FINDINGS OF THE IWSLT 2021 EVALUATION CAMPAIGN

Antonios Anastasopoulos
George Mason U.

Ondřej Bojar
Charles University

Jacob Bremerman
UMD

Roldano Cattoni
FBK

Maha Elbayad
Facebook

Marcello Federico
Amazon AI

Xutai Ma
JHU/Facebook

Satoshi Nakamura
NAIST

Matteo Negri
FBK

Jan Niehues
Maastricht U.

Juan Pino
Facebook

Elizabeth Salesky
JHU

Sebastian Stüker
KIT

Katsuhito Sudoh
NAIST

Marco Turchi
FBK

Alex Waibel
CMU/KIT

Changhan Wang
Facebook

Matthew Wiesner
JHU

Abstract

The evaluation campaign of the International Conference on Spoken Language Translation (IWSLT 2021) featured this year four shared tasks: (i) Simultaneous speech translation, (ii) Offline speech translation, (iii) Multilingual speech translation, (iv) Low-resource speech translation. A total of 22 teams participated in at least one of the tasks. This paper describes each shared task, data and evaluation metrics, and reports results of the received submissions.

1 Introduction

The International Conference on Spoken Language Translation (IWSLT) is the premier annual scientific conference for all aspects of spoken language translation. For 18 years running (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), the conference organizes and sponsors open evaluation campaigns around key challenges in simultaneous and consecutive translation, under real-time/low latency or offline conditions and under low-resource or multilingual constraints. System descriptions and results from participants' systems and scientific papers related to key algorithmic advances and best practice are published in proceedings and presented at the conference. IWSLT is also the venue of the SIGSLT, the Special Interest Group on Spoken Language Translation of ACL, ISCA and ELRA. With its long track record, IWSLT benchmarks and proceedings serve as reference for all researchers and practitioners working on speech translation and related fields.

This paper reports on the evaluation campaign organized by IWSLT 2021, which features four shared tasks:

- **Simultaneous speech translation**, addressing low latency translation of talks, from English to German and English to Japanese, either from a speech file into text, or from a ground-truth transcript into text;
- **Offline speech translation**, proposing speech translation of talks from English into German, using either cascade architectures or end-to-end models, able to directly translate source speech into target text;
- **Multilingual speech translation**, focusing

Proceedings of the 18th International Conference on Spoken Language Translation, pages 1–29
Bangkok, Thailand (Online), August 5–6, 2021. ©2021 Association for Computational Linguistics

Team	Organization
AppTek	AppTek, Germany (Bahar et al., 2021b)
BUT	Brno University of Technology, Czech Republic (Vydana et al., 2021)
ESPnet-ST	ESPnet-ST group, Johns Hopkins University, USA (Inaguma et al., 2021)
FBK	Fondazione Bruno Kessler, Italy (Papi et al., 2021)
FAIR	FAIR Speech Translation (Tang et al., 2021a)
HWN	Huawei Noah's Ark Lab, China (Zeng et al., 2021)
HW-TSC	Huawei Translation Services Center, China
IMS	University of Stuttgart, Germany (Denisov et al., 2021)
KIT	Karlsruhe Institute of Technology, Germany (Nguyen et al., 2021; Pham et al., 2021)
Li	Desheng Li
NAIST	Nara Institute of Science and Technology, Nara, Japan (Fukuda et al., 2021)
NiuTrans	NiuTrans Research, Shenyang, China (Xu et al., 2021b)
ON-TRAC	ON-TRAC Consortium, France (Le et al., 2021)
OPPO	Beijing OPPO Telecommunications Co., Ltd., China
UEdin	University of Edinburgh, UK (Zhang and Sennrich, 2021; Sen et al., 2021)
UM-DKE	Maastricht University, The Netherlands (Liu and Niehues, 2021)
UPC	Universitat Politècnica de Catalunya, Spain (Gállego et al., 2021)
USTC-NESLIP	USTC, iFlytek Research, China (Liu et al., 2021)
USYD-JD	University of Sydney, Peking University, JD Explore Academy (Ding et al., 2021)
VOLCTRANS	ByteDance AI Lab, China (Zhao et al., 2021)
VUS	Voithru, Upstage, Seoul National University, South Korea (Jo et al., 2021)
ZJU	Zhejiang University (Zhang, 2021)

Table 1: List of Participants

on the use of multiple languages to improve supervised and zero-shot speech translation between four Romance languages and English;

- **Low-resource speech translation**, focusing on resource-scarce settings for translating two Swahili varieties (Congolese and Coastal) into English and French.

The shared tasks were attended by 22 participants (see Table 1), including teams from 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 challenge, the data used for training and testing data, the received submissions and the summary of results. Detailed results for each challenge are reported in a corresponding appendix.

2 Simultaneous Speech Translation

Simultaneous translation is the task of translating incrementally with partial text or speech input only. Such capability enables multilingual live communication and access to multilingual multimedia content in real-time. The goal of this challenge, organized for the second consecutive year, is to examine systems that translate text or audio in a source language into text in a target language from the perspective of both translation quality and latency.

2.1 Challenge

Participants were given three parallel tracks to enter and encouraged to enter all tracks:

- text-to-text: translating ground-truth transcripts in real time from English to German and English to Japanese.
- speech-to-text: translating speech into text in real time from English to German.

For the speech-to-text track, participants were encouraged to submit systems either based on cascaded or end-to-end approaches. In addition, the systems were run on a segmented and non-segmented version of the test set, i.e. processing one sound segment corresponding to an input sentence at a time, or processing the whole speech in one sound stream. Participants were required to upload their system as a Docker image so that it could be evaluated by the organizers in a controlled environment. We also provided an example implementation and a baseline system.[1]

2.2 Data and Metrics

For tracks related to English-German, participants were allowed to use the same training and development data as in the Offline Speech Translation track. More details are available in §3.2.

For the English-Japanese text-to-text track, participants could use the parallel data and monolingual data available for the English-Japanese WMT20 news task (Barrault et al., 2020). For development, participants could use the IWSLT 2017 development sets,[2] the IWSLT 2021 development set[3] and the simultaneous interpretation transcripts for the IWSLT 2021 development set.[4] The simultaneous interpretation was recorded as a part of NAIST Simultaneous Interpretation Corpus (Doi et al., 2021).

Systems were evaluated with respect to quality and latency. Quality was evaluated with the standard BLEU metric (Papineni et al., 2002a). Latency was evaluated with metrics developed for simultaneous machine translation, including average proportion (AP), average lagging (AL) and differentiable average lagging (DAL, Cherry and Foster 2019), and later extended to the task of simultaneous speech translation (Ma et al., 2020b).

The evaluation was run with the SIMULEVAL toolkit (Ma et al., 2020a). For the latency measurement of speech input systems, we contrasted computation-aware and non computation-aware latency metrics. The latency was calculated at the word level for English-German systems and at the character level for English-Japanese systems.

The systems were ranked by the translation quality (measured by BLEU) in different latency regimes, low, medium and high. Each regime was determined by a maximum latency threshold measured by AL on the Must-C English-German test set (tst-COMMON) for English-German or on the IWSLT21 dev set for English-Japanese. The thresholds were set to 3, 6 and 15 for the English-German text track, to 1000, 2000 and 4000 for the English-German speech track and to 8, 12 and 16 for English-Japanese text track, and were calibrated by the baseline system. Participants were asked to submit at least one system per latency regime and were encouraged to submit multiple systems for each regime in order to provide more data points for latency-quality trade-off analyses. The organizers confirmed the latency regime by running the systems on tst-COMMON and the IWSLT21 dev set.

2.3 Differences with the First Edition

English-to-Japanese Task This year, we added a new task of English-to-Japanese simultaneous translation. English-Japanese is a challenging language pair for simultaneous translation because of the large word order differences; a simultaneous machine translation model has to wait for the latter part of an English sentence in *Subject-Verb-Object* order to generate a Japanese sentence in *Subject-Object-Verb* order.

SimulEval We standardized the latency evaluation aspect of the task by leveraging the SIMULEVAL toolkit. In addition, speech input systems were run in a controlled environment (a p3.2xlarge AWS instance) in order to be able to fairly compare computation-aware AL.

Unsegmented input Based on feedback from the participants in the first edition of the task, for the speech track, systems were run on both segmented and unsegmented input. The latter setting required participants to implement a segmentation logic in their systems, which is closer to a real-world setting.

2.4 Submissions

The simultaneous task received submissions from 5 teams: 4 teams entered the English-German text track; 3 teams entered the English-Japanese text track and 2 teams entered the English-German speech track. Teams followed the suggestion to submit multiple systems per regime, which resulted in a total of 162 systems overall.

UEDIN (Sen et al., 2021) submitted systems to the text-to-text English-German track. In order to be able to reuse an offline system, UEDIN adapts the re-translation strategy to the simultaneous task. Re-translation is triggered based on a language model applied to the source input. In addition, a

[1] https://github.com/pytorch/fairseq/blob/master/examples/speech_to_text/docs/simulst_mustc_example.md

[2] https://wit3.fbk.eu/2017-01-c

[3] https://drive.google.com/drive/folders/1uSkOT-XqbICMohnvfXdEFffKLdaQX0X7

[4] https://drive.google.com/drive/folders/1bB1s9PKNoRoDFfc567J5zDMcYj_1FFEB

dynamic masking method is employed to stabilize the output translation.

VOLCTRANS (Zhao et al., 2021) submitted systems to the text-to-text English-German and English-Japanese tracks. The participants adopt the efficient wait-k strategy (Elbayad et al., 2020). They augment the training data using back-translation and knowledge distillation. During inference, a look ahead beam search strategy is investigated but the final submission uses greedy search.

USTC-NESLIP (Liu et al., 2021) submitted systems to all tracks, including both end-to-end and cascaded system for the speech tracks. The participants design a novel model architecture, Cross-Attention Augmented Transducer, that modifies RNN-T in order to support reordering between languages. They augment the training data using self-training, back-translation and by synthesizing the source side of the parallel corpora.

AppTek (Bahar et al., 2021b) submitted systems to the English-German speech and text tracks, using a cascaded system for the speech track. Chunks that preserve monotonicity are extracted from a statistical word aligner. A classifier, part of the overall model, is trained on the boundaries in order to control the policy. To better control the latency quality tradeoff, consecutive chunks can be merged according to a probability.

NAIST (Fukuda et al., 2021) submitted systems to the text English-Japanese track. The participants employ the wait-k method and sequence-level knowledge distillation. Because Japanese does not have a strict word order, they randomly shuffle chunks on the target side to augment the training data. An alternative method, next constituent label prediction, was investigated but not submitted to the task.

2.5 Results

We discuss results for the text and speech tracks. More details are available in Appendix A.1.

2.5.1 Text Track

Results for the text track are summarized in the first two tables of Appendix A.1. Four teams (USTC-NESLIP, VOLCTRANS, AppTek, UEdin) submitted systems for English-German and three teams (USTC-NESLIP, VOLCTRANS, NAIST) for English-Japanese. In the table, only the models with the best BLEU score for a given latency regime are reported. In order to obtain a broader sense of latency-quality tradeoffs, we also plot all submitted systems for quality and latency.

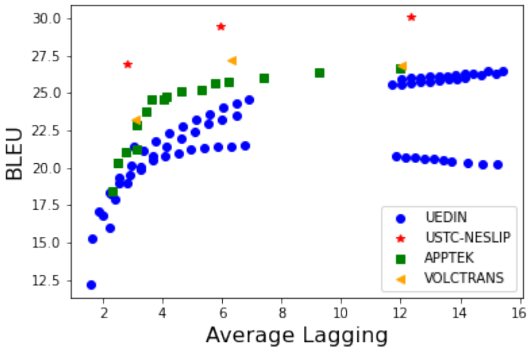

Figure 1: Latency-quality trade-off curves, measured by AL and BLEU, reported on the blind test set, for the systems submitted to the English-German text track.

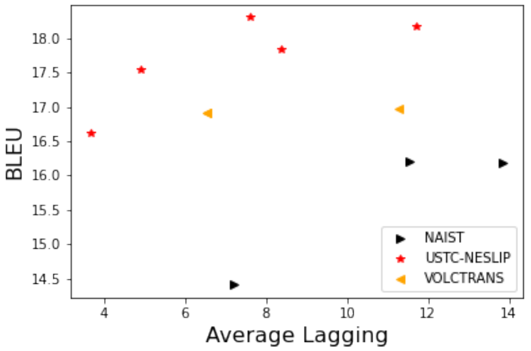

Figure 2: Latency-quality trade-off curves, measured by AL and BLEU, reported on the blind test set, for the systems submitted to the English-Japanese text track.

English-German The ranking is consistent over all the regimes: 1. USTC-NESLIP 2. VOLCTRANS 3. AppTek 4. UEdin. We plot all the submitted English-German systems in Figure 1.

Japanese-English The ranking is consistent over all the regimes: 1. USTC-NESLIP 2. AppTek 3. NAIST. We plot all the submitted English-Japanese systems in Figure 2.

2.5.2 Speech Track (English-German Only)

Results for the speech track are summarized in the third table of Appendix A.1. Two teams (USTC-NESLIP, AppTek) submitted systems, with both segmented and unsegmented speech input. Latency regimes were defined for segmented input systems only. We plan to define latency

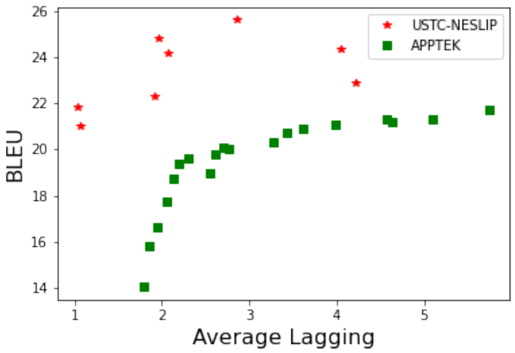

Figure 3: Latency-quality trade-off curves, measured by AL and BLEU, reported on the blind test set, for the systems submitted to the speech track with segmented input. AL is measured in seconds.

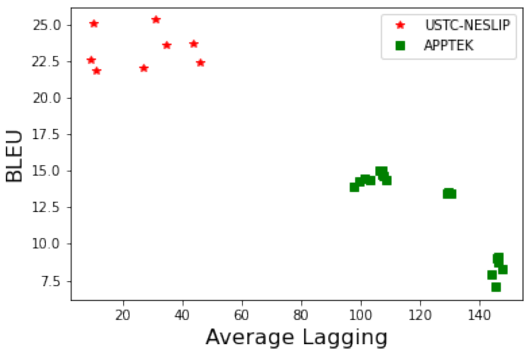

Figure 5: Latency-quality trade-off curves, measured by AL and BLEU, reported on the blind test set, for the systems submitted to the speech track with unsegmented input. AL is measured in seconds.

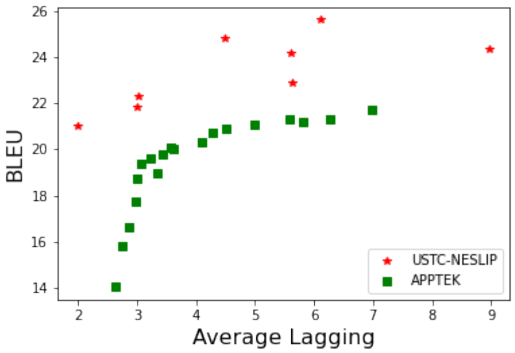

Figure 4: Latency-quality trade-off curves, measured by AL and BLEU, reported on the blind test set, for the systems submitted to the speech track with segmented input. AL is considering the computation time and measured in seconds.

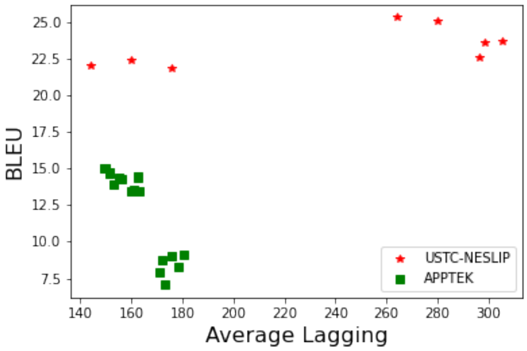

Figure 6: Latency-quality trade-off curves, measured by AL and BLEU, reported on the blind test set, for the systems submitted to the speech track with unsegmented input. AL is considering the computation time and measured in seconds.

regimes for unsegmented input in the next edition. The ranking is consistent over all the regimes in segmented systems and unsegmented systems: 1. USTC 2. AppTek We also report four latency-quality trade-off curves:

- Segmented input systems without considering computation time in Figure 3.
- Segmented input systems considering computation time in Figure 4.
- Unsegmented input systems without considering computation time in Figure 5.
- Unsegmented input systems considering computation time in Figure 6.

3 Offline Speech Translation

Offline speech translation, declined in various forms over the years, is one of the speech tasks with the longest tradition at the IWSLT campaign. Like in the last two evaluation rounds, this year[5] it focused on the translation of English audio data extracted from TED talks[6] into German.

3.1 Challenge

In recent years, offline speech translation (ST) has seen a rapid evolution, characterized by the steady advancement of *direct* end-to-end models (building on a single neural network that directly translates the input audio into target language text) that were able to significantly reduce the performance

[5] http://iwslt.org/2021/offline
[6] http://www.ted.com

gap with respect to the traditional *cascade* approach (integrating ASR and MT components in a pipelined architecture). In light of last year's IWSLT results (Ansari et al., 2020) and of the findings of recent works (Bentivogli et al., 2021) attesting that the gap between the two paradigms has substantially closed, also this year a key element of the evaluation was to set up a shared framework for their comparison. For this reason, and to reliably measure progress with respect to the past rounds, the general evaluation setting was kept unchanged. This stability mainly concerns two aspects: the allowed architectures and the test set provision.

On the architecture side, participation was allowed both with cascade and end-to-end (also known as direct) systems. In the latter case, valid submissions had to be obtained by models that: *i)* do not exploit intermediate symbolic representations (e.g., source language transcription or hypotheses fusion in the target language), and *ii)* rely on parameters that are all jointly trained on the end-to-end task.

On the test set provision side, also this year participants could opt for processing either a precomputed automatic segmentation of the test set or a version of the same test data segmented with their own approach. This option was maintained not only to ease participation (by removing one of the obstacles in audio processing) but also to gain further insights about the importance of a proper segmentation of the input speech. As highlighted in (Ansari et al., 2020), effective preprocessing to reduce the mismatch between the provided training material (often "clean" corpora split into sentence-like segments) and the supplied unsegmented test data is in fact a common trait of top-performing systems.

Multiple submissions were allowed, but participants had to explicitly indicate their "primary" (one at most) and "contrastive" runs, together with the corresponding type of system (cascade/end-to-end), training data condition (constrained/unconstrained), and test set segmentation (own/given).

3.2 Data and Metrics

Training and development data. Also this year, participants had the possibility to train their systems using several resources available for ST, ASR and MT. The major novelty on the data front is that a new TED-derived resource was added to the training corpora usually allowed to satisfy the "constrained" data condition. The new data come from the English-German section of the MuST-C V2 corpus[7] and include training, dev, and test (Test Common), in the same structure of the MuST-C V1 corpus (Cattoni et al., 2021) used last year. Since the 2021 test set was processed using the same pipeline applied to create MuST-C V2, the use of the new training resource was strongly recommended. The main differences with respect to MuST-C v1 are:

- More talks, which results in 20k more audio/text segments;
- Improved cleaning strategies able to better discard low-quality triplets (audio, transcript, translation), in particular when the text is not well-aligned with the audio and the audio is shorter than 50 millisecs;
- The talks were downloaded from the YouTube TED channel,[8] where higher quality audio/videos are available with respect to the TED website used for the previous version of MuST-C. The downloading was performed by means of youtube-dl,[9] the well-known open-source download manager, specifying the "-f bestaudio option". The audios were finally converted from two (stereo) to one (mono) channel and downsampled from 48 to 16 kHz, using FFmpeg.[10] Upon inspection of the spectrograms of the same talks in the two versions of MuST-C, it clearly emerges that the upper limit band in the audios used in MuST-C V1 is 5 kHz, while it is at 8 kHz in the latest version, coherently with the 16 kHz sample rate. This difference does not guarantee the fully compatibility between V1 and V2 of MuST-C.

Besides MuST-C V2, also this year the allowed training corpora include:

- MuST-C V1 (Di Gangi et al., 2019);
- CoVoST (Wang et al., 2020);

[7] http://ict.fbk.eu/must-c/

[8] http://www.youtube.com/c/TED/videos

[9] http://youtube-dl.org/

[10] http://ffmpeg.org/

- WIT^3 (Cettolo et al., 2012) ;
- Speech-Translation TED corpus[11];
- How2 (Sanabria et al., 2018)[12];
- LibriVoxDeEn (Beilharz and Sun, 2019)[13];
- Europarl-ST (Iranzo-Sánchez et al., 2020);
- TED LIUM v2 (Rousseau et al., 2014) and v3 (Hernandez et al., 2018);
- WMT 2019[14] and 2020[15];
- OpenSubtitles 2018 (Lison et al., 2018);
- Augmented LibriSpeech (Kocabiyikoglu et al., 2018)[16]
- Mozilla Common Voice[17] ;
- LibriSpeech ASR corpus (Panayotov et al., 2015).

The list of allowed development data includes the dev set from IWSLT 2010, as well as the test sets used for the 2010, 2013, 2014, 2015 and 2018 IWSLT campaigns. Using other training/development resources was allowed but, in this case, participants were asked to mark their submission as an "unconstrained" one.

Test data. This year's new test set was built from 17 TED talks that are not included yet in the public release of the corpus. Similar to last year, participants were presented with the option of processing either an unsegmented version (to be split with their preferred segmentation method) or an automatically segmented version of the audio data. For the segmented version, the resulting number of segments is 2,336 (corresponding to about 4h15m of translated speech from 17 talks). To measure technology progress with respect to last year's round, participants were asked to process also the undisclosed 2020 test set that, in the segmented version, consists of 2,263 segments (corresponding to about 4.1 hours of translated speech from 22 talks).

Metrics. Systems' performance was evaluated with respect to their capability to produce translations similar to the target-language references. Differently from previous rounds, where such similarity was measured in terms of multiple automatic metrics,[18] this year only the BLEU metric (computed with SacreBLEU (Post, 2018) with default settings) has been considered. Instead of multiple metrics, the attention focused on considering 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.[19] This makes them less literal compared to standard, unconstrained translations;
- Unconstrained translations. These references were created from scratch[20] by adhering to the usual translation guidelines. They are hence exact (more literal) translations, without paraphrasing and with proper punctuation.

Lang	Sentences	Words
EN	2,037	41,214
DE - Orig	2,037	33,925
DE - Uncon.	2,037	40,239

Table 2: Statistics of the official test set for the offline speech translation task (*tst2021*).

As shown in Table 2, the different approaches to generate the human translations lead to significantly different references. While the unconstrained translation has a similar length (counted in words) compared to the corresponding source sentence, the original is ~15% shorter in order to fulfil the additional constraints for subtitling.

Besides considering separate scores for the two types of references, results were also computed by considering both of them in a multi-reference setting. Similarly to last year, the submitted runs

[11] http://i13pc106.ira.uka.de/~mmueller/iwslt-corpus.zip

[12] only English - Portuguese

[13] only German - English

[14] http://www.statmt.org/wmt19/

[15] http://www.statmt.org/wmt20/

[16] only English - French

[17] http://voice.mozilla.org/en/datasets – English version en_1488h_2019-12-10

[18] These were: case-sensitive/insensitive BLEU (Papineni et al., 2002b), case-sensitive/insensitive TER (Snover et al., 2006), BEER (Stanojevic and Sima'an, 2014), and CharacTER (Wang et al., 2016)

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

[20] We would like to thank Facebook, and in particular Juan Pino, for providing us with this new set of references.

were ranked based on case-sensitive BLEU calculated on the test set by using automatic re-segmentation of the hypotheses based on the reference translations by mwerSegmenter.[21]

3.3 Submissions

We received submissions from 12 teams, which is a slight increase (+2) over last year's round. Also this year, participants come from the industry (the majority), the academia and other research institutions. In terms of ST paradigms, though quite evenly distributed, architectural choices show a slight preference for the cascade approach, which highlights a countertrend strategy with respect to the 2020 round, in which half of the participants opted for end-to-end submissions only. In detail:

- 5 teams (BUT, HW-TSC, LI, OPPO, VUS) participated only with cascade systems;
- 3 teams (FBK, NIUTRANS, UPC) participated only with end-to-end systems;
- 4 teams (APPTEK, VOLCTRANS, ESPNET-ST,KIT) participated with both cascade and end-to-end systems.

In total, 55 runs were evaluated: 24 obtained from cascade systems and 31 obtained from end-to-end systems. Concerning the segmentation of the test data (own/given), most of the primary submissions (7 out of 12) were obtained with "own" segmentation strategies aimed to improve the given automatic audio splits provided to participants like in last year's round of the task. As regards the data condition (constrained/unconstrained), all participants but two (BUT and UPC) opted for "constrained" submissions obtained by building their ST models only using the provided training resources.

In the following, we provide a bird's-eye description of each participant's approach.

APPTEK (Bahar et al., 2021b) participated with both cascade and end-to-end speech translation systems fed with "own" automatic segmentation of the test data. The primary cascade system is akin to the conventional cascade systems where source transcriptions are generated as an intermediate representation. ASR exploits an attention-based model (Bahdanau et al., 2015; Vaswani et al., 2017) trained following Zeyer et al. (2018), while the MT component is based on the big Transformer model model. Passing on the renormalized ASR posteriors into the MT model, the model is trained in an end-to-end fashion (inspired by the posterior tight integrated model by Bahar et al. 2021a) using all ASR, MT, and ST available training data. The system uses an improved automatic segmentation based on voice activity detection (VAD) and endpoint detection (EP). The primary end-to-end system also processes the speech input with "own" automatic segmentation. It is based on an ensemble of 4 models combining an LSTM speech encoder and a big Transformer decoder, as well as a pure Transformer model for both the encoder and the decoder. The models are trained using CTC attention loss, spectrogram augmentation, pretraining, synthetic data using forward translation, and fine-tuned on the in-domain TED talks. Following Gaido et al. (2020a), the direct model is also fine-tuned on automatically segmented data to increase its robustness against sub-optimal non-homogeneous utterances.

BUT (Vydana et al., 2021) participated with a cascade system fed with the "given" automatic segmentation of the test data. The primary submission is obtained from a system exploiting joint training of the ASR and MT components, model ensembling and tight ASR-MT coupling. Both ASR and MT are pre-trained on pre-processed clean data and rely on Transformer-based components. Two different ASR models are respectively trained to generate normalized and punctuated text, the latter leading to better results. In the proposed joint training procedure, the context vectors from the final layer of the ASR-decoder are used as inputs by the MT module, and both models are jointly optimized using a multi-task loss. At inference time, beam search is used to obtain the ASR hypotheses, and the corresponding context vectors obtained from the ASR model are used by the MT model for generating translations. The MT model also uses a beam search to produce the hypothesis and the final ST hypothesis is obtained by a coupled search using the joint likelihood from ASR and MT.

ESPNET-ST (Inaguma et al., 2021) participated with both cascade and end-to-end speech translation systems, with primary focus on the direct approach. Both systems are fed with "own" automatic segmentation of the test data. The primary

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

cascade system exploits an ASR component based on Conformer (Gulati et al., 2020a) and an MT component built with Transformer-base trained without case information and punctuation marks. The primary end-to-end system is based on the Conformer encoder, a stacked multi-block architecture including a multi-head self-attention module, a convolution module, and a pair of position-wise feed-forward modules in the Macaron-Net style (Lu et al., 2019). The baseline conformer is improved by training with sequence-level knowledge distillation and by adopting a Multi-Decoder architecture (which equips dedicated decoders for speech recognition and translation tasks in a unified encoder-decoder model enabling search in both source and target language spaces during inference), model ensembling and improved VAD-based audio segmentation (a "bottom-up" variant of (Potapczyk and Przybysz, 2020; Gaido et al., 2021)).

FBK (Papi et al., 2021) participated with an end-to-end-system fed with "own" automatic segmentation of the test data. The primary submission is obtained from a Transformer-based architecture trained with a pipeline involving data augmentation (SpecAugment (Park et al., 2019) and MT-based synthetic data) and characterized by knowledge distillation and a two-step fine-tuning procedure. Both knowledge distillation and the first fine-tuning step (optimized by combining label smoothed cross entropy and the CTC scoring function described in Gaido et al. 2020b) are carried out on manually segmented real and synthetic data. The second fine-tuning step is carried out on a random segmentation of the MuST-C v2 En-De dataset, aimed to make the system robust to automatically segmented test audio data (Gaido et al., 2020a). For the same purpose, a custom hybrid segmentation procedure (Gaido et al., 2021) is applied to the test data before passing them to the system.

HW-TSC participated with a cascade system fed with "own" automatic segmentation of the test data. The ASR component is a Transformer-large model, which is trained on the combination of LibriSpeech, MUST-C v2 and COVOST, where transcriptions are pre-pended by a label indicating the source corpus to make them distinguishable. During inference, the model is forced to decode in the MUST-C alike style by setting the first token as the MUST-C tag. The MT model is a Transformer-large model trained on the WMT19 corpus and fine-tuned on IWSLT-2017 text translation corpus.

KIT (Nguyen et al., 2021) participated with both cascade and end-to-end speech translation systems fed with "own" automatic segmentation of the test data (obtained from the WerRTCVAD toolkit[22]). The primary cascade system exploits sequence-to-sequence ASR models trained with three architectures (LSTM, Transformer and Conformer). Before MT, a Transformer-based segmentation module is in charge to (monolingually) translate disfluent, broken, uncased ASR outputs into more fluent, written-style text with punctuation in order to match the data conditions of the translation system. This is done in a shifting window manner, in which decisions are drawn by means of a simple voting mechanism. For MT, the systems relies on an ensemble of Transformer-large models trained on both clean and noisy synthetic (TED-derived) data. The primary end-to-end system is an improved version of last year's Speech Relative Transformer architecture (Pham et al., 2020c). Its encoder self-attention layer uses Bidirectional relative attention (Pham et al., 2020a) to model the relative distance between one position and other positions in the sequence. Three models, trained with SpecAugment (Park et al., 2019) and different activation functions (GeLU, SiLU and ReLU), are eventually combined in an ensemble.

LI participated with a cascade system fed with the "given" automatic segmentation of the test data. Both the ASR (three models) and the MT components (two models) are based on fairseq (Ott et al., 2019)[23] and were trained on MuST-C data.

NIUTRANS (Xu et al., 2021b) participated with an end-to-end-system fed with "own" automatic segmentation of the test data. The primary submission relies on a deep Transformer model implemented in fairseq and improved by adding the CTC loss as auxiliary loss on the encoders. The system is also enhanced with Conformer (used to replace the Transformer blocks in the encoder), relative position encoding (to improve acoustic modeling and generalize better for unseen sequence lengths; Shaw et al., 2018), and stacked acoustic and textual encoding (to better encode the

[22] http://github.com/wiseman/py-webrtcvad

[23] http://github.com/pytorch/fairseq.git

speech features; Xu et al., 2021a). Data augmentation is also applied via spectrogram augmentation, speed perturbation and sequence-level knowledge distillation, as well as by generating new synthetic speech from MT data and by translating into German the English transcriptions of ASR and ST data. Finally, ensemble decoding is applied to integrate the predictions from several models trained with the different datasets.

OPPO participated with a cascade system fed with the "given" automatic segmentation of the test data. The primary submission is based on Transformer for both the ASR and MT components, which are trained on part of allowed training datasets (MUSTC, LibriSpeech, CoVost, and WMT20). Structured dropout is applied to increase the differences between different models, which are eventually combined via average ensembling.

UPC (Gállego et al., 2021) participated with an end-to-end-system fed with "own" automatic segmentation of the test data (inspired by (Potapczyk et al., 2019)). The primary submission combines a Wav2Vec 2.0 encoder and an mBART decoder, which are respectively pre-trained on the ASR and MT tasks. A length adaptor module, consisting of a stack of convolutional layers, alleviates the length discrepancy between the speech and text modalities. Model fine-tuning to the ST task was carried out following the LNA strategy proposed in (Li et al., 2021). In addition, based on the ST improvements reported in (Escolano et al., 2020), an Adapter module was added to extract richer representations from the output of the encoder (Bapna and Firat, 2019). Data augmentation is also performed via randomized on-the-fly perturbations obtained by adding an echo effect and by modifying tempo and pitch, as well as by applying masking to the output of the Wav2Vec 2.0 feature extraction module. Different approaches were explored to combine the fine-tuning of the pre-trained models and the training of the intermediate modules. The best performance was obtained with a two-stage strategy, where: 1) the Wav2Vec and mBART models are frozen and the intermediate modules are forced to learn how to couple them; 2) model fine-tuning follows the LNA strategy, starting from the solid initial point obtained in the previous step.

VOLCTRANS (Zhao et al., 2021) participated with both cascade and end-to-end speech translation systems fed with the "given" automatic segmentation of the test data. The primary cascade system exploits a Transformer-based ASR trained, using spectrogram augmentation, on both clean and filtered noisy data. MT processing relies on Transformer-based models trained with data augmentation (via back-translation, knowledge distillation and ASR output adaptation) and combined with model ensemble techniques. The primary end-to-end system is trained by exploiting knowledge distillation (leveraging ASR datasets and four MT models) for data augmentation. The encoder and the decoder are pre-trained in a progressive multi-task learning framework, also exploiting a fbank2vec network to learn contextualized audio representations from log Mel-filterbank features.

VUS (Jo et al., 2021) participated with a cascade system fed with the "given" automatic segmentation of the test data. For the ASR component, a pretrained wav2vec 2.0 model (Baevski et al., 2020) was used for the embeddings, and the training was conducted with a Transformer augmented on the output layer of the wav2vec module. Following Potapczyk and Przybysz (2020), data pre-processing was made to remove training samples (laughters, applauses and erroneous scripts) that can lower the ASR performance. ASR output post-processing was also carried out to obtain an accurate sentence-level output, such as setting the sentence boundary between the fragment texts and re-aggregating some wrongly merged sentences. The MT component, also based on Transformer, was trained on a pre-processed version (language identification and length-based filtering and written-to-spoken text conversion through lowercasing, punctuation removal and abbreviations' expansion similar to Bahar et al., 2020) of the WMT 20 en-de news task dataset.

3.4 Results

Detailed results for the offline ST task are provided in Appendix A.2. Specifically, two separate tables respectively show the BLEU scores of participants' primary submissions computed on this year's *tst2021* and last year's *tst2020* test sets. In each table, three BLEU scores are reported, namely:

- `BLEU_NewRef` – computed on the new (exact, literal) translations described in Section

3.2;

- BLEU_TEDRef – computed on the original (subtitle-like) TED translations;

- BLEU_MultiRef – computed using both references in a multi-reference setting.

Systems are ranked according to their BLEU_NewRef score. Background colours are used to differentiate between cascade (white background) and end-to-end architectures (gray background). Additionally, the segmentation strategy (Own vs Given) and the training data condition (Constrained vs Unconstrained) characterising each primary submission are shown in separate columns.

Official results. In terms of this year's BLEU_NewRef primary metric, the top-ranked system achieved a BLEU score of 24.6, which is slightly below the one obtained by last year's winning system (25.3). Also the average (19.8) and median scores (21.7) are inferior compared to last year's round of the evaluation (average: 20.15; median: 21.81). These results, however, are not comparable since they are computed on a different test set (built from different TED talks), which also comprises reference translations that are not the original ones. The evaluation of this year's systems on *tst2020*, which is discussed below, is hence more informative if we want to get an idea about the actual evolution of ST technology.

Computing BLEU on the original TED translations (BLEU_TEDRef) results in overall scores that are significantly lower (top submission: 20.3; average: 16.6; median: 18.2). This large drop indicates the difficulty for all systems to generate translations that are similar to the subtitle-like ones characterising the recent TED talks included in this year's test set.

Unsurprisingly, the BLEU_MultiRef results are considerably higher due to the positive effect of combining more references (top submission: 34.0; average: 27.7; median: 30.5). However, it is worth remarking that, in this multi-reference setting, 12 primary submissions out of 16 reached a BLEU score above 30.0.

Cascade vs end-to-end. A major finding from last year (Ansari et al., 2020) was about the complete reduction of the performance gap between cascade and end-to-end systems. In the same direction, the analysis proposed in (Bentivogli et al., 2021) has recently shown through manual analyses and post-editing-based evaluations that the two paradigms are now substantially on par. In apparent contradiction, this year's results depict a different situation: the two top ranked submissions in the official ranking (based on BLEU_NewRef) are in fact produced with cascade systems (respectively scoring 24.6 and 23.4 BLEU). The first end-to-end submission (obtained under the same segmentation and training data conditions) is two BLEU points below (22.6) the top-ranked system. However, it is interesting to note that the type of reference translations used for evaluation makes a big difference in terms of final results. While all systems perform significantly worse when BLEU is computed against the original TED translations, some low-ranked submissions would climb the rankings if BLEU_TEDRef were used as primary metric. Although this year's winner would remain the same, the 12^{th} and 13^{th} submission would jump respectively to the 3^{rd} and 2^{nd} position. Notably, with a ranking based on BLEU_TEDRef, 7 of the top 10 positions would be occupied by the end-to-end submissions.[24]

All in all, in terms of performance distance between the two paradigms, our findings support those of (Bentivogli et al., 2021) about relying on automatic scores computed against independent references. Across metrics, test sets and language directions, they are less coherent than those computed on human post-edits. Different from last year, in this round the clear winner according to all possible rankings is a cascade system. However, its distance from the other end-to-end systems varies considerably depending on the type of reference translations used (down to 0.7 BLEU points in the ranking based on BLEU_TEDRef). In light of this variability, manual analyses and post-editing-based evaluations like the ones presented in (Bentivogli et al., 2021), would help to precisely assess if the observed BLEU score differences (marginal with BLEU_TEDRef) actually make one approach preferable to the other by final users.

[24] System's ranking based on BLEU_NewRef would end up similarly, with 6 end-to-end submissions in the top 10 positions (the top 2 still being the same cascade systems dominating the official ranking).

The importance of input segmentation. Another important finding from last year's evaluation concerned the importance of properly segmenting the input speech at test time, so to feed the systems with inputs that are closer to the sentence-like segments present in the clean corpora on which they are trained. Also this year, the top five primary runs submitted are all obtained by systems operating with "own" segmentation strategies, which prove to be helpful independently of the underlying paradigm. This is confirmed by the fact that the three lowest BLEU scores are achieved by participants opting for the "given" segmentation. Similar trends emerge with all possible rankings (`BLEU_NewRef`, `BLEU_TEDRef`, and `BLEU_MultiRef`). The importance of a proper segmentation of the input speech is even more evident if we look at the results computed on the *tst2020* test set, where the top seven runs are obtained with custom segmentation and the worst 5 with the given one. These findings are in line with last year's observations and motivate further efforts on improving this critical pre-processing step.

Progress wrt 2020. Overall results computed on *tst2020* are higher compared to those obtained on *tst2021*. However, being the two test sets different as discussed above, the scores are not directly comparable to draw reliable conclusions about the ST technology evolution (which might wrongly be considered as an involution by merely comparing raw BLEU scores on the two benchmarks). Rather, more can be said if we only focus on how this year's systems behave on *tst2020*. The improvement is evident both if we look at the average performance (increasing by more than 1 BLEU point from 20.15 to 21.17) and if we concentrate on the best systems. Specifically, with "own" test data segmentation methods, three teams achieved BLEU scores that are higher (up to 0.7 points) than the one obtained by the 2020 winner under this condition (25.3). With the "given" automatic audio splits, two teams improved (up to 1.8 points) the highest score obtained last year under this condition (22.49). Interestingly, similar to last year, the best system is an end-to-end one. The performance distance with respect to the best cascade result on *tst2020* is even larger (0.6 BLEU points) compared to the one observed last year (0.24). On one side, these results confirm that, on last year's test data (and with BLEU scores computed on the original TED translations), the end-to-end paradigm has an edge on the cascade one. On the other side, they confirm the above observations about the variability of automatic evaluation outcomes, which are highly affected by the overall testing conditions.

Final remarks. By inspecting this year's results, we can draw two final observations that, with an eye at the future, provide us with possible indications for the next rounds of the IWSLT offline ST task. One is about the training data condition: additional training resources did not yield visible advantages. Unfortunately, having only two "unconstrained" submissions makes it hard to draw reliable conclusions on this aspect. However, one might wonder if differentiating between constrained and unconstrained submissions still makes sense if the general goal is to boost research on a rapidly evolving technology. *Is it a good source of interesting observations or has it become an irrelevant distinction?* Reasoning on this question might yield indications for future rounds of the task.

The other observation is about how performance is distributed with respect to the two ST paradigms: while the results of cascade systems are spread across the whole performance interval (3.6 – 24.6 for `BLEU_NewRef`), the scores obtained by end-to-end models are concentrated in a two-point interval (20.6 – 22.6). Such a close performance of direct models should stimulate reflection on the fact that either the architectural restrictions posed to define the "end-to-end" setting (i.e. bypass any intermediate symbolic representation), or other limitations of current technology, result in systems that are quite similar to each other. *Is it still reasonable, for the good of ST, limiting participant's freedom with arbitrary, pre-defined architectural constraints?* Setting less restrictive conditions to experiment with, thus opening to participation with alternative approaches (e.g. by avoiding explicit architectural constraints) is a possible direction to promote more innovation in future rounds of the evaluation campaign.

4 Multilingual Speech Translation

While multilingual translation is an established task, until recently, few parallel resources existed for speech translation and most remain only for translation from English speech. Multilingual models enable transfer from related tasks, which is particularly important for low-resource languages; however, parallel data between two otherwise high-resource languages can also often be rare, making multilingual and zero-shot translation important for many resource settings.

In addition to parallel speech and translations, many sources of data may be useful for speech translation: monolingual speech and transcripts, parallel text, and data from other languages or language pairs. While cascades of separately trained automatic speech recognition (ASR) and machine translation (MT) models can leverage all of these data sources, how to most effectively do so with end-to-end models remains an open and exciting research question.

Speech	**Target Languages**				
	en	es	fr	pt	it
es	Supervised	Supervised	Supervised	Supervised	Supervised
fr	Supervised	Supervised	Supervised	Supervised	—
pt	Supervised	Zero-shot	—	Supervised	—
it	Zero-shot	Zero-shot	—	—	Supervised

Table 3: **Multilingual task language pairs**. Languages are represented by their ISO 639-1 code. Speech, transcripts, and translations were provided for all **Supervised** tasks; for **Zero-shot** ST tasks, only speech and transcripts were provided during training, though target language text may be seen with other source languages. Participants were required to submit translations for all official translation directions.

4.1 Challenge

Motivated by the above, the multilingual speech translation task provided data for two conditions: supervised, and zero-shot. We provided speech and transcripts for four languages (Spanish, French, Portuguese, Italian) and translations in a subset of five languages (English, Spanish, French, Portuguese, Italian) as shown in Table 3. For zero-shot language pairs, data for ASR (speech and transcripts) was released for training, but not translations; the target languages could be observed in other language pairs in training. Both translation directions for one source language (Italian) and one of two translation directions for another (Portuguese) were chosen to be zero-shot to enable comparison between supervised and zero-shot conditions with the same source language, and to measure the impact of having no supervised ST data at all. Participants could use the provided resources in any way.

At evaluation time, we provided speech in the four source languages and asked participants to generate translations in both English and Spanish. Both constrained submissions (using the provided data only, e.g., no models pretrained on external data) and unconstrained submissions were encouraged and evaluated separately. Submitting translations for additional optional language pairs as well as generated transcripts (ASR) for evaluation was not mandatory but encouraged as a useful point of analysis.

4.2 Data and Metrics

For this task we use the Multilingual TEDx data (mTEDx) (Salesky et al., 2021). The data is derived from TEDx talks and translations. The mTEDx data is segmented and aligned at the sentence-level (using automatically generated segmentations and alignments). mTEDx is publicly available on OpenSLR.[25] The data released during the training period contained train, validation, and progress test sets. For the purposes of this task, ST data for three language pairs was withheld until after the evaluation period (Zero-shot in Table 3). Use of any of resources beyond Multilingual TEDx made a submission unconstrained. Any publicly available additional data or pretrained models were permitted for training unconstrained systems.

We evaluated translation output using BLEU as computed by SacreBLEU (Post, 2018) and WER for ASR output. We computed all scores using the provided utterance segmentations from Multilingual TEDx. WER was computed on lowercased text with punctuation removed.

4.3 Submissions

We received 15 submissions from 7 teams.

FAIR (Tang et al., 2021a) submitted unconstrained end-to-end models which leverage pretrained multilingual wav2vec 2.0 and mBART models, and finetune on the provided mTEDx MT and ST data as well as additional corpora. They compare different wav2vec 2.0 models trained on different multilingual corpora and either text (Baevski et al., 2020) or IPA targets (Wang et al., 2021), and mBART with BPE (Liu et al., 2020) or IPA representations (Tang et al., 2021b). They combine different joint and speech-only finetuning, and add an adaptor layer (Li et al., 2021) between the two pretrained models for adapta-

[25] http://openslr.org/100/

tion and downsampling. They ultimately ensemble three models for their final submission.

HWN (Zeng et al., 2021) used a unified Transformer architecture in which audio and text data can be featurized separately by a Conv-Transformer (Huang et al., 2020) and text embeddings, before being fused and used as input to a single encoder and decoder. They use curriculum learning by first training the unified model for the ASR and MT tasks, then continue training adding the ST task and finally fine-tuning using the ST task data only. They also use multiple data augmentation techniques and model ensembling.

KIT (Pham et al., 2021) trained deep Transformer models with relative attention for ASR and ST (Pham et al., 2019, 2020b) to create both cascaded and E2E models. They used additional techniques such as distillation, Macaron feed-forward layers, and the creation of synthetic data to significantly improve both models' performance. Their final submission is an ensemble of their cascade and E2E systems.

UM-DKE (Liu and Niehues, 2021) trained multilingual cascade and E2E models with a variety of techniques to improve performance. They start with a multilingual ASR model, which incorporates language embeddings, speed perturbation, and ensembling. They improve their multilingual MT by removing residual connections in the Transformer architecture, and further ensembling. Finally they train an E2E ST system which benefits from joint training with ASR, pseudo-labels for synthetic data to improve zero-shot pairs, and 'multi-view ensembling,' which ensembles probabilities based on three different speed perturbations.

ON-TRAC (Le et al., 2021) used a dual-decoder Transformer architecture (Le et al., 2020), which includes a single encoder for speech data and separate decoders (that interact with each other) for each of the ASR and ST tasks. They trained ASR and MT models to initialize the ST model and used SpecAugment augmentation. No synthetic data was created for zero-shot translation.

UEDIN (Zhang and Sennrich, 2021) trained multilingual Transformer models with adaptive feature selection (Zhang et al., 2020) to reduce data dimensionality by selecting the most informative speech features. They create pseudo-speech translation data which provides significant improvements on all language pairs, not only zero-shot. They additionally use sparsified linear attention, RMSNorm, scheduling language-specific modeling, and multi-task learning to improve their models, and ensemble models of multiple sizes for their final submission.

ZJU (Zhang, 2021) submitted an ensemble of cascaded ST models, using a Conformer (Gulati et al., 2020b) for ASR and a multilingual Transformer MT model. They use back-translation to create data for zero-shot pairs, add noised data to adapt their MT model to ASR output, and include masked training. They additionally compared end-to-end models with data augmentation and multi-task training.

4.4 Results

Results for the Multilingual Task are shown in Appendix A.3. We calculated task results using the average BLEU on all official ST language pairs: all primary submissions are shown in Table 5. The unconstrained submission from FAIR outperformed all other primary submissions on both supervised and zero-shot language pairs. The KIT submission was the strongest constrained system, aided in part by strong ASR pretraining: ASR results are shown in Table 8. All but one primary submission were ensembles of either multiple end-to-end systems, or end-to-end and cascaded models. We saw a mix of end-to-end and cascaded submissions across primary and constrastive submissions (Table 6); in general, the end-to-end models outperformed cascaded submissions, particularly under zero-shot conditions. We discuss different aspects of the task and submissions further below.

Constrained vs unconstrained. Use of additional data beyond mTEDx appeared to be a clear benefit on all ST pairs, as the FAIR system performed best on all language pairs. Interestingly, the performance difference between the best unconstrained and constrained systems across supervised and zero-shot tasks was similar. When we look at the constrastive submissions and ASR, however, the underlying reason appears not to be the additional data but rather how it is used. The FAIR baseline is initialized from the multilingual wav2vec2.0 model XLSR-53 and the mBART decoder, and is outperformed by many constrained systems. The other FAIR submissions used co-training with the text-to-text MT task and IPA representations for ASR and/or MT models for significant improvements.

Zero-shot performance. Overall we saw very encouraging performance on the zero-shot pairs, with very little degradation from the supervised language pairs for many systems. Three language pairs were zero-shot: pt-es, it-en, and it-es. While Portuguese speech was observed in another translation pair, Italian speech was only observed for ASR. The Italian pairs proved more challenging, but most systems nonetheless outperform the supervised end-to-end baselines in Salesky et al. (2021) through some combination of decoder pretraining, auto-encoding ASR data, or back-translation. Comparing supervised and zero-shot performance with the same source language (pt), we saw stronger performance on the zero-shot than supervised condition, likely indicative of the relatedness of the source and target languages, facilitating zero-shot translation. Though much more English target data has been seen (for constrained systems), pt-es and it-es are both more closely-related languages, and all but one system show better results on these two zero-shot language pairs than it-en. For teams which submitted both end-to-end and cascaded models, there were small but consistent improvements on zero-shot with end-to-end; this may suggest that E2E models more easily transfer from observed related languages and pairs, or perhaps that end-to-end models were more optimized. The systems with the greatest relative difference between supervised and zero-shot pairs were FAIR, HWN, and ON-TRAC. HWN had better performance for languages with more ASR data, and ON-TRAC struggled without e.g. auto-encoding text.

ASR performance impact. Interestingly, ASR performance was not necessarily indicative of ST performance; HWN and KIT ASR outperformed the FAIR ASR without additional training data or ensembling, with the exception of French where both systems struggled, particularly KIT. This was shown in ST performance; UEDIN outperformed KIT on language pairs where French was the source language, precisely where UEDIN had better ASR. All submitted ASR systems outperformed the end-to-end ASR in Salesky et al. (2021), in part through better optimization and use of multilingual models, and in particular use of the CTC objective. Their hybrid LF-MMI models remain generally stronger for Portuguese and French; not necessarily correlated with data size.

Ensembling. Most primary systems were ensembles of 2+ models, which provided improvements of up to 2 BLEU compared with the individual systems, some of which were submitted as constrastive (Table 6). We saw different ensembling techniques, using joint decoding or averaging model output probabilities. Ensembled models were alternatively models of different sizes (UEDIN), trained on different data (FAIR), different combinations of fine-tuning and knowledge distillation (HWN), system with back-translations and with ASR noise added (ZJU), speed perturbations of the same input (UM-DKE), or cascaded and end-to-end models (KIT).

Unofficial language pairs. The unofficial language pairs (Table 7) have the same source languages as the official language pairs, but different target languages. The test sets are parallel with the official blind evaluation sets. The relative performance between primary systems on these additional targets remains similar. Performance on more closely related languages (es-pt) was in fact generally higher, and language pairs with less-observed target languages (es-fr, es-it) were lower. The exception was FAIR, where average performance was almost exactly the same as on the official supervised pairs; the additional datasets used for pretraining likely erase some of these resource differences, supported by the differences between their constrastive submissions which use different pretraining sources.

End-to-End vs Cascade. Three groups submitted an end-to-end system and a cascaded system. In all three cases, the end-to-end system outperforms the cascaded approach. Since the tendency in the offline translation task (section 3) is different (there the cascaded approaches typically perform better than the end-to-end models), this opens up several interesting research questions that should be investigated further. There are several differences between the two tasks that could influence the ranking between the end-to-end and cascaded models: First of all, the amount of ASR and MT training data that is available in addition to end-to-end training data is different. In the offline task, there is significantly more data available for the auxiliary tasks (particularly MT), which may benefit cascaded models more. Secondly, the multilingual task uses provided automatic sentence segmentation which is consistent across train and test, while the offline task does not provide segmentation at test time, requiring teams to perform segmentation to translate, similar to online or simultaneous conditions, which cascaded models may be more robust to. And finally, the ability to facilitate multilingual and zero-shot speech translation might be different in end-to-end and cascaded models.

5 Low-Resource Speech Translation

The goal of the low-resource speech translation task is to investigate pathways for developing speech translation systems for currently underserved languages. The majority of the world's languages are predominantly oral, hence the need for speech-based language tools (translation included) is paramount for them to be of any use to the language community. At the same time, most of these languages are also under-resourced, with little to no data being available for speech transcription and translation.

While offline speech translation has a longstanding tradition at the IWSLT campaign and both monolingual and multilingual models offer impressive promises for downstream model deployment, the majority of recent advances in speech translation both require large amounts of data and are typically benchmarked on language pairs with such data abundance. However, for the vast majority of the world's languages there exist little speech-translation parallel data at the scale needed to train modern speech translation models. Instead, in a real-world situation one will have access to limited, disparate resources (e.g. word-level translations, speech recognition, small parallel text data, monolingual text, raw audio, etc). The low-resource track aims to fill this gap, by encouraging and facilitating research on speech-translation for data-scarce language pairs.

5.1 Challenge

As described above, the shared task focused on the problem of developing speech transcription and translation tools for under-resourced languages. This year's iteration in particular focused on speech translation tools that would match the real-world needs of humanitarian organizations.

There were no restrictions on the type of models (e.g. end-to-end vs. cascade) or additional data that were allowed, the goal for the participants being producing the best possible system under these challenging settings. In collaboration with the Translators without Border, we provided newly collected speech and transcripts in two languages, Coastal Swahili (ISO code: swh) and Congolese Swahili (ISO code: swc), as well as translations in English and French respectively. In addition, we provided pointers to other monolingual speech datasets in the source Swahili varieties, as well as textual parallel corpora between the source and target languages.

5.2 Data and Metrics

The Swahili Varieties Speech Translation Dataset For the purposes of the task we created and released a new speech translation dataset for the two Swahili varieties. The new dataset is publicly available.[26]

The training data were derived from the Gamayun minikits that the Translators without Borders had released for Congolese and Coastal Swahili text translation (Öktem et al., 2020), which included sentence-level translations between Coastal Swahili and English as well as Congolese Swahili and French.[27] We additionally collected read versions for 5,000 sentences from this dataset. For each variety the training set includes voices from 6 speakers (3 male and 3 female). The collection was carried out using mobile phones, as opposed to clean studio settings, to better match the real-world use-case scenarios the shared task envisions.

The development and test data are derived from the TICO-19 dataset (Anastasopoulos et al., 2020), which is a multi-parallel evaluation benchmark on the COVID-19 domain in more than 33 languages. The original English sentences were translated into Coastal Swahili and French, and the French translations were then translated into Congolese Swahili. All translations were performed by professional translators and an extensive quality assurance process was followed. For the purposes of the shared task we additionally collected read utterances in the two Swahili varieties for all 3k sentences. We follow the original dev and test splits. The dev set utterances encompass 2 speakers (1

[26] https://drive.google.com/file/d/1lhifoEY0Kzj6s11W_taKoVW_mAvzzZ04/view?usp=sharing

[27] This dataset was previously used for developing text-based translation systems for humanitarian response (Öktem et al., 2021).

Language Pair	Train		Dev		Test	
	#utt.	#speakers	#utt.	#speakers	#utt.	#speakers
swh-eng	4599	6 (3M, 3F)	868	2 (1M, 1F)	1063	3 (2M, 1F)
swc-fra	5000	6 (3M, 3F)	868	2 (1M, 1F)	2124	6 (3M, 3F)

Table 4: Statistics of the newly-released Swahili varieties speech translation corpus.

male, 1 female) in each language, and the test set includes 3 (2M, 1F) and 6 (3M, 3F) speakers for swh and swc respectively.

Statistics on the whole dataset used for the shared task following cleaning and preprocessing are listed in Table 4. The final dataset is 4-way parallel; the English and French sides are translations of each other, creating opportunities for the evaluation of multilingual systems, as well as, in the future, speech-to-speech translation between the two Swahili varieties.

Additional Data Last, we reiterate that we allowed the use of any other available data, such as any data from the Offline and Multilingual Shared Tasks, any speech recognition corpora like the Swahili ALFFA dataset (Gelas et al., 2012) or the Mozilla Common Voice datasets (Ardila et al., 2020), as well as any text translation datasets like the Gamayun minikits (Öktem et al., 2020). We also allowed the use of pre-trained models like wav2vec (Schneider et al., 2019; Baevski et al., 2020) or mBART (Liu et al., 2020) (among others).

Metrics Systems' performance was evaluated with respect to their capability to produce translations similar to the target-language references. We used the BLEU metric computed with SacreBLEU, in a case-insensitive setting. In addition, we invited participants who produced speech transcriptions in the Swahili variety as a by-product of their system (e.g. if they use a ASR+MT cascade approach) to also submit them. These were evaluated using case-insensitive word error rate (WER). The choice of case-insensitivity is due to our focus on producing *usable* output that aids comprehension; we deem that the effect of proper casing is largely minor in such challenging settings.

5.3 Submissions

The shared task received 4 submissions (9 total runs across the $\{$swh,swc$\}\times\{$eng,fra$\}$ pairs) from 3 teams. All teams followed a cascade ASR→MT approach in their primary submission – this indicates that end-to-end learning is still very challenging in such data-scarce settings, and leaves a lot of room for further future exploration.[28]

In the following, we provide an overview of each submission.

USYD-JD (Ding et al., 2021) uses a pipeline approach, focusing in the MT component and its ability to handle ASR errors. The ASR component is trained on the Swahili Varieties dataset, the ALFFA corpus, and the IARPA Babel Swahili Language Pack using the default settings in Kaldi, also lowercasing all sentences and removing punctuation. The final ASR is post-corrected with the SlotRefine method (Wu et al., 2020). The MT component is a Transformer (Vaswani et al., 2017) that operates in a non-autoregressive manner, trained on almost all available OPUS swa-eng datasets, but additionally utilizing denoising pre-training and bidirectional self-training, tagged back-translation, transductive fine-tuning, output reranking and output post-processing. This NMT system is the only that explores extensive strategies for denoising and pre-training, reaching a

IMS (Denisov et al., 2021) uses a pipeline approach. The ASR component for the primary submission is a Conformer (Gulati et al., 2020b) in its ESPnet implementation, trained by fine-tuning a pretrained SPGISpeech model (O'Neill et al., 2021) on both Swahili varieties using the Swahili Varieties dataset, Gamayun samples, the ALFFA corpus, and the IARPA Babel Swahili Language Pack, also applying some preprocessing steps like converting all numbers to words and removing punctuation. The MT system is a Transformer (Vaswani et al., 2017) using multi-task learning by tagging the input (to distinguish clean text vs. ASR output). They also attempted an end-

[28] We note that the shared task received more than 20 initial registrations. We suspect that the limited amount of received submissions was exactly because of how challenging it can be to create a system that produces decent outputs in these extremely low-resource settings.

to-end ST system which however performed significantly worse.

ON-TRAC (Le et al., 2021) used a pipeline approach, using a hybrid HMM/TDNN automatic speech recognition system fed by wav2vec (Schneider et al., 2019) features, with its output then provided to a neural MT system. The ASR system was trained on the Swahili Varieties dataset, the ALFFA corpus, and the IARPA Babel Swahili Language Pack. The NMT system uses LSTMs with attention, with the swa-eng also using subwords, while the swc-fra system operates at the word level. The swa-eng MT system was trained on 2.2M sentence pairs, resulting from the filtering through langID of all data available on OPUS.[29] The swc-fra NMT system was trained on 1.1M parallel sentences.

5.4 Results

Out of the submitted systems, the USYD-JD submission that explored pre-training strategies was the clear winner of the eng-swa task achieving a BLEU score (case insensitive) of 25.3. Notably, they only focused on the MT component of the pipeline, making it robust to ASR errors and utilizing monolingual data effectively through denoising and pre-training. For the swc-fra pair, the IMS system was the best performing submission for the swc-fra pair with a BLEU score of 13.5. The evaluation of all submissions (including optional language pairs and ASR transcription accuracy) is provided in the Appendix.

The difference in accuracy between the two language pairs could potentially be attributed to the lack of data in Congolese Swahili (as most available datasets are in the Coastal variety). However, the pre-training approaches that the USYD-JD submission uses seem very promising towards building robust MT systems also for the Congolese variety. A clear path for future work towards even better ST systems could explore a pipeline of the improved ASR systems of the ON-TRAC or IMS submissions with the NMT system of the USYD-JD submission. The lack of end-to-end approaches in the submissions (and the evidence from the IMS contrastive submission) suggest that additional research is needed in order to achieve competitive results in such data-scarce settings with end-to-end models.

[29] https://opus.nlpl.eu/

References

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, Alessandro Cattelan, Zi-Yi Dou, Marcello Federico, Christian Federmann, Dmitriy Genzel, Franscisco Guzmán, Junjie Hu, Macduff Hughes, Philipp Koehn, et al. 2020. Tico-19: the translation initiative for covid-19. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020.*

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

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

Parnia Bahar, Tobias Bieschke, Ralf Schluter, and Hermann Ney. 2021a. Tight integrated end-to-end training for cascaded speech translation. In *Proceedings of the IEEE Spoken Language Technology Workshop*, pages 950–957, Shenzhen, China.

Parnia Bahar, Patrick Wilken, Tamer Alkhouli, Andreas Guta, Pavel Golik, Evgeny Matusov, and Christian Herold. 2020. Start-Before-End and End-to-End: Neural Speech Translation by AppTek and RWTH Aachen University. In *Proceedings of the 17th International Conference on Spoken Language Translation (IWSLT).*

Parnia Bahar, Patrick Wilken, Mattia di Gangi, and Evgeny Matusov. 2021b. Without Further Ado: Direct and Simultaneous Speech Translation by AppTek in 2021. In *Proceedings of the 18th International Conference on Spoken Language Translation (IWSLT).*

Dzmitry Bahdanau, Kyung Hyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of the 3rd International Conference on Learning Representations, ICLR 2015*.

Ankur Bapna and Orhan Firat. 2019. Simple, scalable adaptation for neural machine translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1538–1548, Hong Kong, China. Association for Computational Linguistics.

Loïc Barrault, Magdalena Biesialska, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, Tom Kocmi, Philipp Koehn, Chi-kiu Lo, Nikola Ljubešić, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Santanu Pal, Matt Post, and Marcos Zampieri. 2020. Findings of the 2020 conference on machine translation (WMT20). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1–55, Online. Association for Computational Linguistics.

Benjamin Beilharz and Xin Sun. 2019. LibriVoxDeEn - A Corpus for German-to-English Speech Translation and Speech Recognition.

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.

Roldano Cattoni, Mattia Antonino Di Gangi, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2021. Must-c: A multilingual corpus for end-to-end speech translation. *Computer Speech & Language*, 66:101155.

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, Christian Girardi, and Marcello Federico. 2012. WIT^3: Web Inventory of Transcribed and Translated Talks. In *Proceedings of the Annual Conference of the European Association for Machine Translation (EAMT)*, Trento, Italy.

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.

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

Pavel Denisov, Manuel Mager, and Ngoc Thang Vu. 2021. IMS' Systems for the IWSLT 2021 Low-Resource Speech Translation Task. In *Proceedings of the 18th International Conference on Spoken Language Translation (IWSLT)*.

Mattia A. Di Gangi, Roldano Cattoni, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2019. MuST-C: a Multilingual Speech Translation Corpus. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2012–2017, Minneapolis, Minnesota.

Liang Ding, Di Wu, and Dacheng Tao. 2021. The USYD-JD's Speech Translation System for IWSLT2021. In *Proceedings of the 18th International Conference on Spoken Language Translation (IWSLT)*.

Kosuke Doi, Katsuhito Sudoh, and Satoshi Nakamura. 2021. Large-Scale English-Japanese Simultaneous Interpretation Corpus: Construction and Analyses with Sentence-Aligned Data. In *Proceedings of the 18th 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.

Maha Elbayad, Laurent Besacier, and Jakob Verbeek. 2020. Efficient Wait-k Models for Simultaneous Machine Translation. In *Proc. Interspeech 2020*, pages 1461–1465.

Carlos Escolano, Marta R. Costa-jussà, José A. R. Fonollosa, and Carlos Segura. 2020. Enabling zero-shot multilingual spoken language translation with language-specific encoders and decoders.

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.

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.

Ryo Fukuda, Yui Oka, Yausumasa Kano, Yuki Yano, Yuka Ko, Hirotaka Tokuyama, Kosuke Doi, Sakriani Sakti, Katsuhito Sudoh, and Satoshi Nakamura. 2021. NAIST English-to-Japanese Simultaneous Translation System for IWSLT 2021 Simultaneous Text-to-text Task. In *Proceedings of the 18th International Conference on Spoken Language Translation (IWSLT)*.

Marco Gaido, Mattia Antonino Di Gangi, Matteo Negri, Mauro Cettolo, and Marco Turchi. 2020a. Contextualized translation of automatically segmented speech. In *Proceedings of Interspeech 2020, 21st Annual Conference of the International Speech Communication Association*, pages 1471—-1475, Shanghai, China.

Marco Gaido, Mattia Antonio Di Gangi, Matteo Negri, and Marco Turchi. 2020b. End-to-End Speech-Translation with Knowledge Distillation: FBK@IWSLT2020. In *Proceedings of the 17th International Conference on Spoken Language Translation (IWSLT)*.

Marco Gaido, Matteo Negri, Mauro Cettolo, and Marco Turchi. 2021. Beyond voice activity detection: Hybrid audio segmentation for direct speech translation.

Gerard I. Gállego, Ioannis Tsiamas, Carlos Escolano, and Marta R. Costa-jussà José A. R. Fonollosa. 2021. UPC's Speech Translation System for IWSLT 2021. In *Proceedings of the 18th International Conference on Spoken Language Translation (IWSLT)*.

Hadrien Gelas, Laurent Besacier, and Francois Pellegrino. 2012. Developments of Swahili resources for an automatic speech recognition system. In *SLTU - Workshop on Spoken Language Technologies for Under-Resourced Languages*, Cape-Town, Afrique Du Sud.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. 2020a. Conformer: Convolution-augmented transformer for speech recognition. In *Proceedings of Interspeech 2020, 21st Annual Conference of the International Speech Communication Association*, pages 5036—-5040, Shanghai, China.

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

François Hernandez, Vincent Nguyen, Sahar Ghannay, Natalia A. Tomashenko, and Yannick Estève. 2018. TED-LIUM 3: twice as much data and corpus repartition for experiments on speaker adaptation. *CoRR*, abs/1805.04699.

Wenyong Huang, Wenchao Hu, Yu Ting Yeung, and Xiao Chen. 2020. Conv-transformer transducer: Low latency, low frame rate, streamable end-to-end speech recognition. *INTERSPEECH*.

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).

Yong Rae Jo, Young Ki Moon, Minji Jung, Jungyoon Choi, Jihyung Moon, and Won Ik Cho. 2021. VUS at IWSLT 2021: A Finetuned Pipeline for Offline Speech Translation. In *Proceedings of the 18th International Conference on Spoken Language Translation (IWSLT)*.

Ali Can Kocabiyikoglu, Laurent Besacier, and Olivier Kraif. 2018. Augmenting Librispeech with French Translations: A Multimodal Corpus for Direct Speech Translation Evaluation. In *Proceedings of LREC 2018*, Miyazaki, Japan.

Hang Le, Florentier Barbier, Ha Nguyen, Natalia Tomanshenko, Salima Mdhaffar, Souhir Gahbiche, Fethi Bougares, Benjamin Lecouteux, Didier Schwab, and Yannick Esteve. 2021. ON-TRAC's systems for the IWSLT 2021 low-resource speech translation and multilingual speech translation shared tasks. In *Proceedings of the 18th International Conference on Spoken Language Translation (IWSLT)*.

Hang Le, Juan Pino, Changhan Wang, Jiatao Gu, Didier Schwab, and Laurent Besacier. 2020. Dual-decoder transformer for joint automatic speech recognition and multilingual speech translation. In *COLING*, pages 3520–3533.

Xian Li, Changhan Wang, Yun Tang, Chau Tran, Yuqing Tang, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. 2021. Multilingual speech translation with efficient finetuning of pretrained models.

Pierre Lison, Jörg Tiedemann, and Milen Kouylekov. 2018. OpenSubtitles2018: Statistical rescoring of sentence alignments in large, noisy parallel corpora. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Dan Liu, Mengge Du, Xiaoxi Li, Yuchen Hu, and Lirong Dai. 2021. The USTC-NELSLIP Systems for Simultaneous Speech Translation Task at IWSLT 2021. In *Proceedings of the 18th International Conference on Spoken Language Translation (IWSLT)*.

Danni Liu and Jan Niehues. 2021. Maastricht University's Multilingual Speech Translation System for IWSLT 2021. In *Proceedings of the 18th 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.

Yiping Lu, Zhuohan Li, Di He, Zhiqing Sun, Bin Dong, Tao Qin, Liwei Wang, and Tie-Yan Liu. 2019. Understanding and improving transformer from a multi-particle dynamic system point of view. *arXiv preprint arXiv:1906.02762*.

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.

Tuan-Nam Nguyen, Thai-Son Nguyen, Christian Huber, Maximilian Awiszus, Ngoc-Quan Pham, Thanh-Le Ha, Felix Schneider, Sebastian Stuker, and Alexander Waibel. 2021. KIT's IWSLT 2021 Offline Speech Translation System. In *Proceedings of the 18th International Conference on Spoken Language Translation (IWSLT)*.

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.

Alp Öktem, Eric DeLuca, Rodrigue Bashizi, Eric Paquin, and Grace Tang. 2021. Congolese swahili machine translation for humanitarian response. arXiv:2103.10734.

Alp Öktem, Muhannad Albayk Jaam, Eric DeLuca, and Grace Tang. 2020. Gamayun-language technology for humanitarian response. In *2020 IEEE Global Humanitarian Technology Conference (GHTC)*, pages 1–4. IEEE.

Patrick K O'Neill, Vitaly Lavrukhin, Somshubra Majumdar, Vahid Noroozi, Yuekai Zhang, Oleksii Kuchaiev, Jagadeesh Balam, Yuliya Dovzhenko, Keenan Freyberg, Michael D Shulman, et al. 2021. Spgispeech: 5,000 hours of transcribed financial audio for fully formatted end-to-end speech recognition. arXiv:2104.02014.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An asr corpus based on public domain audio books. *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210.

Sara Papi, Marco Gaido, Matteo Negri, and Marco Turchi. 2021. Dealing with training and test segmentation mismatch: FBK@IWSLT2021. In *Proceedings of the 18th International Conference on Spoken Language Translation (IWSLT)*.

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 (ACL)*, pages 311–318, Philadelphia, USA.

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.

Ngoc-Quan Pham, Thanh-Le Ha, Tuan-Nam Nguyen, Thai-Son Nguyen, Elizabeth Salesky, Sebastian Stüker, Jan Niehues, and Alex Waibel. 2020a. Relative positional encoding for speech recognition and direct translation. In *INTERSPEECH*, pages 31–35. ISCA.

Ngoc-Quan Pham, Thanh-Le Ha, Tuan-Nam Nguyen, Thai-Son Nguyen, Elizabeth Salesky, Sebastian Stüker, Jan Niehues, and Alex Waibel. 2020b. Relative positional encoding for speech recognition and direct translation. In *INTERSPEECH*, pages 31–35. ISCA.

Ngoc-Quan Pham, Dan He, Tuan-Nam Nguyen, Thanh-Le Ha, Sebastian Stuker, and Alexander Waibel. 2021. Multilingual Speech Translation KIT @ IWSLT2021. In *Proceedings of the 18th International Conference on Spoken Language Translation (IWSLT)*.

Ngoc-Quan Pham, Thai-Son Nguyen, Jan Niehues, Markus Müller, Sebastian Stüker, and Alexander Waibel. 2019. Very deep self-attention networks for end-to-end speech recognition.

Ngoc-Quan Pham, Felix Schneider, Tuan-Nam Nguyen, Thanh-Le Ha, Thai-Son Nguyen, Maximilian Awiszus, Sebastian Stüker, and Alexander Waibel. 2020c. KIT's IWSLT 2020 SLT Translation System. In *Proceedings of the 17th International Conference on Spoken Language Translation (IWSLT)*.

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, Belgium, Brussels. Association for Computational Linguistics.

Tomasz Potapczyk and Pawel Przybysz. 2020. SRPOL's System for the IWSLT 2020 End-to-End Speech Translation Task. In *Proceedings of the 17th International Conference on Spoken Language Translation (IWSLT)*.

Tomasz Potapczyk, Pawel Przybysz, Marcin Chochowski, and Artur; Szumaczuk. 2019. Samsung's System for the IWSLT 2019 End-to-End Speech Translation Task. In *Proceedings of 16th International Workshop on Spoken Language Translation (IWSLT)*, Hong Kong.

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, Matthew Wiesner, Jacob Bremerman, Roldano Cattoni, Matteo Negri, Marco Turchi, Douglas W. Oard, and Matt Post. 2021. Multilingual tedx corpus for speech recognition and translation.

Ramon Sanabria, Ozan Caglayan, Shruti Palaskar, Desmond Elliott, Loïc Barrault, Lucia Specia, and Florian Metze. 2018. How2: a large-scale dataset for multimodal language understanding. In *Proceedings of the Workshop on Visually Grounded Interaction and Language (ViGIL)*. NeurIPS.

Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. 2019. wav2vec: Unsupervised pre-training for speech recognition. In *INTERSPEECH*.

Sukanta Sen, Ulrich Germann, and Barry Haddow. 2021. The University of Edinburgh's Submission to the IWSLT21 Simultaneous Translation Task. In *Proceedings of the 18th International Conference on Spoken Language Translation (IWSLT)*.

Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. 2018. Self-attention with relative position representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 464–468, New Orleans, Louisiana. Association for Computational Linguistics.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A Study of Translation Edit Rate with Targeted Human Annotation. In *Proceedings of the The Seventh Conference of the Association for Machine Translation in the Americas (AMTA)*, pages 223–231, Cambridge, USA.

Milos Stanojevic and Khalil Sima'an. 2014. BEER: BEtter evaluation as ranking. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*. Association for Computational Linguistics.

Yun Tang, Hongyu Gong, Xian Li, Changhan Wang, Juan Pino, Holger Schwenk, and Naman Goyal. 2021a. FST: the FAIR Speech Translation System for the IWSLT21 Multilingual Shared Task. In *Proceedings of the 18th International Conference on Spoken Language Translation (IWSLT)*.

Yun Tang, Juan Pino, Changhan Wang, Xutai Ma, and Dmitriy Genzel. 2021b. A general multi-task learning framework to leverage text data for speech to text tasks.

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*.

Hari Krishna Vydana, Martin Karafiát, Lukáš Burget, and Honza Černocky. 2021. The IWSLT 2021 BUT Speech Translation Systems. In *Proceedings of the 18th International Conference on Spoken Language Translation (IWSLT)*.

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

Changhan Wang, Morgane Rivière, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. 2021. Voxpopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. *arXiv preprint arXiv:2101.00390*.

Weiyue Wang, Jan-Thorsten Peter, Hendrik Rosendahl, and Hermann Ney. 2016. CharacTer: Translation edit rate on character level. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*. Association for Computational Linguistics.

Di Wu, Liang Ding, Fan Lu, and Jian Xie. 2020. Slotrefine: A fast non-autoregressive model for joint intent detection and slot filling. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1932–1937.

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

Chen Xu, Xiaoqian Liu, Xiaowen Liu, Laohu Wang, Canan Huang, Tong Xiao, and Jingbo Zhu. 2021b. The NiuTrans End-to-End Speech Translation System for IWSLT 2021 Offline Task. In *Proceedings of the 18th International Conference on Spoken Language Translation (IWSLT)*.

Xingshan Zeng, Liangyou Li, and Qun Liu. 2021. Multilingual Speech Translation with Unified Transformer: Huawei Noah's Ark Lab at IWSLT 2021. In *Proceedings of the 18th International Conference on Spoken Language Translation (IWSLT)*.

Albert Zeyer, Kazuki Irie, Ralf Schlüter, and Hermann Ney. 2018. Improved training of end-to-end attention models for speech recognition. In *Interspeech*, Hyderabad, India.

Biao Zhang and Rico Sennrich. 2021. Edinburgh's End-to-End Multilingual Speech Translation System for IWSLT 2021. In *Proceedings of the 18th International Conference on Spoken Language Translation (IWSLT)*.

Biao Zhang, Ivan Titov, Barry Haddow, and Rico Sennrich. 2020. Adaptive feature selection for end-to-end speech translation. In *Findings of EMNLP*, pages 2533–2544.

Linlin Zhang. 2021. ZJU's IWSLT 2021 Speech Translation System. In *Proceedings of the 18th International Conference on Spoken Language Translation (IWSLT)*.

Chengqi Zhao, Zhicheng Liu, Jian Tong, Tao Wang, Mingxuan Wang, Rong Ye, Qianqian Dong, Jun Cao, and Lei Li. 2021. The Volctrans Neural Speech Translation System for IWSLT 2021. In *Proceedings of the 18th International Conference on Spoken Language Translation (IWSLT)*.

Appendix A. Evaluation Results and Details

A.1. Simultaneous Speech Translation

· Summary of the results of the simultaneous speech translation **text track.**
· Results are reported on the blind test set and systems are grouped by latency regime (set on tst-COMMON v2 or IWSLT21 dev set)
· Raw system logs are also provided on the task web site.[30]

English-German	**tst-COMMON v2**				**Blind Test Set**			
	BLEU	AL	AP	DAL	BLEU	AL	AP	DAL
				Low Latency				
USTC-NESLIP	33.16	2.66	0.64	4.38	26.89	2.81	0.63	4.72
VOLCTRANS	28.76	2.86	0.69	4.22	23.24	3.08	0.68	4.25
APPTEK	30.03	2.94	0.68	4.40	22.84	3.12	0.66	4.66
UEDIN	25.06	2.33	0.63	3.69	22.30	4.22	0.71	5.54
				Medium Latency				
USTC-NESLIP	34.82	5.80	0.80	8.89	29.40	5.94	0.78	9.29
VOLCTRANS	32.88	5.80	0.83	9.05	27.22	6.30	0.81	9.24
APPTEK	31.73	5.89	0.80	9.57	25.70	6.22	0.78	10.40
UEDIN	30.58	5.89	0.80	7.20	24.56	6.92	0.81	8.20
				High Latency				
USTC-NESLIP	35.47	12.21	0.95	15.18	30.03	12.35	0.93	16.33
VOLCTRANS	33.23	11.03	0.93	11.40	26.82	12.03	0.92	12.39
APPTEK	33.16	11.19	0.92	14.44	26.62	12.00	0.91	16.05
UEDIN	33.10	14.69	0.98	15.17	26.50	15.41	0.96	16.04

English-Japanese	**IWSLT 21 DEV**				**Blind Test Set**			
	BLEU	AL	AP	DAL	BLEU	AL	AP	DAL
				Low Latency				
USTC-NESLIP	16.36	4.90	0.79	10.30	17.54	4.92	0.78	8.18
VOLCTRANS	15.80	6.34	0.89	13.57	16.91	6.54	0.89	11.26
NAIST	13.77	7.29	0.88	8.07	14.41	7.21	0.88	7.97
				Medium Latency				
USTC-NESLIP	17.53	8.42	0.92	11.81	18.30	7.61	0.90	10.59
VOLCTRANS	15.80	6.34	0.89	13.57	16.91	6.54	0.89	11.26
NAIST	15.22	11.48	0.97	11.98	16.20	11.54	0.97	11.98
				High Latency				
USTC-NESLIP	17.28	11.67	0.97	11.14	18.17	11.71	0.97	13.72
VOLCTRANS	15.85	11.19	0.97	0.97	16.97	11.27	0.97	11.90
NAIST	15.57	13.70	0.99	13.91	16.19	13.83	0.99	14.01

[30] https://iwslt.org/2021/simultaneous

· Summary of the results of the simultaneous speech translation (segmented and unsegmented) **speech track**
· Results are reported on the blind test set and systems are grouped by latency regime (set on tst-COMMON v2, only segmented input.)
· Raw logs are also provided on the task web site.

English-German	**tst-COMMON v2**						
	BLEU	AL	AP	DAL	AL(CA)	AP(CA)	DAL(CA)
Low Latency							
USTC-NESLIP	27.40	0.92	0.68	1.42	2.33	1.33	4.38
Medium Latency							
USTC-NESLIP	29.68	1.86	0.82	2.65	3.66	1.48	5.36
AppTek	24.88	1.96	0.88	3.08	3.37	1.17	4.10
High Latency							
USTC-NESLIP	30.75	2.74	0.90	3.63	5.05	1.56	6.23
AppTek	26.77	3.00	0.99	5.48	6.66	1.32	6.93

English-German	**Blind Test Set**						
	BLEU	AL	AP	DAL	AL(CA)	AP(CA)	DAL(CA)
Low Latency							
USTC-NESLIP	21.85	1.04	0.66	1.47	2.99	1.52	6.41
Medium Latency							
USTC-NESLIP	24.83	1.96	0.80	2.79	4.49	1.63	7.15
AppTek	16.60	1.95	0.80	2.73	2.86	1.06	3.86
High Latency							
USTC-NESLIP	25.62	2.86	0.88	3.85	6.10	1.68	7.93
AppTek	21.08	3.99	0.94	5.06	5.00	1.16	6.12
Unsegmented							
USTC-NESLIP	25.31	30.91	0.51	26.47	264.28	1.10	536.54
AppTek	15.03	107.11	0.44	32.92	149.52	0.63	175.79

A.2. Offline Speech Translation

Speech Translation: TED English-German tst 2021

· Systems are ordered according to BLEU_NewRef: *BLEU* score computed on the NEW reference set (literal translations).
· *BLEU* scores are given as percent figures (%).
· End-to-end systems are indicated by gray background.
· The "segm." column indicates the segmentation strategy (Own vs **Given**).
· The "data condition" indicates the training data condition (Constrained vs **Unconstrained**).
· The † symbol indicates an end-to-end submission exploiting pre-trained models (not all parameters are jointly trained).

System	segm.	data condition	BLEU_NewRef	BLEU_TEDRef	BLEU_MultiRef
HW-TSC	Own	Constrained	24.6	20.3	34.0
KIT	Own	Constrained	23.4	19.0	32.0
APPTEK	Own	Constrained	22.6	18.3	31.0
KIT	Own	Constrained	22.0	18.1	30.3
APPTEK	Own	Constrained	21.9	18.1	30.4
VOLCTRANS	**Given**	Constrained	21.8	17.1	29.5
UPC†	Own	**Unconstrained**	21.8	18.3	30.6
VOLCTRANS	**Given**	Constrained	21.7	18.7	31.3
ESPNET-ST	Own	Constrained	21.7	18.2	30.6
FBK	Own	Constrained	21.6	18.4	30.6
OPPO	**Given**	Constrained	21.5	17.8	30.2
ESPNET-ST	Own	Constrained	21.2	19.3	31.4
NIUTRANS	Own	Constrained	20.6	19.6	30.3
VUS	**Given**	Constrained	15.3	12.4	20.9
BUT	**Given**	**Unconstrained**	11.7	9.8	16.1
LI	**Given**	Constrained	3.6	2.7	4.8

Speech Translation: TED English-German tst 2020

· Systems are ordered according to BLEU_TEDRef: *BLEU* score computed on the ORIGINAL reference set.
· *BLEU* scores are given as percent figures (%).
· End-to-end systems are indicated by gray background.
· The "segm." column indicates the segmentation strategy (Own vs **Given**).
· The "data condition" indicates the training data condition (Constrained vs **Unconstrained**).
· The † symbol indicates an end-to-end submission exploiting pre-trained models (not all parameters are jointly trained).

System	segm.	data condition	BLEU_TEDRef
ESPNET-ST	Own	Constrained	26.0
HW-TSC	Own	Constrained	25.4
KIT	Own	Constrained	25.4
ESPNET-ST	Own	Constrained	24.7
FBK	Own	Constrained	24.7
UPC†	Own	**Unconstrained**	24.6
APPTEK	Own	Constrained	24.5
VOLCTRANS	**Given**	Constrained	24.3
KIT	Own	Constrained	23.2
APPTEK	Own	Constrained	23.1
NIUTRANS	Own	Constrained	22.8
OPPO	**Given**	Constrained	22.6
VOLCTRANS	**Given**	Constrained	22.2
VUS	**Given**	Constrained	13.7
BUT	**Given**	**Unconstrained**	11.4
LI	**Given**	Constrained	0.2

A.3. Multilingual Speech Translation

· Submissions are ordered according to average ST performance across all official language pairs.
· ST systems are scored using the BLEU↑ metric as computed by SacreBLEU (Post, 2018).
· ASR systems are scored using WER↓ computed on lowercased text with punctuation removed.

Official Results:

	Condition			Supervised				Zero-shot			Avg
System	Constrained	E2E	Ensemble	es-en	fr-en	fr-es	pt-en	pt-es	it-en	it-es	
FAIR		✓	✓	42.2	38.7	36.5	31.0	38.2	29.4	37.3	36.2
KIT	✓		✓	39.3	27.1	29.2	30.7	37.3	26.5	32.4	31.8
UEdin	✓	✓	✓	36.2	26.4	29.5	27.0	34.5	23.0	31.1	29.7
UM-DKE	✓	✓	✓	33.9	25.4	27.6	25.7	33.7	22.8	29.4	28.4
ZJU	✓		✓	34.5	25.2	27.4	25.7	31.6	20.8	27.3	27.5
HWN	✓	✓	✓	35.4	26.7	27.0	26.7	27.0	17.6	15.4	25.1
ON-TRAC	✓	✓		20.2	14.4	15.0	13.2	3.0	4.2	4.6	10.7

Table 5: **Multilingual ST:** Results of primary submissions on official language pairs in BLEU↑

All Submissions:

System		Condition			Supervised				Zero-shot			Avg
System		Constrained	E2E	Ensemble	es-en	fr-en	fr-es	pt-en	pt-es	it-en	it-es	
FAIR	primary		✓	✓	42.2	38.7	36.5	31.0	38.2	29.4	37.3	36.2
FAIR	joint_U_W		✓		41.5	37.4	35.2	29.2	36.8	29.1	36.8	35.1
FAIR	joint_U		✓		40.4	36.4	34.4	29.0	38.2	28.4	34.6	33.9
FAIR	joint_X		✓		40.6	36.5	34.7	28.2	38.2	27.8	33.3	33.5
KIT	contrastive	✓	✓		38.9	28.5	29.7	30.2	37.1	25.8	33.0	31.9
KIT	primary	✓		✓	39.3	27.1	29.2	30.7	37.3	26.5	32.4	31.8
UEdin	primary	✓	✓	✓	36.2	26.4	29.5	27.0	34.5	23.0	31.1	29.7
UEdin	contrastive	✓	✓		35.0	25.5	28.8	26.2	33.3	22.4	30.1	28.8
UM-DKE	primary	✓	✓	✓	33.9	25.4	27.6	25.7	33.7	22.8	29.4	28.4
ZJU	primary	✓		✓	34.5	25.2	27.4	25.7	31.6	20.8	27.3	27.5
UEdin	contrastive	✓			33.3	23.7	26.9	23.6	30.0	19.7	26.7	26.3
UM-DKE	contrastive	✓		✓	34.5	21.9	24.3	24.3	29.3	21.7	26.8	26.1
FAIR	baselines_R		✓		34.1	28.4	29.3	19.8	25.3	20.0	25.8	26.1
HWN	primary	✓	✓	✓	35.4	26.7	27.0	26.7	27.0	17.6	15.4	25.1
ON-TRAC	primary	✓	✓		20.2	14.4	15.0	13.2	3.0	4.2	4.6	10.7

Table 6: **Multilingual ST:** Results of all submissions (primary and contrastive) on official language pairs in BLEU↑

Additional Results (Unofficial Language Pairs and ASR):

	Condition			Supervised			
System	Const.	E2E	Ens.	es-fr	es-it	es-pt	fr-pt
FAIR		✓	✓	33.7	33.0	46.5	35.5
KIT	✓		✓	32.4	32.3	46.6	28.8
UEdin	✓	✓	✓	30.3	32.9	44.5	30.1
HWN	✓	✓	✓	27.0	30.8	43.2	26.9
ON-TRAC	✓	✓		8.2	11.1	25.6	14.9

Table 7: **Multilingual ST:** Results of primary submissions on unofficial language pairs in BLEU↑ (optional)

	Condition			ASR				Avg
System	Const.	E2E	Ens.	es	fr	it	pt	
HWN	✓	✓		11.1	22.2	16.2	23.8	18.3
KIT	✓	✓		10.0	26.5	15.5	22.1	18.5
FAIR		✓	✓	11.2	18.7	19.6	27.4	19.2
UEdin	✓	✓		12.0	23.4	18.7	25.9	20.0

Table 8: **ASR:** Results of primary submissions on ASR in WER↓ (optional), sorted by average WER

A.4. Low-Resource Speech Translation

Official Results:

System	swh-eng	swc-fra	swc-eng
IMS.primary	14.9	**13.5**	**7.7**
IMS.contrastive	6.7	2.7	3.9
ON-TRAC	12.9	9.1	–
USYD-JD	**25.3**	–	–

Table 9: **Low-Resource ST:** Results of all speech translation submissions (case-insensitive BLEU↑). The swc-eng and swa-fra pairs were optional.

System	Coastal Swahili (swh)	Congolese Swahili (swc)
ON-TRAC	31.2	36.8
USYD-JD	34.4	–

Table 10: **ASR:** Results of all (optional) speech transcriptions submissions (case-insensitive WER↓).