UTTER

Unified Transcription and Translation for Extended Reality (UTTER)

Horizon Europe Research and Innovation Action Number: 101070631

D15 – First report on XR models

Nature	Report	**Work Package**	WP3
Due Date	28/03/2024	**Submission Date**	28/03/2024
Main authors	Marcely Zanon Boito (NAV), Laurent Besacier (NAV)		
Co-authors	Barry Haddow (UEDIN), José Souza (UNB)		
Reviewers	Vlad Niculae (UVA)		
Keywords	pre-trained models, XR models, efficient downstream speech models		
Version Control			
v0.1	**Status**	Draft	14/03/2024
v1.0	**Status**	Final	22/03/2024

This project has received funding from the European Union's Horizon Europe Research and Innovation programme under Grant Agreement No 101070631 and from the UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee (Grant No 10039436).

Contents

List of Figures

Abstract

In this report we present four different projects related to *Multimodal, Multilingual Pre-trained XR Models* from UTTER project's WP3. These include two foundation models, `mHuBERT-147` and `TowerLM`, which cover respectively speech and textual modalities. We also report on two task-specific efficient models: the Multilingual DistilWhisper for automatic speech recognition, and an approach for efficient CTC regularization for speech translation.

1 Contributors

Task	Who is reporting	Paper
T3.1	Boito and Besacier §3.2	Marcely Zanon Boito, Vivek Iyer, Nikolaos Lagos, Laurent Besacier, Ioan Calapodescu. *"mHuBERT-147: A Compact Multilingual HuBERT Model"*. Under review.
T3.1	de Souza §3.1	Duarte M. Alves, José Pombal, Nuno M. Guerreiro, Pedro H. Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, Pierre Colombo, José G. C. de Souza, André F. T. Martins. *"Tower: An open multilingual large language model for translation-related tasks"*. arXiv, 2024.
T3.2	Boito §4.1	Thomas Palmeira Ferraz, Marcely Zanon Boito, Caroline Brun, Vassilina Nikoulina. *"Multilingual DistilWhisper: Efficient Distillation of Multi-task Speech Models via Language-Specific Experts"*. IEEE ICASSP 2024.
T3.2	Haddow §4.2	Biao Zhang, Barry Haddow, Rico Sennrich. *"Efficient CTC Regularization via Coarse Labels for End-to-End Speech Translation"*. EACL 2023.

Table 1: List of publications to be discussed

2 Introduction

2.1 Objectives

Proposal

"The purpose of this package is to develop pretrained models that can be effectively leveraged for multi-modal (written and spoken language) translation. The objective can be structured into two components:

- **Model pretraining and availability:** Collect speech and language models on selected and large datasets. A large language model developed for the BigScience project will be adapted and integrated for usage in the project. An analysis and evaluation of existing pretrained speech/textual models and their relevance for the target task will be done. All the selected speech and text models will be integrated and provided in a unified interface for usage by the other working groups.
- **Architectural investigations:** Investigate how speech and text pretrained models can be efficiently combined together for translation. We will analyze the performances of several combinations of the pretrained speech and text architectures made available in the previous tasks for end-to-end spoken and/or written translation. We will formulate recommendations for optimally combining pretrained models."

Work Completed

Since the project's proposal and the BigScience project, more competitive pre-trained models for text were released. We thus shift focus towards the adaptation of these, notably proposing `TowerLM`, a derived model from the popular open-source large language model (LLM) LLaMA-2 (Touvron et al., 2023b). For speech, observations from the self-supervised learning (SSL) SUPERB benchmark (wen Yang et al., 2021) convinced us of the potential of building a multilingual speech model based on the Hidden Units BERT (HuBERT, Hsu et al. (2021)) architecture. We trained and will soon release the first general-purpose multilingual HuBERT: `mHuBERT-147` (expected release date: 06/06/2024).

Regarding efficient training, we proposed a parameter efficient distillation approach for the speech architecture Whisper, which allows us to bridge the gap between large and small models without much cost for inference. Moreover, we proposed a CTC regularization approach via coarse labels for improving the performance of speech translation models.

3 Task 3.1: Investigating and pretraining text and speech models

Proposal

"We will provide access to very large pretrained language models created during the BigScience research workshop. Being trained on large-scale, diverse, and multi-lingual datasets, the provided checkpoints are well-suited for the target task. On the speech side, an analysis of available pretrained models, such as Conformer, Wav2Vec2, HuBERT will be done. We will also provide an

interface to access the best performing, open-sourced pretrained speech models. Depending on the available budget for pretraining, we will either continue pretraining from existing checkpoints or run pretraining from scratch on the provided audio data. Pretraining a competitive speech model that has between 1 and 10 billion parameters from scratch would require between 250,000 and 500,000 V100 GPU hours cf. Wav2Vec2, HuBERT, BigSSL. Continuing pretraining on existing open-sourced checkpoints would require significantly less GPU hours."

Work Completed

For text, we perform continuous pretraining (i.e. training from existing checkpoint) using as base the LLaMA-2 model. This is followed by instruction tuning, creating a very competitive model for translation-related tasks. For speech, we train the first general-purpose multilingual HuBERT model (from scratch), covering 147 languages.

3.1 TOWER: An Open Multilingual Large Language Model for Translation-Related Tasks

Many important tasks within multilingual NLP, such as quality estimation, automatic post-edition, or grammatical error correction, involve analyzing, generating or operating with text in multiple languages, and are relevant to various translation workflows — we call these **translation-related tasks**. Recently, general-purpose large language models (LLMs) challenged the paradigm of *per-task* dedicated systems, achieving state-of-the-art performance on several recent WMT shared tasks (Kocmi et al., 2023; Freitag et al., 2023; Neves et al., 2023). Unfortunately, strong capabilities for *multiple* translation-related tasks have so far been exhibited by *closed* LLMs only (Hendy et al., 2023; Kocmi and Federmann, 2023; Fernandes et al., 2023; Raunak et al., 2023). Perhaps because most *open* LLMs are English-centric, approaches leveraging these models still lag behind, having thus far achieved competitive results only when specializing on a *single* task (Xu et al., 2024a, 2023; Iyer et al., 2023). Here, *closed* means models available only behind an API such as GPT-3.5 and GPT-4 and *open* relates to models whose weights are available for download and allowed to derive from.

Our backbone language model is LLaMA-2, which is very competitive on a wide range of tasks (Touvron et al., 2023b) and achieves the best zero-shot translation quality across available open LLMs (Xu et al., 2024a). Nevertheless, the LLaMA-2 family was trained on relatively little non-English data, limiting its potential for multilingual tasks, such as machine translation.

3.1.1 TowerBase

We extend LLaMA-2's training on a highly-multilingual dataset comprising 20 billion tokens — measured with the model's tokenizer — for 10 languages: English (en), German (de), French (fr), Dutch (nl), Italian (it), Spanish (es), Portuguese (pt), Korean (ko), Russian (ru), and Chinese (zh). While previous work exclusively leverages monolingual data (Xu et al., 2024b), we draw inspiration from work including parallel data during pretraining (Anil et al., 2023; Briakou et al., 2023), and *mix parallel sentences* (one-third) along with monolingual data (two-thirds).

Monolingual data. We collect monolingual data from mC4 (Xue et al., 2021) uniformly sampling across our languages. Additionally, we *improve data quality* with standard cleaning proced-

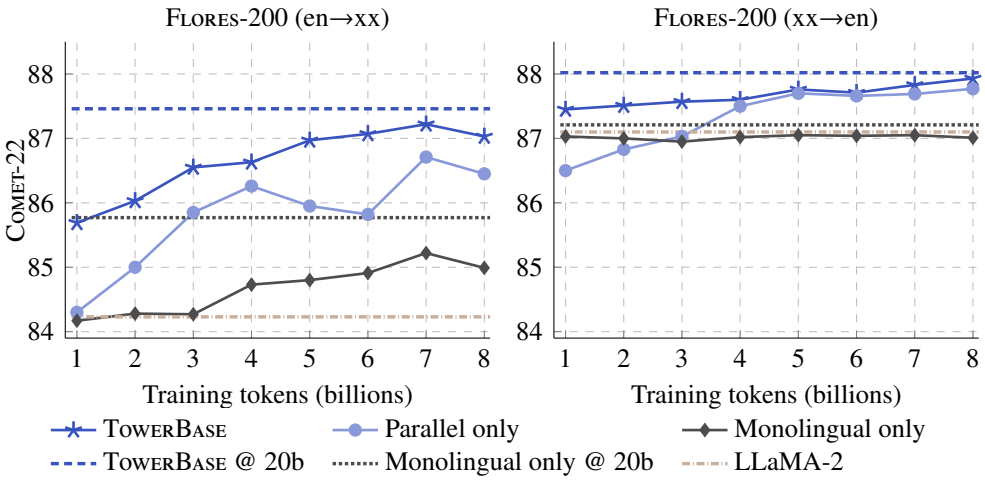

Figure 1: Translation quality on Flores dataset for continue pretraining data recipes. The TOWERBASE recipe, outlined in Section 3.1.1, mixtures monolingual with parallel data. The "Parallel only" recipe only processed 8 billion tokens due to compute constraints.

ures (Wenzek et al., 2019; Touvron et al., 2023a): deduplication, language identification, and perplexity filtering with KenLM (Heafield, 2011).

Parallel Data. We uniformly sample to-English (xx→en) and from-English (en→xx) language pairs from various public sources. Additionally, we *ensure translation quality* by removing sentence pairs below quality thresholds for Bicleaner (Sánchez-Cartagena et al., 2018; Ramírez-Sánchez et al., 2020) and COMETKIWI-22 (Rei et al., 2022).

Model Training. We train our models with a codebase based on Megatron-LLM (Cano et al., 2023) on 8 A100-80GB GPUs, an effective batch size of 1.57 million tokens per gradient step, and a cosine scheduler with initial and final learning rates of 3×10^{-5} and 3×10^{-6}, respectively. The training times for TOWERBASE 7B and 13B were 10 and 20 days.

3.1.2 Experiments

Parallel data during continued pretraining improves translation quality. Figure 1 reports 5-shot translation quality on FLORES-200 for multiple continued pretraining data recipes. Mixing monolingual and parallel data achieves the highest quality, outperforming both monolingual only and parallel only data. In general, improvements are more noticeable on en→xx directions, likely due to the English-centric nature of LLaMA-2's training. Nevertheless, while monolingual only data improves over the base LLaMA-2 by 0.1 COMET-22 points on xx→en directions, our recipe gains nearly a full point.[1]

Parallel data during continued pretraining is sample efficient, but quality continues to improve with more tokens. At the 2 billion tokens mark, combining parallel sentences with

[1] While 0.1 COMET-22 points translates to 54.9% human agreement, one COMET-22 point translates to 90.9% (Kocmi et al., 2024).

monolingual data (i) yields more than 50% of the improvement over the base model, and (ii) surpasses the recipe leveraging solely monolingual data. Additionally, while training on more tokens has diminishing returns — 85% of the total performance gains appear by the 5 billion tokens mark — it continues to improve translation quality.

For more details about the different models in the Tower family can be found in the paper (Alves et al., 2024). Models are available in the Tower HuggingFace collection.[2]

3.2 mHuBERT-147: A Compact Multilingual HuBERT Model

Self-supervised Learning (SSL) approaches for speech representation learning are now the essential foundation blocks for speech processing solutions. These models leverage large amounts of unlabeled speech data during training, and learn from different pretext tasks in order to build exploitable contextualized speech representations that can be leveraged for different downstream tasks. For English, several models have been proposed (Pascual et al., 2019; Ravanelli et al., 2020; Baevski et al., 2020; Hsu et al., 2021; Chen et al., 2022; Liu et al., 2020), but most multilingual models available to the community are based on wav2vec 2.0 (Conneau et al., 2021; Babu et al., 2022; Pratap et al., 2023), with the only current exception being WavLabLM (Chen et al., 2023b).

Meanwhile, the Hidden units BERT model (HuBERT, Hsu et al. (2021)) – where pre-training is performed in 2 or 3 iterations, and target training labels are externally obtained via k-means clustering – presents superior performance on the English SSL benchmark SUPERB (wen Yang et al., 2021), outperforming wav2vec 2.0. The English version of this model also shows decent cross-lingual adaptation capabilities on the multilingual benchmark ML-SUPERB (Shi et al., 2023a). Recently, HuBERT has also emerged as a popular choice for producing discrete speech units for multimodal LLMs (Wang et al., 2023b; Chang et al., 2023; Chou et al., 2023).

However, despite recent efforts to reduce hardware costs (Li et al., 2022; Lin et al., 2023; Chen et al., 2023a), HuBERT's superior performance comes with higher training costs, a characteristic that is accentuated in multilingual settings requiring larger amounts of speech data. The divergence in costs arises from HuBERT's multi-iteration training process, which contrasts with wav2vec 2.0's single iteration training on unlabeled speech data alone. HuBERT requires high-dimensional feature extraction across the entire training dataset to generate discrete labels, along with a minimum of two model training and clustering steps, resulting in increased disk and CPU/GPU resource demands. We are aware of only small-scale multilingual HuBERT models that train using a single dataset: there is an mHuBERT trained on 3 languages (Lee et al., 2022); and a HuBERT collection covering between 5 and 12 languages (Duquenne et al., 2023).

In contrast, in this work, we tackle the challenge of training the first general-purpose massively multilingual HuBERT speech representation model. This model is trained using 90,430 hours of clean open-license data across 147 languages. For reducing pre-processing costs, we downsample large popularly used speech datasets, hypothesizing that source diversity is more important than quantity. We also propose to replace the original HuBERT clustering implementation with *faiss*-based clustering (Douze et al., 2024), increasing label assignment speed by 5.2 times. Finally, for training, we employ a two-level multilingual up-sampling approach factoring in both linguistic and dataset diversity for increased overall multilingual performance.

[2] https://huggingface.co/collections/Unbabel/tower-659eaedfe36e6dd29eb1805c

Despite being trained with considerably less data than popular multilingual models, and being a compact second-iteration model, `mHuBERT-147` is competitive with larger multilingual SSL models, reaching the second position at ML-SUPERB leaderboard, and setting new SOTA for two metrics. Complementary few-shot ASR evaluation on FLEURS-102 shows that our model matches MMS-300M on average, while having 70% less parameters (1.8 times faster training; 1.4 times faster decoding). Our findings suggest that `mHuBERT-147` is a promising model for multilingual speech downstream tasks, offering an unprecedented balance between high performance and parameter efficiency. Our future work will include the evaluation of the third iteration 95M parameter `mHuBERT-147` that is currently under training, and that we believe will further improve our performance on ML-SUPERB and FLEURS-102.

`mHuBERT-147` data: We gathered 90,430 hours of speech from datasets with permissive licences in 147 languages. For this multilingual collection, our goal was to prioritize linguistic diversity over data quantity alone. In total, our training set spans 19 language families (sorted in decreasing order of data quantity): Indo-European, Niger-Congo, Uralic, Afro-Asiatic, Constructed (Esperanto), Turkic, Dravidian, Sino-Tibetan, Austronesian, Koreanic, Kra-Dai, Japonic, Language isolate (Basque), Kartvelian, Austroasiatic, Mongolic, Northwest Caucasian, Creole and Tupian. The report D2.1 on data and resources details the dataset.

3.2.1 `mHuBERT-147` Training

Our training follows the multi-iteration pre-training objective of HuBERT (Hsu et al., 2021). We detail how `mHuBERT-147` differs from HuBERT training, including a) a more complex data up-sampling strategy, and b) replacing the original k-means implementation with the efficient *faiss* (Douze et al., 2024) Inverted File Index (IVF) – drastically increasing labeling speed.

Two-level language-source up-sampling: To optimize for exposure to different languages and data sources during training, we employ two-level up-sampling during multilingual batching. Let N be the total number of examples in the training set, with n_l corresponding to the count for language l. The sampling probability for l is $P_l \propto \left(\frac{n_l}{N}\right)^{\alpha}$, where α is a hyper-parameter in $[0, 1]$; with $\alpha = 1$ resulting in no up-sampling, and lower values resulting in higher probabilities for under-represented languages. For each epoch, we sample N times from the probability distributions P_l. In this way we reach a quantity B_l of examples selected per language l. We sample B_l utterances by considering varied data sources (datasets). The probability of sampling an utterance of l from data source x is given by $P_x \propto \left(\frac{n_l(x)}{n_l}\right)^{\beta}$, where $n_l(x)$ corresponds to the number of examples of language l from data source x, and β is a hyper-parameter in $[0, 1]$. We sort the N selected utterances by length before batching, to minimize random cropping.

Faiss clustering: We use *faiss*-based (Douze et al., 2024) k-means clustering for faster label assignment. This library facilitates efficient similarity search and clustering of dense vectors. We cluster using Inverted File Index (IVF) with the following setting: `OPQM_D,IVFK_HNSW32,PQMx4fsr`, that we now detail. [3]

[3] More information can be found at: https://github.com/facebookresearch/faiss/wiki/The-index-factory

- "`OPQM_D,..,PQMx4fsr`" is used for indexing RAM usage. This option optimizes vector representation by rotation and then performs input vector projection into dimension D. Product quantization (PQ) is then applied to hash the vectors into M 4-bit codes, resulting in the storage of M/2 bytes per vector. We use $M = 16, D = 64$.
- "`..,IVFK_HNSW32,..`" denotes the indexing itself. IVF performs coarse quantization via an efficient implementation of k-means. This option uses Hierarchical Navigable Small Worlds (HNSW) graphs for cluster assignment, with 32 being the number of links per vertex. In practice, this greatly speeds up clustering (5.2x faster than Hsu et al. (2021)).

3.2.2 Experiments

ML-SUPERB Evaluation

For evaluating the quality of the multilingual representations learned by `mHuBERT-147`, we use the ML-SUPERB benchmark (Shi et al., 2023a). This benchmark comprises two settings: 10min and 1h; and four tasks: monolingual ASR; multilingual ASR, LID, and joint multilingual ASR and LID. The monolingual setting constitutes 13 (language, domain) pairs, while the multilingual one has 240 pairs across 143 languages in total – of which 123 and 20 constitute the *normal* (10min/1h per language) and *few-shot* (5 utterances per language) evaluation settings respectively. The architecture consists of learnable weights over the frozen SSL features, a CNN for reducing feature dimensionality by a half, and two Transformer layers ($dim = 256$; 8 attention heads). In line with the official guidelines,[4] we only tune the learning rates (best of $10^3; 10^4; 10^5$) on the validation set. Due to the high compute costs of evaluating on this benchmark, we do not retrain the other submissions we compare against and instead reuse the leaderboard scores from Shi et al. (2023b). Therefore, we are unfortunately unable to provide confidence intervals, as these were not originally requested by the benchmark organizers. We also recompute the global SUPERB scores for all models following Shi et al. (2023a), since this metric leverages SOTA scores for normalization, and our model achieves new SOTA for two tasks (bold values in Table 2).

Table 2 presents results for the 10min/1h settings. The current SOTA models from (Shi et al., 2023b) are shown at the top (NWHC models are variants of MMS-300M). Other relevant multilingual SSL models are displayed in the bottom portion: MMS-300M; XLS-R-300M, and the best WavLabLM model from (Chen et al., 2023b) (136 languages; 40K hours).[5] We omit XLSR-53 results as these are consistently worse than XLS-R (Shi et al., 2023a). Looking at the bottom rows, we note that `mHuBERT-147` outperforms the XLS-R and WavLabLM models across all tasks. Compared to MMS-300M, our model is only surpassed in the Monolingual ASR 10min, and few-shot ASR+LID 10min/1h tasks. Against the current SOTA (top rows), we see that `mHuBERT-147` is again very competitive, despite being trained on much lesser data. We also set new SOTA ACC scores for LID 1h and Multilingual ASR+LID 10min, ranking second on the ML-SUPERB leaderboard, while being much more compact than all other models.

SSL	# Params	Monolingual ASR normal CER (↓)	Multilingual ASR normal CER (↓)	Multilingual ASR few-shot CER (↓)	LID normal ACC (↑)	Multilingual ASR + LID normal ACC (↑)	Multilingual ASR + LID normal CER (↓)	Multilingual ASR + LID few-shot CER (↓)	$SUPERB_s$ (↑)
MMS-1B	965M	**33.3 / 25.7**	**21.3 / 18.1**	**30.2 / 30.8**	**84.8** / 86.1	73.3 / 74.8	**26.0** / 25.5	**25.4 / 24.8**	**985. 8 / 950.7** (1st)
NWHC1	317M	39.5 / 30.5	28.9 / 21.5	41.4 / 38.6	67.1 / 87.4	77.1 / **90.6**	28.8 / **21.5**	40.3 / 38.2	776.3 / 879.6
NWHC2	317M	39.5 / 30.5	29.3 / 21.6	42.0 / 39.3	64.4 / 88.1	77.4 / **90.6**	28.4 / 21.8	41.5 / 38.8	761.8 / 876.0
mHuBERT-147	**95M**	35.9 / 27.6	25.4 / 22.5	34.2 / 33.8	74.8 / **90.1**	**81.0** / 89.0	26.3 / 23.6	33.9 / 34.4	897.2 / 928.5 (2nd)
MMS-300M	317M	33.8 / 30.5	28.7 / 24.0	36.5 / 36.5	62.3 / 84.3	71.9 / 74.3	31.5 / 30.0	30.9 / 29.2	826.7 / 846.8 (3rd)
XLS-R-300M	317M	39.7 / 30.6	29.2 / 22.0	40.9 / 39.3	66.9 / 87.9	55.6 / 85.6	28.4 / 22.9	42.1 / 42.4	732.4 / 853.1
WavLabLM-large-MS	317M	40.5 / 32.8	37.8 / 31.9	43.8 / 42.8	71.7 / 81.1	70.8 / 80.0	37.0 / 32.2	43.4 / 41.2	709.4 / 743.4

Table 2: ML-SUPERB 10min/1h results. Current SOTA is shown on the top portion of the table, our submission is shown in the middle, other relevant multilingual models are presented below it. Updated SOTA scores for each metric are presented in **bold**.

SSL	General Avg (98)	WE (24)	EE (16)	CMN (11)	SSA (20)	SA (12)	SEA (11)	CJK (4)
MMS-1B	8.4	6.6	4.9	7.4	10.6	8.1	10.5	19.4
MMS-300M	**24.1**	**18.0**	**8.1**	34.9	24.5	**22.0**	26.9	92.4
XLS-R-300M	25.1	17.6	9.1	30.6	28.1	25.4	30.7	**87.0**
mHuBERT-147	**24.1**	20.6	12.1	**14.7**	**24.4**	30.9	**25.8**	93.2

Table 3: FLEURS-102 CER (↓) geographic group averages, with number of languages between parentheses.

FLEURS-102 Evaluation

We complement ML-SUPERB evaluation by training monolingual ASR models on the FLEURS-102 dataset (Conneau et al., 2023), competing with XLS-R and MMS (300M/1B). In this full fine-tuning few-shot setting, we add a feedforward NN as the CTC layer on top of the pre-trained stack, optimizing the resulting model for ASR using approximately 10 hours of speech. Thus, unlike the experiments in the previous section, here, larger models will have the advantage of having more parameters for adaptation. With these experiments we want to illustrate that in this unfavorable setting, `mHuBERT-147` can still be competitive, while being faster at training and inference time. We implement monolingual ASR using the *transformers* library (Wolf et al., 2020a). All models were trained for 30 epochs on *fp16* using V100-32GB GPUs. Since individual language optimization would be prohibitively costly for this dataset, we select a subset of 29 languages, covering the different geographic groups, and optimize parameters using XLS-R. We use 10^5 as learning rate, warm-up ratio of 0.1 and dropout of 0.1 (300M and 1B) or 0.3 (95M). The increased dropout for the latter is due to the ASR models being considerably smaller (70% less params.). We train two models per language with different seeds, adding up to four runs in cases of high variability in scores ($\geq$ 20 CER). We apply MMS transcript normalization (Pratap et al., 2023), reporting CER averages over the two best runs. We exclude four languages for which all models failed to converge: Hebrew, Tamil, Sindhi, Welsh.

Table 3 presents results grouped by FLEURS-102 geographic groups: Western Europe (WE); Eastern Europe (EE); Central-Asia/Middle-East/North-Africa (CMN); Sub-Saharan Africa (SSA); South-Asia (SA); South-East Asia (SEA); Isolates (CJK). The results illustrate the impressive im-

[4] https://github.com/espnet/espnet/tree/master/egs2/ml_superb/asr1

[5] We highlight that although XLS-R and MMS are referred in the literature as "300M", the correct parameter count is 317M.

pact of capacity on few-shot adaptation. While MMS-300M and 1B models score similarly on ML-SUPERB, here we observe a considerable gap between them. This impact is more evident when looking at results for languages with very large vocabularies (e.g. CJK: Chinese, Cantonese, Japanese and Korean), in which all *smaller* models failed to properly train. We find that in these cases, adding capacity (more hidden layers) or better language-specific hyper-parameters tuning were solutions that would allow us to reduce most of this performance gap between 300M and 1B models. For instance, for MMS-300M, by increasing the learning rate, we were able to reach an average CER of 46.2 for CJK (2x better). Since we could not perform this hyper-parameter search for all other languages and models, we do not present these results here, only highlighting the challenge of hyper-parameter search for FLEURS-102.

Finally, across FLEURS-102, we observe that `mHuBERT-147` is competitive against models 3 times larger: its general CER average over the 98 languages is the same as MMS-300M, and lower than XLS-R (-1 CER). We measured training and inference efficiency across three runs and two languages (English and Kannada) in exclusive node execution mode. We find that `mHuBERT-147`, which has 70% less parameters, is respectively 1.8 and 3 times faster to train on average than 300M and 1B models. For test inference, our model is respectively 1.4 and 2 times faster on average than 300M and 1B models. Moreover, we highlight that, due to its reduced parameter size, our model is the only one able to perform training for languages with larger vocabulary (≥ 1000) on a 24GB GPU. These results highlight our model as a compact but powerful solution for speech processing applications.

This model is part of an Interspeech 2024 paper submission. Due to their anonymity rules, we do not share the models and scripts for now. Everything will be made publicly available after the 06/06/2024.

4 Task T3.2: Efficient Training

Proposal

"Neural network training requires significant computational resources, especially for the extremely large models trained on huge amounts of data such as GPT-3. We will explore several venues for reducing the computational requirements. We will continualy improve the codebase that we use to make sure we take advantage of the most recent hardware improvements. We will experiment with lower precision training in order to make more efficient use of the available hardware. Finally, we will look at more efficient use of the available training data as noisy data not only increases training time but also decreases model accuracy."

Work Completed

For the first half of this project, we concentrated our efforts on training efficiency for speech models. In this report we present work on distillation for ASR, and on a regularization approach for efficient speech translation training.

4.1 Multilingual DistilWhisper: Efficient Distillation of Multi-task Speech Models via Language-Specific Experts

Whisper (Radford et al., 2023) is a multitask and multilingual speech model covering 99 languages. It yields commendable automatic speech recognition (ASR) results in a subset of its covered languages, but the model still underperforms on a non-negligible number of under-represented languages, a problem exacerbated in smaller model versions. In this work, we propose *DistilWhisper*, an approach able to bridge the performance gap in ASR for these languages while retaining the advantages of multitask and multilingual capabilities.

Our approach involves two key strategies: lightweight modular ASR fine-tuning of `whisper-small` using language-specific experts, and knowledge distillation from `whisper-large-v2`. This dual approach allows us to effectively boost ASR performance while keeping the robustness inherited from the multitask and multilingual pre-training.

Through extensive experiments on a diverse set of languages we demonstrate the effectiveness of *DistilWhisper* compared to standard fine-tuning or LoRA (Hu et al., 2021) adapters. Our lightweight ASR fine-tuning approach based on CLSR modules generalizes better than LoRA, and the introduction of KD further boosts results in both in- and out-of-domain test sets. We perform additional ablation studies showing our approach can cope with different amounts of training data. Finally, we demonstrate that the flexibility introduced by the gating mechanism equips *DistilWhisper* with an efficient adaptation approach, leveraging the LS modules only when those are relevant.

We believe that such architecture would make usage of Whisper models accessible to a larger amount of researchers and practitioners since it allows to boost the performance of a low-inference cost model by 35.2% using only 14 h of training data. We make available the models' weights[6] and code[7] developed in this work.

4.1.1 DistilWhisper Architecture

With the goal of increasing performance for different languages in models of limited capacity, we propose the *DistilWhisper* approach: we plug conditional language-specific routing (CLSR) modules (Zhang et al., 2021) into a small Whisper (`whisper-small`), and optimize these modules jointly on ASR fine-tuning and KD from a larger Whisper (`whisper-large-v2`).[8]

CLSR module. We extend CLSR modules for the first time to the speech domain. This module learns a hard binary gate g(·) for each input token by using its hidden embedding z^l. These decisions enable a layer to selectively guide information through either a LS path denoted as h^{lang} or a shared path referred to as h^{shared}, as in Eq 1. In contrast to the original CLSR, in this work we use LS gates as shown in Figure 2, instead of sharing them across languages. This allows us to train LS components individually (i.e. in parallel), and then only load the relevant modules at inference. Moreover, our approach also differs from the original CLSR by the positioning: supported by previous work (Zhang et al., 2021; Pfeiffer et al., 2022), we limit CLSR to the feed-forward, which we also replace entirely by the CLSR module, reducing further the number of parameters. Gating follows (Zhang et al., 2021): each gate $g(.)$ is made by a two-layer bottleneck network, which is summed to an increasing zero-mean Gaussian noise during training in order to discretize it. At

[6] Available at: https://huggingface.co/naver.
[7] Code available at: https://github.com/naver/multilingual-distilwhisper.
[8] We highlight that at the time of this publication, `whisper-large-v3` was not yet released.

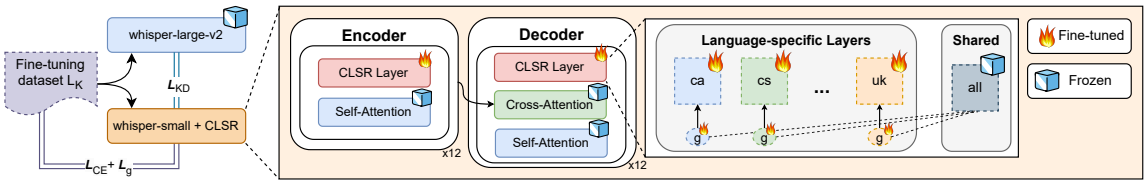

Figure 2: The *DistilWhisper* optimization approach (left), and its architecture (right). The feed-forward is replaced by a CLSR module, where the LS gates (g) learn to alternate between the pre-trained frozen multilingual representation and the LS layer.

inference time, we adopt hard gating.

$$\text{CLSR}(z^l) = g(z^l) \cdot h^{lang}(z^l) + (1 - g(z^l)) \cdot h^{shared}(z^l). \quad (1)$$

***DistilWhisper* approach** is detailed at Figure 2. Our student is enriched with CLSR modules at each feed-forward for each language. These CLSR layers are initialized from the frozen weights of the corresponding feed-forward layer. At training time, for each language the model updates only the corresponding LS layers and gates. At inference time, the model loads the shared layers (multilingual) and the LS modules and gates for the languages of interest, resulting in a limited parameter overhead. We highlight that the use of CLSR modules brings more flexibility to our architecture when compared to adapters, as it allows for routing at the token-level. This makes this approach more capable of leveraging pre-existing knowledge (shared frozen module) via LS gating activation.

4.1.2 Experiments

Experimental Setup

Datasets: We downsample the train and validation sets of the CommonVoice 13.0 (CV-13) dataset (Ardila et al., 2020), using equal amounts of training data for each selected language: 10k utterances for training (approx. 14 h), 1k for validation. Data selection depends on the amount of upvotes utterances received by annotators. We do not downsample the test set. The FLEURS (Conneau et al., 2023) dataset is used for out-of-domain evaluation, as it provides both a good language overlap with CV-13, and an effective out-of-domain setting for ASR evaluation. For instance, average number of tokens per sample for CV-13 is 36, and 97 for FLEURS.

Language Selection: We consider all Whisper languages with a WER gap of more than 11 between large and small models on CV-13. We then narrow this list considering: 1) minimum amount of utterances (10k); 2) overlap with the FLEURS dataset. The final list of languages is: Catalan (ca), Czech (cs), Galician (gl), Hungarian (hu), Polish (pl), Thai (th), Tamil (ta) and Ukranian (uk).[9] These languages encompass 5 language sub-families and vary widely in terms of coverage in the Whisper training set, spanning from 4,300 h (pl) to just 9 h (gl).

Models: We compare our approach to both `whisper-small` (pre-trained student) and `whisper-large-v2` (teacher) models, as well as two approaches of fine-tuning (FT) for the student: standard fine-tuning (all weights are updated), and LoRA adaptation on top of the feed-forward layer.

[9] Although Arabic would also qualify considering our criteria, we find that the dialect from FLEURS differs from the ones present on CV-13.

		FLEURS	CV-13	FLEURS (out-of-domain)								CV-13 (in-domain for FT only)							
	#params	avg	avg	ca	cs	gl	hu	pl	ta	th	uk	ca	cs	gl	hu	pl	ta	th	uk
whisper-large-v2	1.5B	12.5	14.9	5.6	14.3	16.6	17.9	5.9	19.3	12.2	8.1	16.9	14.4	18.9	18.7	8.0	17.3	9.2	15.5
whisper-small	244M	28.3	31.4	14.6	40.4	32.7	43.0	16.7	36.0	22.8	20.5	30.1	38.4	35.5	45.6	18.6	30.0	20.3	32.3
whisper-small+FT	244M	23.3±0.06	16.3±0.09	15.5	31.0	16.9	**36.7**	22.0	22.7	15.6	25.9	**13.7**	20.5	**11.3**	24.1	16.3	13.6	**7.4**	23.4
whisper-small+LoRA-FT	379M	24.9±0.07	18.2±0.02	17.6	36.9	18.2	41.6	25.9	**15.2**	**11.7**	31.8	14.0	23.7	12.7	28.0	21.2	**12.0**	7.9	26.4
whisper-small+CLSR-FT	369M	23.4±0.19	16.3±0.08	15.7	30.5	17.2	**36.9**	22.8	22.7	15.6	25.8	14.1	20.3	**11.6**	24.3	16.1	13.3	7.4	23.4
DistilWhisper	369M	**22.8**±0.21	**16.0**±0.04	**15.3**	**30.2**	**16.7**	**36.9**	**21.4**	21.8	15.1	**24.9**	**13.8**	**20.0**	11.8	**24.0**	**15.9**	12.6	**7.2**	**23.1**

Table 4: WER (↓) with dataset averages (avg) for baselines (top), adaptation approaches (middle), and our method (bottom) for in-domain (CV-13, FT only) and out-of-domain (FLEURS, all) test sets. Best results for `whisper-small` in **bold**.

	Train	FLEURS	CV-13	FLEURS			CV-13		
	size	avg	avg	ca	ta	th	ca	ta	th
whisper-small+CLSR-FT	3k	20.5±0.17	15.0±0.07	17.9	25.6	18.0	19.0	16.4	9.8
DistilWhisper	3k	**20.2**±0.13	**14.6**±0.08	**17.4**	25.5	**17.7**	**18.7**	**15.7**	**9.6**
whisper-small+CLSR-FT	10k	18.0±0.25	11.6±0.01	15.7	22.7	15.6	14.1	13.3	7.4
DistilWhisper	10k	**17.4**±0.13	**11.2**±0.08	**15.3**	**21.8**	**15.1**	**13.8**	**12.6**	**7.2**
whisper-small+CLSR-FT	28k	15.7±0.15	9.5±0.13	13.5	19.8	13.9	11.3	11.3	6.0
DistilWhisper	28k	**15.5**±0.03	**9.3**±0.06	**13.3**	**19.3**	**13.7**	11.3	**11.0**	**5.7**

Table 5: Average WER (↓) for different training data sizes (3k, 10k, and 28k utterances) for in-domain (CV-13) and out-of-domain (FLEURS) test sets. Best results in **bold**.

Finally, we also investigate the impact of the CLSR layer without the use of KD (CLSR-FT), decoupling the effect of KD from the flexibility offered by the routing mechanism on the consequent robustness of the model.

Implementation: We train all models using the Transformers library (Wolf et al., 2020b), and make use of `whisper-small` and `whisper-large-v2` pre-trained weights from HuggingFace.[10] All models are trained for 10 epochs using learning rate 10^{-4} with linear decay, one epoch warm-up, batch size 16, and label smoothing factor 0.1. For LoRA, we use the hyperparameters proposed by (Wang et al., 2023a). For CLSR training we set gate budget $b = 0.5$ and skip-gate probability $s = 0.2$. For KD we employ JS divergence with temperature $\tau = 1$, weighted such as the learning objective is $\mathcal{L} = \mathcal{L}_{\text{CE}} + \mathcal{L}_{\text{g}} + 2\mathcal{L}_{\text{KD}}$. We report normalized WER using the Whisper normalization with a slight modification to avoid splitting numbers and latin-scripted text into individual characters in languages that do not use space delimitation (th). In all cases, the best model is chosen based on WER on the down-sampled CV-13 validation set.

Results

We conduct training for each setting using three distinct seeds and present the average scores. Table 4 presents our results. The top portion presents `whisper-large-v2` (upper bound) and `whisper-small` (lower bound) pre-trained scores. The middle portion presents standard fine-tuning (FT) and LoRA adaptation at the feed-forward layers (LoRA-FT). Our results are presented in the bottom: CLSR-FT corresponds to the setting without $\mathcal{L}_{\text{KD}}$, while *DistilWhisper* is the com-

[10] https://huggingface.co/openai/

plete setting in which both CLSR and KD losses are leveraged.

***DistilWhisper* versus other adaptation approaches.** For `whisper-small`, we observe that both FT and LoRA-FT approaches (middle portion of Table 4) are able to improve performance on both in- and out-of-domain test sets. However, for FT this boost in performance comes with the cost of language specialization. In contrast to that, LoRA-FT is a light adaptation technique that does not modify the pre-trained representation. This method increases performance on both in-domain (avg -13.1) and out-of-domain (avg -3.5) test sets compared to `whisper-small`. *DistilWhisper* further improves performance over `whisper-small` (avg -15.3) and LoRA-FT (avg -2.2) for in-domain data. It also presents better out-of-domain adaptation capabilities compared to LoRA-FT (avg -2.1).

Impact of knowledge distillation. We observe that *DistilWhisper* on average outperforms all other adaptation approaches (FT, LoRA-FT) for in- and out-of-domain test sets (bottom portion of Table 4). Comparing our models (CLSR-FT and *DistilWhisper*), we observe that the version with KD (*DistilWhisper*) exhibits a slight increase in average in-domain performance (-0.3). In out-of-domain settings, this model consistently outperforms CLSR-FT across all languages (avg -0.6), which confirms our initial hypothesis that the KD loss leverages the robustness from the teacher into the final model. Overall, these results highlight the effectiveness of our proposed architecture: we were able to reduce the out-of-domain performance gap between `whisper-large-v2` and `whisper-small` by 35.2% (avg -5.5) with a parameter overhead at inference time of only 10% (25 M).

Effect of training data size. We now show the effectiveness of our approach on lower and higher data resource settings. For this, we select a subset of languages for which we find more training data available on CV-13 (ca, th, ta). Table 5 presents results for our approach in low (3k utterances; ∼4 h), and higher-resource settings (28k utterances; ∼40 h), compared to the 10k results from Table 4. We observe that, as expected, increasing the amount of trainable examples leads to superior ASR performance for both approaches, with the leveraging of KD (*DistilWhisper*) being consistently superior to CLSR-FT. For the 28k setup (ca, th, ta), we are able to reduce the out-of-domain WER gap between `whisper-large-v2` and `whisper-small` by 75% (from 12 to 3 WER).[11] For the 3k setup, we reduce the WER gap by 35.8% using only 4 h of training data. This implies that our approach has the potential to improve ASR performance across low-resource languages for which less training data is available.

A full description of the work can be found in our paper (Ferraz et al., 2024). It will be published at IEEE ICASSP 2024 conference (April 2024). Model weights[12] and code for training and inference[13] are made available to the community.

4.2 Efficient CTC Regularization via Coarse Labels for End-to-End Speech Translation

Developing techniques to support the translation from a source-language audio to a target-language text directly, or end-to-end (E2E) speech translation (ST), has attracted increasing attention recently due to its potential of reducing translation latency and avoiding error propagation (Duong et al., 2016; Bérard et al., 2016). However, solving this task is non-trivial because of the speech-text

[11] `whisper-large-v2` and `whisper-small` avg FLEURS scores for ca, th, ta are respectively 12.5 and 24.5.
[12] https://huggingface.co/collections/naver/multilingual-distilwhisper-6576ecae8d209fc6a767d9e7
[13] https://github.com/naver/multilingual-distilwhisper

modality gap: one word corresponds to a stochastic sequence of speech signals that vary greatly across speakers and over contexts, which increases the learning difficulty. Recent progress on E2E ST mainly focuses on bridging this gap through the encoder-decoder framework from diverse perspectives (Di Gangi et al., 2019; Salesky et al., 2019; Zhang et al., 2020; Wang et al., 2020; Han et al., 2021; Zheng et al., 2021).

CTC regularization is such an approach that facilitates the modeling of translation by aligning speech representations from the encoder with discrete labels dynamically via the lens of the Connectionist Temporal Classification (CTC) objective (Graves et al., 2006). Bahar et al. (2019) first examined the use of the source transcript as discrete labels, improving translation quality consistently across various ST settings; Zhang et al. (2022) further discovered that using the target translation as labels instead can also be surprisingly effective although speech-translation pairs arguably violates CTC's monotonicity prerequisite. Nevertheless, these successes come at the cost of increased computational overheads and model parameters because CTC demands an extra prediction layer over its label space for probability estimation and this space is often huge – traditionally the source or target vocabulary size (Gaido et al., 2020). We thus explore strategies to achieve the best of both worlds, i.e., improving the efficiency of CTC regularization without hurting its performance.

We address this problem by reexamining the need for genuine vocabulary labels for CTC. In contrast to CTC-based generation (Graves et al., 2006), the prediction layer in CTC regularization of ST is discarded at inference. In other words, sticking to genuine labels is computationally unnecessary. Since the large label space of CTC is a crucial bottleneck hindering training efficiency, we explore ways of reducing it. We propose **Co**arse **La**beling for CTC (CoLaCTC)) that manipulates this space by merging vocabulary labels based on simple heuristic rules. Concretely, we map the source or target vocabulary to a pseudo label space subject to some predefined size using simple operations, such as *truncation*, *modulo*, *division* and *log-scaling*.

Despite the label space being transformed, the generated coarse labels still maintain a strong correlation with their vocabulary counterparts, ensuring their informativeness for representation learning. We rigorously examined our method on the MuST-C (Di Gangi et al., 2019) and the Multilingual TEDx (Salesky et al., 2021) benchmarks, covering 4 source languages and 8 target languages. Across diverse settings, CoLaCTC successfully achieves comparable or even better translation performance than the CTC baseline but with significantly improved training efficiency (up to 1.77× speedup depending on the original vocabulary size). Our main contributions are summarized below:

- We propose coarse labeling for CTC regularization which offers a mechanism to decouple the CTC label size from the vocabulary size; with CoLaCTC, a CTC-regularized model can be trained nearly as fast as a non-CTC model.
- We compare two types of CTC regularization for ST, i.e., using transcript or translation for labeling, and show that transcript performs better *when it is available*.
- CoLaCTC delivers promising performance on 4 source and 8 target languages, and also generalizes to both types of CTC regularization.
- Our empirical analysis reveals that CoLaCTC benefits translation similarly to the CTC baseline on different aspects, including homophone translation, and seems to improve the contextualization of speech representations.

A full description of the work can be found in our paper (Zhang et al., 2023).

5 Conclusion

In this report we presented the work we conducted during the first one year and a half of UTTER project in the context of WP3. Future work includes the multimodal integration of the speech representations learned by `mHuBERT-147` into `TowerLM`. We will also continue working on the pre-training of useful general-purpose speech and text foundation models. Finally, we will also continue to investigate methods for efficient training for both speech and textual tasks.

References

Duarte M. Alves, José Pombal, Nuno M. Guerreiro, Pedro H. Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, Pierre Colombo, José G. C. de Souza, and André F. T. Martins. Tower: An open multilingual large language model for translation-related tasks. 2024. URL https://arxiv.org/pdf/2402.17733.pdf.

Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, Eric Chu, Jonathan H. Clark, Laurent El Shafey, Yanping Huang, Kathy Meier-Hellstern, Gaurav Mishra, Erica Moreira, Mark Omernick, Kevin Robinson, Sebastian Ruder, Yi Tay, Kefan Xiao, Yuanzhong Xu, Yujing Zhang, Gustavo Hernandez Abrego, Junwhan Ahn, Jacob Austin, Paul Barham, Jan Botha, James Bradbury, Siddhartha Brahma, Kevin Brooks, Michele Catasta, Yong Cheng, Colin Cherry, Christopher A. Choquette-Choo, Aakanksha Chowdhery, Clément Crepy, Shachi Dave, Mostafa Dehghani, Sunipa Dev, Jacob Devlin, Mark Díaz, Nan Du, Ethan Dyer, Vlad Feinberg, Fangxiaoyu Feng, Vlad Fienber, Markus Freitag, Xavier Garcia, Sebastian Gehrmann, Lucas Gonzalez, Guy Gur-Ari, Steven Hand, Hadi Hashemi, Le Hou, Joshua Howland, Andrea Hu, Jeffrey Hui, Jeremy Hurwitz, Michael Isard, Abe Ittycheriah, Matthew Jagielski, Wenhao Jia, Kathleen Kenealy, Maxim Krikun, Sneha Kudugunta, Chang Lan, Katherine Lee, Benjamin Lee, Eric Li, Music Li, Wei Li, YaGuang Li, Jian Li, Hyeontaek Lim, Hanzhao Lin, Zhongtao Liu, Frederick Liu, Marcello Maggioni, Aroma Mahendru, Joshua Maynez, Vedant Misra, Maysam Moussalem, Zachary Nado, John Nham, Eric Ni, Andrew Nystrom, Alicia Parrish, Marie Pellat, Martin Polacek, Alex Polozov, Reiner Pope, Siyuan Qiao, Emily Reif, Bryan Richter, Parker Riley, Alex Castro Ros, Aurko Roy, Brennan Saeta, Rajkumar Samuel, Renee Shelby, Ambrose Slone, Daniel Smilkov, David R. So, Daniel Sohn, Simon Tokumine, Dasha Valter, Vijay Vasudevan, Kiran Vodrahalli, Xuezhi Wang, Pidong Wang, Zirui Wang, Tao Wang, John Wieting, Yuhuai Wu, Kelvin Xu, Yunhan Xu, Linting Xue, Pengcheng Yin, Jiahui Yu, Qiao Zhang, Steven Zheng, Ce Zheng, Weikang Zhou, Denny Zhou, Slav Petrov, and Yonghui Wu. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*, 2023. URL https://arxiv.org/abs/2305.10403.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. Common voice: A massively-multilingual speech corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4218–4222, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.520.

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale. In *Proc. Interspeech 2022*, pages 2278–2282, 2022. doi: 10.21437/Interspeech.2022-143.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 2020.

Parnia Bahar, Tobias Bieschke, and Hermann Ney. A comparative study on end-to-end speech to text translation. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 792–799, 2019. doi: 10.1109/ASRU46091.2019.9003774.

Alexandre Bérard, Olivier Pietquin, Christophe Servan, and Laurent Besacier. Listen and translate: A proof of concept for end-to-end speech-to-text translation. In *NIPS Workshop on End-to-end Learning for Speech and Audio Processing*, Barcelona, Spain, 2016.

Eleftheria Briakou, Colin Cherry, and George Foster. Searching for needles in a haystack: On the role of incidental bilingualism in palm's translation capability. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, 2023. URL https://aclanthology.org/2023.acl-long.524.pdf.

Alejandro Hernández Cano, Matteo Pagliardini, Andreas Köpf, Kyle Matoba, Amirkeivan Mohtashami, Xingyao Wang, Olivia Simin Fan, Axel Marmet, Deniz Bayazit, Igor Krawczuk, Zeming Chen, Francesco Salvi, Antoine Bosselut, and Martin Jaggi. epfllm megatron-llm, 2023. URL https://github.com/epfLLM/Megatron-LLM.

Xuankai Chang, Brian Yan, Kwanghee Choi, Jeeweon Jung, Yichen Lu, Soumi Maiti, Roshan Sharma, Jiatong Shi, Jinchuan Tian, Shinji Watanabe, et al. Exploring speech recognition, translation, and understanding with discrete speech units: A comparative study. *arXiv preprint arXiv:2309.15800*, 2023.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 2022.

William Chen, Xuankai Chang, Yifan Peng, Zhaoheng Ni, Soumi Maiti, and Shinji Watanabe. Reducing Barriers to Self-Supervised Learning: HuBERT Pre-training with Academic Compute. In *Proc. INTERSPEECH 2023*, pages 4404–4408, 2023a. doi: 10.21437/Interspeech.2023-1176.

William Chen, Jiatong Shi, Brian Yan, Dan Berrebbi, Wangyou Zhang, Yifan Peng, Xuankai Chang, Soumi Maiti, and Shinji Watanabe. Joint prediction and denoising for large-scale multilingual self-supervised learning. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–8. IEEE, 2023b.

Ju-Chieh Chou, Chung-Ming Chien, Wei-Ning Hsu, Karen Livescu, Arun Babu, Alexis Conneau, Alexei Baevski, and Michael Auli. Toward joint language modeling for speech units and text. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6582–6593, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.438. URL https://aclanthology.org/2023.findings-emnlp.438.

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. Unsupervised Cross-Lingual Representation Learning for Speech Recognition. In *Proc. Interspeech 2021*, pages 2426–2430, 2021. doi: 10.21437/Interspeech.2021-329.

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

Mattia A. Di Gangi, Roldano Cattoni, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 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, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1202. URL https://www.aclweb.org/anthology/N19-1202.

Mattia A. Di Gangi, Matteo Negri, and Marco Turchi. Adapting Transformer to End-to-End Spoken Language Translation. In *Proc. Interspeech 2019*, pages 1133–1137, 2019. doi: 10.21437/Interspeech.2019-3045. URL http://dx.doi.org/10.21437/Interspeech.2019-3045.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. The faiss library. 2024.

Long Duong, Antonios Anastasopoulos, David Chiang, Steven Bird, and Trevor Cohn. An attentional model for speech translation without transcription. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 949–959, San Diego, California, June 2016. Association for Computational Linguistics. doi: 10.18653/v1/N16-1109. URL https://www.aclweb.org/anthology/N16-1109.

Paul-Ambroise Duquenne, Hongyu Gong, Ning Dong, Jingfei Du, Ann Lee, Vedanuj Goswami, Changhan Wang, Juan Pino, Benoît Sagot, and Holger Schwenk. SpeechMatrix: A large-scale mined corpus of multilingual speech-to-speech translations. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16251–16269, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.899. URL https://aclanthology.org/2023.acl-long.899.

Patrick Fernandes, Daniel Deutsch, Mara Finkelstein, Parker Riley, André Martins, Graham Neubig, Ankush Garg, Jonathan Clark, Markus Freitag, and Orhan Firat. The devil is in the errors: Leveraging large language models for fine-grained machine translation evaluation. In *Proceedings of the Eighth Conference on Machine Translation*, Singapore, December 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.wmt-1.100.

Thomas Palmeira Ferraz, Marcely Zanon Boito, Caroline Brun, and Vassilina Nikoulina. Multilingual distilwhisper: Efficient distillation of multi-task speech models via language-specific experts. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024.

Markus Freitag, Nitika Mathur, Chi-kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Tom Kocmi, Frederic Blain, Daniel Deutsch, Craig Stewart, Chrysoula Zerva, Sheila Castilho, Alon Lavie, and George Foster. Results of wmt23 metrics shared task: Metrics might be guilty but references are not innocent. In *Proceedings of the Eighth Conference on Machine Translation*, Singapore, December 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.wmt-1.51.

Marco Gaido, Mattia A. Di Gangi, Matteo Negri, and Marco Turchi. End-to-end speech-translation with knowledge distillation: FBK@IWSLT2020. In *Proceedings of the 17th International Conference on Spoken Language Translation*, pages 80–88, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.iwslt-1.8. URL https://aclanthology.org/2020.iwslt-1.8.

Alex Graves, Santiago Fernández, and Faustino Gomez. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In *In Proceedings of the International Conference on Machine Learning, ICML 2006*, pages 369–376, 2006.

Chi Han, Mingxuan Wang, Heng Ji, and Lei Li. Learning shared semantic space for speech-to-text translation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2214–2225, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.195. URL https://aclanthology.org/2021.findings-acl.195.

Kenneth Heafield. KenLM: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, Edinburgh, Scotland, July 2011. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W11-2123.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. How good are gpt models at machine translation? a comprehensive evaluation. *arXiv preprint arXiv:2302.09210*, 2023. URL https://arxiv.org/abs/2302.09210.

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

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2021.

Vivek Iyer, Pinzhen Chen, and Alexandra Birch. Towards effective disambiguation for machine translation with large language models. In *Proceedings of the Eighth Conference on Machine Translation*, Singapore, December 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.wmt-1.44.

Tom Kocmi and Christian Federmann. GEMBA-MQM: Detecting translation quality error spans with GPT-4. In *Proceedings of the Eighth Conference on Machine Translation*, Singapore, December 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.wmt-1.64.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Philipp Koehn, Benjamin Marie, Christof Monz, Makoto Morishita, Kenton Murray, Makoto Nagata, Toshiaki Nakazawa, Martin Popel, Maja Popović, and Mariya Shmatova. Findings of the 2023 conference on machine translation (WMT23): LLMs are here but not quite there yet. In *Proceedings of the Eighth Conference on Machine Translation*, Singapore, December 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.wmt-1.1.

Tom Kocmi, Vilém Zouhar, Christian Federmann, and Matt Post. Navigating the metrics maze: Reconciling score magnitudes and accuracies. *arXiv prepring arXiv:2401.06760*, 2024. URL https://arxiv.org/abs/2401.06760.

Ann Lee, Hongyu Gong, Paul-Ambroise Duquenne, Holger Schwenk, Peng-Jen Chen, Changhan Wang, Sravya Popuri, Yossi Adi, Juan Pino, Jiatao Gu, and Wei-Ning Hsu. Textless speech-to-speech translation on real data. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz, editors, *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 860–872, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.63. URL https://aclanthology.org/2022.naacl-main.63.

Pengcheng Li, Genshun Wan, Fenglin Ding, Hang Chen, Jianqing Gao, Jia Pan, and Cong Liu. Improved speech pre-training with supervision-enhanced acoustic unit. *arXiv preprint arXiv:2212.03482*, 2022.

Tzu-Quan Lin, Hung-yi Lee, and Hao Tang. Melhubert: A simplified hubert on mel spectrograms. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–8. IEEE, 2023.

Andy T Liu, Shu-wen Yang, Po-Han Chi, Po-chun Hsu, and Hung-yi Lee. Mockingjay: Unsupervised speech representation learning with deep bidirectional transformer encoders. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6419–6423. IEEE, 2020.

Mariana Neves, Antonio Jimeno Yepes, Aurélie Névéol, Rachel Bawden, Giorgio Maria Di Nunzio, Roland Roller, Philippe Thomas, Federica Vezzani, Maika Vicente Navarro, Lana Yeganova, Dina Wiemann, and Cristian Grozea. Findings of the WMT 2023 biomedical translation shared task: Evaluation of ChatGPT 3.5 as a comparison system. In *Proceedings of the Eighth Conference on Machine Translation*, Singapore, 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.wmt-1.2.

Santiago Pascual, Mirco Ravanelli, Joan Serrà, Antonio Bonafonte, and Yoshua Bengio. Learning problem-agnostic speech representations from multiple self-supervised tasks. In *Proceedings of the 20th Annual Conference of the International Speech Communication Association*, pages 161–165, 2019.

Jonas Pfeiffer, Naman Goyal, Xi Lin, Xian Li, James Cross, Sebastian Riedel, and Mikel Artetxe. Lifting the curse of multilinguality by pre-training modular transformers. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3479–3495, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.255. URL https://aclanthology.org/2022.naacl-main.255.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, et al. Scaling speech technology to 1,000+ languages. *arXiv preprint arXiv:2305.13516*, 2023.

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

Gema Ramírez-Sánchez, Jaume Zaragoza-Bernabeu, Marta Bañón, and Sergio Ortiz Rojas. Bifixer and bicleaner: two open-source tools to clean your parallel data. In *Proceedings of the*

22nd Annual Conference of the European Association for Machine Translation, Lisboa, Portugal, 2020. European Association for Machine Translation. URL https://aclanthology.org/2020.eamt-1.31.

Vikas Raunak, Amr Sharaf, Yiren Wang, Hany Awadalla, and Arul Menezes. Leveraging GPT-4 for automatic translation post-editing. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, Singapore, 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.findings-emnlp.804.

Mirco Ravanelli, Jianyuan Zhong, Santiago Pascual, Pawel Swietojanski, João Monteiro, Jan Trmal, and Yoshua Bengio. Multi-task self-supervised learning for robust speech recognition. In *Proceedings of the 2020 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 6989–6993, 2020.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, Abu Dhabi, United Arab Emirates (Hybrid), 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.wmt-1.60.

Elizabeth Salesky, Matthias Sperber, and Alan W Black. Exploring phoneme-level speech representations for end-to-end speech translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1835–1841, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1179. URL https://www.aclweb.org/anthology/P19-1179.

Elizabeth Salesky, Matthew Wiesner, Jacob Bremerman, Roldano Cattoni, Matteo Negri, Marco Turchi, Douglas W. Oard, and Matt Post. The Multilingual TEDx Corpus for Speech Recognition and Translation. In *Proc. Interspeech 2021*, pages 3655–3659, 2021. doi: 10.21437/Interspeech.2021-11.

Víctor M. Sánchez-Cartagena, Marta Bañón, Sergio Ortiz-Rojas, and Gema Ramírez. Prompsit's submission to WMT 2018 parallel corpus filtering shared task. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, Belgium, Brussels, 2018. Association for Computational Linguistics. URL https://aclanthology.org/W18-6488.

Jiatong Shi, Dan Berrebbi, William Chen, En-Pei Hu, Wei-Ping Huang, Ho-Lam Chung, Xuankai Chang, Shang-Wen Li, Abdelrahman Mohamed, Hung yi Lee, and Shinji Watanabe. ML-SUPERB: Multilingual Speech Universal PERformance Benchmark. In *Proc. INTERSPEECH 2023*, pages 884–888, 2023a. doi: 10.21437/Interspeech.2023-1316.

Jiatong Shi, William Chen, Dan Berrebbi, Hsiu-Hsuan Wang, Wei-Ping Huang, En-Pei Hu, Ho-Lam Chuang, Xuankai Chang, Yuxun Tang, Shang-Wen Li, et al. Findings of the 2023 ml-superb challenge: Pre-training and evaluation over more languages and beyond. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–8. IEEE, 2023b.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez,

Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023a. URL https://arxiv.org/abs/2302.13971.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023b. URL https://arxiv.org/abs/2307.09288.

Chengyi Wang, Yu Wu, Shujie Liu, Ming Zhou, and Zhenglu Yang. Curriculum pre-training for end-to-end speech translation. *arXiv preprint arXiv:2004.10093*, 2020.

Minghan Wang, Yinglu Li, Jiaxin Guo, Xiaosong Qiao, Zongyao Li, Hengchao Shang, Daimeng Wei, Shimin Tao, Min Zhang, and Hao Yang. WhiSLU: End-to-End Spoken Language Understanding with Whisper. In *Proc. INTERSPEECH 2023*, pages 770–774, 2023a. doi: 10.21437/Interspeech.2023-1505.

Tianrui Wang, Long Zhou, Ziqiang Zhang, Yu Wu, Shujie Liu, Yashesh Gaur, Zhuo Chen, Jinyu Li, and Furu Wei. Viola: Unified codec language models for speech recognition, synthesis, and translation. *arXiv preprint arXiv:2305.16107*, 2023b.

Shu wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Yist Y. Lin, Andy T. Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, Tzu-Hsien Huang, Wei-Cheng Tseng, Ko tik Lee, Da-Rong Liu, Zili Huang, Shuyan Dong, Shang-Wen Li, Shinji Watanabe, Abdelrahman Mohamed, and Hung yi Lee. SUPERB: Speech Processing Universal PERformance Benchmark. In *Proc. Interspeech 2021*, pages 1194–1198, 2021. doi: 10.21437/Interspeech.2021-1775.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. Ccnet: Extracting high quality monolingual datasets from web crawl data. *arXiv preprint arXiv:1911.00359*, 2019. URL https://arxiv.org/abs/1911.00359.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-art natural language processing. In Qun Liu and David Schlangen, editors, *Proceedings of the*

2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 38–45, Online, October 2020a. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-demos.6. URL https://aclanthology.org/2020.emnlp-demos.6.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45, 2020b.

Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. A paradigm shift in machine translation: Boosting translation performance of large language models. In *The Twelfth International Conference on Learning Representations*, 2024a. URL https://openreview.net/forum?id=farT6XXntP.

Haoran Xu, Amr Sharaf, Yunmo Chen, Weiting Tan, Lingfeng Shen, Benjamin Van Durme, Kenton Murray, and Young Jin Kim. Contrastive preference optimization: Pushing the boundaries of llm performance in machine translation. *arXiv preprint arXiv:2401.08417*, 2024b. URL https://arxiv.org/abs/2401.08417.

Wenda Xu, Danqing Wang, Liangming Pan, Zhenqiao Song, Markus Freitag, William Wang, and Lei Li. INSTRUCTSCORE: Towards explainable text generation evaluation with automatic feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Singapore, 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.emnlp-main.365.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mt5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, Online, 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.naacl-main.41.

Biao Zhang, Ivan Titov, Barry Haddow, and Rico Sennrich. Adaptive feature selection for end-to-end speech translation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2533–2544, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.230. URL https://aclanthology.org/2020.findings-emnlp.230.

Biao Zhang, Ankur Bapna, Rico Sennrich, and Orhan Firat. Share or not? learning to schedule language-specific capacity for multilingual translation. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=Wj4ODo0uyCF.

Biao Zhang, Barry Haddow, and Rico Sennrich. Revisiting end-to-end speech-to-text translation from scratch. In *Proceedings of ICML*, 2022.

Biao Zhang, Barry Haddow, and Rico Sennrich. Efficient CTC regularization via coarse labels for end-to-end speech translation. In Andreas Vlachos and Isabelle Augenstein, editors, *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2264–2276, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.eacl-main.166. URL https://aclanthology.org/2023.eacl-main.166.

Renjie Zheng, Junkun Chen, Mingbo Ma, and Liang Huang. Fused acoustic and text encoding for multimodal bilingual pretraining and speech translation. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 12736–12746. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/zheng21a.html.

ENDPAGE

UTTER

HORIZON-CL4-2021-HUMAN-01 101070631

D15 First report on XR models