are hence exact translations (i.e. literal and with proper punctuation).

For the ACL presentation scenario, paper presentations from ACL 2022 were transcribed and translated into the target languages. A detailed description of the data set can be found in Salesky et al. (2023). There are 5 presentations in each of the dev and test sets with a total duration 1h per split. Talks were selected to include diverse paper topics and speaker backgrounds. This test set is shared with the Multilingual task (§5).

For the press conferences and interviews scenario, the test set comprises 10 EPTV videos of variable duration (6m on average), amounting to a total of 1h:1m. The details of the new test sets are reported in Table 2.

Metrics. Systems were evaluated with respect to their capability to produce translations similar to the target-language references. The similarity was measured in terms of BLEU and COMET (Rei et al., 2020a) metrics. The submitted runs were ranked based on the BLEU calculated on the concatenation of the three test sets by using automatic resegmentation[6] of the hypotheses based on the reference translations. For the BLEU computed on the concatenation of the three test sets, the new unconstrained ones have been used for the TED data. As observed on IWSLT 2022 manual evaluation of simultaneous speech-to-text translation (Macháček et al., 2023), COMET is correlating with human judgments best and BLEU correlation is also satisfactory. Moreover, to meet the requests of last year's participants, a human evaluation was performed on the best-performing submission of each participant.

	Talks / Videos	Duration
English-German		
TED	42	3h:47m:53s
ACL	5	59m:22s
EPTV	10	1h:1m
English-Chinese		
TED	37	3h:2m:22s
ACL	5	59m:22s
English-Japanese		
TED	38	3h:19m:34s
ACL	5	59m:22s

Table 2: Statistics of the official test sets for the IWSLT 2023 offline speech translation task.

2.3 Submissions

This year, 10 teams participated in the offline task, submitting a total of 37 runs. Table 3 provides a breakdown of the participation in each sub-task showing, for each training data condition, the number of participants, the number of submitted runs and, for each training data condition (constrained, constrained^{+LLM}, unconstrained), the number of submitted runs obtained with cascade and direct systems.

- BIGAI (Xie, 2023) participated both with cascade and direct models for en-de, en-ja, and en-zh translations, which were trained under the constrained^{+LLM} condition. The cascade is the concatenation of an ASR model and an MT system. The ASR consists of the first 12 Transformer layers

[3] This set of 42 TED talks is also referred to as the "Common" test set (not to be confused with MuST-C "tst-COMMON") because it serves in both Offline and Simultaneous https://iwslt.org/2023/simultaneous tasks.

[4] http://www.ted.com/participate/translate/subtitling-tips

[5] We would like to thank Meta for providing us with this new set of references.

[6] Performed with mwerSegmenter - https://www-i6.informatik.rwth-aachen.de/web/Software/mwerSegmenter.tar.gz

English-German											
Participants	**Runs**	**Constrained**			**Constrained**$^{+LLM}$			**Unconstrained**			
6	16	2	Cascade	1	12	Cascade	1	2	Cascade	2	
			Direct	1		Direct	11		Direct	-	
English-Chinese											
Participants	**Runs**	**Constrained**			**Constrained**$^{+LLM}$			**Unconstrained**			
7	16	5	Cascade	3	3	Cascade	1	8	Cascade	7	
			Direct	2		Direct	2		Direct	1	
English-Japanese											
Participants	**Runs**	**Constrained**			**Constrained**$^{+LLM}$			**Unconstrained**			
3	5	2	Cascade	1	2	Cascade	1	1	Cascade	1	
			Direct	1		Direct	1		Direct	-	

Table 3: Breakdown of the participation in each sub-task (English→German, English→Chinese, English→Japanese) of the IWSLT offline ST track. For each language direction, we report the number of participants, the number of submitted runs and, for each training data condition (constrained, constrained^{+LLM}, unconstrained), the number of submitted runs obtained with cascade and direct systems.

from wav2vec2-large-960h-lv60-self and an adapter model to compress the feature vectors. Transcripts are obtained through a CTC greedy decoding step. The MT is based on mbart-large-50-one-to-many-mmt. The direct model consists of two separate encoders for speech and text, followed by a shared decoder. The speech and text encoders are respectively based on the cascade ASR and MT encoders. An adapter model is introduced to connect the two encoders. The direct model combines the cross entropy loss for MT and the CTC loss for ASR, together with a hyperparameter to balance the weights between the two losses. The training procedure involves dedicated fine-tuning steps, data filtering and audio re-segmentation into shorter segments.

- I2R (Huzaifah et al., 2023) participated with a direct approach for en-de translation, which was trained under the constrained^{+LLM} condition. The model consists of two separate encoders for speech and text, followed by a shared encoder and a decoder. The speech encoder is initialised with WavLM large, while DeltaLM base is used to initialise the text encoder, the shared encoder and the decoder. To leverage both text and speech sources, the shared encoder is induced to learn a joint multimodal representation obtained through forced alignment of speech and text data. The resulting mixed speech-text representation is passed to the shared encoder initially pre-trained on text data only. A DeltaLM-based MT model incrementally trained on in-domain and out-of-domain data is used as a teacher during fine-tuning of the ST system. The ST model is built on a mix of ASR, ST and synthetic data. Additional techniques applied include on-the-fly audio augmentation to increase robustness to variable audio quality, domain tagging to condition the ST output to the different output styles of the test data, and ST model ensembling.

- HW-TSC (Li et al., 2023) participated with cascade systems for all language directions and in all three training data conditions. The ASR model used for the constrained training condition is the Conformer. For the constrained^{+LLM} condition, the encoder of wav2vec2 and the decoder of mBART50 are combined to fine-tune on all data an ASR model trained on MuST-C. Whisper (Radford et al., 2022), fine-tuned on MuST-C, is instead used for the unconstrained training condition. All models are built using audio inputs augmented with SpecAugment and CTC. The MT component is a Transformer-based model trained in a one-to-many multilingual fashion. It exploits data filtering and data augmentation techniques, combined with dropout regularization and domain adaptation methods, as well as solutions

to increase robustness to ASR noise (through synthetic noise generation and data augmentation).

- MINETRANS (Du et al., 2023) participated with en-zh cascade systems trained under constrained and unconstrained conditions. The submitted runs are obtained with a pipeline of ASR, punctuation recognition, and MT components. The ASR is an RNN-Transducer. For the unconstrained condition, GigaSpeech is added to the training data allowed in the constrained setting. In both conditions, pre-processing and filtering techniques are applied to improve data quality, while SpecAugment is used for data augmentation. Before being passed to the MT component, the unpunctuated ASR output is processed by means of a BERT-based punctuation recognition model. For the MT component, two strategies are implemented. The first one relies on different Transformer-based models for supervised training. A base Transformer and an M2M_100 model are used for the constrained condition. A translation model trained on additional in-house corpora is used for the unconstrained condition. The second strategy adopted for the MT component relies on a large language model (Chat-GPT) for prompt-guided translation.

- NIUTRANS (Han et al., 2023) participated with a direct en-zh system trained under the constrained condition. It consists of two separate encoders for speech and text with an adapter in between, followed by a decoder. The speech encoder is pre-trained with an ASR encoder, while the textual encoder and the decoder with pre-trained MT components. Different architectures with variable size were tested both for ASR (enhanced with CTC loss and inter-CTC loss to speed up convergence) and MT (used to generate pseudo-references so as to increase the size of the SLT data). The final system is an ensemble aiming at maximizing the diversity between models.

- NEURODUB[7] participated with a cascade en-de system trained under the unconstrained condition. It consists of a 4-staged process including the ASR, the punctuation module performing both sentence extraction and punctuation placement, the speaker- and gender distinction component, and the translation model. Every stage is trained on the crawled data from the web.

- NEMO (Hrinchuk et al., 2023) participated with direct systems for all language directions in the constrained training data condition. Pre-trained models and synthetic training data are exploited in different ways to cope with the scarcity of direct ST data. A Conformer-based ASR model trained on all allowed speech-to-text data is used to initialize the SLT encoder. A Transformer-based NMT model trained on all allowed parallel data and fine-tuned on TED talks is used to generate synthetic translation alternatives for all available speech-to-text and text-to-text data. A TTS model based on Fast Pitch (Łańcucki, 2021) and trained on the English transcripts of all TED-derived data is used to generate the synthetic speech version of English texts in the available text corpora. The submitted SLT systems are based on a Conformer-based encoder followed by a Transformer decoder trained on this mix of (gold and synthetic) speech-to-text and text-to-text data.

- XIAOMI (Huang et al., 2023) participated with a direct en-zh system trained under the constrained^{+LLM} condition. It consists of a speech encoder, a text encoder, and a text decoder, with all parameters initialized using the pre-trained HuBERT and mBART models. The speech encoder is composed of a feature extractor based on convolutional neural networks and a Transformer encoder. In addition to the cross-entropy loss, ASR, MT, and a contrastive loss, which tries to learn an encoder that produces similar representations for similar instances independently from the modalities, are added. Self-training is also used to leverage unlabelled data. In addition to the allowed datasets, a large set of pseudo references are generated translating the

[7] Unofficial participant, as no system paper is available.

transcripts of the ASR corpora. During training, a second fine-tuning is performed on MuST-C as in-domain data. The final system is an ensemble of the two best-performing models.

- UPC (Tsiamas et al., 2023) participated with a direct en-de system trained under the constrained^{+LLM} condition. It consists of a speech encoder, a textual encoder, and a text decoder. The speech encoder includes a semantic encoder to align speech and text encoder representations. The coupling modules include the CTC and Optimal Transport (OT) losses to the outputs of the acoustic and semantic encoders, and the addition of a second auxiliary OT loss for the inputs of the semantic encoder. The speech encoder is based on wav2vec 2.0, while the textual encoder uses mBART50. Knowledge distillation is used to generate additional data to fine-tune part of the SLT model architecture (the feature extractor, the acoustic encoder, and the CTC module are frozen during fine-tuning).

  USTC (Zhou et al., 2023) participated with cascade and direct en-zh models trained under the unconstrained condition. For the ASR of the cascade, two approaches are implemented. The first one exploits a fusion models trained on the allowed data expanded with speed perturbation, oversampling, concatenation of adjacent voices and synthetic data generation via TTS. The second approach is based on Whisper large (Radford et al., 2022) and SHAS for audio segmentation. The MT component of the cascade system exploits an ensemble of Transformer-based models enhanced with knowledge distillation, domain adaptation and robust training strategies. For direct SLT, two approaches are implemented. The first one is an encoder-decoder initialized with the ASR and MT models of the cascade. The second approach is a Stacked Acoustic-and-Textual Encoding extension of SATE (Xu et al., 2021). The final submissions also include ensembles obtained by combining cascade and direct systems.

2.4 Results

Also this year, the submissions to the IWSLT Offline translation task were evaluated both with automatic metrics and through human evaluation. The results for each sub-task are shown in detail in the Appendix.

2.4.1 Automatic Evaluation

The results for each of the language pairs are shown in the tables in Appendix B.1. We present results for English-German (Table 14), English-Chinese (Table 16) and English-Japanese (Table 15). The evaluation was carried out in terms of BLEU (the primary metric, in continuity with previous years), and COMET. We report individual scores for the three (or two, as in the case of en-ja and en-zh) different test sets as well as metrics calculated on the concatenation of the different test sets. For each sub-task, systems are ranked based on the BLEU score computed on the concatenated test sets.

End-to-End vs Cascaded This year the cascaded systems performed in general better than the end-to-end systems. For English-to-German, for nearly all metrics, the cascaded systems are always ranked best. For English-to-Japanese, the results show a similar situation to English-to-German, with the cascade systems outperforming the end-to-end model. The supremacy of the cascade models is confirmed by all the metrics, with a clear gap in performance between the worst cascade and the best end-to-end models. For English-to-Chinese, the picture is not as clear. However, the only participant who submitted a primary system using the cascaded and one using the end-to-end paradigm (USTC), the cascaded performed better in all metrics.

Metrics For English-to-German, in general, the results of the BLEU metric correlate quite well with the scores of the COMET metric. Except for relatively small changes, e.g. the order is different for the different HW-TSC systems. One exception is the submissions by UPC and NeMo that are ranked differently in the two metrics. Therefore, a comparison to the human evaluation will be interesting. In the English-to-Japanese task, the scores of the HW-TSC systems are very close to each other and some swaps are visible between BLEU and COMET. However, the changes are only related to the HW-TSC systems and do not mod-

ify the overall evaluation of the systems. In the English-to-Chinese task, there are two situations where the metrics differ significantly. The ranking for USTC end-to-end compared to the HW-TSC systems is different with respect to COMET, which rewards the HW-TSC submissions. A similar situation is visible for NiuTrans and Xiaomi, where BLEU favors the NiuTrans translations, while COMET assigns higher scores, and ranking, to the Xiaomi submissions.

Data conditions For the different data conditions, the gains by using additional large language models or additional data are not clear. HW-TSC submitted three primary systems for each data condition and they all perform very similarly. However, for en-zh the unconstrained system by USTC was clearly the best and for en-de the best system except HW-TSC was also an unconstrained one. The additional benefit of the pre-trained models is even less clear. There is no clear picture that the systems with or without this technology perform better.

Domains One new aspect this year is the evaluation of the systems on three different test sets and domains. First of all, the absolute performance on the different domains is quite different. The systems perform clearly worse on the EPTV test sets. For the relationship between ACL and TED, the picture is not as clear. While the BLEU scores on ACL are higher, the COMET scores are lower. Only for English-to-Japanese, both metrics are higher on the ACL test set. One explanation could be that the references for the ACL talks are generated by post-editing an MT output. This could indicate that the post-edited references inflate the BLEU score, while the COMET score seems to be more robust to this phenomenon. When comparing the different systems, the tendency is for all cases the same. However, some perform slightly better in one condition. For example, the end-to-end system from USTC performs very well on TED compared to other systems but less well on ACL.

2.4.2 Human Evaluation

At the time of writing, human evaluation is still in progress. Its results will be reported at the conference and they will appear in the updated version of this paper in Appendix A.

3 Simultaneous SLT

Simultaneous speech translation means the system starts translating before the speaker finishes the sentence. The task is essential to enable people to communicate seamlessly across different backgrounds, in low-latency scenarios such as translation in international conferences or travel.

This year, the task included two tracks: speech-to-text and speech-to-speech, covering three language directions: English to German, Chinese and Japanese.

3.1 Challenge

There are two major updates compared with previous years:

- Removal of the text-to-text track. The task focuses on the real-world live-translation setting, where the speech is the input medium.
- Addition of a speech-to-speech track. Translation into synthetic speech has gained increasing attention within the research community, given its potential application to real-time conversations.

To simplify the shared task, a single latency constraint is introduced for each track: 2 seconds of Average Lagging for speech-to-text, and 2.5 seconds of starting offset for speech-to-speech. The participants can submit no more than one system per track / language direction, as long as the latency of the system is under the constraint. The latency of the system is qualified on the open MuST-C tst-COMMON test set (Di Gangi et al., 2019a).

The participants made submissions in a format of docker images, which were later run by organizers on the blind-test set in a controllable environment. An example of implementation was provided with the SimulEval toolkit (Ma et al., 2020a).

3.2 Data

The training data condition of the simultaneous task follows "constrained with large language models" setting in the Offline translation task, as described in Section 2.2

The test data has two parts:

Common TED talks. It's the the same as in the Offline task, as described in Section 2.2 .For English to German, Chinese and Japanese

Non-Native see Appendix A.1.1. For English to German.

3.3 Evaluation

Two attributes are evaluated in the simultaneous task: quality and latency.

For quality, we conducted both automatic and human evaluation. BLEU score (Papineni et al., 2002a) is used for automatic quality evaluation. For speech output, the BLEU score is computed on the transcripts from Whisper (Radford et al., 2022) ASR model. The ranking of the submission is based on the BLEU score on the Common blind test set. Furthermore, we conducted BLASER (Chen et al., 2022) evaluation on the speech output. We also conducted human evaluation on speech-to-text translation quality, including general human evaluation for all three language pairs, and task specific human evaluation on German and Japanese outputs.

For latency, we only conducted automatic evaluation. We report the following metrics for each speech-to-text systems.

- Average Lagging (AL; Ma et al., 2019, 2020b)
- Length Adaptive Average Lagging (LAAL; Polák et al., 2022; Papi et al., 2022)
- Average Token Delay (ATD; Kano et al., 2023)
- Average Proportion (AP; Cho and Esipova, 2016)
- Differentiable Average Lagging (DAL; Cherry and Foster, 2019)

We also measured the computation aware version of the latency metrics, as described by Ma et al. (2020b). However, due to the new synchronized SimulEval agent pipeline design, the actual computation aware latency can be smaller with carefully designed parallelism.

For speech-to-speech systems, we report start-offset and end-offset. The latency metrics will not be used for ranking.

3.4 Submissions

The simultaneous shared task received submissions from six teams, whereas all the teams participated in at least one language direction in speech-to-text translation. Among the teams, five teams entered the English-to-German track; four teams entered the English-to-Chinese track; three teams entered the English-to-Japanese track. Even though this year is our first time introducing the simultaneous speech-to-speech track, four teams out of six, submitted speech-to-speech systems.

- CMU(Yan et al., 2023) participated in both the speech-to-text and speech-to-speech tracks for English-German translation. Their speech-to-text model combined self-supervised speech representations, a Conformer encoder, and an mBART decoder. In addition to the cross-entropy attentional loss, the translation model was also trained with CTC objectives. They used machine translation pseudo labeling for data augmentation. Simultaneous decoding was achieved by chunking the speech signals and employing incremental beam search. For their speech-to-speech system, they incorporated a VITS-based text-to-speech model, which was trained separately.
- HW-TSC (Guo et al., 2023; Shang et al., 2023) participated in both the speech-to-text and speech-to-speech tracks for all three language directions. Their model was a cascaded system that combined an U2 ASR, a Transformer-based machine translation model, and a VITS-based text-to-speech model for speech-to-speech translation. The MT model was multilingual and offered translation in all three directions by conditioning on language embeddings. For data augmentation, they adopted data diversification and forward translation techniques. Their simultaneous decoding policy employed chunk-based incremental decoding with stable hypotheses detection. They also utilized additional TTS models for the speech-to-speech track.
- NAIST(Fukuda et al., 2023) participated in the speech-to-text translation direction for all three language directions and English-to-Japanese speech-to-speech translation. Their system consisted of a HuBERT encoder and an mBART decoder. They employed three techniques to improve translation quality: inter-connection to combine pre-trained representations, prefix alignment fine-tuning for simultaneous decoding, and local agreement

to find stable prefix hypotheses. They also utilized an additional Tacotron2-based TTS model for speech-to-speech translation with the wait-k decoding policy.

- FBK(Papi et al., 2023) participated in the English-to-German speech-to-text translation track, using an end-to-end Conformer-based speech-to-text model. Considering computational latency, their focus was on efficient usage of offline models. They employed three simultaneous policies, including local agreement, encoder-decoder attention, and EDATT v2, to achieve this.

- CUNI-KIT(Polák et al., 2023) participated in the English-to-German speech-to-text translation track. Their system utilized WavLM and mBART as the base framework. The key highlights of their system were in the decoding strategy and simultaneous policies. They applied empirical hypotheses filtering during decoding and adopted CTC to detect the completion of block inference.

- XIAOMI(Huang et al., 2023) participated in both the speech-to-text and speech-to-speech tracks for English-Chinese translation. Their end-to-end system utilized HuBERT and mBART with a wait-k decoding strategy and an Information-Transport-based architecture. They further enhanced their system by applying data filtering on long sentences and misaligned audio/text, data augmentation with pseudo labeling, and punctuation normalization. They also incorporated contrastive learning objectives.

3.5 Automatic Evaluation

We rank the system performance based on BLEU scores. The detailed results can be found in Appendix B.2.

3.5.1 Speech-to-Text

English-German On the Common test set, the ranking is HW-TSC, CUNI-KIT, FBK, NAIST, CMU, as shown in Table 17. Meanwhile, on the Non-Native test set, the ranking differs considerably. While HW-TSC performs best on Common test set, they end up second to last on Non-Native. The situation is reversed for NAIST and CMU who end up at the tail of Common scoring but reach the best scores on the Non-Native set. We attribute this to better robustness of NAIST and CMU towards the noise in Non-Native test set.

English-Chinese The ranking is HW-TSC, CUNI-KIT, XIAOMI, NAIST, as shown in Table 18.

English-Japanese The ranking is HW-TSC, CUNI-KIT, NAIST, as shown in Table 19.

3.5.2 Speech-to-Speech

Despite the great novelty and difficulty of speech-to-speech track, there are 5 submissions in total: 2 in German, 2 in Chinese and 1 in Japanese. The full results can be seen in table Table 20. For English-to-German, the ranking is CMU, HW-TSC. For English-to-Chinese, HW-TSC is the only participant. For English-to-Japanese, the ranking is HW-TSC, NAIST.

We also provide the BLASER scores, which directly predict the quality of translations based on speech embeddings. We note that since reference audios are not available in our datasets, we use text LASER (Heffernan et al., 2022) to embed reference text to compute the scores. While the BLASER scores indicate the same quality ranking for English to German as BLEU scores, on the Japanese output they are similar. It's possible that BLASER is adequately developed on Japanese outputs

3.6 Human Evaluation

In the Simultaneous task, speech-to-text track, English-German and English-Japanese were manually evaluated, each with a different scoring method.

3.6.1 English-German

For English-to-German, we used the same human evaluation method as last year, originally inspired by Javorský et al. (2022). We evaluated (1) the best system selected by BLEU score, and (2) transcription of human interpretation, the same as used in last year evaluation (more details can be found in Anastasopoulos et al. (2022a), Section 2.6.1).

Figure 1 plots automatic and manual evaluation in relation with each other. We confirm the generally good correlation with BLEU (Pearson .952 across the two test set parts), as observed by Macháček et al. (2023), although individual system results are rather interesting this year.

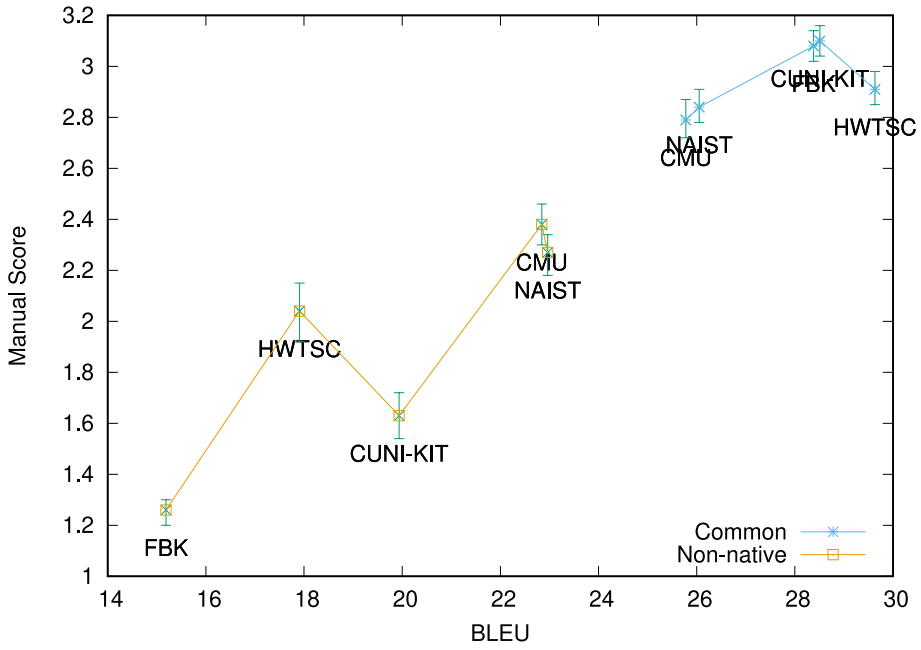

Figure 1: Manual and automatic evaluation of Simulatenous speech-to-text English-to-German translation on the Common (TED talks) and Non-Native test sets. The error bars were obtained by bootstrap resampling, see the caption of Table 22.

On the Common test set, HWTSC performed best in terms of BLEU but the manual scoring seems to prefer CUNI-KIT and FBK. CMU and NAIST are worst in BLEU but on par with HWTSC in terms of manual scores.

The situation is very different on the Non-Native test set: CMU and NAIST score best both in manual scores and in BLEU while CUNI-KIT and esp. FBK get much worse scores, again, both manual and automatic.

The Non-Native test set is substantially harder with respect to sound conditions, and the striking difference drop observed for both CUNI-KIT and FBK can be an indication of some form of overfitting towards the clean input of Common (TED talks).

Appendix A.1.1 presents details of the human evaluation and results are shown in Table 22.

3.6.2 English-Japanese

For English-to-Japanese, we also followed the methodology in the last year. We hired a professional interpreter for human evaluation using JTF Translation Quality Evaluation Guidelines (JTF, 2018) based on Multidimensional Quality Metrics (MQM; Lommel et al., 2014). We applied the error weighting by Freitag et al. (2021a). Appendix A.1.2 presents details of the human evaluation.

The human evaluation results are shown in Table 23. The error score almost correlates with BLEU against the additional reference, but the difference in the error scores was very small between HW-TSC and CUNI-KIT in spite of the 0.8 BLEU difference.

3.7 Final remarks

This year, we simplified the conditions by focusing solely on low-latency systems to reduce the burden of submission and evaluation. We also introduced the novel and challenging speech-to-speech track, and were happy to receive 5 submissions.

We note potential modifications for future editions:

- Providing further simplified submission format.
- Ranking with better designed metrics to address the overfitting towards BLEU scores.
- Aligning more with offline tasks on more test domains and evaluation metrics.

4 Automatic Subtitling

In recent years, the task of automatically creating subtitles for audiovisual content in another language has gained a lot of attention, as we have

seen a surge in the amount of movies, series and user-generated videos which are being streamed and distributed all over the world.

For the first time, this year IWSLT proposed a specific track on automatic subtitling, where participants were asked to generate subtitles of audio-visual documents, belonging to different domains with increasing levels of complexity.

4.1 Challenge

The task of automatic subtitling is multi-faceted: starting from speech, not only the translation has to be generated, but it must be segmented into subtitles compliant with constraints that ensure high-quality user experience, like a proper reading speed, synchrony with the voices, the maximum number of subtitle lines and characters per line, etc. Most audio-visual companies define their own subtitling guidelines, which can differ slightly from each other. Participants were asked to generate subtitles according to some of the tips listed by TED, in particular:

- the maximum subtitle reading speed is 21 characters / second;
- lines cannot exceed 42 characters, white spaces included;
- never use more than two lines per subtitle.

It was expected that participants used only the audio track from the provided videos (dev and test sets), the video track being of low quality and provided primarily as a means to verify time synchronicity and other aspects of displaying subtitles on screen.

The subtitling track requires to automatically subtitle in German and/or Spanish audio-visual documents where the spoken language is always English, and which were collected from the following sources:

- TED talks from the MuST-Cinema[8] corpus;
- press interviews from the Multimedia Centre of the European Parliament (EPTV)[9];
- physical training videos offered by Peloton[10]
- TV series from ITV Studios.[11]

[8]https://ict.fbk.eu/must-cinema
[9]https://multimedia.europarl.europa.eu
[10]https://www.onepeloton.com
[11]https://www.itvstudios.com

domain	set	AV docs	hh:mm	ref subtitles de	ref subtitles es
TED	dev	17	04:11	4906	4964
	test	14	01:22	1375	1422
EPTV	dev	12	01:03	960	909
	test	10	01:01	891	874
Peloton	dev	9	03:59	4508	4037
	test	8	02:43	2700	2661
ITV	dev	7	06:01	4489	4763
	test	7	05:08	4807	4897

Table 4: Statistics of the dev and test sets for the subtitling task.

4.2 Data and Metrics

Data. This track proposed two **training** conditions to participants: **constrained**, in which only a pre-defined list of resources is allowed, and **unconstrained**, without any data restrictions. The constrained setup allowed to use the same training data as in the Offline Speech Translation task (see Section 2.2 for the detailed list), with the obvious exclusion of the parallel resources not involving the English-{German, Spanish} pairs. In addition, two monolingual German and Spanish text corpora built on OpenSubtitles, enriched with subtitle breaks, document meta-info on genre and automatically predicted line breaks, have been released.

For each language and domain, a **development** set and a **test** set were released. Table 4 provides some information about these sets.

The evaluation was carried out from three perspectives, subtitle quality, translation quality and subtitle compliance, through the following automatic measures:

- Subtitle quality vs. reference subtitles:
  - **SubER**, primary metric, used also for ranking (Wilken et al., 2022)[12];
  - **Sigma** (Karakanta et al., 2022b)[13].
- Translation quality vs. reference translations:
  - **BLEU**[14] and **CHRF**[15] via sacreBLEU
  - **BLUERT** (Sellam et al., 2020)

[12]https://github.com/apptek/SubER
[13]https://github.com/fyvo/EvalSubtitle
[14]sacreBLEU signature: nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.0.0
[15]sacreBLEU signature: nrefs:1|case:mixed|eff:yes|nc:6|nw:0|space:no|version:2.0.0

Automatic subtitles are realigned to the reference subtitles using mwerSegmenter (Matusov et al., 2005a)[16] before running sacreBLEU and BLEURT.

- Subtitle compliance:[17]
  - rate of subtitles with reading speed higher than 21 char / sec (**CPS**);
  - rate of lines longer than 42 char (**CPL**);
  - rate of subtitles with more than two lines (white spaces included) (**LPB**).

4.3 Submissions

Three teams submitted automatically generated subtitles for the test sets of this task.

- APPTEK (Bahar et al., 2023) submitted runs in the constrained setup for both language pairs. The primary submissions came from a cascade architecture composed of the following modules: neural encoder-decoder ASR, followed by a neural Machine Translation model trained on the data allowed in the constrained track, with the source (English) side lowercased and normalized to resemble raw ASR output, as well as adapted to the IWSLT subtitling domains, followed by a subtitle line segmentation model (intelligent line segmentation by APPTEK). A contrastive run was generated for the en→de pair only by a direct speech translation system with CTC-based timestamp prediction, followed by the intelligent line segmentation model of APPTEK. The system was trained on the constrained allowed data plus forward translated synthetic data (translations of allowed ASR transcripts) and synthetic speech data for selected sentences from the allowed parallel data. For the en→de pair, APPTEK also submitted a run in the unconstrained setup, where a cascade architecture was employed consisting of: neural encoder-decoder CTC ASR, followed by a neural punctuation prediction model and inverse text normalization model, followed by an MT model adapted to the IWSLT domains (sentences similar in embedding similarity space to the development sets of the four domains TED, EPTV, ITV, Peloton), followed by a subtitle line segmentation model (intelligent line segmentation by APPTEK).

- FBK (Papi et al., 2023) submitted primary runs for the two language pairs, generated by a direct neural speech translation model, trained in the constrained setup, that works as follows: i) the audio is fed to a Subtitle Generator that produces the (un-timed) subtitle blocks; ii) the computed encoder representations are passed to a Source Timestamp Generator to obtain the caption blocks and their corresponding timestamps; iii) the subtitle timestamps are estimated by the Source-to-Target Timestamp Projector from the generated subtitles, captions, and source timestamps.

- MATESUB (Perone, 2023) submitted primary runs for the two language pairs, automatically generated by the back-end subtitling pipeline of MATESUB, its web-based tool that supports professionals in the creation of high-quality subtitles (https://matesub.com/). The MATESUB subtitling pipeline is based on a cascade architecture, composed of ASR, text segmenter and MT neural models, which allows covering any pair from about 60 languages and their variants, including the two language pairs of the task. Since MATESUB is a production software, its neural models are trained on more resources than those allowed for the constrained condition, therefore the submissions fall into the unconstrained setup.

4.4 Results

Scores of all runs as computed by automatic metrics are shown in Tables 24 and 25 in the Appendix. Averaged over the 4 domains, APPTEK achieved the lowest SubER scores with their primary submission for en→de in the constrained and unconstrained condition, with the overall best results for the latter. For en→es, MATESUB obtained the overall lowest SubER with their unconstrained system.

We observe that in terms of domain difficulty, the TV series (from ITV) pose the most challenges for automatic subtitling. This has to do with diverse acoustic conditions in which speech is found in movies and series - background music, noises,

[16] https://www-i6.informatik.rwth-aachen.de/web/Software/mwerSegmenter.tar.gz

[17] https://github.com/hlt-mt/FBK-fairseq/blob/master/examples/speech_to_text/scripts/subtitle_compliance.py

shouts, and cross-talk. All of this makes the task of recognizing speech quite challenging, which results in error accumulation in the downstream components. Unconstrained systems by APPTEK and MATESUB perform significantly better on this domain, which shows the importance of training on additional data that is more representative of real-life content.

The second-hardest domain are the fitness videos from Peloton. Here, despite a generally clear single-speaker audio with reduced background noise, the challenge is the MT: some of the fitness- and sports-specific terminology and slang pose significant challenges in translation to their German and Spanish equivalents.

Surprisingly, even the EPTV interviews pose significant challenges for subtitling, despite the fact that the topics discussed in the interviews are found in abundance in the allowed speech-to-text and text-to-text parallel data for the constrained condition (Europarl, Europarl-ST). Here, the issues such as spontaneous speech with many pauses, as well as speaker separation may have been cause of some of the errors.

The TED talks which have been the main domain for the IWSLT evaluations in the past years are the easiest to be automatically subtitled. Whereas the current level of subtitle quality for TED talks may require minimal human corrections or can even be shown unedited on the screen, for the other three domains the automatic subtitles will require significant post-editing. This shows the importance of running evaluations not only under very controlled conditions as in the case of TED talks, but on a variety of real-life content where multiple research challenges in speech translation are yet to be overcome.

This year's direct speech translation systems seem to be too weak to compete with the cascaded approaches. In particular, a full end-to-end approach like the one from FBK that directly generates subtitle boundaries is currently inferior in comparison with the systems that adopt a specific solution for segmenting the text (intelligent line segmentation by APPTEK and a neural text segmenter by MATESUB). Such specific solutions lead to almost perfect subtitle compliance. But even in terms of pure speech translation quality as measured e.g. with BLEU and BLEURT the cascaded systems currently provide better translations even under constrained training data conditions.

Regarding the automatic metrics used in the evaluation, we observed that the metric Sigma provides scores which are not consistent with the other measures: for example, German subtitles from MATESUB seem to be the worst as measured by Sigma, but this is unlikely based on the values of the other metrics. Yet the pure MT quality metrics also exhibit some discrepancies in how the performance of the same system on the four domains is ranked. This ranking sometimes differs depending on whether you choose BLEU, ChrF, or BLEURT as the "primary" metric. The two most striking cases are:

- the en→de APPTEK unconstrained primary submission, for which the BLEU score for the ITV test data was 14.43 and for Peloton 10.47, but the BLEURT scores were very similar: 0.4069 and 0.4028;
- the en→de FBK constrained primary system, for which the BLEU score was 7.73 on the Peloton part of the test data vs. 8.05 on the ITV part, but the BLEURT scores showed a better quality for Peloton translations: 0.3137 vs. 0.2255.

All of these discrepancies highlight the importance of human evaluation, which we have not conducted this time. One of the reasons for this is that in most prior research (Matusov et al., 2019; Karakanta et al., 2022a) the automatic subtitling quality is evaluated in post-editing scenarios, which are too expensive to be run on significant amounts of data as they require professional subtitle translators. On the other hand, as mentioned above, for 3 out of 4 domains the quality of the automatically generated subtitle translations is low, so that an evaluation of user experience when watching subtitles would be also challenging, especially if the users would have to assign evaluation scores to individual subtitles or sentences. With all of this in mind, we decided to postpone any human evaluation to the next edition of the subtitling track at IWSLT.

Overall, this first edition of the subtitling track emphasised the crucial role of the following components related to speech processing: noise reduction and/or speech separation, speaker diarization, and sentence segmentation. So far they have been underestimated in speech translation research. Current automatic solutions do not reach the level of quality that is necessary in subtitling. Therefore, we encourage further research

into these areas, for which subtitle translation is a good test case.

5 Multilingual SLT

The NLP and speech communities are rapidly expanding with increasing focus on broader language coverage and multilinguality. However, despite the community's efforts on ASR and SLT, research is rarely focused on applying these efforts to the data within the scientific domain. It is clear from recent initiatives to caption technical presentations at NLP and speech conferences that transcription and translation in the technical domain is needed, desired, and remains a disproportionate challenge for current ASR and SLT models compared to standard datasets in these spaces. Motivated by the ACL 60-60 initiative[18] to translate the ACL Anthology to up to 60 languages for the 60th anniversary of ACL, which will be reported on at this year's ACL conference co-located with IWSLT, this year's Multilingual Task evaluates the ability of current models to translate technical presentations to a set of ten diverse target languages.

5.1 Challenge

Translating technical presentations combines several challenging conditions: domain-specific terminology, recording conditions varying from close-range microphones to laptop microphones with light background noise or feedback, diverse speaker demographics, and importantly unsegmented speech typically 10-60 minutes in duration. This task focuses on one-to-many translation from English to ten target languages. Providing English ASR was optional though encouraged. In-domain data is scarce, particularly parallel data, though all language pairs are covered by current publicly available corpora; further challenging for current domain adaptation techniques, monolingual data is typically available for the source language (English) only. We present two conditions: constrained (using only the out-of-domain data allowed and provided for other tasks this year) and unconstrained (allowing any additional data, included crawled, which may facilitate e.g., domain adaptation). To evaluate submissions, we use evaluation sets curated from presentations at ACL 2022 which were professionally transcribed and translated with the support of ACL and the 60-60 initiative as described in Salesky et al. (2023).

[18] https://www.2022.aclweb.org/dispecialinitiative

5.2 Data and Metrics

Data. We use the ACL 60-60 evaluation sets created by Salesky et al. (2023) to evaluate this challenge task. The data comes from ACL 2022 technical presentations and is originally spoken in English, and then transcribed and translated to ten target languages from the 60/60 initiative: Arabic, Mandarin Chinese, Dutch, French, German, Japanese, Farsi, Portuguese, Russian, and Turkish. The resulting dataset contains parallel speech, transcripts, and translation for ten language pairs, totaling approximately one hour for the development set and one hour for the evaluation set.

During the evaluation campaign, the only in-domain data provided is the development set. To simulate the realistic use case where recorded technical presentations would be accompanied by a research paper, in addition to the talk audio we provide the corresponding paper title and abstract, which are likely to contain a subset of relevant keywords and terminology and could be used by participants to bias or adapt their systems. Constrained training data follows the Offline task (see Sec. 2.2) with pretrained models and out-of-domain parallel speech and text provided for all 10 language pairs. The unconstrained setting allowed participants to potentially crawl additional in-domain data to assist with adaptation, as was done by one team (JHU). For the official rankings, we use the official evaluation set, which was held blind until after the evaluation campaign.

To mimic realistic test conditions where the audio for technical presentations would be provided as a single file, rather than gold-sentence-segmented, for both the development and evaluation sets we provided the full unsegmented wav files, as well as an automatically generated baseline segmentation using SHAS (Tsiamas et al., 2022) to get participants started. Two teams used the baseline segmentation, while one (JHU) used longer segments which improved the ASR quality of their particular pretrained model. To evaluate translation quality of system output using any input segmentation, we provided gold sentence-segmented transcripts and translations, which system output could be scored with as described below in 'Metrics.'

Metrics. Translation output was evaluated using multiple metrics for analysis: translation output using chrF (Popović, 2015a), BLEU (Papineni et al., 2002b) as computed by SACREBLEU (Post, 2018), and COMET (Rei et al., 2020b) and ASR output using WER. For BLEU we use the recommended language-specific tokenization in SACREBLEU for Chinese, Japanese, Korean, and the metric-default otherwise. Translation metrics were calculated with case and punctuation. WER was computed on lowercased text with punctuation removed. NFKC normalization was applied on submitted systems and references. All official scores were calculated using automatic resegmentation of the hypothesis based on the reference transcripts (ASR) or translations (SLT) by mwerSegmenter (Matusov et al., 2005b), using character-level segmentation for resegmentation for those languages which do not mark whitespace. The official task ranking is based on average chrF across all 10 translation language pairs.

5.3 Submissions

We received 11 submissions from 3 teams, as described below:

- BIT (Wang et al., 2023b) submitted a single constrained one-to-many multilingual model to cover all 10 language pairs, trained using a collection of multiple versions of the MuST-C dataset (Di Gangi et al., 2019b). They use English ASR pre-training with data augmentation from SpecAugment (Park et al., 2019), and multilingual translation finetuning for all language pairs together. The final model is an ensemble of multiple checkpoints. No adaptation to the technical domain is performed.

- JHU (Xinyuan et al., 2023) submitted two cascaded systems, one constrained and one unconstrained, combining multiple different pretrained speech and translation models, and comparing different domain adaptation techniques. Their unconstrained system uses an adapted Whisper (Radford et al., 2022) ASR model combined with NLLB (NLLB Team et al., 2022), M2M-100 (Fan et al., 2020), or mBART-50 (Tang et al., 2020) MT models depending on the language pair, while the constrained system uses wav2vec2.0 (Baevski et al., 2020a) and mBART-50 or M2M-100. They compare using talk abstracts to prompt Whisper to training in-domain language models on either the small amount of highly-relevant data in the talk abstract or larger LMs trained on significantly more data they scraped from the ACL Anthology and release with their paper. They see slight improvements over the provided SHAS (Tsiamas et al., 2022) segments using longer segments closer what Whisper observed in training. They show that prompting Whisper is *not* competitive with in-domain language models, and provide an analysis of technical term recall and other fine-grained details.

- KIT (Liu et al., 2023) submitted multiple constrained multilingual models, both end-to-end and cascaded, which combine several techniques to adapt to the technical domain given the absence of in-domain training data, using pretrained speech and translation models as initializations (WavLM: Chen et al. 2021, DeltaLM: Ma et al. 2021, mBART-50: Tang et al. 2020). These include kNN-MT to bias generated output to the technical domain; data diversification to enrich provided parallel data; adapters for lightweight finetuning to the language pairs for translation (though they note that this does not necessarily stack with data diversification); and for their cascaded model, adaptation of the ASR model to the target technical domain using n-gram re-weighting, noting that it is typically easier to adapt or add lexical constraints to models with separate LMs, as opposed to encoder-decoder models. Additional techniques (ensembling, updated ASR encoder/decoder settings, knowledge distillation, synthesized speech) are also used for further small improvements.

5.4 Results

All task results are shown in Appendix B.4. The official task ranking was determined by the average chrF across all 10 target languages after resegmentation to the reference translations.Table 26. Scores for all submissions by individual language pairs are shown in Table 28 (chrF), Table 29 (COMET), and Table 30 (BLEU).

Overall, the majority of approaches combined strong pretrained speech and translation models to do very well on the ACL 60-60 evalua-

tion data. For this task, cascaded models performed consistently better than direct/end-to-end approaches; all of the top 6 submissions were cascades, and 4/5 of the lowest-performing systems were direct. Optional English ASR transcripts were submitted for 3 systems (JHU$_{unconstrained}$, KIT$_{primary}$, JHU$_{constrained}$), all of which were cascades; we see that WER aligns with speech translation performance in these cases. The only unconstrained model, from JHU, utilized larger pretrained models and crawled in-domain language modeling data for ASR to great success, and was the top system on all metrics (Table 26). The remaining submissions were all constrained (here meaning, used the white-listed training data and smaller pretrained models). The *KIT$_{primary}$* system was the best performing constrained model. While BIT trained models from scratch on TED to reasonable performance on MuST-C, large pretrained models and domain adaptation were key for high performance on the technical in-domain test set. chrF and BLEU result in the same system rankings, while COMET favors the end-to-end models slightly more, though not affecting the top 3 systems (JHU$_{unconstrained}$, KIT$_{primary}$, KIT$_{constrastive1}$).

Domain adaptation techniques had consistent positive impact on system performance. The KIT team submitted constrained systems only and thus were limited to the dev bitext and talk abstracts for domain adaptation. Despite its small size (<500 sentences) they were able to generate consistent improvements of up to ~1chrF and ~ 1 BLEU using kNN-MT (*primary/contrastive1* vs *contrastive2*); with this method, extending the dev data to include the abstracts for the evaluation set talks (*primary* vs *contrastive1*) had neglible effect on all 3 metrics. The JHU submissions saw that decoding with interpolated in-domain language models outperformed knowledge distillation or prompting pretrained models with information for each talk in this case; small talk-specific LMs did provide slight improvements in WER, but significant improvements of 2-3 WER were gained by extending the limited highly relevant data from talk abstracts and the dev set to the larger domain-general data crawled from the 2021 ACL conference and workshop proceedings.

Without in-domain target-language monolingual data, conventional techniques for adaptation of end-to-end ST models did not apply (finetuning, backtranslation, ...). The data diversification applied by KIT via TTS 'backtranslation' (*contrastive5*, *contrastive7*) did not affect chrF or BLEU, but did provide small (0.5-0.6) improvements on COMET.

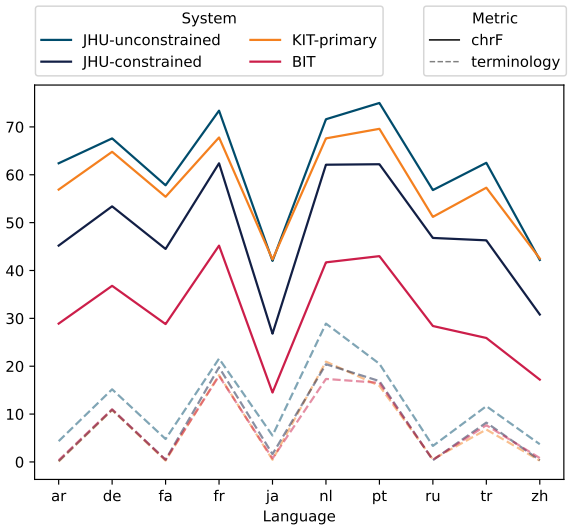

Figure 2: Official task metric performance (chrF) vs terminology recall for teams' primary submissions.

In addition to the overall evaluation set, we look at the recall of specific terminology annotated for the ACL evaluation sets. For the three submissions (JHU$_{unconstrained}$, KIT$_{primary}$, JHU$_{constrained}$) which provided supplementary ASR, we first investigate terminology recall and propagation between ASR and downstream ST. Recall that the overall WER of these systems was 16.9, 23.7, and 34.1, respectively. Of the 1107 labeled terminology words and phrases from the ACL 60-60 evaluation set annotations, 87.8% / 77.3% / 71.7% individual instances were correctly transcribed by these systems, respectively. Of these, 12.0% / 7.4% / 7.9% were then maintained and correctly translated to each target language respectively on average. We plot the official task metric (chrF) against terminology recall in Figure 2 for all primary submissions. We see that there were consistent differences across languages in how terminology was maintained, which generally but not fully corresponds to overall performance (ex: Dutch, Turkish). While the domain adaptation techniques used ensured strong *transcription* performance for the JHU and KIT submissions, this was not generally maintained for *translation* with a significant drop, converging with BIT which did not perform domain adaptation. Additional work is needed to

ensure targeted lexical terms are correctly transcribed and translated, both in general as well as comparably across different languages.

While the JHU submissions finetuned to each target language individually, the KIT systems finetuned multilingually; no contrastive systems were submitted with which to ablate this point, but both teams' papers describe consistently worse performance finetuning multilingually rather than bilingually, which KIT was able to largely mitigate with language adapters in development in isolation but in their final submission on eval language adapters were consistently slightly worse (*contrastive4* 'with' vs *contrastive3* 'without.'). It remains to be seen the degree to which one-to-many models can benefit from multilingual training.

The Offline task additionally used the ACL 60-60 evaluation sets as part of their broader evaluation for 3 language pairs (en→ de, ja, zh), enabling a wider comparison across 25 total systems. We show the Multilingual task submissions compared to the Offline on these languages in Table 27. On these three language pairs, performance is generally higher than the remaining language pairs in the Multilingual task. We again consistently see stronger performance on this task from cascaded models, and unconstrained submissions or those with larger pretrained LLMs, though there are notable outliers such as the HW-TSC constrained model. The Offline submissions did not perform domain adaptation specifically to the technical ACL domain, but appear to be benefit from better domain-general performance in some cases, particularly for submissions targeting only Chinese. We note slight differences in system rankings between metrics (COMET and BLEU) and target languages, particularly for Japanese and Chinese targets, possibly highlighting the difference in metric tokenization for these pairs.

6 Speech-to-Speech Translation

Speech-to-speech translation (S2ST) involves translating audio in one language to audio in another language. In the offline setting, the translation system can assume that the entire input audio is available before beginning the translation process. This differs from streaming or simultaneous settings where the system only has access to partial input. The primary objective of this task is to encourage the advancement of automated methods for offline speech-to-speech translation.

6.1 Challenge

The participants were tasked with creating speech-to-speech translation systems that could translate from English to Chinese using various methods, such as a cascade system (ASR + MT + TTS or end-to-end speech-to-text translation + TTS), or an end-to-end / direct system. They were also allowed to use any techniques to enhance the performance of the system, apart from using unconstrained data.

6.2 Data and Metrics

Data. This task allowed the same training data from the Offline task on English-Chinese speech-to-text translation. More details are available in Sec. 2.2. In addition to the Offline task data, the following training data was allowed to help build English-Chinese speech-to-speech models and Chinese text-to-speech systems:

- **GigaS2S**, target synthetic speech for the Chinese target text of GigaST (Ye et al., 2023) that was generated with an in-house single-speaker TTS system;
- **aishell 3** (Shi et al., 2020), a multi-speaker Chinese TTS dataset.

It's noted that several datasets allowed for the Offline task such as Common Voice (Ardila et al., 2019) actually contain multi-speaker Chinese speech and text data that could help for this task.

Metrics. All systems were evaluated with both automatic and human evaluation metrics.

Automatic metrics. To automatically evaluate translation quality, the speech output was automatically transcribed with a Chinese ASR system[19] (Yao et al., 2021), and then **BLEU**[20] (Papineni et al., 2002a), **chrF**[21] (Popović, 2015b), **COMET**[22] (Rei et al., 2022) and **SEScore2**[23] (Xu et al., 2022) were computed between the generated transcript and the human-produced text reference. BLEU and chrF were computed using SacreBLEU

[19] https://github.com/wenet-e2e/wenet/blob/main/docs/pretrained_models.en.md

[20] sacreBLEU signature: nrefs:1|case:mixed|eff:no|tok:zh|smooth:exp|version:2.3.1

[21] sacreBLEU signature: nrefs:1|case:mixed|eff:yes|nc:6|nw:0|space:no|version:2.3.1

[22] https://huggingface.co/Unbabel/wmt22-comet-da

[23] https://github.com/xu1998hz/SEScore2

(Post, 2018). Furthermore, the output speech could be evaluated directly using BLASER (Chen et al., 2022). More information could be found at `stopes`[24] (Andrews et al., 2022).

Human evaluation. Output speech translations were evaluated with respect to translation quality and speech quality.

- **Translation quality:** Bilingual annotators were presented with the source audio, source transcript and the generated target audio, then gave scores on the translation quality between 1 and 5 (worst-to-best)). There were 4 annotators per sample and we retained the median score.
- **Output speech quality:** In addition to translation quality (capturing meaning), the quality of the speech output was also human-evaluated. The annotators were requested to give an overall score by considering three dimensions: naturalness (voice and pronunciation), clarity of speech (understandability), and sound quality (noise and other artifacts). Each sample was assessed by 4 annotators and scored on a scale of 1-5 (worst-to-best)), with a minimum score interval of 0.5.

The detailed guidelines for output speech quality evaluation were similar to last year (Anastasopoulos et al., 2022a).

6.3 Submissions

We received eight submissions from five teams. The MINETRANS team submitted four systems and each of the other teams submitted one system.

- HW-TSC (Wang et al., 2023a) submitted a cascaded system composed of an ensemble of Conformer and Transformer-based ASR models, a multilingual Transformer-based MT model and a diffusion-based TTS model. Their primary focus in their submission is to investigate the modeling ability of the diffusion model for TTS tasks in high-resource scenarios. The diffusion TTS model takes raw text as input and generates waveform by iteratively denoising on pure Gaussian noise. Based on the result, they conclude that the diffusion model outperforms normal TTS models and brings positive gain to the entire S2ST system.
- KU (Yang et al., 2023) submitted a cascade system composed of a speech-to-text translation (ST) model and a TTS model. Their ST model comprises a ST decoder and an ASR decoder. The two decoders can exchange information with each other with the interactive attention mechanism. For the TTS part, they use FastSpeech2 as the acoustic model and HiFi-GAN as the vocoder.
- NPU-MSXF (Song et al., 2023) submitted a cascaded system of separate ASR, MT, and TTS models. For ASR, they adopt ROVER-based model fusion and data augmentation strategies to improve the recognition accuracy and generalization ability. Then they use a three-stage fine-tuning process to adapt a pre-trained mBART50 model to translate the output of ASR model. The three-stage fine-tuning is based on Curriculum Learning and it involves three sets of data: (1) the original MT data, (2) the MT data in ASR transcription format and (3) the ASR outputs. For TTS, they leverage a two-stage framework, using network bottleneck features as a robust intermediate representation for speaker timbre and linguistic content disentanglement. Based on the two-stage framework, pre-trained speaker embedding is leveraged as a condition to transfer the speaker timbre in the source speech to the translated speech.
- XIAOMI (Huang et al., 2023) submitted a cascade system composed of a speech-to-text translation (ST) model and a TTS model. The ST model is the same as the one they submitted to the Offline SLT track. It is based on an encoder-decoder architecture from the pre-trained HuBERT and mBART models. For the TTS model, they use the Tacotron2 framework. It is first trained with AISHELL-3 dataset and then finetuned with GigaS2S dataset. Furthermore, they implement several popular techniques, such as data filtering, data augmentation, speech segmentation, and model ensemble, to improve the overall performance of the system.
- MINETRANS (Du et al., 2023) submitted three end-to-end S2ST systems (MINE-

[24] https://github.com/facebookresearch/stopes/tree/main/demo/iwslt_blaser_eval

TRANS_E2E, including *primary*, *contrastive1*, and *contrastive2*), and a cascade S2ST system (MINETRANS_Cascade). Their end-to-end systems adopt the speech-to-unit translation (S2UT) framework. The end-to-end S2UT model comprises a speech encoder, a length adapter and an unit decoder. The S2UT model is trained to convert the source speech into units of target speech. A unit-based HiFi-GAN vocoder is finally applied to convert the units into waveform. Based on their results, they conclude that the widely used multi-task learning technique is not important for model convergence once large-scale labeled training data is available, which means that the mapping from source speech to target speech units can be learned directly and easily. Furthermore, they apply other techniques, such as consistency training, data augmentation, speech segmentation, and model ensemble to improve the overall performance of the system. Their cascade system consists of ASR, MT and TTS models. Their ASR and MT replicates those used for the Offline SLT submission. Their TTS model is a combination of FastSpeech2 and HiFi-GAN.

6.4 Results

Results as scored by automatic metrics are shown in Table 31 and human evaluation results are shown in Table 32 in the Appendix.

Overall results. According to the automatic metrics used in the evaluation, XIAOMI obtained the highest score in ASR-BLEU, ASR-chrF, ASR-COMET and ASR-SEScore2. NPU-MSXF obtained the second highest score, followed subsequently by HW-TSC, MINETRANS_E2E, KU and MINETRANS_Cascade. The BLEU, chrF, COMET and SEScore2 rankings were exactly the same. The scores for the test-expanded data were lower than those for the test-primary data, likely due to a domain mismatch with the training data. For human evaluation along the translation quality perspective, XIAOMI obtained the highest score, followed by NPU-MSXF, then HW-TSC and MINETRANS_E2E, then MINETRANS_Cascade, and finally KU. This ranking was mostly consistent with the automatic ranking, showing that automatic metrics were useful in evaluating the translation quality of systems. For human evaluation along the speech quality perspective, NPU-MSXF obtained the highest score, followed by HW-TSC, XIAOMI, MINETRANS_E2E, MINETRANS_Cascade and KU. With a equal weighting of translation quality and speech quality, NPU-MSXF obtained the highest overall score in human evaluation, followed by XIAOMI and the others.

S2ST approaches. This year, all systems but MINETRANS_E2E were cascaded systems, with three systems adopting an ASR + MT + TTS approach and two systems adopting an end-to-end S2T + TTS approach. This showed that cascade approach was still dominant in the community. Although MINETRANS_E2E performed better than MINETRANS_Cascade in all evaluation metrics, we could not draw conclusions on the comparison between cascade and end-to-end given the limited data points. Future challenges can encourage more direct or end-to-end submissions.

6.5 Conclusion

This is the second time that speech-to-speech translation (S2ST) is presented in one of the IWSLT tasks. S2ST is an important benchmark for general AI as other NLP tasks, e.g. dialogue system, question answering and summarization can also be implemented in speech-to-speech manner. Compared to the setting last year, the size of the training data set available to the participants is much larger. The BLEU scores obtained in this challenge is high in general, compared to MT and ST of the same language direction. Although not required by the task, NPU-MSXF is the only team that implemented speaker timbre transfer in their system. We plan to include evaluation metrics addressing this aspect in the next edition.

7 Dialect SLT

The Dialect Speech Translation shared task is a continuation of last year's task. We use the same training data as 2022 and evaluated systems on the 2022 evaluation set to measure progress; in addition, we added a new 2023 evaluation set as blind test. From the organizational perspective, we merged the call for shared task with the the Low-Resource tasks (Section 8) in order to encourage cross-submission of systems.

7.1 Challenge

Diglossic communities are common around the world. For example, Modern Standard Arabic (MSA) is used for formal spoken and written communication in most parts of the Arabic-speaking world, but local dialects such as Egyptian, Moroccan, and Tunisian are used in informal situations. Diglossia poses unique challenges to speech translation because local "low" dialects tend to be low-resource with little ASR and MT training data, and may not even have standardized writing, while resources from "high" dialects like MSA provides opportunities for transfer learning and multilingual modeling.

7.2 Data and Metrics

Participants were provided with the following datasets:

- (a) 160 hours of Tunisian conversational speech (8kHz), with manual transcripts
- (b) 200k lines of manual translations of the above Tunisian transcripts into English, making a three-way parallel data (i.e. aligned audio, transcript, translation) that supports end-to-end speech translation models
- (c) 1200 hours of Modern Standard Arabic (MSA) broadcast news with transcripts for ASR, available from MGB-2
- Approximately 42,000k lines of bitext in MSA-English for MT from OPUS (specifically: Opensubtitles, UN, QED, TED, GlobalVoices, News-Commentary).

In 2022, we constructed three conditions: The basic condition trains on (a) and (b), provided by the Linguistic Data Consortium (LDC); the dialect adaptation condition trains on (a), (b), (c), (d); the unconstrained condition can use any additional data and pre-trained models. In 2023, due to the coordinated organization with other Low-Resource Tasks this year, we renamed basic condition as **"constrained condition"**, and the other two conditions are merged as the **"unconstrained condition"**.

All train and test sets are time-segmented at the utterance level. Statistics are shown in Table 5. There are three test sets for evaluation with BLEU[25].

- **test1**: Participants are encouraged to use this for internal evaluation since references are provided. This is part of LDC2022E01 released to participants for training and development, obtained by applying the standard data split and preprocessing[26].
- **test2**: official evaluation for 2022, from LDC2022E02
- **test3**: official evaluation for 2023, from LDC2023E09

7.3 Submissions

We received submission from four teams:

- GMU (Mbuya and Anastasopoulos, 2023) participated in five language-pairs in the Low-Resource tasks as well as this task. They focused on investigating how different self-supervised speech models (Wav2vec 2.0, XLSR-53, and HuBERT) compare when initialized to an end-to-end (E2E) speech translation architecture.
- JHU (Hussein et al., 2023) submitted both cascaded and E2E systems, using transformer and branchformer architectures. They investigated the incorporation of pretrained text MT models, specifically mBART50 and distilled NLLB-200. Further, they explored different ways for system combination and handling of orthographic variation and channel mismatch.
- ON-TRAC (Laurent et al., 2023) participated in two language-pairs in the Low-Resource task as well as this task. For this task, they focused on using SAMU-XLS-R as the multilingual, multimodal pretrained speech encoder and mBART as the text decoder.
- USTC (Deng et al., 2023) proposed a method for synthesis of pseudo Tunisian-MSA-English paired data. For the cascaded system, they explored ASR with different feature extraction (VGG, GateCNN) and neural architectures (Conformer, Transformer). For E2E, they proposed using SATE and a hybrid SATE architecture to take advantage

[25] SacreBLEU signature for dialect speech translation task: `nrefs:1|case:lc|eff:no|tok:13a|smooth:exp|version:2.0.0`

[26] https://github.com/kevinduh/iwslt22-dialect

Dataset	Speech	Text (#lines)			Use
	(#hours)	Tunisian	MSA	English	
LDC2022E01 train	160	200k	-	200k	Constrained condition
LDC2022E01 dev	3	3833	-	3833	Constrained condition
LDC2022E01 test1	3	4204	-	4204	Participant's internal evaluation
LDC2022E02 test2	3	4288	-	4288	Evaluate progress from 2022
LDC2023E09 test3	3	4248	-	4248	Official evaluation for 2023
MGB2	1100	-	1.1M	-	Unconstrained condition
OPUS	-	-	42M	42M	Unconstrained condition
Any other data	-	-	-	-	Unconstrained condition

Table 5: Datasets for Dialect Shared Task.

of the pseudo Tunisian-MSA-English text data. Additionally, methods for adapting to ASR errors and system combination were examined.

7.4 Results

The full set of BLEU results on the English translations are available in Tables 33 and 34. We also evaluated the WER results for the ASR component of cascaded systems, in Table 35.

In general, there is an improvement compared to 2022. On test2, the best system in 2022 (achieved by the CMU team) obtained 20.8 BLEU; several systems this year improved upon that result, for example USTC's primary system achieved 23.6 BLEU and JHU's primary system achieved 21.2 BLEU. On the official evaluation on test3, the best system achieved 21.1 BLEU in the unconstrained condition and 18.1 BLEU in the constrained condition.

From the system descriptions, it appears the ingredients for strong systems include: (a) effective use of pretrained speech and text models, (b) system combination among both cascaded and E2E systems, and (c) synthetic data generation to increase the size of dialectal data.

We do not plan to continue this shared task next year. Instead, the plan is to make the data available from the LDC. We encourage researchers to continue exploring dialectal and diglossic phenomena in the future.

8 Low-resource SLT

The Low-resource Speech Translation shared task focuses on the problem of developing speech transcription and translation tools for low-resourced languages.

8.1 Challenge

This year, the task introduced speech translation of recorded utterances from Irish to English, Marathi to Hindi, Maltese to English, Pashto to French, Tamasheq to French, and Quechua to Spanish. The different language pairs vary by the amount of data available, but in general, they have in common the dearth of high-quality available resources, at least in comparison to other much higher-resourced settings.

8.2 Data and Metrics

We describe the data available for each language pair below. Table 6 provides an overview of the provided datasets.

Irish–English Irish (also known as Gaeilge) has around 170,000 L1 speakers and 1.85 million people (37% of the population) across the island (of Ireland) claim to be at least somewhat proficient with the language. In the Republic of Ireland, it is the national and first official language. It is also one of the official languages of the European Union (EU) and a recognized minority language in Northern Ireland with the ISO `ga` code.

The provided Irish audio data were compiled from Common Voice (Ardila et al., 2020a),[27] and Living-Audio-Dataset.[28] The compiled data were automatically translated into English and corrected by an Irish linguist. The Irish–English corpus consists of 11.55 hours of Irish speech data (see Table 6), translated into English texts.

Marathi–Hindi Marathi is an Indo-Aryan language which has the ISO code `mr`, and is domi-

[27] https://commonvoice.mozilla.org/en/datasets

[28] https://github.com/Idlak/Living-Audio-Dataset

Language Pairs		Train Set	Dev Set	Test Set	Additional Data
Irish–English	ga–eng	9.46	1.03	0.44	n/a
Marathi–Hindi	mr–hi	15.3	3.7	4.4	monolingual audio with transcriptions (ASR), monolingual text
Maltese–English	mlt–eng	2.5	-	1.35	monolingual audio with transcriptions (ASR), monolingual text
Pashto–French	pus–fra	61	2.5	2	n/a
Tamasheq–French	tmh–fra	17	-	-	untranscribed audio, data in other regional languages
Quechua–Spanish	que–spa	1.60	1.03	1.03	60 hours of monolingual audio with transcriptions (ASR) and MT data (not transcribed)

Table 6: Training, development and test data details (in hours) for the language pairs of the low-resource shared task.

nantly spoken in the state of Maharashtra in India. It is one of the 22 scheduled languages of India and the official language of Maharashtra and Goa. As per the 2011 Census of India, it has around 83 million speakers which covers 6.86% of the country's total population.[29] Marathi is the third most spoken language in India.

The provided Marathi–Hindi corpus consists of 22.33 hours of Marathi speech data (see Table 6) from the news domain, extracted from News On Air[30] and translated into Hindi texts.[31] The dataset was manually segmented and translated by Panlingua.[32] Additionally, the participants were directed that they may use monolingual Marathi audio data (with transcription) from Common Voice (Ardila et al., 2020a),[33] as well as the corpus provided by He et al. (2020)[34] and the Indian Language Corpora (Abraham et al., 2020).[35]

Maltese–English Maltese is a Semitic language, with about half a million native speakers, spoken in the official language of Malta and the EU. It is written in Latin script.

The provided data was divided into three parts. First, around 2.5 hours of audio with Maltese transcription and an English translation were released, along with about 7.5 hours of audio with only Maltese transcriptions. Last, the participants were directed to several monolingual Maltese textual resources. The provided datasets were taken from the MASRI corpus (Hernandez Mena et al., 2020).

Pashto–French Pashto is spoken by approximately forty to sixty million people in the world. It is particularly spoken by the Pashtun people in the south, east and southwest of Afghanistan (it is one of the two official languages), as well as in the north and northwest Pakistan but also in Iran, Tajikistan and India (Uttar Pradesh and Cashmere) and one of the two official languages of Afghanistan.

The corpus was totally provided by ELDA, and is available on the ELRA catalog: *TRAD Pashto Broadcast News Speech Corpus* (ELRA catalogue, 2016b) that consists of audio files and *TRAD Pashto-French Parallel corpus of transcribed Broadcast News Speech - Training data* (ELRA catalogue, 2016a) which are their transcriptions.

This dataset is a collection of about 108 hours of Broadcast News with transcriptions in Pashto and translations into French text. The dataset is built from collected recordings from 5 sources: Ashna TV, Azadi Radio, Deewa Radio, Mashaal Radio and Shamshad TV. Original training data contains 99 hours of speech in Pashto, which corresponds to 29,447 utterances translated into French. Training data corresponds to 61 hours of speech (Table 6).

Tamasheq–French Tamasheq is a variety of Tuareg, a Berber macro-language spoken by nomadic

[29] https://censusindia.gov.in/nada/index.php/catalog/42561

[30] https://newsonair.gov.in

[31] https://github.com/panlingua/iwslt2023_mr-hi

[32] http://panlingua.co.in/

[33] https://commonvoice.mozilla.org/en/datasets

[34] https://www.openslr.org/64/

[35] https://www.cse.iitb.ac.in/~pjyothi/indiccorpora/

tribes across North Africa in Algeria, Mali, Niger and Burkina Faso. It accounts for approximately 500,000 native speakers, being mostly spoken in Mali and Niger. This task is about translating spoken Tamasheq into written French. Almost 20 hours of spoken Tamasheq with French translation are freely provided by the organizers. A major challenge is that no Tamasheq transcription is provided, as Tamasheq is a traditionally oral language.

The provided corpus is a collection of radio recordings from Studio Kalangou[36] translated to French. It comprises 17 hours of clean speech in Tamasheq, translated into the French language. The organizers also provided a 19-hour version of this corpus, including 2 additional hours of data that was labeled by annotators as potentially noisy. Both versions of this dataset share the same validation and test sets. Boito et al. (2022a) provides a thorough description of this dataset.

In addition to the 17 hours of Tamasheq audio data aligned to French translations, and in light of recent work in self-supervised models for speech processing, we also provide participants with unlabeled raw audio data in the Tamasheq language, as well as in other 4 languages spoken from Niger: French (116 hours), Fulfulde (114 hours), Hausa (105 hours), Tamasheq (234 hours) and Zarma (100 hours). All this data comes from the radio broadcastings of Studio Kalangou and Studio Tamani.[37]

Note that this language pair is a continuation of last year's shared task. An additional separate test set was provided this year.

Quechua–Spanish Quechua is an indigenous language spoken by more than 8 million people in South America. It is mainly spoken in Peru, Ecuador, and Bolivia where the official high-resource language is Spanish. It is a highly inflective language based on its suffixes which agglutinate and are found to be similar to other languages like Finnish. The average number of morphemes per word (synthesis) is about two times larger than in English. English typically has around 1.5 morphemes per word and Quechua has about 3 morphemes per word.

There are two main regional divisions of Quechua known as Quechua I and Quechua II. This data set consists of two main types of Quechua spoken in Ayacucho, Peru (Quechua Chanka ISO: `quy`) and Cusco, Peru (Quechua Collao ISO: `quz`) which are both part of Quechua II and, thus, considered a "southern" languages. We label the data set with `que` - the ISO norm for Quechua II mixtures.

The constrained setting allowed a Quechua-Spanish speech translation dataset along with the additional parallel (text-only) data for machine translation compiled from previous work (Ortega et al., 2020). The audio files for training, validation, and test purposes consisted of excerpts of the Siminchik corpus (Cardenas et al., 2018) that were translated by native Quechua speakers. For the unconstrained setting, participants were directed to another larger data set from the Siminchik corpus which consisted of 60 hours of fully transcribed Quechua audio (monolingual).

[36] https://www.studiokalangou.org/
[37] https://www.studiotamani.org/

8.2.1 Metrics

We use standard lowercase BLEU as well as charF++ to automatically score all submissions. Additional analyses for some language pairs are provided below.

Due to the exceptionally hard setting, which currently leads to generally less competent translation systems, we did not perform the human evaluation of the outputs.

8.3 Submissions

Below we discuss all submissions for all language pairs, given that there were several overlaps. A brief summary per language is below:

- Irish–English received four submissions from one team (GMU);
- Marathi–Hindi received submissions from four teams (ALEXA AI, BUT, GMU, and SRI-B);
- Maltese–English received five submissions from one team (UM-DFKI);
- Pashto–French received submissions from two teams (GMU, ON-TRAC);
- Tamasheq–French received submissions from four teams (ALEXA AI, GMU, NAVER, and ON-TRAC);
- Quechua-Spanish received three submissions (GMU, NAVER, and QUESPA).

Below we discuss each team's submission in detail:

- ALEXA AI (Vishnu et al., 2023) submitted one primary and three contrastive systems, all of these are in the unconstrained condition (Table 44) for Tamasheq-French, and one primary and five contrastive systems on the unconstrained condition for Marathi–Hindi. For Marathi–Hindi, their systems relied on an end-to-end speech translation approach, using the wav2vec 2.0 base model finetuned on 960 hours of English speech (Baevski et al., 2020b) as encoder baseline and it was also finetuned on 94 hours of Marathi audio data. The team focused on evaluating three strategies including data augmentation, an ensemble model and post-processing techniques. For Tamasheq–French, they reuse the same end-to-end AST model proposed by the ON-TRAC Consortium in the last year's IWSLT edition (Boito et al., 2022b). This model consists of a speech encoder that is initialized by the wav2vec 2.0 (Baevski et al., 2020a) base model pre-trained on 243 hours of Tamasheq audio data released by the ON-TRAC Consortium [38]. The decoder of this model is a shallow stack of 2 transformer layers with 4 attention heads. A feed-forward layer is put in between the encoder and the decoder for matching the dimension of the encoder output and that of the decoder input. In this work, they focus on leveraging different data augmentation techniques including audio stretching, back translation, paraphrasing, and weighted loss. Another important endeavor of their work is experimenting with different post-processing approaches with LLMs, such as re-ranking, sentence correction, and token masking. Besides, they also ensemble AST models trained with different seeds and data augmentation methods, which is proven to improve the performance of their systems. Their primary system scores 9.30 BLEU on the 2023 test set.

- BUT (Kesiraju et al., 2023) submitted one primary and one contrastive system using the ESPnet (Inaguma et al., 2021) toolkit. The primary system was built with the end-to-end and bilingual ASR model while the contrastive was built with a cascade which uses various backbone models including ASR, the bilingual ASR, transformer-based seq2seq MT, LM for re-scoring and XLM.

- GMU (Mbuya and Anastasopoulos, 2023) focused on end-to-end speech translation systems. End-to-end (E2E) transformer-based encoder-decoder architecture (Vaswani et al., 2017) was used for primary constrained submission. For unconstrained submissions, they explored self-supervised pre-trained speech models and used wav2vec 2.0 (Baevski et al., 2020a) and HuBERT (Hsu et al., 2021) for the low resource task. They used wav2vec 2.0 - with removing the last three layers - for their primary submission. HuBERT was used for the contrastive1 submission - without removing any layer. For contrastive2, End-to-end with ASR (E2E-ASR) architecture uses the same architecture as the E2E. The difference is that a pre-trained ASR model was used to initialize its encoder.

- ON-TRAC (Laurent et al., 2023) participated in the Pashto–French (one primary and three contrastive systems, both for constrained and unconstrained settings) and Tamasheq–French (one primary and five contrastive systems, all of which are unconstrained (c.f. Table 44). For Pashto–French, the primary cascaded system is based on a convolutional model (Gehring et al., 2017) upgraded, while contrastive3 is based on small basic transformers. For Primary and contrastive1 systems, SAMU-XLS-R (Khurana et al., 2022) was used with pre-trained encoder with 100 and 53 languages. The two constrained contrastive E2E systems share the same encoder-decoder architecture using transformers (Vaswani et al., 2017). The difference lies in the use or not of a transformer language model trained from scratch on the provided dataset.

  All of their systems for Tamasheq–French are based on the same end-to-end encoder-decoder architecture. In this architecture, the encoder is initialized by a pre-

[38] https://huggingface.co/LIA-AvignonUniversity/IWSLT2022-tamasheq-only

trained semantic speech representation learning model named SAMU-XLS-R (Khurana et al., 2022), while the decoder is initialized with the decoder of the pre-trained mBART model. Their work heavily relies on different versions of the SAMU-XLS-R model, which are pre-trained on different combinations of multilingual corpora of 53, 60, and 100 languages. In addition, they leverage training data from higher resource corpora, such as CoVoST-2 (Wang et al., 2020a) and Europarl-ST (Iranzo-Sánchez et al., 2020), for training their end-to-end models. Their primary system, which scores 15.88 BLEU on the Tamasheq–French 2023 test set, was trained on the combination of (CoVoST-2, Europarl-ST and the IWSLT 2022's test set), with the encoder is initialized by the SAMU-XLS-R model trained on the data gathered from 100 languages.

- NAVER (Gow-Smith et al., 2023) submitted one primary and two contrastive systems to the Tamasheq–French track, as well as one primary and two contrastive systems for the unconstrained condition in the Quechua–Spanish track. In their work for the Tamasheq–French track, they concentrate on parameter-efficient training methods that can perform both ST and MT in a multilingual setting. In order to do so, they initialize their models with a pre-trained multilingual MT model (mBART (Liu et al., 2020) or NLLB (NLLB Team et al., 2022)), which is then fine-tuned on the ST task by inputting features extracted with a frozen pre-trained speech representation model (wav2vec 2.0 or HuBERT (Hsu et al., 2021)). The encoder of their translation model is slightly modified where they stack several modality-specific layers at the bottom. In addition, adapter layers are also inserted in between layers of the pre-trained MT model at both the encoder and decoder sides. While these new components get fine-tuned during the training process, the pre-trained components of the MT model are frozen. One of the appealing characteristics of their approach is that it allows the same model to do both speech-to-text and text-to-text translation (or transcription). Furthermore, their method maximizes knowledge transfer to improve low-resource performance. Their primary system, which is ensembled from 3 different runs on the combination of both ST and ASR data, scores 23.59 BLEU on the 2023 test set.

  For the Quechua–Spanish track, the overall architecture for their systems consists of first initializing a PLM which was then fine-tuned on the speech translation task by inputting features from a frozen pre-trained speech representation. Similar adaptations were done with an MT model to control domain and length mismatch issues. One of the interesting takeaways from their approaches is that their contrastive 2 system (1.3 billion parameters (NLLB Team et al., 2022)) outperformed their contrastive 1 system (3.3 billion parameters (NLLB Team et al., 2022)) despite it having less parameters. NAVER's primary submission was an ensemble approach that included the use of PLMs for both the ASR (Baevski et al., 2020a) and MT systems ((NLLB Team et al., 2022)) and included training on both Tamasheq and Quechua data. Their submissions to QUE–SPA did not include the use of mBART or HuBERT (Hsu et al., 2021) as was done for other language pairs that NLE submitted.

- QUESPA (Ortega et al., 2023) submitted to both conditions (constrained and unconstrained) a total of six systems including a primary, contrastive 1, and contrastive 2 for each condition. They also claim to have tried several other combinations but did not submit those systems. For the constrained condition, their primary system scored second best, slightly less than team GMU with a BLEU score of 1.25 and chrF2 of 25.35. They also scored third best for the constrained condition with 0.13 BLEU and 10.53 chrF2 using their contrastive 1 system. It is worthwhile to note that chrF2 was used by the organizers when BLEU scores were below five. For their constrained systems, a direct speech translation system was submitted similar to the GMU team's primary approach that used Fairseq (Wang et al., 2020b). QUESPA extracted mel-filter bank (MFB) features similar to the S2T approach in previous work Wang et al. (2020b). The main difference between QUESPA's submission and GMU's submissions was that the GMU team

increased the number of decoder layers to 6 which resulted in a slightly better system for GMU. The other systems submitted for the constrained setting were cascade systems where ASR and MT were combined in a pipeline setting. Their contrastive 1 and 2 system submissions for the constrained task respectively used wav2letter++ (Pratap et al., 2019) and a conformer architecture similar to previous work (Gulati et al., 2020) along with an OpenNMT (Klein et al., 2017) translation system trained on the constrained ST and MT data. Both of those systems performed poorly scoring less than 1 BLEU. For the unconstrained condition, the three systems that were presented by QUESPA consisted of pipeline approaches of PLMs that were fine-tuned on the additional 60 hours of Siminchik audio data along with the constrained data. Their primary and contrastive 1 unconstrained ASR systems were trained using the 102-language FLEURS (Conneau et al., 2023) model and used the MT system that was based on NLLB (NLLB Team et al., 2022) which just so happens to include Quechua as one of its languages. Their contrastive 2 ASR system was based on wav2letter++ (Pratap et al., 2019) while their contrastive 2 MT system was identical to the MT systems used for their Primary and Contrastive 1 submissions.

- SRI-B (Radhakrishnan et al., 2023) submitted four systems. For Marathi–English, they submitted one primary and one contrastive system in the constrained setting and one primary and one contrastive system in the unconstrained setting. They used end-to-end speech translation networks comprising a conformer encoder and a transformer decoder for both constrained and unconstrained.

- UM-DFKI (Williams et al., 2023) submitted five systems. It included one primary and four contrastive systems in unconstrained settings. They used a pipeline approach for all of their submissions. For ASR, their system builds upon (Williams, 2022) on fine-tuning XLS-R based system. mBART-50 was used for fine-tuning the MT part of the pipeline.

8.4 Results

Irish–English As discussed earlier, only the GMU team participated in the GA–ENG translation track and submitted one primary system to constrained, one primary system to unconstrained and the rest of the two systems to contrastive on unconstrained conditions. The end-to-end and end-to-end with ASR models submitted primary constrained and contrastive2 unconstrained systems. Both the systems achieved 15.1 BLEU scores. They did not perform well in comparison to the wav2vec 2.0 and HuBERT models. The detail of the results of this track can be found in Table 36 and 37.

Marathi–Hindi The results of this translation track can be found in Table 38 and 39. Overall we see varying performances among the systems submitted to this track, with some performing much better on the test set. Out of the 16 submissions, the SRI-B team's primary system achieved the best result of 31.2 and 54.8 in BLEU and in charF++ respectively on the constrained condition while the BUT team's primary system achieved the best results of 39.6 in BLEU and 63.3 in charF++ on the unconstrained condition. In both constrained and unconstrained conditions, the GMU systems achieved the lowest results of 3.3 and 5.9 in BLEU and 16.8 and 20.3 in charF++ respectively.

Maltese–English The results of this translation track can be found in Table 42. UM-DFKI used contrastive approaches in training their ASR system. For their contrastive1 system, their fine-tuning consisted of using Maltese, Arabic, French and Italian corpora. Their contrastive2, contrastive3, and contrastive4 approaches respectively use a subset from Arabic, French and Italian ASR corpus along with Maltese data. The best result of 0.7 BLEU was achieved with their contrastive1 system.

Pashto–French The detailed results can be found in Table 41 and Table 40 of the Appendix. We rank the system performance based on test BLEU scores. The best score BLEU was achieved by ON-TRAC primary system (SAMU-XLS-R model trained on 100 languages). For the constrained condition, the cascaded approach based on convolutional models, gives the best performance.

Tamasheq-French The results of this translation track can be found in Table 43 and 44. Compared to the last year's edition, this year has witnessed a growing interest in this low-resource translation track in terms of both quantity and quality of submissions. Almost all submissions achieve relatively better results than the last year's best system (5.7 BLEU on test2022 (Boito et al., 2022b)). Furthermore, it is notable that cascaded systems are not favorable in this track while none of the submitted systems is of this kind.

This year, this language pair remains a challenging low-resource translation track. There is only one submission to the constrained condition from GMU with an end-to-end model scoring 0.48 BLEU on this year's test set. For this reason, all the participants are in favor of exploiting pre-trained models, hence being subject to the unconstrained condition. Among these pre-trained models, self-supervised learning (SSL) from speech models remains a popular choice for speech encoder initializing. Using a wav2vec2.0 model pre-trained on unlabelled Tamasheq data for initializing their speech encoder, GMU gains +7.55 BLEU score in comparison with their Transformer-based encoder-decoder model training from scratch (their primary constrained system). At the decoder side, pre-trained models such as mBART or NLLB are commonly leveraged for initializing the decoder of the end-to-end ST model. Besides, data augmentation and ensembling are also beneficial as shown by ALEXA AI when they consistently achieve ~ 9 BLEU in all of their settings.

Outstanding BLEU scores can be found in the work of the ON-TRAC team. An interesting pre-trained model named SAMU-XLS-R is shown to bring significant improvements. This is a multilingual multimodal semantic speech representation learning framework (Khurana et al., 2022) which fine-tunes the pre-trained speech transformer encoder XLS-R (Babu et al., 2021) using semantic supervision from the pre-trained multilingual semantic text encoder LaBSE (Feng et al., 2022). Exploiting this pre-trained model and training end-to-end ST models on the combinations of different ST corpora, they achieve more than 15 BLEU in all of their settings.

NAVER tops this translation track by a multilingual parameter-efficient training solution that allows them to leverage strong pre-trained speech and text models to maximize performance in low-resource languages. Being able to be trained on both ST and ASR data due to the multilingual nature, all of their submissions heavily outperform the second team ON-TRAC by considerable margins. Their primary system, which is ensembled from 3 different runs, uses NLLB1.3B as the pre-trained MT system, and wav2vec2.0 Niger-Mali [39] as the speech presentation extractor. After being trained on a combination of both ST corpora (Tamasheq-French, mTEDx fr-en, mTEDx es-fr, mTEDx es-en, mTEDx fr-es (Salesky et al., 2021)) and AST corpora (TED-LIUM v2 (Rousseau et al., 2014), mTEDx fr, mTEDx es), this system establishes an impressive state-of-the-art performance of the Tamasheq-French language pair, scoring 23.59 BLEU on the 2023 test set.

Quechua–Spanish The QUE–SPA results for all systems submitted to this low-resource translation track can be found in Table 45 and 46 of the appendix. To our knowledge, this first edition of the QUE–SPA language pair in the low-resource track of IWSLT has witnessed the best BLEU scores achieved by any known system in research for Quechua. The two best performing systems: 1.46 BLEU (constrained) and 15.70 (unconstrained) show that there is plenty of room to augment approaches presented here. Nonetheless, submissions from the three teams: GMU, NAVER, and QUESPA have shown that it is possible to use PLMs to create speech-translation systems with as little as 1.6 hours of parallel speech data. This is a notable characteristic of this task and surpasses previous work in the field.

We have found that the NLLB (NLLB Team et al., 2022) system's inclusion of Quechua in recent years has had a greater impact than expected for ease-of-use. Similarly, the use of Fairseq (Wang et al., 2020b) seems to be the preferred toolkit for creating direct S2T systems, cascaded or not. The QUE–SPA submissions for the unconstrained conditions preferred the use of a cascading system in a pipeline approach where pre-trained models were fine-tuned first for ASR and then for MT.

The constrained setting leaves much room for improvement. Nonetheless, GMU and QUESPA's near identical submissions have shown that the in-

[39] https://huggingface.co/LIA-AvignonUniversity/IWSLT2022-Niger-Mali

crease of 3 layers during decoding can be powerful and should be explored further. It would be worthwhile for the organizers of the QUE–SPA track to obtain more parallel data including translations for future iterations of this task.

The unconstrained setting clearly can benefit from an ensembling technique and training with multiple languages – in these submissions, the training of a model with an additional language like Tamasheq alongside Quechua does not seem to have a negative impact on performance. Although, it is hard to ascertain whether the slight performance gain of less than 1 BLEU point of the NLE team's submission compared to QUESPA's submission was due to the ensembling, freezing of the models, or the language addition.

As a final takeaway, the NLE team's submissions scored quite well under the unconstrained condition. It should be noted that for other language pairs NLE's high system performance was also due to the ensembling of systems that were executed using different initialization parameters on at least three unique runs. As an aside, small gains were achieved under the constrained condition when comparing the GMU submission to the QUESPA system due to the increase in decoding layers. QUESPA's inclusion of a language model on top of a state-of-the-art dataset (Fleurs) allowed them to achieve scores similar to NAVER's without additional tuning or ensembling. State-of-the-art performance was achieved by all three teams that submitted systems.

General Observations As in previous years, the low-resource shared task proved particularly challenging for the participants, but there are several encouraging signs that further reinforce the need for more research in the area.

First, more teams than ever participated in the shared task, showing a continued interest in the field. Second, we note that for the language pair that was repeated from last year (Tamasheq–French), almost all submissions outperformed last year's best submission, with an accuracy increase of more than 17 BLEU points in the unconstrained setting. Last, we highlight the breadth of different approaches employed by the participants, ranging from the use of finetuned pre-trained models to pre-training from scratch, to parameter efficient dine-tuning as well as cascaded pipeline systems, all of which seem to have benefits to offer, to a certain extent, to different language pairs.

Limitations As noted by some participants, the Irish–English and Maltese–English translation track data has limitations. For Irish–English, the speech translation systems can achieve very high BLEU scores on the test set if the built systems have used wav2vec 2.0 and/or the Irish ASR model which is trained on the Common Voice (Ardila et al., 2020b) dataset. Similarly, the GMU team has achieved high BLEU scores especially when they used wav2vec 2.0 and HuBERT models. We plan to continue this translation track next year by updating the test and training data to thoroughly investigate the data quality as well as the reason to obtain the high BLEU scores. For Maltese–English, some participants reported issues with the data quality, which we hope to resolve in future iterations of the shared task.

9 Formality Control for SLT

Different languages encode formality distinctions in different ways, including the use of honorifics, grammatical registers, verb agreement, pronouns, and lexical choices. While machine translation (MT) systems typically produce a single generic translation for each input segment, SLT requires adapting the translation output to be appropriate to the context of communication and target audience. This shared task thus challenges machine translation systems to generate translations of different formality levels.

9.1 Challenge

Task Given a source text, X in English, and a target formality level, $l \in \{F, IF\}$, the goal in formality-sensitive machine translation (Niu et al., 2017) is to generate a translation, Y, in the target language that accurately preserves the meaning of the source text and conforms to the desired formality level, l. The two formality levels typically considered are "F" for formal and "IF" for informal, resulting in two translations: Y_F and Y_{IF} respectively. For example, the formal and informal translations for the source text "Yeah Did your mom know you were throwing the party?" (originally informal) in Korean are shown in the table below:

This shared task builds on last year's offering, which evaluated systems' ability to control formality on the following translation tasks: translation from English (EN) into Korean (KO) and Vietnamese (VI) in the *supervised* setting, and from English (EN) into Portugal Portuguese (PT)

Source: Yeah Did your mom know you were throwing the party?

Korean Informal: 그, 어머님은 [F]네가[/F] 그 파티 연 거 [F]아셔[/F]?

Korean Formal: 그, 어머님은 [F]님이[/F] 그 파티 연 거 [F]아세요[/F]?

Table 7: Contrastive formal and informal translations into Korean. Grammatical formality markers are annotated with [F]text[/F].

and Russian (RU) in the *zero-shot setting*. Results showed that formality-control is challenging in zero-shot settings and for languages with many grammatical and lexical formality distinctions. This year's edition invited participants to advance research in effective methods for bridging the gap in formality control for zero-shot cases and for languages with rich grammatical and lexical formality distinctions.

9.2 Data and Metrics

Participants were provided with test data, as well as MT quality and formality control metrics. In addition, we provided training data, consisting of formal and informal translation of texts for the supervised language pairs (EN-KO, EN-VI).

9.2.1 Formality Annotated Dataset

We provide targeted datasets comprising source segments paired with two contrastive reference translations, one for each formality level (informal and formal) for two EN-VI, EN-KO in the *supervised* setting and EN-RU, EN-PT in the *zero-shot* setting (see Example 7)[40]. The sizes and properties of the released datasets for all the language pairs are listed in Table 8. Formal translations tend to be longer than informal texts for Vietnamese compared to other language pairs. The number of phrasal formality annotations ranges from 2 to 3.5 per segment, with Korean exhibiting a higher diversity between the formal and informal translations as indicated by the TER score.

9.2.2 Training Conditions

We allowed submissions under the constrained and unconstrained data settings described below:

[40] https://github.com/amazon-science/contrastive-controlled-mt/tree/main/IWSLT2023

Constrained (C) Participants were allowed to use the following resources: Textual MuST-C v1.2 (Di Gangi et al., 2019b), CCMatrix (Schwenk et al., 2021), OpenSubtitles (Lison and Tiedemann, 2016) and dataset in the constrained setting from the Formality Control track at IWSLT22 (Anastasopoulos et al., 2022a).

Unconstrained (U) Participants could use any publicly available datasets and resources: the use of pre-trained language models was also allowed. Additionally, using additionally automatically annotated bitext with formality labels was also allowed.

9.3 Formality Classifier

We release a multilingual classifier (MC) trained to predict the formality of a text for all the language pairs: EN-KO, EN-VI, EN-RU, and EN-PT. We finetune an `xlm-roberta-base` (Conneau et al., 2020) model on human-written formal and informal translations following the setup from Briakou et al. (2021). Our classifier achieves an accuracy of $> 98\%$ in detecting the formality of human-written translations for the four target languages (Table 10). Participants were allowed to use the classifier both for model development and for evaluation purposes as discussed below.

9.4 Automatic Metrics

We evaluate the submitted system outputs along the following two dimensions:

1. Overall translation quality, evaluated using SacreBLEU v2.0.0 (Papineni et al., 2002b; Post, 2018), and COMET (Rei et al., 2020b) on both the shared task-provided test sets based on topical chat (Gopalakrishnan et al., 2019) and on the FLORES devtest (NLLB Team et al., 2022; Goyal et al., 2022).

2. Formality control, evaluated using:

   - Matched-Accuracy (mACC), a reference-based corpus-level automatic metric that leverages phrase-level formality markers from the references to classify a system-generated hypothesis as formal, informal, or neutral (Nadejde et al., 2022).
   - Classifier-Accuracy (cACC), a reference-free metric that uses the multilingual formality classifier discussed above to label a system-generated hypothesis as formal or informal.

Language	Type	Size	Length			# Phrasal Annotations		TER(F, IF)
			Source	Formal	Informal	Formal	Informal	
EN-VI	Train	400	20.35	28.52	25.48	2.71	1.49	23.70
	Test	600	21.82	29.59	26.77	2.79	1.55	23.00
EN-KO	Train	400	20.00	13.41	13.40	3.35	3.35	24.52
	Test	600	21.22	13.56	13.55	3.51	3.51	25.32
EN-RU	Test	600	21.02	18.03	18.00	2.06	2.05	13.59
EN-PT	Test	600	21.36	20.22	20.27	1.93	1.93	10.46

Table 8: Formality Track Shared Task Data Statistics.

Participant	Settings	Classifier Use	Languages	Model type	Formality
UMD-baseline	U	✓	All	Multilingual	Exemplars
CoCoA-baseline	C	✗	EN-{VI, KO}	Bilingual	Side-constraint
AppTek	U	✗	EN-{PT, RU}	Bilingual	Side-constraint
HW-TSC	U+C	✓	All	Bilingual	Side-constraint
KUxUpStage	U	✓	All	Bilingual	N/A
UCSC	U	✗	EN-{VI, KO}	Multilingual	Style-Embedding

Table 9: Formality Track Submissions Summary. Most participants train bilingual systems but leverage a diverse set of formality encoding mechanisms for control.

Target Language	Accuracy
Korean	99.9%
Vietnamese	99.3%
Russian	99.9%
Portuguese	98.6%

Table 10: The multilingual classifier can identify the target formality for human written text across all languages with > 98% accuracy.

The final corpus-level score for each of the two metrics described above is the percentage of system outputs that matches the desired formality level. For example, the cAcc for the target formality, Formal (F), is given by, $c\text{Acc}(F) = \frac{1}{M}\sum_{i=1}^{M} \mathbb{1}[MC(Y) == F]$, where M is the number of system outputs.

9.5 Submissions

We provide methodology descriptions and a summary of the two baseline systems and four submissions received for the shared task below and in Table 9. Three out of six submissions made use of the formality classifier released for system development. We received two multilingual and four bilingual systems. We refer the reader to the system description papers for more details.

- CoCoA (baseline) uses a supervised method where a generic neural MT model is fine-tuned on labeled contrastive translation pairs (Nadejde et al., 2022). For the constrained, supervised setting, the generic neural MT model was trained on parallel data allowed for the constrained task and fine-tuned on formal and informal data released for the shared task. Following Nadejde et al. (2022), contrastive pairs were upsampled with a fixed upsampling factor of five for all language pairs.

- UMD (baseline) uses 16 few-shot target formality-specific exemplars to prompt XGLM-7.5B (Lin et al., 2021) to generate style-controlled translations. For the supervised setting, these examples are drawn from the official training data, whereas for the zero-shot setup, the examples from the Tatoeba corpus (Artetxe and Schwenk, 2019) are filtered and marked with target formality using the provided formality classifier.

- AppTek (Bahar et al., 2023) submitted outputs using their production quality translation systems that support formality-controlled translation generation for EN-PT and EN-

RU. These are Transformer-Big models trained on a large public dataset from the OPUS collection (Tiedemann, 2012), automatically marked with formality using a sequence of regular expressions. The formality level is encoded with a pseudo-token at the beginning of each training source sentence with one of 3 values: formal, informal, or no style.

- HW-TSC (Wang et al., 2023a) describes a system that uses a multi-stage pre-training strategy on task-provided data to train strong bilingual models. Using these bilingual models, they employ beam re-ranking on the outputs generated using the test source. The generated hypothesis are ranked using the formality classifier and phrasal annotations, iteratively fine-tuning the model on this data until test performance convergences. Initial formality control is enabled by a special token and re-affirmed through classifier output and annotations from training.

- KUxUpStage (Lee et al., 2023) uses large-scale bilingual transformer-based MT systems trained on high-quality datasets and mBART for the supervised and zero-shot settings respectively. They generate a formality-controlled translation dataset for supervision in the zero-shot setting using GPT-4 and filter the generated source-translation pairs using the formality classifier. All bilingual models are then finetuned independently for the two target formality directions to generate formality-controlled outputs, resulting in #(Language-pairs) × 2 (Formal/Informal) models.

- UCSC (Vakharia et al., 2023) focused on using a single multilingual translation model for all the language pairs under the unconstrained setting. They finetune the pre-trained model, `mBART-large-50` (Tang et al., 2020), using the provided contrastive translations (§ 9.2.1) with an added *style embedding intervention* layer.

9.6 Results

Tables 47 and 48 in the Appendix show the main automatic evaluation results for the shared task.

Overall Results For the supervised language pairs in both constrained and unconstrained settings, most submitted systems were successfully able to control formality. The average `mAcc` scores ranged from 78-100. Controlling formality in Korean was found to be more challenging than translating with formality control in Vietnamese as reflected by the relatively lower `mAcc` scores which we believe to be due to the variation in formality expression of Korean honorific speech reflected in pretraining data.

HW-TSC consistently achieves the best scores across the board for all language pairs and both settings due to the use of transductive learning. Interestingly, the constrained submission by HW-TSC achieves better or competitive results compared to their unconstrained system suggesting that the use of a pre-trained language model or additional resources is not necessary to generate high-quality formality-controlled translations. Generally, the systems generate higher quality outputs in the formal setting relative to the informal setting for both supervised language pairs according to BLEU and COMET, which might be due to the bias of the dataset used during pre-training which is typically news and hence more formal.

In the zero-shot unconstrained setting, this formality bias is even more prominent. We observe a much wider distribution in the formality scores for English-Portuguese (`mAcc`: F 90-100, IF: 58-100), possibly due to the high ambiguity in the informal language and the confounding dialectal influence of Brazilian Portuguese dominant in the pre-training corpora, which is known to use formal register even in typically informal contexts (Costa-jussà et al., 2018). HW-TSC and AppTek achieve the best translation quality for English-Portuguese and English-Russian respectively. The lowest scoring submission in both quality and formality control (UCSC) did not include any fine-tuning or adaptation of the base mBART model to the two zero-shot language pairs: English-Russian and English-Portuguese. This suggests that formality information is not transferred from the unrelated language pairs, EN-KO and EN-VI, and that some language-specific supervision is needed to mark grammatical formality appropriately in Russian and Portuguese.

How well do systems match the desired target formality? We show the distribution of the scores generated using the formality classifier for

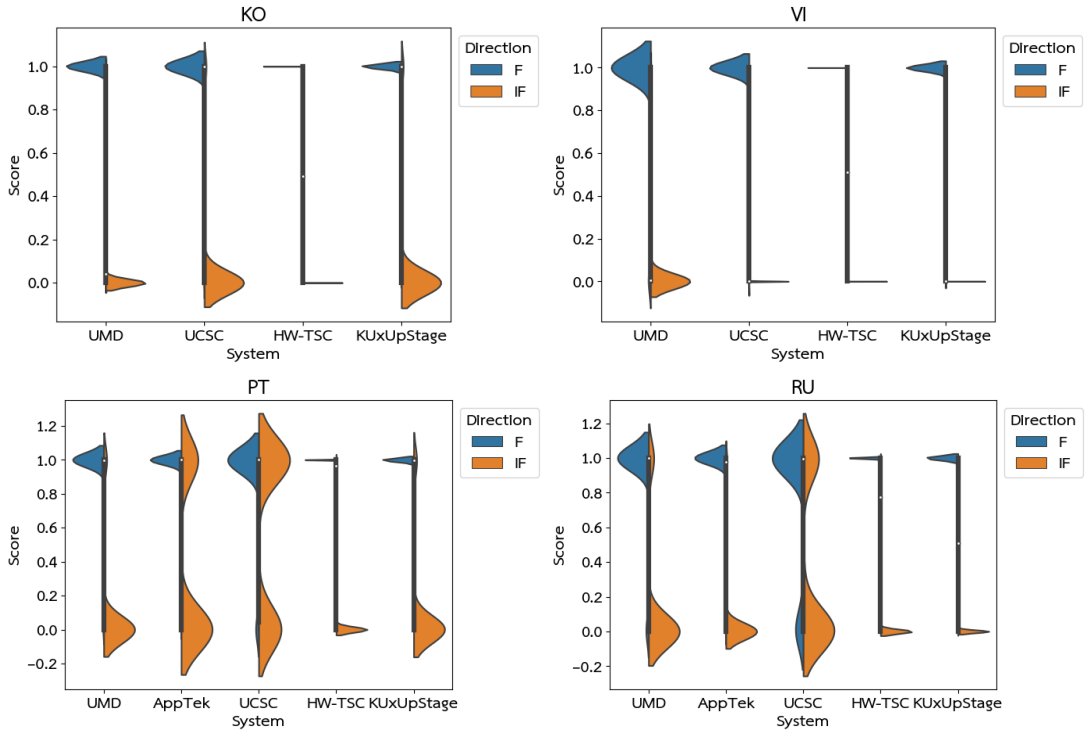

Figure 3: Formality Classifier Scores' Distribution on the submitted system outputs in the Unconstrained setting: HW-TSC can precisely match the target formality as depicted by the peaky distribution.

all the systems submitted to all language pairs under the unconstrained setting in Figure 3. For supervised language pairs, formal (blue) and informal (orange) output scores peak at 1.0 and 0.0 respectively. In the zero-shot setting, for both Portuguese (APPTEK, UCSC) and Russian (UCSC) translations, the informal outputs have a bimodal distribution, highlighting that these models generate many formal translations under informal control.

How contrastive are the generated translations? We show the Translation Edit Rate (TER) between the formal and informal outputs for all submitted systems across all language pairs in Figure 4. While the references are designed to be minimally contrastive, the formal and informal system outputs exhibit a much larger edit distance. HW-TSC has the lowest TER rate for all language pairs except English-Korean.

Discussion Overall, the shared task results show that finetuning a strong supervised general-purpose MT system with as low as 400 in-domain contrastive samples seems to be sufficient in generating high-quality contrastive formality-controlled translations. However, several avenues for improvement remain open. The languages that

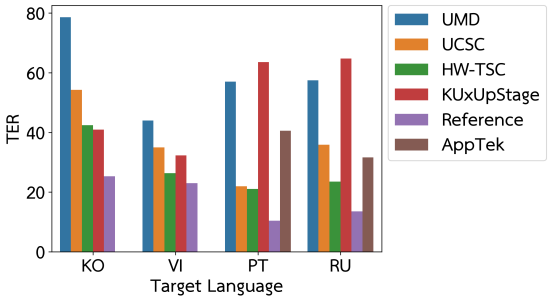

Figure 4: TER between the Formal (F) and Informal (IF) Outputs for all submitted systems across all language pairs.

exhibit an ambiguous or richer formality distinction either due to close dialectal variations (like Portuguese) or due to multiple levels of honorifics (like Korean and Japanese) still remain challenging. Unsupervised transfer of formality knowledge between related languages remains relatively unexplored (Sarti et al., 2023). Furthermore, this year's task only considered two levels of formality distinctions with minimal edits. It remains unclear whether the models are also capable of modeling multiple levels of formality potentially with minimal edits in the generated translations. Finally, no submissions have explored monolingual editing of translations as a potential solution for

formality-controlled MT, despite the edit-focused nature of the contrastive translations. We recommend that future work on formality-controlled machine translation targets these challenges.

10 Automatic Dubbing

10.1 Challenge

This task focuses on automatic dubbing: translating the speech in a video into a new language such that the new speech is natural when overlayed on the original video (see Figure 5).

Participants were given German videos, along with their text transcripts, and were asked to produced dubbed videos where the German speech has been translated in to English speech.

Automatic dubbing is a very difficult/complex task (Brannon et al., 2023), and for this shared task we focus on the characteristic which is perhaps most characteristic of dubbing: isochrony. Isochrony refers to the property that the speech translation is time aligned with the original speaker's video. When the speaker's mouth is moving, a listener should hear speech; likewise, when their mouth isn't moving, a listener should not hear speech.

To make this task accessible for small academic teams with limited training resources, we make some simplifications: First, we assume the input speech has already been converted to text using an ASR system and the desired speech/pause times have been extracted from the input speech. Second, to alleviate the challenges of training a TTS model, the output is defined to be phonemes and their durations. These phonemes and durations are played through an open-source FastSpeech2 (Ren et al., 2022) text-to-speech model to produce the final speech.[41]

10.2 Data and Metrics

Official training and test data sets were provided[42] by the organizers. The training data was derived from CoVoST2 (Wang et al., 2021) and consists of:

1. Source (German) text
2. Desired target speech durations (e.g. 2.1s of speech, followed by a pause, followed by 1.3s of speech)
3. Target (English) phonemes and durations corresponding to a translation which adheres to the desired timing

The test data was produced by volunteers and consists of videos of native German speakers reading individual sentences from the German CoVoST-2 test set.[43] This test set was divided in to two subsets; *Subset 1* where there are no pauses in the speech and *Subset 2* where there is one or more pause in the speech. More details on this data are presented in (Chronopoulou et al., 2023).

10.3 Submissions

Despite high initial interest, we received only one submission, which was from the Huawei Translation Services Center (HW-TSC) (Rao et al., 2023). However, we had two systems (Chronopoulou et al., 2023; Pal et al., 2023) built for the task for which we had not yet performed human evaluation, so we still had enough systems for a interesting comparison.

- Interleaved (Baseline): Our first baseline and the basis for this shared task is from Chronopoulou et al. (2023). They propose to jointly model translations and speech timing, giving the model the freedom to change the translation to fit the timing, or and make scarifies in translation quality to meet timing constraints or relax timing constraints to improve translation quality. This is achieved by simply binning target phoneme durations and interleaving them with target phonemes during training and inference. To avoid teaching the model that speech durations should be prioritized over translation quality[44], noise with standard deviation 0.1 is added to the target phrase durations to simulate the source durations used at inference.

- Factored (Baseline): Pal et al. (2023) build on the first baseline by using target factors (García-Martínez et al., 2016), where alongside predicting phoneme sequences as the target, we also predict durations for each phoneme as a target factor. Additionally, they propose auxiliary counters, which are similar to target factors except the model is not

[41] https://github.com/mtresearcher/FastSpeech2

[42] https://github.com/amazon-science/iwslt-autodub-task/tree/main/data

[43] Each volunteer provided their consent to use this data for automatic dubbing task.

[44] Median speech overlap is just 0.731 in a large corpus of human dubs (Brannon et al., 2023)

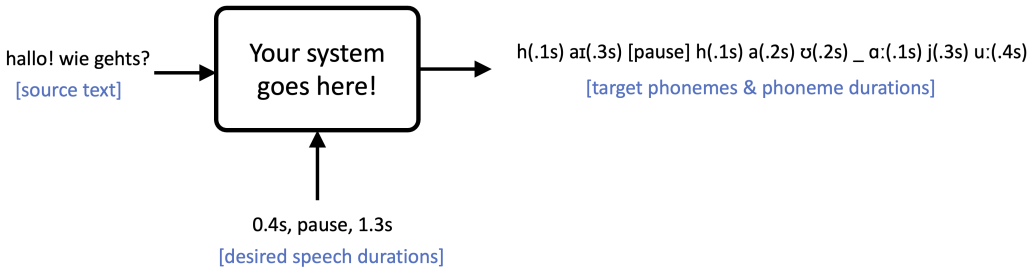

Figure 5: To illustrate, here's an example in which "hallo! wei gehts?" is translated to "hi! how are you?" such that the output will fit in the desired target speech durations of 0.4s and 1.3s, with a pause in between

trained to predict them. Instead, they providing additional information to the decoder consisting of (1) the total number of frames remaining, (2), the number of pauses remaining, and (3) the number of frames remaining in the current phrase. As in the first baseline, noise of standard deviation 0.1 is added to the target phrase durations during training to simulate source durations.

- Text2Phone (Baseline): As a sanity check, we added a third, non-isochronic baseline trained to take in German text and produce English phonemes, without any duration information. We train on the same data as the first two baselines, but exclude duration information from training and instead predict phoneme durations using the duration model from the FastSpeech2 model.

- HW-TSC: In contrast to our three baselines, (Rao et al., 2023) took a more traditional approach to dubbing and followed the prior works on verbosity control (Lakew et al., 2021, 2019) to first generate a set of translation candidates and later re-rank them. Their system consists of four parts: 1) voice activity detection followed by pause alignment, 2) generating a list of translation candidates, 3) phoneme duration prediction, followed by 4) re-ranking/scaling the candidates based on the durations (see Figure 6). With the last step in the pipeline, the top scored candidate is ensured to have the best speech overlap with the source speech amongst all candidate translations.

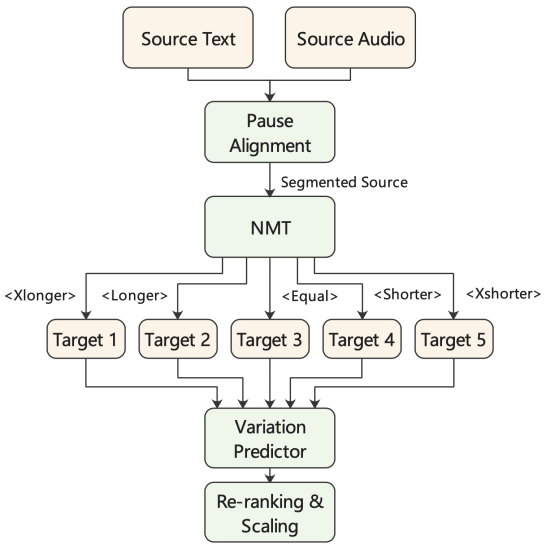

Figure 6: System diagram for HW-TSC dubbing system. Image from Rao et al. (2023).

10.4 Evaluation & Metric

The dubbed English videos were judged by a mixture of native and non-native speakers, all of which were researchers in automatic dubbing. For each video in the the test set, one judge was shown the four system outputs in random order and asked to rate them from 1-6. The judges were not given a defined rubric or guidelines to follow but were asked to be consistent.

As a metric we opted for mean opinion score (MOS) methodology where the scores for a system as judged by humans are averaged in one score.[45]

Feedback from the judges indicate that the baseline and submitted systems often produce poor translations (perhaps due to the small amount of training data used by each system), and the voice quality from the FastSpeech 2 model was far from perfect. However, they felt that having all systems share the same voice made it much easier to compare across dubbing systems.

When we looked at the distribution of scores per

[45] https://en.wikipedia.org/wiki/Mean_opinion_score

annotator (judge) level, the numbers showed that each annotator had a bias towards dubbing, some liked dubbing more than others which is intuitive but has not been studied before in the context of automatic dubbing. As shown in Table 11, it is clear that annotator A2 had a significantly higher preference for dubbing as compared to annotator A4 in terms of MOS.

Annotator	MOS↑	CI
A1	3.34	±0.16
A2	3.74	±0.19
A3	3.53	±0.13
A4	3.07	±0.15

Table 11: MOS (on a scale of 1-6) with confidence interval (CI) at 95% per annotator showing the biases towards general purpose dubbed content.

We also looked at MOS for the two different subsets to understand whether it was difficult for the submitted systems to dub the videos. As it turns out, *Subset 1* has an significantly higher MOS of 3.54 (± 0.11) compared to *Subset 2* with a MOS of 3.31 (± 0.11). This shows it is significantly more difficult for all systems to dub *Subset 2* than *Subset 1*.

10.5 Results

Results are shown in Table 12. All three dubbing systems outperform the non-isochronic Text2Phone baseline (Chronopoulou et al., 2023), as expected. The factored baseline improves over the interleaved baseline, consistent with the automatic metric results reported by Pal et al. (2023).

The HW-TSC system (Rao et al., 2023) outperforms all the baselines in terms of mean opinion score, making it the clear winner of the IWSLT 2023 dubbing shared task. Unfortunately, since HW-TSC system was unconstrained (it trains on additional bitext compared to the baselines) and uses fundamentally different approaches than the baselines, it is not possible to attribute it's performance to any single factor.

Lip-sync is an important feature of dubbing, it is important that the final generated audio is in sync with the lip movements of the on-screen speaker in the original video. As an analysis, we looked at Lip-Sync Error Distance (LSE-D) (Chung and Zisserman, 2016) following the evaluation methodology in Hu et al. (2021). LSE-D is not a perfect metric but it is an indication to

		MOS↑	
System	Constrained?	Mean	CI
Text2Phone	Yes	3.16	±0.19
Interleaved	Yes	3.33	±0.18
Factored	Yes	3.43	±0.19
HW-TSC	No	3.77	±0.19

Table 12: Mean opinion score for baselines 1) Text2Phone 2) Interleaved (Chronopoulou et al., 2023) 3) Factored (Pal et al., 2023) and 4) submitted system of HW-TSC (Rao et al., 2023).

	LSE-D↓	
System	Subset1	Subset2
Original	7.39	7.67
Text2Phone	11.64	13.31
Interleaved	11.71	12.35
Factored	11.73	12.48
HW-TSC	12.11	12.77

Table 13: Results of Lip-Sync Error Distance (LSE-D) via Syncnet pre-trained model (Chung and Zisserman, 2016). Lower the better.

the amount of Lip-Sync errors in the video. From Table 13, *Subset 1* consistently has a lower lip-sync error than *Subset 2* in all cases pointing that its difficult to generate lip-synced dubs for *Subset 2*. This result is also in line with the MOS scores we obtained for two subsets where the annotators preferred dubs for *Subset 1*. Secondly, original videos show significantly lower lip-sync error distance (12.x v/s 7.x) than dubbed videos showing that automatic dubbing research still has a long way to go to reach lip-sync quality in original videos.

Acknowledgements

Claudia Borg, Thierry Declerck, Rishu Kumar and John Judge acknowledge H2020 LT-Bridge Project (GA 952194). Rishu Kumar would also like to thank the EMLCT[46] programme. Atul Kr. Ojha and John P. McCrae would like to thank Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289_P2 Insight_2, and Panlingua Language Processing LLP for providing the Marathi-Hindi speech translation data and for their support. John Judge would also like to acknowledge the support of SFI under grant SFI/13/RC/2106_P2 ADPAT. Ondřej Bojar would like to acknowledge the grant 19-26934X

[46]https://mundus-web.coli.uni-saarland.de/

(NEUREM3) of the Czech Science Foundation. Antonios Anastasopoulos and Milind Agarwal are supported by the US National Science Foundation CCRI-Planning 2234895 award, as well as a National Endowment for the Humanities PR-276810-21 award.

References

Basil Abraham, Danish Goel, Divya Siddarth, Kalika Bali, Manu Chopra, Monojit Choudhury, Pratik Joshi, Preethi Jyoti, Sunayana Sitaram, and Vivek Seshadri. 2020. Crowdsourcing speech data for low-resource languages from low-income workers. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2819–2826.

Yasuhiro Akiba, Marcello Federico, Noriko Kando, Hiromi Nakaiwa, Michael Paul, and Jun'ichi Tsujii. 2004. Overview of the IWSLT04 Evaluation Campaign. In *Proceedings of the International Workshop on Spoken Language Translation*, pages 1–12, Kyoto, Japan.

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. 2022a. 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, 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. 2022b. 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.

Pierre Andrews, Guillaume Wenzek, Kevin Heffernan, Onur Çelebi, Anna Sun, Ammar Kamran, Yingzhe Guo, Alexandre Mourachko, Holger Schwenk, and Angela Fan. 2022. stopes-modular machine translation pipelines. In *Proceedings of the The 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 258–265.

Ebrahim Ansari, Amittai Axelrod, Nguyen Bach, Ondrej Bojar, Roldano Cattoni, Fahim Dalvi, Nadir Durrani, Marcello Federico, Christian Federmann, Jiatao Gu, Fei Huang, Kevin Knight, Xutai Ma, Ajay Nagesh, Matteo Negri, Jan Niehues, Juan Pino, Elizabeth Salesky, Xing Shi, Sebastian Stüker, Marco Turchi, and Changhan Wang. 2020. Findings of the IWSLT 2020 Evaluation Campaign. In *Proceedings of the 17th International Conference on Spoken Language Translation (IWSLT 2020)*, Seattle, USA.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M Tyers, and Gregor Weber. 2019. Common voice: A massively-multilingual speech corpus. *arXiv preprint arXiv:1912.06670*.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M Tyers, and Gregor Weber. 2020a. Common voice: A massively-multilingual speech corpus. In *LREC*.

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

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

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika

Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, et al. 2021. XLS-R: Self-supervised cross-lingual speech representation learning at scale. *arXiv preprint arXiv:2111.09296*.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020a. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems*, volume 33, pages 12449–12460.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020b. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, 33:12449–12460.

Parnia Bahar, Patrick Wilken, Javier Iranzo-Sánchez, Mattia Di Gangi, Evgeny Matusov, and Zoltán Tüske. 2023. Speech Translation with Style: AppTek's Submissions to the IWSLT Subtitling and Formality Tracks in 2023. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT)*.

Luisa Bentivogli, Mauro Cettolo, Marco Gaido, Alina Karakanta, Alberto Martinelli, and Marco Turchi Matteo Negri. 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*, Bangkok, Thailand. Association for Computational Linguistics.

Marcely Zanon Boito, Fethi Bougares, Florentin Barbier, Souhir Gahbiche, Loïc Barrault, Mickael Rouvier, and Yannick Estéve. 2022a. Speech resources in the tamasheq language. *Language Resources and Evaluation Conference (LREC)*.

Marcely Zanon Boito, John Ortega, Hugo Riguidel, Antoine Laurent, Loïc Barrault, Fethi Bougares, Firas Chaabani, Ha Nguyen, Florentin Barbier, Souhir Gahbiche, and Yannick Estève. 2022b. ON-TRAC Consortium Systems for the IWSLT 2022 Dialect and Low-resource Speech Translation Tasks. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT)*.

William Brannon, Yogesh Virkar, and Brian Thompson. 2023. Dubbing in Practice: A Large Scale Study of Human Localization With Insights for Automatic Dubbing. *Transactions of the Association for Computational Linguistics*, 11:419–435.

Eleftheria Briakou, Sweta Agrawal, Joel Tetreault, and Marine Carpuat. 2021. Evaluating the evaluation metrics for style transfer: A case study in multilingual formality transfer. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1321–1336, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ronald Cardenas, Rodolfo Zevallos, Reynaldo Baquerizo, and Luis Camacho. 2018. Siminchik: A speech corpus for preservation of southern quechua. *ISI-NLP 2*, page 21.

Mauro Cettolo, Marcello Federico, Luisa Bentivogli, Jan Niehues, Sebastian Stüker, K. Sudoh, K. Yoshino, and Christian Federmann. 2017. Overview of the IWSLT 2017 Evaluation Campaign. In *Proceedings of the 14th International Workshop on Spoken Language Translation (IWSLT 2017)*, pages 2–14, Tokyo, Japan.

Mauro Cettolo, Jan Niehues, Sebastian Stüker, Luisa Bentivogli, Roldano Cattoni, and Marcello Federico. 2015. The IWSLT 2015 Evaluation Campaign. In *Proceedings of the 12th International Workshop on Spoken Language Translation (IWSLT 2015)*, Da Nang, Vietnam.

Mauro Cettolo, Jan Niehues, Sebastian Stüker, Luisa Bentivogli, and Marcello Federico. 2013. Report on the 10th IWSLT Evaluation Campaign. In *Proceedings of the Tenth International Workshop on Spoken Language Translation (IWSLT 2013)*, Heidelberg, Germany.

Mauro Cettolo, Jan Niehues, Sebastian Stüker, Luisa Bentivogli, and Marcello Federico. 2014. Report on the 11th IWSLT Evaluation Campaign, IWSLT 2014. In *Proceedings of the Eleventh International Workshop on Spoken Language Translation (IWSLT 2014)*, Lake Tahoe, USA.

Mauro Cettolo, Jan Niehues, Sebastian Stüker, Luisa Bentivogli, and Marcello Federico. 2016. The IWSLT 2016 Evaluation Campaign. In *Proceedings of the 13th International Workshop on Spoken Language Translation (IWSLT 2016)*, Seattle, USA.

Mingda Chen, Paul-Ambroise Duquenne, Pierre Andrews, Justine Kao, Alexandre Mourachko, Holger Schwenk, and Marta R. Costa-jussà. 2022. Blaser: A text-free speech-to-speech translation evaluation metric.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Micheal Zeng, and Furu Wei. 2021. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16:1505–1518.

Colin Cherry and George Foster. 2019. Thinking slow about latency evaluation for simultaneous machine translation. *arXiv preprint arXiv:1906.00048*.

Kyunghyun Cho and Masha Esipova. 2016. Can neural machine translation do simultaneous translation? *arXiv preprint arXiv:1606.02012*.

Alexandra Chronopoulou, Brian Thompson, Prashant Mathur, Yogesh Virkar, Surafel M. Lakew, and Marcello Federico. 2023. Jointly Optimizing Translations and Speech Timing to Improve Isochrony in Automatic Dubbing. ArXiv:2302.12979.

J. S. Chung and A. Zisserman. 2016. Out of time: automated lip sync in the wild. In *Workshop on Multi-view Lip-reading, ACCV*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. 2023. Fleurs: Few-shot learning evaluation of universal representations of speech. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 798–805. IEEE.

Marta R. Costa-jussà, Marcos Zampieri, and Santanu Pal. 2018. A neural approach to language variety translation. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018)*, pages 275–282, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Pan Deng, Shihao Chen, Weitai Zhang, Jie Zhang, and Lirong Dai. 2023. The USTC's Dialect Speech Translation System for IWSLT 2023. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT)*.

Mattia A. Di Gangi, Roldano Cattoni, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2019a. 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.

Mattia A. Di Gangi, Roldano Cattoni, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2019b. 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.

Yichao Du, Guo Zhengsheng, Jinchuan Tian, Zhirui Zhang, Xing Wang, Jianwei Yu, Zhaopeng Tu, Tong Xu, and Enhong Chen. 2023. The MineTrans Systems for IWSLT 2023 Offline Speech Translation and Speech-to-Speech Translation Tasks. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT)*.

Matthias Eck and Chiori Hori. 2005. Overview of the IWSLT 2005 evaluation campaign. In *Proceedings of the International Workshop on Spoken Language Translation*, pages 1–22, Pittsburgh, PA.

ELRA catalogue. 2016a. Trad pashto broadcast news speech corpus. https://catalogue.elra.info/en-us/repository/browse/ELRA-S0381/. ISLRN: 918-508-885-913-7, ELRA ID: ELRA-S0381.

ELRA catalogue. 2016b. Trad pashto-french parallel corpus of transcribed broadcast news speech - training data. http://catalog.elda.org/en-us/repository/browse/ELRA-W0093/. ISLRN: 802-643-297-429-4, ELRA ID: ELRA-W0093.

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.

Marcello Federico, Luisa Bentivogli, Michael Paul, and Sebastian Stüker. 2011. Overview of the IWSLT 2011 Evaluation Campaign. In *Proceedings of the International Workshop on Spoken Language Translation*, pages 11–27, San Francisco, USA.

Marcello Federico, Mauro Cettolo, Luisa Bentivogli, Michael Paul, and Sebastian Stüker. 2012. Overview of the IWSLT 2012 Evaluation Campaign. In *Proceedings of the International Workshop on Spoken Language Translation*, pages 11–27, Hong Kong, HK.

F. Feng, Y. Yang, D. Cer, N. Arivazhagan, and W. Wang. 2022. Language-agnostic BERT Sentence Embedding. In *Proceedings of the 60th ACL*.

Cameron Shaw Fordyce. 2007. Overview of the IWSLT 2007 evaluation campaign. In *Proceedings of the International Workshop on Spoken Language Translation*, pages 1–12, Trento, Italy.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021a. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, George Foster, Alon Lavie, and Ondřej Bojar. 2021b. Results of the WMT21 metrics shared task: Evaluating metrics with expert-based human evaluations on TED and news domain. In *Proceedings of the Sixth Conference on Machine Translation*, pages 733–774, Online. Association for Computational Linguistics.

Ryo Fukuda, Yuta Nishikawa, Yasumasa Kano, Yuka Ko, Tomoya Yanagita, Kosuke Doi, Mana Makinae, Sakriani Sakti, Katsuhito Sudoh, and Satoshi Nakamura. 2023. NAIST Simultaneous Speech-to-speech Translation System for IWSLT 2023. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT)*.

Mercedes García-Martínez, Loïc Barrault, and Fethi Bougares. 2016. Factored neural machine translation architectures. In *Proceedings of the 13th International Conference on Spoken Language Translation*, Seattle, Washington D.C. International Workshop on Spoken Language Translation.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin. 2017. Convolutional sequence to sequence learning.

Karthik Gopalakrishnan, Behnam Hedayatnia, Qinlang Chen, Anna Gottardi, Sanjeev Kwatra, Anu Venkatesh, Raefer Gabriel, and Dilek Hakkani-Tür. 2019. Topical-Chat: Towards knowledge-grounded open-domain conversations. In *Proc. Interspeech 2019*, pages 1891–1895.

Edward Gow-Smith, Alexandre Berard, Marcely Zanon Boito, and Ioan Calapodescu. 2023. NAVER LABS Europe's Multilingual Speech Translation Systems for the IWSLT 2023 Low-Resource Track. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT)*.

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. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. 2020. Conformer: Convolution-augmented transformer for speech recognition. *Interspeech*, pages 5036–5040.

Jiaxin Guo, Daimeng Wei, Zhanglin Wu, Zongyao Li, Zhiqiang Rao, Minghan Wang, Hengchao Shang, Xiaoyu Chen, Zhengzhe Yu, Shaojun Li, Yuhao Xie, Lizhi Lei, and Hao Yang. 2023. The HW-TSC's Simultaneous Speech-to-Text Translation system for IWSLT 2023 evaluation. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT)*.

Yuchen Han, Xiaoqian Liu, Hao Chen, Yuhao Zhang, Chen Xu, Tong Xiao, and Jingbo Zhu. 2023. The NiuTrans End-to-End Speech Translation System for IWSLT23 English-to-Chinese Offline Task. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT)*.

Fei He, Shan-Hui Cathy Chu, Oddur Kjartansson, Clara Rivera, Anna Katanova, Alexander Gutkin, Isin Demirsahin, Cibu Johny, Martin Jansche, Supheakmungkol Sarin, and Knot Pipatsrisawat. 2020. Open-source multi-speaker speech corpora for building Gujarati, Kannada, Malayalam, Marathi, Tamil and Telugu speech synthesis systems. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6494–6503, Marseille, France. European Language Resources Association.

Kevin Heffernan, Onur Çelebi, and Holger Schwenk. 2022. Bitext mining using distilled sentence representations for low-resource languages. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2101–2112, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Carlos Daniel Hernandez Mena, Albert Gatt, Andrea DeMarco, Claudia Borg, Lonneke van der Plas, Amanda Muscat, and Ian Padovani. 2020. MASRI-HEADSET: A Maltese corpus for speech recognition. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6381–6388, Marseille, France. European Language Resources Association.

Oleksii Hrinchuk, Vladimir Bataev, Evelina Bakhturina, and Boris Ginsburg. 2023. NVIDIA NeMo Offline Speech Translation Systems for IWSLT 2023. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT)*.

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 and Lang. Proc.*, 29:3451–3460.

Chenxu Hu, Qiao Tian, Tingle Li, Wang Yuping, Yuxuan Wang, and Hang Zhao. 2021. Neural dubber: Dubbing for videos according to scripts. In *Thirty-Fifth Conference on Neural Information Processing Systems*.

Wuwei Huang, Mengge Liu, Xiang Li, Yanzhi Tian, Fengyu Yang, Wen Zhang, Jian Luan, Bin Wang, Yuhang Guo, and Jinsong Su. 2023. The Xiaomi AI Lab's Speech Translation Systems for IWSLT 2023 Offline Task, Simultaneous Task and Speech-to-Speech Task. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT)*.

Amir Hussein, Cihan Xiao, Neha Verma, Matthew Wiesner, Thomas Thebaud, and Sanjeev Khudanpur. 2023. JHU IWSLT 2023 Dialect Speech Translation System Description. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT)*.

Muhammad Huzaifah, Kye Min Tan, and Richeng Duan. 2023. I2R's End-to-End Speech Translation System for IWSLT 2023 Offline Shared Task. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT)*.

Hirofumi Inaguma, Brian Yan, Siddharth Dalmia, Pengcheng Guo, Jiatong Shi, Kevin Duh, and Shinji Watanabe. 2021. ESPnet-ST IWSLT 2021 Offline Speech Translation System. In *Proceedings of the 18th International Conference on Spoken Language Translation (IWSLT)*.

Javier Iranzo-Sánchez, Joan Albert Silvestre-Cerdà, Javier Jorge, Nahuel Roselló, Adrià Giménez, Albert Sanchis, Jorge Civera, and Alfons Juan. 2020. Europarl-st: A multilingual corpus for speech translation of parliamentary debates. In *Proc. of 45th Intl. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2020)*, pages 8229–8233, Barcelona (Spain).

Dávid Javorský, Dominik Macháček, and Ondřej Bojar. 2022. Continuous rating as reliable human evaluation of simultaneous speech translation. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 154–164, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Japan Translation Federation JTF. 2018. JTF Translation Quality Evaluation Guidelines, 1st Edition (in Japanese).

Yasumasa Kano, Katsuhito Sudoh, and Satoshi Nakamura. 2023. Average Token Delay: A Latency Metric for Simultaneous Translation. In *Proceedings of Interspeech 2023*. To appear.

Alina Karakanta, Luisa Bentivogli, Mauro Cettolo, Matteo Negri, and Marco Turchi. 2022a. Post-editing in automatic subtitling: A subtitlers' perspective. In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pages 261–270, Ghent, Belgium. European Association for Machine Translation.

Alina Karakanta, François Buet, Mauro Cettolo, and François Yvon. 2022b. Evaluating subtitle segmentation for end-to-end generation systems. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3069–3078, Marseille, France. European Language Resources Association.

Santosh Kesiraju, Karel Beneš, Maksim Tikhonov, and Jan Černocký. 2023. BUT Systems for IWSLT 2023 Marathi - Hindi Low Resource Speech Translation Task. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT)*.

Sameer Khurana, Antoine Laurent, and James Glass. 2022. Samu-xlsr: Semantically-aligned multimodal utterance-level cross-lingual speech representation. *IEEE Journal of Selected Topics in Signal Processing*, pages 1–13.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72, Vancouver, Canada. Association for Computational Linguistics.

Surafel M Lakew, Yogesh Virkar, Prashant Mathur, and Marcello Federico. 2021. Isometric mt: Neural machine translation for automatic dubbing. *arXiv preprint arXiv:2112.08682*.

Surafel Melaku Lakew, Mattia Di Gangi, and Marcello Federico. 2019. Controlling the output length of neural machine translation. In *Proc. IWSLT*.

Antoine Laurent, Souhir Gahbiche, Ha Nguyen, Haroun Elleuch, Fethi Bougares, Antoine Thiol, Hugo Riguidel, Salima Mdhaffar, Gaëlle Laperrière, Lucas Maison, Sameer Khurana, and Yannick Estève. 2023. ON-TRAC consortium systems for the IWSLT 2023 dialectal and low-resource speech translation tasks. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT)*.

Seugnjun Lee, Hyeonseok Moon, Chanjun Park, and Heuiseok Lim. 2023. Improving Formality-Sensitive Machine Translation using Data-Centric Approaches and Prompt Engineering. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT)*.

Zongyao Li, Zhanglin Wu, Zhiqiang Rao, Xie YuHao, Guo JiaXin, Daimeng Wei, Hengchao Shang, Wang Minghan, Xiaoyu Chen, Zhengzhe YU, Li Shao-Jun, Lei LiZhi, and Hao Yang. 2023. HW-TSC at IWSLT2023: Break the Quality Ceiling of Offline Track via Pre-Training and Domain Adaptation. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT)*.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, et al. 2021. Few-shot learning with multilingual language models. *arXiv preprint arXiv:2112.10668*.

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929, Portorož, Slovenia. European Language Resources Association (ELRA).

Danni Liu, Thai Binh Nguyen, Sai Koneru, Enes Yavuz Ugan, Ngoc-Quan Pham, Tuan Nam Nguyen, Tu Anh Dinh, Carlos Mullov, Alexander Waibel, and Jan Niehues. 2023. KIT's Multilingual Speech Translation System for IWSLT 2023. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT)*.

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.

Arle Lommel, Hans Uszkoreit, and Aljoscha Burchardt. 2014. Multidimensional Quality Metrics (MQM): A Framework for Declaring and DescribingTranslation Quality Metrics. *Revista Tradumàtica: tecnologies de la traducció*, 12:455–463.

Mingbo Ma, Liang Huang, Hao Xiong, Renjie Zheng, Kaibo Liu, Baigong Zheng, Chuanqiang Zhang, Zhongjun He, Hairong Liu, Xing Li, Hua Wu, and Haifeng Wang. 2019. STACL: Simultaneous translation with implicit anticipation and controllable latency using prefix-to-prefix framework. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3025–3036, Florence, Italy. Association for Computational Linguistics.

Shuming Ma, Li Dong, Shaohan Huang, Dongdong Zhang, Alexandre Muzio, Saksham Singhal, Hany Hassan Awadalla, Xia Song, and Furu Wei. 2021. DeltaLM: Encoder-decoder pre-training for language generation and translation by augmenting pretrained multilingual encoders. *arXiv*.

Xutai Ma, Mohammad Javad Dousti, Changhan Wang, Jiatao Gu, and Juan Pino. 2020a. SIMULEVAL: An evaluation toolkit for simultaneous translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 144–150, Online. Association for Computational Linguistics.

Xutai Ma, Juan Pino, and Philipp Koehn. 2020b. SimulMT to SimulST: Adapting simultaneous text translation to end-to-end simultaneous speech translation. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 582–587, Suzhou, China. Association for Computational Linguistics.

Dominik Macháček, Ondřej Bojar, and Raj Dabre. 2023. MT Metrics Correlate with Human Ratings of Simultaneous Speech Translation. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT)*.

Evgeny Matusov, Gregor Leusch, Oliver Bender, and Hermann Ney. 2005a. Evaluating machine translation output with automatic sentence segmentation. In *Proc. of the International Workshop on Spoken Language Translation (IWSLT)*, pages 138–144.

Evgeny Matusov, Gregor Leusch, Oliver Bender, and Hermann Ney. 2005b. Evaluating machine translation output with automatic sentence segmentation. In *Proceedings of the Second International Workshop on Spoken Language Translation*, Pittsburgh, Pennsylvania, USA.

Evgeny Matusov, Patrick Wilken, and Yota Georgakopoulou. 2019. Customizing neural machine translation for subtitling. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 82–93, Florence, Italy. Association for Computational Linguistics.

Jonathan Mbuya and Antonios Anastasopoulos. 2023. GMU Systems for the IWSLT 2023 Dialect and Low-resource Speech Translation Tasks. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT)*.

Maria Nadejde, Anna Currey, Benjamin Hsu, Xing Niu, Marcello Federico, and Georgiana Dinu. 2022. CoCoA-MT: A dataset and benchmark for contrastive controlled MT with application to formality. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 616–632, Seattle, United States. Association for Computational Linguistics.

J. Niehues, R. Cattoni, S. Stüker, M. Negri, M. Turchi, T. Ha, E. Salesky, R. Sanabria, L. Barrault, L. Specia, and M. Federico. 2019. The IWSLT 2019 Evaluation Campaign. In *Proceedings of the 16th International Workshop on Spoken Language Translation (IWSLT 2019)*, Hong Kong, China.

Jan Niehues, Roldano Cattoni, Sebastian Stüker, Mauro Cettolo, Marco Turchi, and Marcello Federico. 2018. The IWSLT 2018 Evaluation Campaign. In *Proceedings of the 15th International Workshop on Spoken Language Translation (IWSLT 2018)*, pages 2–6, Bruges, Belgium.

Xing Niu, Marianna Martindale, and Marine Carpuat. 2017. A study of style in machine translation: Controlling the formality of machine translation output. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2814–2819, Copenhagen, Denmark. Association for Computational Linguistics.

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. *arXiv preprint*.

John E Ortega, Richard Castro Mamani, and Kyunghyun Cho. 2020. Neural machine translation with a polysynthetic low resource language. *Machine Translation*, 34(4):325–346.

John E. Ortega, Rodolfo Zevallos, and William Chen. 2023. QUESPA Submission for the IWSLT 2023 Dialect and Low-resource Speech Translation Tasks. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT)*.

Proyag Pal, Brian Thompson, Yogesh Virkar, Prashant Mathur, Alexandra Chronopoulou, and Marcello Federico. 2023. Improving isochronous machine translation with target factors and auxiliary counters.

Sara Papi, Marco Gaido, and Matteo Negri. 2023. Direct Models for Simultaneous Translation and Automatic Subtitling: FBK@IWSLT2023. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT)*.

Sara Papi, Marco Gaido, Matteo Negri, and Marco Turchi. 2022. Over-generation cannot be rewarded: Length-adaptive average lagging for simultaneous speech translation. In *Proceedings of the Third Workshop on Automatic Simultaneous Translation*, pages 12–17, Online. Association for Computational Linguistics.

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

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002b. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. 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. *Interspeech 2019*.

Michael Paul. 2006. Overview of the IWSLT 2006 Evaluation Campaign. In *Proceedings of the International Workshop on Spoken Language Translation*, pages 1–15, Kyoto, Japan.

Michael Paul. 2008. Overview of the IWSLT 2008 Evaluation Campaign. In *Proceedings of the International Workshop on Spoken Language Translation*, pages 1–17, Waikiki, Hawaii.

Michael Paul. 2009. Overview of the IWSLT 2009 Evaluation Campaign. In *Proceedings of the International Workshop on Spoken Language Translation*, pages 1–18, Tokyo, Japan.

Michael Paul, Marcello Federico, and Sebastian Stüker. 2010. Overview of the IWSLT 2010 Evaluation Campaign. In *Proceedings of the International Workshop on Spoken Language Translation*, pages 3–27, Paris, France.

Simone Perone. 2023. Matesub: the Translated Subtitling Tool at the IWSLT2023 Subtitling task. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT)*.

Peter Polák, Danni Liu, Ngoc-Quan Pham, Jan Niehues, Alexander Waibel, and Ondřej Bojar. 2023. Towards Efficient Simultaneous Speech Translation: CUNI-KIT System for Simultaneous Track at IWSLT 2023. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT)*.

Peter Polák, Ngoc-Quan Pham, Tuan Nam Nguyen, Danni Liu, Carlos Mullov, Jan Niehues, Ondřej Bojar, and Alexander Waibel. 2022. CUNI-KIT system for simultaneous speech translation task at IWSLT 2022. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, pages 277–285, Dublin, Ireland (in-person and online). Association for Computational Linguistics.

Maja Popović. 2015a. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Maja Popović. 2015b. chrf: character n-gram f-score for automatic mt evaluation. In *Proceedings of the tenth workshop on statistical machine translation*, pages 392–395.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Vineel Pratap, Awni Hannun, Qiantong Xu, Jeff Cai, Jacob Kahn, Gabriel Synnaeve, Vitaliy Liptchinsky, and Ronan Collobert. 2019. Wav2letter++: A fast open-source speech recognition system. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6460–6464. IEEE.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision.

Balaji Radhakrishnan, Saurabh Agrawal, Raj Prakash Gohil, Kiran Praveen, Advait Vinay Dhopeshwarkar, and Abhishek Pandey. 2023. SRI-B's systems for IWSLT 2023 Dialectal and Low-resource track: Marathi-Hindi Speech Translation. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT)*.

Zhiqiang Rao, Hengchao Shang, Jinlong Yang, Daimeng Wei, Zongyao Li, Lizhi Lei, and Hao Yang. 2023. Length-Aware NMT and Adaptive Duration for Automatic Dubbing. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT)*.

Ricardo Rei, José GC de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André FT Martins. 2022. Comet-22: Unbabel-ist 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585.

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

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020b. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. 2022. Fastspeech 2: Fast and high-quality end-to-end text to speech.

Anthony Rousseau, Paul Deléglise, and Yannick Esteve. 2014. Enhancing the ted-lium corpus with selected data for language modeling and more ted talks. In *LREC*.

Elizabeth Salesky, Kareem Darwish, Mohamed Al-Badrashiny, Mona Diab, and Jan Niehues. 2023. Evaluating Multilingual Speech Translation Under Realistic Conditions with Resegmentation and Terminology. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*. Association for Computational Linguistics.

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 *Proc. Interspeech 2021*, pages 3655–3659.

Gabriele Sarti, Phu Mon Htut, Xing Niu, Benjamin Hsu, Anna Currey, Georgiana Dinu, and Maria Nadejde. 2023. RAMP: Retrieval and attribute-marking enhanced prompting for attribute-controlled translation.

Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, Armand Joulin, and Angela Fan. 2021. 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 (Volume 1: Long Papers)*, pages 6490–6500, Online. Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Hengchao Shang, Zhiqiang Rao, Zongyao Li, Zhanglin Wu, Jiaxin Guo, Minghan Wang, Daimeng Wei, Shaojun Li, Zhengzhe Yu, Xiaoyu Chen, Lizhi Lei, and Hao Yang. 2023. The HW-TSC's Simultaneous Speech-to-Speech Translation system for IWSLT 2023 evaluation. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT)*.

Yao Shi, Hui Bu, Xin Xu, Shaoji Zhang, and Ming Li. 2020. Aishell-3: A multi-speaker mandarin tts corpus and the baselines. *arXiv preprint arXiv:2010.11567*.

Kun Song, Yi Lei, Peikun Chen, Yiqing Cao, Kun Wei, Yongmao Zhang, Lei Xie, Ning Jiang, and Guoqing Zhao. 2023. The NPU-MSXF Speech-to-Speech Translation System for IWSLT 2023 Speech-to-Speech Translation Task. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT)*.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning. *arXiv preprint arXiv:2008.00401*.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

Ioannis Tsiamas, Gerard I. Gállego, Jose Fonollosa, and Marta R. Costa-jussà. 2023. Speech Translation with Foundation Models and Optimal Transport: UPC at IWSLT23. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT)*.

Ioannis Tsiamas, Gerard I. Gállego, José A. R. Fonollosa, and Marta R. Costa-jussà. 2022. SHAS: Approaching optimal Segmentation for End-to-End Speech Translation. In *Proc. Interspeech 2022*, pages 106–110.

Priyesh Vakharia, Shree Vignesh S, Pranjali Basmatkar, and Ian Lane. 2023. Low-Resource Formality Controlled NMT Using Pre-trained LM. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT)*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All You Need. In *Proceedings of NIPS 2017*.

Akshaya Vishnu, Kudlu Shanbhogue, Ran Xue, Soumya Saha, Daniel Zhang, and Ashwinkumar Ganesan. 2023. Amazon Alexa AI's Low-Resource Speech Translation System for IWSLT2023. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT)*.

Changhan Wang, Juan Pino, Anne Wu, and Jiatao Gu. 2020a. Covost: A diverse multilingual speech-to-text translation corpus. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 4197–4203.

Changhan Wang, Yun Tang, Xutai Ma, Anne Wu, Dmytro Okhonko, and Juan Pino. 2020b. fairseq s2t: Fast speech-to-text modeling with fairseq. *arXiv preprint arXiv:2010.05171*.

Changhan Wang, Anne Wu, Jiatao Gu, and Juan Pino. 2021. CoVoST 2 and Massively Multilingual Speech Translation. In *Proc. Interspeech 2021*, pages 2247–2251.

Minghan Wang, Yinglu Li, Jiaxin Guo, Zongyao Li, Hengchao Shang, Daimeng Wei, Min Zhang, Shimin Tao, and Hao Yang. 2023a. The HW-TSC's Speech-to-Speech Translation System for IWSLT 2023. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT)*.

Zhipeng Wang, Yuhang Guo, and Shuoying Chen. 2023b. BIT's System for Multilingual Track. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT)*.

Patrick Wilken, Panayota Georgakopoulou, and Evgeny Matusov. 2022. SubER - a metric for automatic evaluation of subtitle quality. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, pages 1–10, Dublin, Ireland (in-person and online). Association for Computational Linguistics.

Aiden Williams. 2022. The applicability of Wav2Vec 2.0 for low-resource Maltese ASR. B.S. thesis, University of Malta.

Aiden Williams, Kurt Abela, Rishu Kumar, Martin Bär, Hannah Billinghurst, Kurt Micallef, Ahnaf Mozib Samin, Andrea DeMarco, Lonneke van der Plas, and Claudia Borg. 2023. UM-DFKI Maltese Speech Translation. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT)*.

Zhanglin Wu, Zongyao Li, Daimeng Wei, Hengchao Shang, Jiaxin Guo, Xiaoyu Chen, Zhiqiang Rao, Zhengzhe YU, Jinlong Yang, Shaojun Li, Yuhao Xie, Bin Wei, Jiawei Zheng, Ming Zhu, Lizhi Lei, Hao Yang, and Yanfei Jiang. 2023. Improving Neural Machine Translation Formality Control with Domain Adaptation and Reranking-based Transductive Learning. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT)*.

Zhihang Xie. 2023. The BIGAI Offline Speech Translation Systems for IWSLT 2023 Evaluation. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT)*.

Henry Li Xinyuan, Neha Verma, Bismarck Bamfo Odoom, Ujvala Pradeep, Matthew Wiesner, and Sanjeev Khudanpur. 2023. JHU IWSLT 2023 Multilingual Speech Translation System Description. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT)*.

Chen Xu, Bojie Hu, Yanyang Li, Yuhao Zhang, shen huang, Qi Ju, Tong Xiao, and Jingbo Zhu. 2021. Stacked acoustic-and-textual encoding: Integrating the pre-trained models into speech translation encoders.

Wenda Xu, Xian Qian, Mingxuan Wang, Lei Li, and William Yang Wang. 2022. Sescore2: Retrieval augmented pretraining for text generation evaluation. *arXiv preprint arXiv:2212.09305*.

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)*.

Zhengdong Yang, Shuichiro Shimizu, Sheng Li Wangjin Zhou, and Chenhui Chu. 2023. The Kyoto Speech-to-Speech Translation System for IWSLT 2023. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT)*.

Zhuoyuan Yao, Di Wu, Xiong Wang, Binbin Zhang, Fan Yu, Chao Yang, Zhendong Peng, Xiaoyu Chen, Lei Xie, and Xin Lei. 2021. Wenet: Production oriented streaming and non-streaming end-to-end speech recognition toolkit. *arXiv preprint arXiv:2102.01547*.

Rong Ye, Chengqi Zhao, Tom Ko, Chutong Meng, Tao Wang, Mingxuan Wang, and Jun Cao. 2023. Gigast: A 10,000-hour pseudo speech translation corpus. In *Interspeech 2023*.

Xinyuan Zhou, Jianwei Cui, Zhongyi Ye, Yichi Wang, Luzhen Xu, Hanyi Zhang, Weitai Zhang, and Lirong Dai. 2023. Submission of USTC's system for the IWSLT 2023 - Offline Speech Translation Track. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT)*.

Adrian Łańcucki. 2021. Fastpitch: Parallel text-to-speech with pitch prediction. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6588–6592, Toronto, Canada. IEEE.

Appendix A. Human Evaluation

A Human Evaluation

Human evaluation was carried out for the Simultaneous and Offline SLT shared tasks. At the time of writing, only the former evaluation has been completed which is reported here. The human evaluation of the Offline Task will be recounted during the conference and possibly in an update version of this report.

A.1 Simultaneous Speech Translation Task

Simultaneous Speech Translation Task ran two different types of manual evaluation: "continuous rating" for English-to-German and MQM for English-to-Japanese.

A.1.1 Human Evaluation for the English-to-German Simultaneous Task

We used a variant of "continuous rating" as presented by Javorský et al. (2022). The evaluation process and the guidelines presented to annotators were the same as during the last year evaluation (consult Section A.1.1 in Anastasopoulos et al. (2022a) for more details).

Time Shift for Better Simultaneity Last year, we reduced the delay by shifting the subtitles ahead in time to ease the memory overload of the evaluators. Since this year only a low latency regime was used, we left the subtitles intact for the system outputs. For interpreting, we used the same shift as last year.

Two Test Sets: Common and Non-Native The main part of the test set for the English-to-German task was the Common test set. The Common test set is a new instance (different from previous years) consisting of selected TED talks and it serves both in the Offline Speech Translation task as well as in the Simultaneous Translation task. Following the last year, we also added the Non-Native part that was created and is in use since IWSLT 2020 Non-Native Translation Task. The Non-Native part is described in Ansari et al. (2020) Appendix A.6.

We show the size of the corpus, as well as the amount of annotation collected in Table 21.

Processing of Collected Rankings Once the results are collected, they are processed as follows. We first inspect the timestamps on the ratings, and remove any ratings that have timestamps more than 20 seconds greater than the length of the audio. Because of the natural delay (even with the time-shift) and because the collection process is subject to network and computational constraints, there can be ratings that are timestamped greater than the audio length. If the difference is however too high, we judge it to be an annotation error. We also remove any annotated audio where there is fewer than one rating per 20 seconds, since the annotators were instructed to annotate every 5-10 seconds.

Obtaining Final Scores To calculate a score for each system, we average the ratings across each annotated audio,[47] then average across the multiple annotations for each audio to obtain a system score for that audio. Finally we average across all audios to obtain a score for each system. This type of averaging renders all input speeches equally important and it is not affected by the speech length.

We show the results in Table 22. We observe that all systems perform better on the Common part of the test set than on the Non-Native one. The difference in scores between the best and the worst system is not so significant: It makes only ~0.3. When examining the evaluation of Non-Native audios, we can see that best systems on the Common part are worst on Non-Native. Given that the quality of the recordings in the non-native part is low on average and the speakers are not native, we hypothesize that systems with worse performance on Common part are more robust. Such systems then achieve an increased performance given noisy inputs.

A.1.2 Human Evaluation for the English-to-Japanese Simultaneous Task

For the English-to-Japanese Simultaneous Translation Task, we conducted a human evaluation using a variant of Multidimensional Quality Metrics (MQM; Lommel et al., 2014). MQM has been used in recent MT evaluation studies (Freitag et al., 2021a) and WMT Metrics shared task (Freitag et al., 2021b). For the evaluation of Japanese translations, we used *JTF Translation Quality Evaluation Guidelines* (JTF,

[47] Note that the ratings could be also weighted with respect to the duration of time segments between the ratings but Macháček et al. (2023) documented on 2022 data that the difference is negligible.

2018), distributed by Japan Translation Federation (JTF). The guidelines are based on MQM but include some modifications in consideration of the property of the Japanese language.

We hired a Japanese-native professional interpreter as the evaluator, while the evaluator was a translator in the last year (Anastasopoulos et al., 2022a). The evaluator checked translation hypotheses along with their source speech transcripts and chose the corresponding error category and severity for each translation hypothesis using a spreadsheet. Here, we asked the evaluator to focus only on *Accuracy* and *Fluency* errors, because other types of errors in Terminology, Style, and Locale convention would not be so serious in the evaluation of simultaneous translation. Finally, we calculated the cumulative error score for each system based on the error weighting presented by Freitag et al. (2021a), where *Critical* and *Major* errors are not distinguished.

Appendix B. Automatic Evaluation Results and Details

B.1 Offline SLT

· Systems are ordered according to the BLEU score computed on the concatenation of the three test sets (Joint BLEU, third column).
· The "D" column indicates the data condition in which each submitted run was trained, namely: Constrained (C), constrained^{+LLM} (C$^+$), Unconstrained (U).
· For the BLEU scores computed on the TED test set, "Orig" and "New" respectively indicate the results computed on the original (subtitle-like) TED translations and the unconstrained (exact, more literal) translations as references.
· Direct systems are indicated by gray background.
· "*" indicates a late submission.
· "+" indicates an unofficial submission.

System	D	Joint		TED					ACL		EPTV	
		BLEU	COMET	BLEU			COMET		BLEU	COMET	BLEU	COMET
Ref				New	Orig	Both	New	Orig				
HW-TSC	C	32.4	0.8213	34.8	30.2	42.1	0.8327	0.8208	38.1	0.8090	16.7	0.3829
HW-TSC	U	32.3	0.8209	34.9	30.9	42.4	0.8331	0.8223	36.9	0.8073	16.9	0.3819
HW-TSC	C$^+$	31.9	0.8210	34.4	30.6	41.9	0.8332	0.8230	37.2	0.8063	16.8	0.3823
NeuroDub$^+$	U	30.4	0.8089	31.8	25.8	38.5	0.8205	0.8082	41.1	0.7956	15.4	0.3784
NeMo	C	28.5	0.7759	30.5	26.4	37.7	0.7977	0.7871	31.9	0.7171	15.6	0.3680
UPC	C$^+$	27.9	0.7892	29.8	25.5	36.6	0.8098	0.7985	32.1	0.7473	15.6	0.3746
I2R	C$^+$	22.4	0.7070	24.0	20.3	29.5	0.7248	0.7172	23.9	0.6841	13.3	0.3506
BIGAI*	C$^+$	20.3	0.6945	22.3	19.3	27.4	0.7128	0.7055	19.6	0.6295	11.5	0.3555

Table 14: Official results of the automatic evaluation for the Offline Speech Translation Task, **English to German**.

System	D	Joint		TED					ACL	
		BLEU	COMET	BLEU			COMET		BLEU	COMET
Ref				New	Orig	Both	New	Orig		
HW-TSC	U	21.0	0.8177	18.8	22.6	29.1	0.8111	0.8029	30.7	0.8473
HW-TSC	C	20.9	0.8181	18.7	22.7	29.0	0.8123	0.8042	30.1	0.8443
HW-TSC	C$^+$	20.9	0.8177	18.7	22.6	28.9	0.8114	0.8034	30.7	0.8463
NeMo	C	18.1	0.7741	16.5	20.4	25.6	0.7734	0.7666	24.9	0.7769
BIGAI*	C$^+$	10.7	0.7122	10.7	13.2	16.8	0.7201	0.7228	10.4	0.6769

Table 15: Official results of the automatic evaluation for the Offline Speech Translation Task, **English to Japanese**.

System	D	Joint		TED					ACL	
		BLEU	COMET	BLEU			COMET		BLEU	COMET
Ref				New	Orig	Both	New	Orig		
USTC	U	54.7	0.8627	53.9	36.8	62.1	0.8648	0.7992	58.0	0.8535
USTC	U	52.8	0.8357	52.9	35.5	60.6	0.8439	0.7798	52.5	0.7999
HW-TSC	C	51.1	0.8499	50.6	34.5	57.8	0.8521	0.7876	53.0	0.8404
HW-TSC	C$^+$	51.1	0.8494	50.6	34.5	57.9	0.8514	0.7870	53.0	0.8406
HW-TSC	U	51.0	0.8497	50.6	34.5	57.8	0.8519	0.7874	52.8	0.8401
NiuTrans	C	49.4	0.8255	50.0	34.3	57.9	0.8376	0.7740	47.1	0.7733
Xiaomi	C$^+$	47.1	0.8279	47.2	32.4	54.1	0.8375	0.7773	46.5	0.7866
NeMo	C	45.6	0.8032	46.5	31.8	53.8	0.8177	0.7575	41.8	0.7404
MineTrans	U	45.0	0.7920	46.3	32.0	53.2	0.8134	0.7546	39.9	0.6997
BIGAI*	C$^+$	31.9	0.7260	33.0	23.3	38.6	0.7428	0.7014	27.4	0.6534
MineTrans	C	28.7	0.6371	27.7	18.6	32.2	0.6375	0.5976	31.8	0.6354

Table 16: Official results of the automatic evaluation for the Offline Speech Translation Task, **English to Chinese**.

B.2 Simultaneous SLT

Team	BLEU	LAAL	AL	AP	DAL	ATD
Common						
HW-TSC	29.63	2.26 (3.93)	2.11 (3.86)	0.83 (1.59)	3.17 (8.99)	2.28 (6.77)
CUNI-KIT	28.51	2.35 (3.63)	2.24 (3.56)	0.79 (1.11)	2.88 (4.50)	2.26 (2.96)
FBK	28.38	2.25 (2.99)	2.09 (2.88)	0.84 (1.03)	2.70 (3.65)	2.15 (2.48)
NAIST	26.05	2.36 (3.30)	2.22 (3.21)	0.82 (1.07)	3.05 (4.45)	2.25 (3.06)
CMU	25.78	1.99 (3.39)	1.92 (3.33)	0.82 (1.31)	3.78 (6.56)	2.46 (4.63)
Non-Native						
NAIST	22.96	2.43 (3.52)	1.95 (3.22)	0.845 (1.02)	3.37 (4.71)	3.13 (3.92)
CMU	22.84	2.47 (3.74)	2.36 (3.63)	0.798 (1.16)	4.54 (6.77)	3.77 (5.47)
CUNI-KIT	19.94	3.42 (5.00)	3.24 (4.87)	0.744 (1.04)	4.14 (5.87)	3.82 (4.84)
HW-TSC	17.91	3.57 (6.67)	3.44 (6.61)	0.705 (1.65)	4.39 (12.91)	4.04 (11.13)
FBK	15.19	4.10 (5.34)	3.94 (5.22)	0.89 (1.12)	4.53 (5.85)	3.76 (4.65)

Table 17: Simultaneous Speech-to-Text Translation, English to German. Except for AP, the latency is measured in seconds. Numbers in brackets are computation aware latency.

Team	BLEU	LAAL	AL	AP	DAL	ATD
HW-TSC	44.95	2.13 (3.80)	2.06 (3.76)	0.78 (1.48)	3.21 (8.66)	0.99 (5.31)
CUNI-KIT	44.16	2.13 (3.30)	2.06 (3.25)	0.77 (1.08)	2.78 (4.38)	0.89 (1.54)
XIAOMI	43.69	2.30 (3.03)	2.23 (2.98)	0.80 (1.08)	2.93 (4.08)	0.90 (1.47)
NAIST	36.80	2.00 (2.80)	1.88 (2.74)	0.76 (1.03)	2.66 (4.22)	0.77 (1.49)

Table 18: Simultaneous Speech-to-Text Translation, English to Chinese. Except for AP, the latency is measured in seconds. Numbers in brackets are computation aware latency.

Team	BLEU	LAAL	AL	AP	DAL	ATD
HW-TSC	16.63	2.60 (4.38)	2.56 (4.36)	0.71 (1.31)	3.62 (9.07)	0.83 (5.12)
CUNI-KIT	14.92	2.20 (3.55)	2.16 (3.53)	0.68 (1.06)	2.74 (5.17)	0.53 (1.50)
NAIST	14.66	2.52 (3.43)	2.45 (3.39)	0.75 (1.03)	3.24 (5.16)	0.60 (1.57)

Table 19: Simultaneous Speech-to-Text Translation, English to Japanese. Except for AP, the latency is measured in seconds. Numbers in brackets are computation aware latency.

Target Language	Team	ASR BLEU	BLASER	Start Offset	End Offset	ATD
German	CMU	22.62	0.122	2.37	5.21	4.22
	HW-TSC	19.74	-0.442	2.04	5.09	3.75
Japanese	HW-TSC	15.53	-1.70	2.37	3.48	3.56
	NAIST	10.19	-1.68	2.58	4.32	3.49
Chinese	HW-TSC	31.68	-0.696	1.92	3.12	3.23

Table 20: Simultaneous Speech-to-Speech from English Speech. The latency is measured in seconds. The BLEU scores are computed based on transcript from the default Whisper (Radford et al., 2022) ASR model for each language direction.

	Common	Non-native
Number of audios	42	43
Mean audio length (seconds)	400.3	208.8
Mean ratings per audio	65.6	36.5

Table 21: Human evaluation for the English-to-German task on two test sets: the Common one (used also in automatic scoring) and the Non-native one. We show the size of the test sets, and the number of ratings collected. On average, our annotators provide a quality judgement ever 6 seconds.

	Common	Non-native
CUNI-KIT	$3.10_{3.04\rightarrow3.16}$	$1.63_{1.54\rightarrow1.72}$
FBK	$3.08_{3.02\rightarrow3.14}$	$1.26_{1.20\rightarrow1.30}$
HWTSC	$2.91_{2.85\rightarrow2.98}$	$2.04_{1.92\rightarrow2.15}$
NAIST	$2.84_{2.78\rightarrow2.91}$	$2.27_{2.18\rightarrow2.34}$
CMU	$2.79_{2.72\rightarrow2.87}$	$2.38_{2.30\rightarrow2.46}$
Interpreter	–	$2.79_{2.71\rightarrow2.87}$

Table 22: Human evaluation results for English-to-German Simultaneous task on the 1–5 (worst-to-best) scale, with 95% confidence intervals. We calculate a mean score for each annotated audio file, then a mean across annotators (for each audio), then a mean across all audio files for each system. To compute confidence intervals, we take the scores for annotated audios, perform 10,000x bootstrap resampling, compute the mean score for each resample, then compute $[2.5, 97.5]$ percentiles across the resampled means.

Team	BLEU (on two talks)		Error score	Number of errors		
	TED ref.	Additional ref.		Critical	Major	Minor
HW-TSC	26.59	18.71	383	1	56	98
CUNI-KIT	24.21	17.95	384	0	56	104
NAIST	25.10	16.75	398	0	61	93
Baseline	7.69	6.27	1,074	3	205	34

Table 23: Human evaluation results on two talks (107 lines) in the English-to-Japanese Simultaneous speech-to-text translation task. Error weights are 5 for Critical and Major errors and 1 for Minor errors.

B.3 Automatic Subtitling

team	condition	system	domain	Subtitle quality		Translation quality			Subtitle compliance		
				SubER	Sigma	Bleu	ChrF	Bleurt	CPS	CPL	LPB
APPTEK	U	prmry	ALL	70.64	73.35	15.38	38.36	.4376	87.74	100.00	100.00
			ted	59.72	74.33	23.74	49.14	.5683	92.58	100.00	100.00
			eptv	73.98	67.09	15.81	45.21	.5229	86.65	100.00	100.00
			pltn	77.63	72.79	10.47	33.18	.4069	88.98	100.00	100.00
			itv	69.83	74.48	14.43	35.27	.4028	86.01	100.00	100.00
MATESUB	U	prmry	ALL	75.41	65.22	14.81	39.50	.4591	84.97	99.25	100.00
			ted	67.70	62.01	20.37	50.05	.5500	90.55	98.61	100.00
			eptv	87.04	57.73	12.08	43.59	.4705	88.59	99.20	100.00
			pltn	79.72	68.27	10.06	34.46	.4264	89.17	99.29	100.00
			itv	73.11	67.04	14.92	37.13	.4501	80.21	99.47	100.00
APPTEK	C	prmry	ALL	77.05	72.50	12.74	34.31	.3420	93.35	100.00	100.00
			ted	59.61	74.29	26.78	50.93	.5539	97.33	100.00	100.00
			eptv	76.25	68.49	14.43	42.37	.4604	95.76	100.00	100.00
			pltn	80.72	69.56	9.40	31.20	.3419	93.45	100.00	100.00
			itv	80.87	72.62	9.08	27.74	.2612	91.14	100.00	100.00
FBK	C	prmry	ALL	79.70	75.73	11.22	33.32	.3172	69.98	83.50	99.98
			ted	63.85	76.79	21.48	50.31	.5511	71.39	79.83	100.00
			eptv	79.76	69.04	13.20	42.69	.4722	74.95	82.08	99.91
			pltn	83.71	74.02	7.73	30.17	.3137	70.02	84.20	99.96
			itv	82.67	77.17	8.05	26.10	.2255	67.75	85.12	100.00
APPTEK	C	cntrstv	ALL	83.53	70.39	9.73	30.51	.2914	89.60	100.00	100.00
			ted	68.47	72.97	19.07	46.17	.4921	90.53	100.00	100.00
			eptv	81.69	66.36	11.46	39.25	.4150	94.57	100.00	100.00
			pltn	86.37	69.79	7.08	27.89	.2780	91.50	100.00	100.00
			itv	87.25	68.29	6.70	23.85	.2204	86.85	100.00	100.00

Table 24: Automatic evaluation results for the Subtitling Task: en→de. *C* and *U* stand for *constrained* and *unconstrained* training condition, respectively; *prmry* and *cntrstv* for *primary* and *contrastive* systems.

team	condition	system	domain	Subtitle quality		Translation quality			Subtitle compliance		
				SubER	Sigma	Bleu	ChrF	Bleurt	CPS	CPL	LPB
MATESUB	U	prmry	ALL	68.11	68.37	22.34	47.38	.5059	86.07	99.52	100.00
			ted	45.94	66.85	40.36	65.72	.7047	92.62	99.48	100.00
			eptv	74.47	59.59	21.06	54.11	.5728	90.15	99.44	100.00
			pltn	74.87	70.99	15.96	41.86	.4666	88.27	99.60	100.00
			itv	71.25	71.06	18.50	41.07	.4592	81.93	99.51	100.00
APPTEK	C	prmry	ALL	71.68	74.99	18.67	40.21	.3637	95.42	100.00	100.00
			ted	45.81	74.50	39.37	62.11	.6562	97.20	100.00	100.00
			eptv	66.60	73.31	23.57	51.94	.5379	96.27	100.00	100.00
			pltn	76.00	74.63	14.03	36.95	.3664	95.18	100.00	100.00
			itv	80.20	75.90	11.37	29.75	.2487	94.67	100.00	100.00
FBK	C	prmry	ALL	73.31	74.44	17.79	39.54	.3419	77.00	91.34	99.99
			ted	45.68	74.31	40.21	65.09	.6737	78.95	88.14	100.00
			eptv	68.47	69.63	23.92	52.19	.5490	79.81	88.05	100.00
			pltn	78.45	75.78	12.84	35.89	.3513	77.79	92.67	99.96
			itv	82.00	76.16	9.33	27.14	.2063	74.67	92.94	100.00

Table 25: Automatic evaluation results for the Subtitling Task: en→es. Legenda in Table 24.

B.4 Multilingual Speech Translation

Below we show the Multilingual task (§ 5) results and overall rankings, ordered according to the average chrF across all 10 target languages after resegmentation to the reference translations.

We also compare to the Offline submissions on the ACL 60-60 evaluation set on the 3 language pairs used for the Offline task.

Finally, we show the scores for each metric (chrF, COMET, BLEU) per language pair for all systems.

	System	Constrained?	chrF	COMET	BLEU	English WER
1	JHU$_{unconstrained}$		61.1	82.3	39.3	16.9
2	KIT$_{primary}$	✓ + LLM	57.5	77.0	34.9	23.7
3	KIT$_{contrastive1}$	✓ + LLM	57.5	76.8	34.8	—
4	KIT$_{contrastive2}$	✓ + LLM	56.4	76.5	34.0	—
5	KIT$_{contrastive4}$	✓ + LLM	56.2	76.4	33.7	—
6	KIT$_{contrastive3}$	✓ + LLM	55.9	76.3	33.5	—
7	KIT$_{contrastive5}$	✓ + LLM	54.5	76.7	31.7	—
8	KIT$_{contrastive7}$	✓ + LLM	53.9	76.6	31.1	—
9	KIT$_{contrastive6}$	✓ + LLM	53.7	75.9	30.9	—
10	JHU$_{constrained}$	✓ + LLM	48.1	65.3	24.5	34.1
11	BIT$_{primary}$	✓	31.0	51.7	11.7	—

Table 26: **Overall task ranking** with metrics averaged across **all ten** language pairs on the evaluation set. We show the official task metric (chrF) as well as the unofficial metrics (COMET, BLEU, and English WER). All metrics are calculated after resegmentation to reference transcripts and translations. Direct / end-to-end systems are highlighted in gray.

			de		ja		zh	
System	Task	Constrained?	COMET	BLEU	COMET	BLEU	COMET	BLEU
USTC	Off.						85.4 (1)	58.0 (1)
HW-TSC	Off.	✓	80.9 (2)	38.1 (3)	84.4 (3)	30.1 (7)	84.0 (2)	53.0 (2)
JHU	Mult.		81.3 (1)	41.2 (1)	84.7 (1)	33.9 (4)	82.0 (3)	46.5 (11)
HW-TSC	Off.		80.7 (3)	36.9 (6)	84.7 (1)	30.7 (6)	84.0 (2)	52.8 (3)
HW-TSC	Off.	✓ + LLM	80.6 (4)	37.2 (5)	84.6 (2)	30.7 (6)	84.0 (2)	53.0 (2)
NeuroDub	Off.		79.6 (5)	41.1 (2)				
USTC	Off.						80.0 (4)	52.5 (4)
KIT$_{pr}$	Mult.	✓ + LLM	74.9 (6)	37.5 (4)	82.0 (4)	35.7 (1)	79.3 (5)	49.4 (6)
KIT$_{c1}$	Mult.	✓ + LLM	74.6 (8)	36.5 (7)	82.0 (4)	35.2 (2)	79.3 (5)	49.7 (5)
KIT$_{c2}$	Mult.	✓ + LLM	74.3 (9)	36.5 (7)	81.6 (6)	34.0 (3)	78.6 (10)	49.4 (6)
KIT$_{c3}$	Mult.	✓ + LLM	74.7 (7)	36.1 (9)	81.4 (7)	33.3 (5)	78.4 (11)	48.6 (7)
KIT$_{c4}$	Mult.	✓ + LLM	74.2 (10)	36.4 (8)	81.7 (5)	33.9 (4)	78.4 (11)	48.2 (8)
KIT$_{c5}$	Mult.	✓ + LLM	74.9 (6)	33.8 (10)	80.3 (8)	27.3 (8)	79.1 (6)	46.7 (10)
UPC	Off.	✓ + LLM	74.7 (7)	32.1 (12)				
KIT$_{c6}$	Mult.	✓ + LLM	73.9 (11)	32.9 (11)	80.0 (9)	26.6 (9)	78.9 (7)	45.7 (13)
KIT$_{c7}$	Mult.	✓ + LLM	73.9 (11)	32.9 (11)	80.3 (8)	25.6 (10)	78.8 (8)	46.0 (12)
Xiaomi	Off.	✓ + LLM					78.7 (9)	46.5 (11)
NiuTrans	Off.	✓					77.3 (12)	47.1 (9)
NeMo	Off.	✓	71.7 (12)	31.9 (13)	77.7 (10)	24.9 (11)	74.0 (13)	41.8 (14)
I2R	Off.	✓ + LLM	68.4 (13)	23.9 (14)				
JHU	Mult.	✓ + LLM	59.0 (15)	23.7 (15)	69.3 (11)	18.9 (12)	67.9 (15)	37.4 (16)
MINE-Trans	Off.						70.0 (14)	39.9 (15)
BIGAI*	Off.	✓ + LLM	63.0 (14)	19.6 (16)	67.7 (12)	10.4 (13)	65.3 (16)	27.4 (18)
MINE-Trans	Off.	✓					63.5 (17)	31.8 (17)
BIT	Mult.	✓	47.2 (16)	11.1 (17)	56.2 (13)	8.0 (14)	55.7 (18)	19.8 (19)

Table 27: Submissions from all tracks on the ACL 60-60 evaluation sets on the **three** language pairs shared across tracks (En → De, Ja, Zh), ordered by average metric ranking. Direct / end-to-end systems are highlighted in gray.

Submission	ar	de	fa	fr	ja	nl	pt	ru	tr	zh	Avg.
JHU$_{unconstrained}$	62.4	67.6	57.8	73.4	42.0	71.6	75.0	56.8	62.5	42.2	61.1
KIT$_{primary}$	56.9	64.8	55.4	67.8	42.3	67.6	69.6	51.2	57.3	42.5	57.5
KIT$_{contrastive1}$	56.9	64.6	55.6	67.8	42.0	67.6	69.6	51.2	56.7	42.7	57.5
KIT$_{contrastive2}$	56.1	63.6	52.9	67.3	40.8	66.5	69.2	50.6	55.6	41.3	56.4
KIT$_{contrastive4}$	56.2	63.3	53.0	67.2	40.7	66.5	68.8	50.4	55.1	40.3	56.2
KIT$_{contrastive3}$	55.5	63.7	52.1	66.9	40.3	66.0	68.9	50.0	55.2	40.6	55.9
KIT$_{contrastive5}$	55.3	61.3	53.8	65.2	35.9	63.7	67.3	48.6	54.9	39.2	54.5
KIT$_{contrastive7}$	54.7	60.3	54.0	64.4	34.5	63.4	67.2	47.8	54.2	38.2	53.9
KIT$_{contrastive6}$	54.6	60.3	52.7	64.3	35.5	62.7	66.4	48.2	53.8	38.4	53.7
JHU$_{constrained}$	45.2	53.4	44.5	62.4	26.8	62.1	62.2	46.8	46.3	30.8	48.1
BIT	28.9	36.8	28.8	45.2	14.5	41.7	43.0	28.4	25.9	17.2	31.0

Table 28: chrF with resegmentation for each target language on the evaluation set, sorted by the system average. Direct / end-to-end systems are highlighted in gray.

Submission	ar	de	fa	fr	ja	nl	pt	ru	tr	zh	Avg.
JHU$_{unconstrained}$	82.7	81.3	80.6	81.4	84.7	84.1	84.9	78.9	82.5	82.0	82.3
KIT$_{primary}$	78.0	74.9	75.8	74.4	82.0	77.7	78.4	72.5	76.6	79.3	77.0
KIT$_{constrastive1}$	77.7	74.6	75.7	74.5	82.0	77.6	78.4	72.2	76.4	79.3	76.8
KIT$_{constrastive5}$	78.5	74.9	75.9	74.6	80.3	76.8	78.5	71.6	76.9	79.1	76.7
KIT$_{constrastive7}$	78.2	73.9	76.3	74.2	80.3	76.7	80.3	71.3	76.2	78.8	76.6
KIT$_{constrastive2}$	77.3	74.3	74.9	74.3	81.6	77.3	78.4	72.1	75.8	78.6	76.5
KIT$_{constrastive4}$	77.2	74.2	75.0	74.3	81.7	77.3	78.2	72.0	75.5	78.4	76.4
KIT$_{constrastive3}$	76.9	74.7	74.6	74.2	81.4	76.9	78.2	71.8	75.7	78.4	76.3
KIT$_{constrastive6}$	77.8	73.9	75.2	73.3	80.0	75.4	77.7	70.8	75.7	78.9	75.9
JHU$_{constrained}$	67.9	59.0	66.1	63.2	69.3	66.2	67.8	62.0	64.0	67.9	65.3
BIT	52.8	47.2	48.7	52.2	56.2	53.8	54.8	47.7	48.0	55.7	51.7

Table 29: COMET with resegmentation for each target language on the evaluation set, sorted by the system average. Direct / end-to-end systems are highlighted in gray.

	ar	de	fa	fr	ja	nl	pt	ru	tr	zh	Avg.
JHU$_{unconstrained}$	33.4	41.2	35.0	50.0	33.9	44.8	51.7	27.9	28.1	46.5	39.3
KIT$_{primary}$	25.9	37.5	29.8	41.3	35.7	40.4	44.3	22.4	21.8	49.4	34.9
KIT$_{constrastive1}$	25.6	37.5	30.1	41.1	35.2	40.6	44.5	22.6	21.3	49.7	34.8
KIT$_{constrastive2}$	24.7	36.5	28.0	42.4	34.0	38.8	43.8	21.9	20.6	49.4	34.0
KIT$_{constrastive4}$	24.4	36.4	28.4	42.1	33.9	38.9	43.0	21.6	20.3	48.2	33.7
KIT$_{constrastive3}$	24.0	36.1	27.6	41.9	33.3	38.2	43.6	21.5	20.1	48.6	33.5
KIT$_{constrastive5}$	23.7	33.8	28.7	39.6	27.3	35.9	40.7	19.6	20.6	46.7	31.7
KIT$_{constrastive7}$	23.4	32.9	28.6	38.8	25.6	36.0	40.9	19.1	20.1	46.0	31.1
KIT$_{constrastive6}$	23.0	32.9	28.3	38.9	26.6	35.0	39.7	19.7	19.1	45.7	30.9
JHU$_{constrained}$	15.0	23.7	21.9	33.1	18.9	31.3	33.2	17.2	12.8	37.4	24.5
BIT	5.7	11.1	7.4	19.7	8.0	16.3	18.6	6.3	4.1	19.8	11.7

Table 30: BLEU with resegmentation for each target language on the evaluation set, sorted by the system average. BLEU scores in grey are calculated using language-specific tokenization (ja) or at the character-level (zh); see §5.2 for specific tokenization details. Direct / end-to-end systems are highlighted in gray.

B.5 Speech-to-Speech Translation

System	Test-primary				Test-expanded				Overall			
Ref	BLEU	chrF	COMET	SEScore2	BLEU	chrF	COMET	SEScore2	BLEU	chrF	COMET	SEScore2
Cascade Systems												
XIAOMI	47.9	41.0	79.91	-12.27	34.5	29.2	79.07	-20.15	38.4	32.3	79.35	-17.48
NPU-MSXF	47.4	40.7	79.90	-12.21	34.0	28.5	78.68	-20.23	37.7	31.8	79.09	-17.52
HW-TSC	43.2	36.9	76.96	-14.23	32.4	27.7	76.43	-21.61	35.3	30.1	76.61	-19.12
KU	36.7	31.3	69.09	-17.07	25.0	21.7	67.94	-25.68	28.2	24.3	68.33	-22.77
MINETRANS_Cascade	33.9	28.6	67.49	-17.68	24.7	21.5	64.71	-26.34	27.2	23.4	65.65	-23.41
E2E Systems												
MINETRANS_E2E (contrastive2)	45.0	38.3	74.83	-13.62	31.1	26.4	73.28	-22.03	34.9	29.6	73.81	-19.18
MINETRANS_E2E (contrastive1)	44.5	38.0	74.14	-13.92	31.0	26.4	72.90	-22.20	34.8	29.5	73.32	-19.40
MINETRANS_E2E (primary)	44.4	38.0	74.40	-13.86	31.1	26.4	73.00	-22.12	34.7	29.5	73.47	-19.32

Table 31: Official results of the **automatic evaluation** for the English to Chinese Speech-to-Speech Translation Task.

System	Translation Quality Score	Speech Quality Score	Overall
Cascade Systems			
NPU-MSXF	3.70	3.98	3.84
XIAOMI	3.72	3.67	3.70
HW-TSC	3.58	3.75	3.67
MINETRANS_Cascade	3.16	3.26	3.21
KU	2.92	3.01	2.97
E2E Systems			
MINETRANS_E2E (contrastive2)	3.58	3.50	3.54

Table 32: Official results of the **human evaluation** for the English to Chinese Speech-to-Speech Translation Task.

B.6 Dialectal SLT

Tunisian Arabic→English (Unconstrained Condition)

		test2					test3				
Team	System	BLEU	bp	pr1	chrF	TER	BLEU	bp	pr1	chrF	TER
USTC	primary	23.6	1.0	52.7	46.7	64.6	21.1	1.0	49.0	43.8	69.0
USTC	contrastive1	22.8	1.0	51.7	45.7	65.7	20.2	1.0	47.7	42.9	70.7
JHU	contrastive5	21.6	.99	50.7	45.0	66.9	19.1	1.0	46.6	41.9	72.3
JHU	primary	21.2	1.0	50.0	44.8	67.7	18.7	1.0	46.0	41.9	73.1
JHU	contrastive4	20.7	1.0	49.3	44.2	68.4	18.3	1.0	45.5	41.3	73.7
JHU	contrastive3	19.9	.98	49.0	43.0	68.7	18.2	1.0	45.5	40.5	73.1
JHU	contrastive1	19.4	.99	48.2	42.4	69.8	17.1	1.0	44.3	39.7	74.9
JHU	contrastive2	18.7	.97	48.4	41.8	69.4	17.1	1.0	44.7	39.2	74.1
ON-TRAC	post-eval	18.2	1.0	45.9	42.7	73.8	16.3	1.0	41.6	40.3	79.6
GMU	contrastive1	15.0	1.0	41.4	38.4	78.2	13.4	1.0	37.2	36.1	83.9
GMU	contrastive2	14.1	1.0	40.1	37.5	79.8	12.9	1.0	36.6	35.4	84.7
GMU	primary	16.6	1.0	44.5	39.7	74.1	14.6	1.0	40.4	37.6	79.6
ON-TRAC	primary	7.0	1.0	27.3	36.4	86.9	6.2	1.0	24.2	34.3	92.0
2022 best:CMU		20.8	.93	53.1	44.3	64.5	-	-	-	-	-

Table 33: Automatic evaluation results for the Dialect Speech Translation task, Unconstrained Condition. Systems are ordered in terms of the official metric BLEU on test3. We also report brevity penalty (bp) and unigram precision (pr1) of BLEU, chrF, and TER.

Tunisian Arabic→English (Constrained Condition)

		test2					test3				
Team	System	BLEU	bp	pr1	chrF	TER	BLEU	bp	pr1	chrF	TER
USTC	primary	20.5	.99	49.9	43.6	67.6	18.1	1.0	45.7	40.8	73.1
JHU	primary	19.1	.94	50.5	42.4	67.2	17.6	.96	46.6	39.9	71.9
GMU	primary	5.0	1.0	20.3	21.9	102.2	4.5	1.0	18.4	20.7	105.5
2022 best:CMU		20.4	.94	52.2	43.8	65.4	-	-	-	-	-
baseline		11.1	.88	40.0	31.9	77.8	10.4	.90	36.6	29.9	81.4

Table 34: Automatic evaluation results for the Dialect Speech Translation task, Constrained Condition.

Tunisian Arabic ASR Automatic Evaluation Results

ASR System	test2 WER↓		test2 CER↓		test3 WER↓		test3 CER↓	
	Orig	Norm	Orig	Norm	Orig	Norm	Orig	Norm
JHU / constrained / primary	70.3	43.7	30.7	22.7	74.0	44.9	33.1	24.8
JHU / unconstrained / primary	69.3	40.6	29.0	20.7	72.9	41.6	31.5	22.9
USTC / constrained / primary	49.5	40.8	24.2	20.9	52.3	43.2	27.1	23.8
USTC / unconstrained / primary	47.4	39.3	23.1	20.0	49.2	40.5	25.2	22.1
2022best:ON-TRAC/unconstrained	65.7	41.5	28.1	21.1	-	-	-	-

Table 35: Word Error Rate (WER) and Character Error Rate (CER) of the ASR component of submitted cascaded systems on test2 and test3. The original version (Orig) matches the minimal text pre-processing provided by the organizer's data preparation scripts, and results in relatively high WER. As diagnosis, we ran additional Arabic-specific normalization (Norm) for e.g. Alif, Ya, Ta-Marbuta on the hypotheses and transcripts before computing WER/CER. We are grateful to Ahmed Ali for assistance on this.

B.7 Low-Resource SLT

Irish→English (Constrained Condition)

Team	System	BLEU	chrF2
GMU	primary	15.1	26.5

Table 36: Automatic evaluation results for the Irish to English task, Constrained Condition.

Irish→English (Unconstrained Condition)

Team	System	BLEU	chrF2
GMU	primary	68.5	74.5
GMU	contrastive1	77.4	81.6
GMU	contrastive2	15.1	26.5

Table 37: Automatic evaluation results for the Irish to English task, Unconstrained Condition.

Marathi→Hindi (Constrained Condition)

Team	System	BLEU	chrF2
GMU	primary	3.3	16.8
SRI-B	primary	31.2	54.8
SRI-B	contrastive	25.7	49.4

Table 38: Automatic evaluation results for the Marathi to Hindi task, Constrained Condition.

Marathi→Hindi (Unconstrained Condition)

Team	System	BLEU	chrF2
Alexa AI	primary	28.6	49.4
Alexa AI	contrastive1	25.6	46.3
Alexa AI	contrastive2	23	41.9
Alexa AI	contrastive3	28.4	49.1
Alexa AI	contrastive4	25.3	46.3
Alexa AI	contrastive5	19.6	39.9
BUT	primary	39.6	63.3
BUT	contrastive	28.6	54.4
GMU	primary	7.7	23.8
GMU	contrastive1	8.6	24.7
GMU	contrastive2	5.9	20.3
SRI-B	primary	32.4	55.5
SRI-B	contrastive	29.8	53.2

Table 39: Automatic evaluation results for the Marathi to Hindi task, Unconstrained Condition.

Pashto→French (Unconstrained Condition)

		BLEU	
Team	System	valid	test
ON-TRAC	primary	24.82	24.87
ON-TRAC	contrastive1	23.38	23.87
GMU	primary	11.99	16.87
GMU	contrastive1	11.27	15.24
ON-TRAC	contrastive2	12.26	15.18
ON-TRAC	contrastive3	12.16	15.07
GMU	contrastive2	9.72	13.32

Table 40: Automatic evaluation results for the Pashto to French task, Unconstrained Condition.

Pashto→French (Constrained Condition)

		BLEU	
Team	System	valid	test
ON-TRAC	primary	14.52	15.56
ON-TRAC	contrastive1	11.06	15.29
ON-TRAC	contrastive2	11.11	15.06
ON-TRAC	contrastive3	10.5	9.2
GMU	primary	2.66	5.92

Table 41: Automatic evaluation results for the Pashto to French task, Constrained Condition.

Maltese→English (Unconstrained Condition)

Team	System	BLEU
UM-DFKI	primary	0.6
UM-DFKI	contrastive1	0.7
UM-DFKI	contrastive2	0.4
UM-DFKI	contrastive3	0.3
UM-DFKI	contrastive4	0.4

Table 42: Automatic evaluation results for the Maltese to English task, Unconstrained Condition.

Tamasheq→French (Constrained Condition)

Team	System	BLEU	chrF2	TER
GMU	primary	0.48	19.57	106.23

Table 43: Automatic evaluation results for the Tamasheq to French task, Constrained Condition.

Tamasheq→French (Unconstrained Condition)

Team	System	BLEU	chrF2	TER
NAVER	primary	23.59	49.84	64.00
NAVER	contrastive1	21.31	48.15	66.41
NAVER	contrastive2	18.73	46.11	70.32
ON-TRAC	primary	15.88	43.88	73.85
ON-TRAC	contrastive1	16.35	44.22	74.26
ON-TRAC	contrastive2	15.46	43.59	75.30
ON-TRAC	contrastive3	15.49	43.74	75.07
ON-TRAC	contrastive4	16.25	44.11	74.26
ON-TRAC	contrastive5	15.54	43.91	75.08
Alexa AI	primary	9.30	32.29	81.25
Alexa AI	contrastive1	8.87	32.04	81.03
Alexa AI	contrastive2	9.50	33.67	80.85
Alexa AI	contrastive3	9.28	32.86	82.33
GMU	primary	8.03	33.03	87.81
GMU	contrastive1	1.30	23.63	96.72
GMU	contrastive2	2.10	24.33	94.58

Table 44: Automatic evaluation results for the Tamasheq to French task, Unconstrained Condition.

Quechua→Spanish (Constrained Condition)

Team	System	BLEU	chrF2
GMU	primary	1.46	21.46
QUESPA	primary	1.25	25.35
QUESPA	contrastive1	0.13	10.53
QUESPA	contrastive2	0.11	10.63

Table 45: Automatic evaluation results for the Quechua to Spanish task, Constrained Condition. ChrF2 scores were only taken into account for those systems that scored less than 5 points BLEU.

Quechua→Spanish (Unconstrained Condition)

Team	System	BLEU
GMU	primary	1.78
GMU	contrastive1	1.86
GMU	contrastive2	1.63
NAVER	primary	15.70
NAVER	contrastive1	13.17
NAVER	contrastive2	15.55
QUESPA	primary	15.36
QUESPA	contrastive1	15.27
QUESPA	contrastive2	10.75

Table 46: Automatic evaluation results for the Quechua to Spanish task, Unconstrained Condition. ChrF2 scores were only taken into account for those systems that scored less than 5 points BLEU.

B.8 Formality Control for SLT

	Model		EN-KO				EN-VI			
			BLEU	COMET	mACC	cACC	BLEU	COMET	mACC	cACC
CONSTRAINED	CoCoA (baseline)	F	11.1	0.5044	28.5	55	43.2	0.6189	99	99
		IF	11.1	0.5125	80.4	58	41.5	0.6021	98	99
	HW-TSC	F	**25.6**	**0.7512**	89	**100**	**51.3**	**0.7522**	**100**	**100**
		IF	**26.1**	**0.7367**	**100**	**100**	**49.8**	**0.7209**	**100**	**100**
UNCONSTRAINED	UMD (baseline)	F	4.9	0.2110	78	99	26.7	0.3629	96	95
		IF	4.9	0.1697	98	99	25.3	0.3452	97	98
	HW-TSC	F	25.4	**0.7347**	87	**100**	**48.2**	**0.7214**	**100**	**100**
		IF	26.2	**0.7218**	**100**	**100**	**48.3**	**0.7102**	**100**	**100**
	KUxUpStage	F	**26.6**	0.7269	87	**100**	47.0	0.6685	99	**100**
		IF	**27.1**	0.7145	98	95	45.6	0.6373	99	**100**
	UCSC	F	23.3	0.5210	86	98	44.6	0.6771	99	98
		IF	22.8	0.4724	98	96	43.5	0.6281	99	**100**

Table 47: Results for the Formality Track (Supervised Setting). Most systems perform well in this setting, though MT quality on formal (F) tends to be higher than informal (IF)

	Model		EN-PT				EN-RU			
			BLEU	COMET	mACC	cACC	BLEU	COMET	mACC	cACC
CONSTRAINED	HW-TSC	F	**47.4**	**0.7337**	**100**	**100**	**36.5**	**0.6472**	**100**	**100**
		IF	**47.9**	**0.7442**	**100**	**100**	**35.6**	**0.6442**	**100**	**100**
UNCONSTRAINED	UMD (baseline)	F	27.3	0.4477	96	98	21.3	0.3492	96	92
		IF	30.9	0.4161	93	91	21.0	0.3475	84	85
	AppTek	F	34.6	0.6089	99	99	**35.4**	**0.6165**	99	98
		IF	42.4	0.6776	64	65	**33.3**	**0.6026**	98	97
	HW-TSC	F	**45.4**	**0.7737**	**100**	**100**	33.7	0.5804	**100**	**100**
		IF	**49.1**	**0.7845**	**100**	**100**	32.4	0.5558	**100**	**100**
	KUxUpStage	F	31.0	0.5251	**100**	**100**	25.8	0.4446	**100**	**100**
		IF	19.9	0.2486	68	90	26.3	0.4181	**100**	**100**
	UCSC	F	26.6	0.4048	90	91	18.4	-0.1713	99	79
		IF	28.4	0.4252	58	42	14.9	-0.2766	52	67

Table 48: Results for the Formality Track (Zero-shot Setting). Appreciable differences in formality control exist between formal (F) and informal (IF), suggesting that formality bias exists in participant systems.