Recent Highlights in Multilingual and Multimodal Speech Translation

Danni Liu and **Jan Niehues**
Karlsruhe Institute of Technology, Germany
{danni.liu, jan.niehues}@kit.edu

Abstract

Speech translation has witnessed significant progress driven by advancements in modeling techniques and the growing availability of training data. In this paper, we highlight recent advances in two ongoing research directions in ST: scaling the models to **1)** many translation directions (multilingual ST) and **2)** beyond the text output modality (multimodal ST). We structure this review by examining the sequential stages of a model's development lifecycle: determining training resources, selecting model architecture, training procedures, evaluation metrics, and deployment considerations. We aim to highlight recent developments in each stage, with a particular focus on model architectures (dedicated speech translation models and LLM-based general-purpose model) and training procedures (task-specific vs. task-invariant approaches). Based on the reviewed advancements, we identify and discuss ongoing challenges within the field of speech translation.

1 Introduction

Speech translation (ST) is the task of automatically converting speech in a source language into its equivalent in a target language. Recently, there has been significant interest in *multilingual* models (Di Gangi et al., 2019; Inaguma et al., 2019; Li et al., 2021; Le et al., 2021; Radford et al., 2023) that serve a broad range of translation directions, as well as *multimodal* models (Inaguma et al., 2023; Rubenstein et al., 2023; Seamless Communication et al., 2023b) that not only generate text translations but can also synthesize speech output.[1] Both developments are crucial steps towards making ST technologies more inclusive. By expanding language coverage and offering diverse output modalities, these advancements make ST models accessible to a wider range of users, allowing them to interact with the technology in their preferred language and format. Besides the practical relevance, multilingual and multimodal translation are instances of multi-task learning (Caruana, 1997), a central machine learning challenge.

In this paper, we aim to review recent advancements in multilingual and multimodal ST. We structure the review by the stages in a model's development lifecycle, as illustrated in Figure 1. These stages consist of model coverage and architecture selection, training procedures, evaluation methodologies, and deployment considerations. In the review of current model architectures (§3), besides discussing dedicated models for translation, we review emerging models in adapting text-based large language models (LLMs) for speech processing. Given the inherent multi-task learning nature of both multilingual and multimodal ST, we put special emphasis on the learning procedure (§4). Specifically, we take two perspectives from task-specific and task-invariant modeling, and discuss their roles in terms of the trade-off between interference and transfer.

While prioritizing direct ST, we also review related multilingual and multimodal techniques in automatic speech recognition (ASR) and text-to-text machine translation (MT), as they often are extendable to ST tasks. We also note that this work is not an exhaustive survey, but rather aims to highlight directions of recent developments and provide context for open challenges.

2 Training Resources

Determining training resources is one of the initial steps when building a speech translation model. This section provides a brief overview of the language and modality coverage (§2.1) in existing training resources, followed by discussions on scaling datasets by augmentation or mining (§2.2).

[1] Here we restrict our discussion to the two modalities of speech and text. We acknowledge the relevance of additional modalities, such as vision, and leave them for open questions.

Proceedings of the 21st International Conference on Spoken Language Translation (IWSLT 2024), pages 283–301
August 15-16, 2024 ©2024 Association for Computational Linguistics

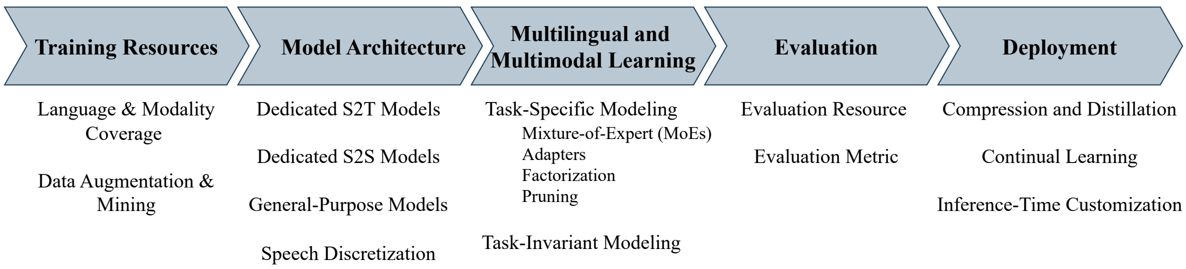

Figure 1: Overall structure of the paper, following sequential stages of model development lifecycle.

Dataset	Directions	Modality & Type	# Lang. Pairs	Total Hours
MuST-C (Di Gangi et al., 2019; Cattoni et al., 2021)	en→X	S2T	14	0.4k
Europarl-ST (Iranzo-Sánchez et al., 2020)	X→X	S2T	12	0.5k
CoVoST 2 (Wang et al., 2021b)	en→X, X→en	S2T	36	3k
mTEDx (Salesky et al., 2021)	X→X	S2T	13	0.4k
VoxPopuli (Wang et al., 2021a)	X→X	S2T/S, interpretation	210	17k
CVSS (Jia et al., 2022b)	X→en	S2T/S, synthesized	21	2k
SpeechMatrix (Duquenne et al., 2023a)	X→X	S2T/S, mined	136	418k

Table 1: Overview of popuplar speech translation training resources.

2.1 Language and Modality Coverage

Curating datasets for speech translation is labor-intensive. Popular training resources often rely on contributions from volunteers on platforms like TED and Common Voice, or are sourced from governmental bodies. Table 1 provides an overview of commonly used speech translation datasets. A trend towards broader language coverage is evident, with datasets like Europarl-ST and mTEDx covering non-English translation directions. Moreover, there has also been growing availability of translation resources with speech output, exemplified by VoxPopuli, CVSS, and SpeechMatrix.

2.2 Augmenting and Mining Data

Speech translation models suffer from the scarcity of parallel data. To address this challenge, several data augmentation approaches have emerged. One approach is to leverage pretrained MT models to convert ASR data into synthetic speech translation pairs (Pino et al., 2020). Text-to-speech (TTS) systems can also be employed to create augmented training data from existing text resources (Jia et al., 2019a, 2022b). Another way to tackle data scarcity is to mine parallel data in large unpaired data collections. In general, these approaches typically invovle learning a multilingual or multimodal sentence embedder, where distances within the embedding space can be used to identify potential parallel data points (Schwenk, 2018). The effectiveness of this method on ST was demonstrated by Duquenne et al. (2021), who showed that mined speech-to-text data can improve the performance of direct speech translation models. This line work was extended with the creation of SpeechMatrix (Duquenne et al., 2023a), a large-scale speech-to-speech translation corpus built using mined data.

2.3 Outlook

Understanding the Impact of Data Quality and Style The increasing volume of ST training resources comes with a risk on data quality. While scaling up training data volume offers obvious benefits, noisy data could hinder model performance. To the best of our knowledge, there is currently no established best practice for data filtering in speech translation. Current research presents conflicting findings on the impact of data quality. For example, Ouyang et al. (2022) observed no improvement in model performance when removing misaligned parallel data from the training set, while Gaido et al. (2022) demonstrated gains by filtering out such misalignments. Meanwhile it also remains unclear whether data filtering best practices are language-specific. Besides data quality, a deeper understanding of training data style's impact on ST performance is also beneficial. In the related field of MT, Maillard et al. (2023) showed gains by using small amounts of professionally-translated data. In ST, Ko et al. (2023) observed that interpretation-style data facilitates simultaneous translation models. Inspired by this finding, Sakai et al. (2024) pro-

Model	# Param	S2T X→en (21 lang.)	S2T en→X (15 lang.)	S2S X→en (21 lang.)	Learning
Speech-to-Text					
XLS-R (Babu et al., 2022)	2B	22.1	27.8	–	self-supervised + supervised FT
MAESTRO (Chen et al., 2022b)	0.6B	25.2	–	–	self-supervised + supervised FT
Whisper Large (Radford et al., 2023)	1.6B	29.7	–	–	(weakly) supervised
ComSL Large (Le et al., 2023)	1.3B	31.5	–	–	(weakly) supervised
AudioPaLM (Rubenstein et al., 2023)	8B	35.4	–	–	supervised FT
↪ + PaLM 2 (Anil et al., 2023)	8B	37.8	–	–	supervised FT
ZeroSWOT Large (Tsiamas et al., 2024)	1.7B	–	31.2	–	zero-shot combination pretrained ASR & MT
Speech-to-Text/Speech					
AudioPaLM S2ST (Rubenstein et al., 2023)	8B	36.2	–	32.5	supervised FT
SeamlessM4T Large (Seamless Communication et al., 2023b)	2.3B	34.1	30.6	36.5	self-supervised + supervised FT
↪ v2 (Seamless Communication et al., 2023a)	2.3B	36.6	31.7	39.2	self-supervised + supervised FT

Table 2: Performance overview of selected recent models for speech-to-text (S2T; BLEU↑; on **CoVoST 2**) and speech-to-speech translation (S2S; ASR-BLEU↑; on **CVSS**).

pose augmenting existing datasets with synthetic targets that mimic the style of interpretation data. Overall, exploring other data styles relevant to specific speech translation tasks could be promising for further performance improvements.

Targeted Resources for Low-Resource Languages The training resources in Table 1 primarily cover high-resource languages. For truly low-resource languages, readily available internet data may be scarce or non-existent. In such cases, collaboration with local communities becomes essential for data collection. The AmericasNLP speech translation shared task (Ebrahimi et al., 2021) is a successful example of this approach. The initiative focused on gathering speech translation data for indigenous languages of the Americas, demonstrating the feasibility of community-driven data collection for low-resource languages.

3 Model Architecture

In this section, we first review dedicated model architectures for speech-to-text (S2T; §3.1) and speech-to-speech (S2S; §3.2) translation, with a focus on the use of foundation models. Afterwards, we discuss recent developments in adapting general-purpose LLMs (§3.3) for encoding or generating speech.

3.1 Dedicated S2T Translation Models

Integrating Foundation Models Foundation models have become essential resources for training. Reflecting this trend, since 2022, a selection of (often massively multilingual) audio and text foundation models are allowed in the constrained data condition[2] in IWSLT (Anastasopoulos et al., 2022). However, as most current speech foundation models are either unsupervised/encoder-only (Baevski et al., 2020; Chung et al., 2021a; Chen et al., 2022a) or supervised with a limited translation directions (Radford et al., 2023), further adaptation is typically needed on specific speech translation tasks. A promising direction has been to pair pretrained audio encoders with text decoders, as frequently used in recent IWSLT system submissions (Gállego et al., 2021; Pham et al., 2022; Huang et al., 2023). In this process, additional lightweight adapters often are injected to bridge the audio and text representations (Li et al., 2021; Gállego et al., 2021; Zhao et al., 2022). For a focused survey of foundation models in S2T translation, we refer the readers to Gaido et al. (2024).

Representative Models and Trends Table 2 presents a chronological overview of some recent S2T translation models. Examining benchmark results on the CoVoST 2 dataset, a substantial performance improvement (+15.7 BLEU) is observed for X→en directions over the last two years. However, the picture for en→X directions remains less clear due to the limited number of data points. Nonetheless, when also considering the speech-

[2] as opposed the unconstrained data condition with no restrictions on training data and resources

to-text/speech results, we clearly see the progress in en→X is far behind X→en (22.1→36.6 BLEU vs. 27.8→31.7 BLEU). Regarding the learning paradigm, a trend emerges from developing new self-supervised representation learning schemes (XLS-R, MAESTRO) towards directly using pretrained models (ComSL, AudioPaLM), in particular the plug-and-play combination of pretrained modules (Tsiamas et al., 2024) in zero-shot conditions.

3.2 Dedicated S2S Translation Models

Challenges of Generating Speech Speech generation presents unique challenges compared to text generation. First, the inherent longer length of audio signals poses significant computational demands for conventional autoregressive approaches. Moreover, capturing long-range dependencies within these extended sequences becomes more difficult for the model. Second, speech generation is often an under-specified problem. Unlike text, speech can be produced with various voice characteristics for the same content. This ambiguity creates a larger space of possible outputs that the model must handle.

Textless Models An advantage of speech-to-speech translation is the possibility to circumvent intermediate written text. Indeed, there has been growing interest in textless models (Jia et al., 2019b; Tjandra et al., 2019; Zhang et al., 2021b; Lee et al., 2022; Jia et al., 2022a), which do not rely on intermediate text representations and are especially suitable for S2ST of languages without standard writing systems. In general, these approaches first create discrete representations with unsupervised acoustic unit discovery by clustering or autoencoding (Tjandra et al., 2019; Zhang et al., 2021b; Hsu et al., 2021). The learned inventory of acoustic units could be viewed as learned phonemes. The input speech are then mapped to the discrete units, after which a unit-to-speech model is responsible for creating the output speech. Discretization of speech is further discussed in §3.4. Another advantage of textless models is the potential of preserving source voice characteristics. In particular, SeamlessExpressive (Seamless Communication et al., 2023a) is a recent model dedicated to voice characteristic preservation. Expressivity embeddings are extracted from the source speech and integrated in the output speech generation. Specifically, the model disentangles semantic and expressivity components from the source speech by learning speech reconstruction.

Representative Models and Trends In the lower section of Table 2, we list recent models supporting both S2T and S2S translation: AudioPaLM S2ST (Rubenstein et al., 2023) and SeamlessM4T (Seamless Communication et al., 2023b,a). AudioPaLM S2ST, in contrast to its variant lacking speech generation capabilities, is additionally trained on TTS and S2S translation data. The inclusion of additional modalities not only enables speech generation as an output, but also improves S2T translation performance (35.4→36.2 BLEU). Similar to its text generation counterpart, AudioPaLM S2ST fuses AudioLM (Borsos et al., 2023a) and the text-based PaLM model (Anil et al., 2023). The model has a joint vocabulary for both audio and text inputs. The audio tokens are created by an upgraded version of the USM encoder (Zhang et al., 2023b), which discretizes and downsamples the speech input. Speech tokenization is further discussed in (§3.1). Unlike AudioPaLM, **SeamlessM4T** utilizes an encoder-decoder architecture primarily fine-tuned from NLLB (NLLB Team et al., 2022). Its encoder additionally can additionally process speech inputs based on w2v-BERT representations (Chung et al., 2021b). Both AudioPaLM S2ST and SeamlessM4T achieve speech generation by optionally chaining a speech generation module after the text generation stage. AudioPaLM S2ST first converts audio tokens to SoundStream tokens (Zeghidour et al., 2022), which are then used by a vocoder to synthesize audio waveforms. SeamlessM4T, on the other hand, employs a text-to-unit encoder-decoder model followed by a vocoder.

3.3 General-Purpose Models

Adapting LLMs to Encode and Generate Speech Driven by the recent advancements in LLMs, there has been a surge of interest in adapting them for speech translation tasks. However, most publicly available LLMs, such as those in the LLaMA family (Touvron et al., 2023a,b), only support the text-to-text modality. To enable speech translation, these models require additional adaptation for both speech encoding and generation. A common approach for speech encoding involves discretizing and downsampling the audio input. This process transforms the continuous audio signal into a sequence of discrete tokens that the LLM can readily ingest. On the output side, typically discrete audio

Model	Speech Tokenization	Backbone LLM	Generation Module	Evaluated on ST
AudioPaLM (Rubenstein et al., 2023)	USM encoder (variant)	PaLM (8B)	SoundStorm	✓
PolyVoice (Dong et al., 2024)	HuBERT	GPT-2 (1.6B)	SoundStream (variant)	✓
SALMONN (Tang et al., 2024)	Window-level Q-Former	Vicuna (13B)	–	✓
NExT-GPT (Wu et al., 2023)	ImageBind	Vicuna (7B)	AudioLDM	✗
CoDi-2 (Tang et al., 2023)	ImageBind	LLaMA 2 (7B)	AudioLDM 2	✗
AnyGPT (Zhan et al., 2024)	SpeechTokenizer	LLaMA 2 (7B)	SoundStorm (variant)	✗

Table 3: Selected recent works adapting LLMs for speech processing and their components (speech tokenziation module, backbone LLM, and speech generation module).

tokens are generated similarly to text tokens. Afterwards, a synthesizer, for instance SoundStorm (Borsos et al., 2023b), converts these tokens to speech waveforms.

Representative Models and Trends In Table 3, we summarize recent works in LLMs for encoding and generating speech. Regarding the *speech tokenization* modules, common choices include ImageBind (Girdhar et al., 2023), SpeechTokenizer (Zhang et al., 2023a), HuBERT (Hsu et al., 2021), and the encoder of USM (Zhang et al., 2023b). For the *backbone LLMs*, the surveyed models mostly choose use small LLM variants (<10B parameters). For the *audio generation* module, popular choices are diffusion-based AudioLDM (Liu et al., 2023a), vector-quantization-based SoundStream (Zeghidour et al., 2022) and SoundStorm (Borsos et al., 2023b). As many of the reviewed models in Table 3 are not evaluated on speech translation, currently it is still difficult conclusively compare them to more conventional architectures.

3.4 Speech Tokenization

As introduced earlier, speech tokenization offers benefits in various applications, including textless translation and integration with text-based LLMs. Table 4 provides an overview of prominent approaches for speech tokenization and their underlying techniques. A common thread among these methods is the use of residual vector quantization (RVQ) (Barnes et al., 1996), which partitions the latent space into a finite number of subsets. While HuBERT employs k-means clustering, similar to RVQ in its objective of latent space partitioning, it differs in its implementation of offline clustering in a separate stage. In contrast to the other methods, ImageBind (Girdhar et al., 2023) directly encodes audio by transforming the spectrogram by Vision Transformer (ViT) (Dosovitskiy et al., 2021). It is worth exploring whether this approach carries sufficient fine-grained information for speech transcription or translation. The window-level Q-Former used in SALMONN (Tang et al., 2024) is also inspired by image processing. A sliding window of fixed size is applied on the speech features, where each window is processed by a Q-Former (Li et al., 2023), which creates a fixed number of token embeddings. These audio tokens embeddings are later ingested by the backbone LLM.

Model	Technique
HuBERT (Hsu et al., 2021)	k-means clustering
SoundStream (Zeghidour et al., 2022)	RVQ
SoundStorm (Borsos et al., 2023b)	RVQ
SpeechTokenizer (Zhang et al., 2023a)	RVQ
ImageBind (Girdhar et al., 2023)	spectrogram + ViT
Win.-level Q-Former (Tang et al., 2024)	sliding-window + Q-Former

Table 4: Common speech tokenization techniques.

3.5 Outlook

More Unified Speech and Text Generation As reviewed in this section, current speech and text generation approaches primarily rely on sequential processing or separate model branches. This raises the question of whether a more unified approach could be beneficial. Circumventing sequential processing could be particularly beneficial under real-time constraints.

Comparison between Architecture Paradigms Given the recency of some reviewed model types, especially those leveraging LLMs for general-purpose tasks (§3.3), a clear understanding of their performance compared to established architectures is still missing. Comprehensive benchmarking efforts targeting these recently emerged approaches could bridge this gap.

Identifying Scaling Law Prior works have examined how increasing model size affects model performance in MT (Fernandes et al., 2023). As the reviewed approaches in this work primarily focus on smaller LLMs, similar investigations for ST, particularly considering the foundation model size, could yield valuable practical insights.

How far will Transformers take us? A broader open question is whether alternative architectures can challenge the dominance of Transformers. State-space models (Gu et al., 2022a; Gu and Dao, 2023) could be a promising candidate, as their strength lies in capturing long-range dependencies, a crucial aspect for effective ST due to the inherent sequential nature of speech.

4 Multilingual and Multimodal Learning

Both multilingual and multimodal speech translation are instances of multi-task learning, where each translation direction in one input-output modality pair corresponds to one task. As also observed in general multi-task learning (Caruana, 1997), a key goal here is to maximize the transfer while minimizing the interference between tasks, while maintaining an efficient trade-off (Arivazhagan et al., 2019b). Given a defined model architecture (§3), different training procedures control the learned representations. In this section, we will discuss the relevant approaches in detail, taking two perspectives from task-specific (§4.1) and task-invariant modeling (§4.2).

4.1 Task-Specific Modeling

A central question when adding task-specific capacity is determining the optimal allocation between shared and task-specific components. Early works use hand-picked sharing strategies of sub-networks, such as language-specific decoders (Dong et al., 2015), attention heads (Zhu et al., 2020), and layer norm/linear transformation (Zhang et al., 2020). Recently, research interests shifted towards learning to balance between task-specific and shared capacity. We summarize representative approaches in the following categories: **1)** mixture-of-experts, **2)** adapters, **3)** factorization, and **4)** pruning, as illustrated in Figure 2. While these approaches may share similar end goals, the categorization helps to outline their specific computational approaches.

Mixture-of-Experts (MoEs) Compared to their dense counterparts, MoE networks (Eigen et al., 2014; Shazeer et al., 2017; Lepikhin et al., 2021) incorporate multiple expert subnets and use a gating mechanism to selectively activate the expert modules. Besides increasing model capacity, this approach also provides a neat framework for balancing between task-specific and task-agnostic modules. MoEs can be seen as neural architecture search (Baker et al., 2017), where the search space is the combination of the parallel expert modules.

For multilingual applications, a common configuration of MoE is to reserve one universal expert shared by all languages, while keeping the remaining experts language-specific. The importance of each expert module is learned by a gating mechanism. The final output is a mix between language-specific and shared ones. The overall amount of language-specific capacity can be controlled by a budget (Zhang et al., 2021a). There have been works applying MoEs in both multilingual ASR (Gaur et al., 2021; Kwon and Chung, 2023; Hu et al., 2023; Wang et al., 2023b) and MT (Zhang et al., 2021a; NLLB Team et al., 2022; Pires et al., 2023). In direct ST, there are fewer works using MoE. One work (Berrebbi et al., 2022) uses the MoE gating mechanism to balance different acoustic features to improve ST robustness.

Adapters Like MoEs, adapters (Rebuffi et al., 2017; Houlsby et al., 2019; Bapna and Firat, 2019) is another of form conditionally activated network. They can be seen as a restricted case of MoE with hard gating and fixed routing[3]. In this case, how the adapters are allocated to tasks needs to be decided a priori. A variety of allocation schemes have been explored, for example by language pairs (Bapna and Firat, 2019), single languages (Philip et al., 2020), and language families (Chronopoulou et al., 2023). In multilingual ST, language-specific adapters have been shown to improve over monolithic multilingual models and achieve comparable results to full fine-tuning (Le et al., 2021). Besides adding capacity, a more common use-case of adapters in speech translation is to bridge speech and text representations (Li et al., 2021; Escolano et al., 2021; Zhao et al., 2022), especially when coupling pretrained ASR and MT models (Gállego et al., 2021; Tsiamas et al., 2024). Further discussions on this are in §4.2.

Factorization Another perhaps less explored line of work uses factorization to balance language-specific and shared parameters. By decomposing originally shared parameters into (low-rank) factors that are either language-specific or shared, factorization enables a learned task allocation of

[3]Fusion between adapters (Pfeiffer et al., 2021) is an exception.

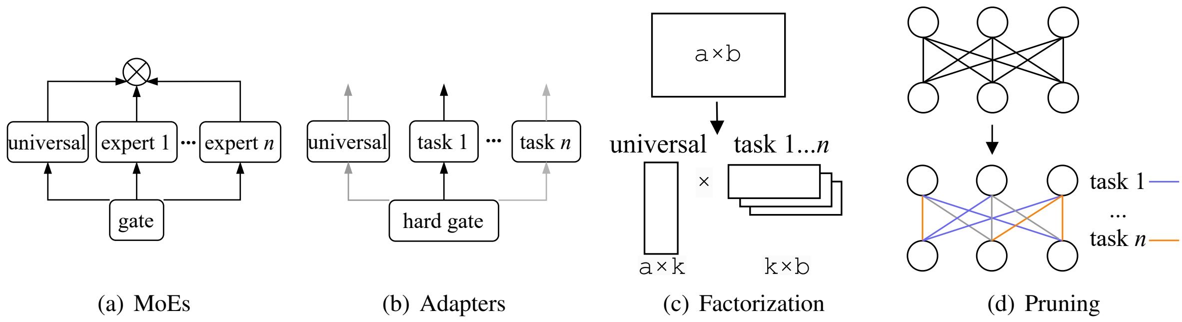

Figure 2: Representative approaches for task-specific modeling.

parameters. This approach has seen applications in multilingual ASR (Pham et al., 2021) and MT (Xu et al., 2023). Compared to MoEs or adapters, an advantage of factorized models is their fewer total parameters, especially under large language coverage (Pham et al., 2021; Xu et al., 2023).

Pruning Pruning also leads to sparse sub-networks, similar to with MoEs. The difference is that pruning starts with a trained model, and then finetunes the selected sub-network. This therefore does not increase model capacity like MoEs. For multilingual models, per-language pruning results in a partially shared network, fostering a learned distribution of language-specific and shared capacities. This approach has demonstrated effectiveness in multilingual ASR (Lu et al., 2022; Yang et al., 2023b) and MT (Lin et al., 2021; Koishekenov et al., 2023; He et al., 2023). The pruned sub-networks are shown to correspond to language relatedness (Lin et al., 2021; He et al., 2023), suggesting the validity of the learned sharing patterns.

4.2 Task-Invariant Modeling

As introduced in §4.1, task-specific modeling often helps to alleviate interference in supervised conditions. One the other hand, language- or modality-invariant representations are often beneficial in zero-shot or low-resource data conditions as well as retrieval tasks.

Aligning Speech and Text Representations Many prior works (Liu et al., 2020b; Dinh et al., 2022; Ye et al., 2022; Wang et al., 2022; Ouyang et al., 2023; Duquenne et al., 2022, 2023b) seek to align speech and text representations, such that semantically similar sentences are represented similarly irrespective of their source modality (speech or text). A semantically-aligned multimodal latent space has at least the following benefits: **1)** It could facilitate the plug-and-play use of pretrained unimodal models (Duquenne et al., 2023b; Yang et al., 2023a; Tsiamas et al., 2024). **2)** Text representations are often more robust than speech due to more training data, where cross-modal alignment can help distill from the resource-richer text-based task (Liu et al., 2020b; Tang et al., 2021). Indeed, multiple works showed that enforcing cross-modal universal representations improves low-resource (Dinh et al., 2022; Ouyang et al., 2023) and zero-shot ST (Wang et al., 2022; Duquenne et al., 2022; Tsiamas et al., 2024). A major challenge in the alignment of speech and text is the length mismatch, where speech sequences are often factors longer than text. Therefore some shrinking mechanism is often necessary, e.g., by CTC-based downsampling (Liu et al., 2020b; Gaido et al., 2021), CNN-based length adapters (Gállego et al., 2021), or learning to aggregate the representations from both modalities to fixed sizes (Duquenne et al., 2022, 2023b).

Language-Invariant Modeling Another form of task-invariant modeling is to enforce similar representations for different languages, thereby establishing a language-agnostic semantic latent space. In multilingual MT, such approaches (Arivazhagan et al., 2019a; Pham et al., 2019; Liu et al., 2021) are shown effective on zero-shot translation of new language pairs not included in training. Another application where language-invariant modeling helps is similarity search, where multilingual sentence encoders (Artetxe and Schwenk, 2019; Duquenne et al., 2023b) are used to mine parallel data (Schwenk et al., 2021; Duquenne et al., 2023a) for translation training corpora.

4.3 Outlook

Synergy between Languages and Modalities Multi-task learning inherently faces a tradeoff between knowledge sharing and negative interference. This becomes particularly challenging to investigate in recent LLM-based models capable of handling a wide range of modalities (§3.3). A deeper understanding of the interactions between tasks will enable targeted solutions to mitigate interference and promote knowledge sharing.

Efficiently Adding Languages and Modalities While in this paper we primarily focuses on the two modalities of speech and text, expanding modality coverage is a natural next step. For new modalities, vision offers significant potential for real-world applications, including sign language translation (Müller et al., 2023) and lip reading (Afouras et al., 2020). Recent foundation models like Audio-Visual BERT (Shi et al., 2022) demonstrates the feasibility of multimodal processing that incorporates vision. An additional interesting direction is the continual learning of trained ST systems. The key challenge would be to integrate additional languages or modalities into the model without compromising its existing performance.

5 Evaluation

The evaluation of multilingual and multimodal ST models relies on more resources than their bilingual and unimodal counterparts. Here we outline relevant developments in evaluation resources (§5.1) and metrics (§5.2).

5.1 Evaluation Resources

The evaluation of multilingual and multimodal ST models heavily rely on multiway parallel evaluation data, such as the FLoRes evaluation set (Goyal et al., 2022; NLLB Team et al., 2022) and its speech-based extension FLEURS (Conneau et al., 2022). Meanwhile, the increasing training data scale of large foundation models introduces significant risks of data contamination. A very alarming example is the inclusion of the FLoRes-200 evaluation data (NLLB Team et al., 2022) in the training corpus of BLOOMZ (Muennighoff et al., 2023), leading to highly inflated performance scores on this specific set (Zhu et al., 2023), and rendering downstream models based on BLOOMZ untestable by this benchmark. As any Internet content could be ingested in LLM training, developing new, unpublished test sets becomes even more essential. The recent initiative of test suites in WMT (Kocmi et al., 2023) as well as in IWSLT is a significant step forward in addressing this challenge.

5.2 Evaluation Metric

Speech-to-Text Evaluation While the translation community is gradually moving beyond BLEU (Papineni et al., 2002) to neural metrics better calibrated to human ratings (Freitag et al., 2022) such as COMET (Rei et al., 2020), language coverage remains a challenge for very low-resource languages. For instance, COMET supports 109 languages at the time of writing[4], whereas evaluation on extremely low-resource languages often rely on match-based scores like chrF (Popović, 2015). Noteworthy are initiatives like AfriCOMET (Wang et al., 2023a) to scale neural metrics to lower-resource languages.

Speech-to-Speech Evaluation For evaluation of speech-to-speech translation, the emergence of similar neural metrics like BLASER (Chen et al., 2023) as replacement of ASR-BLEU is also encouraging. For expressive speech, evaluation on voice preservation primarily has been relying on basic acoustic features such as the fundamental frequency (Akuzawa et al., 2018) or pitch and energy (Jeuris and Niehues, 2022), which do not account for speech naturalness. Recently, Seamless Communication et al. (2023a) propose AutoPCP and a rhythm evaluation toolkit to measure prosody.

5.3 Outlook

Reliably Measuring Progress As discussed in §5.1, the advent of LLM also introduces higher risks of test data leakage. Besides calling for more rigorous documentation by model developers and critical evaluation by practitioners applying these models to downstream tasks, this also presents a crucial research question: how to effectively create representative testing scenarios to properly measure progress. Recent targeted evaluation datasets (Salesky et al., 2023) and community-driven creation of test suites (Kocmi et al., 2023) are excellent examples of such efforts. Only with such robust testing methodologies can we ensure the generalizability of observed performance improvements.

6 Deployment

In this section, we review three aspects relevant to model deployment: compression and distillation for serving the models (§6.1), continual learning of new capabilities (§6.2), and inference-time customization (§6.3).

[4] https://github.com/Unbabel/COMET?tab=readme-ov-file#languages-covered

6.1 Compression and Distillation

While tight-integrated multi-task models offer the advantage of a compact and unified structure that simplifies deployment, the growing trend of incorporating large pretrained components can negate part of this initial benefit. Recent works in pruning massively multilingual MT models (Mohammadshahi et al., 2022; Koishekenov et al., 2023) show successful model compression while maintaining translation quality. Another related direction is to distill larger models into to smaller student models (NLLB Team et al., 2022).

6.2 Continual Learning

Given a deployed model, one use-case is to add more languages or modalities to the existing system. A trade-off here is maintaining performance on existing tasks and achieving optimal adaptation to the new task. While continual learning for adding languages has been explored in multilingual ASR (Li et al., 2022; Pham et al., 2023) and MT (Gu et al., 2022b; Sun et al., 2023; Liu et al., 2023b) its application in direct ST remains less investigated. Recent advancements in parameter-efficient finetuning approaches, such as LoRA (Hu et al., 2022), offer an alternative modular approach. By training only the newly added parameters, inherently, one can naturally decouple the new knowledge from previously acquired information.

6.3 Inference-Time Customization

Deployed models sometimes require customization to meet additional constraints specific to the use case. An example is real-time applications, such as simultaneous translation, where speech input needs to be decoded before it is complete. While other approaches involve designing separate models for online scenarios, repurposing offline models for online use cases (Liu et al., 2020a; Papi et al., 2022, 2023) has been shown to be a competitive alternative. This is particularly advantageous on foundation models (Papi et al., 2024) where retraining the model for specific use-cases is infeasible.

6.4 Outlook

Retrieval-Augmented Generation For both continual learning and inference-time customization as reviewed above, retrieval-augmented generation could be a promising approach. For instance, a separate data store could house continual learning data points, allowing for model updates without modifying the deployed model itself. Retrieval-augmented translation has demonstrated success in the text domain (Zhang et al., 2018; Xu et al., 2020; Cai et al., 2021; Hoang et al., 2023; Hao et al., 2023). In the context of ST, Du et al. (2022) explored kNN-MT (Khandelwal et al., 2021) for domain adaption using a joint speech and text input model with a text-based data store. However, it remains unclear how speech-based retrieval can benefit ST performance. Methods for efficiently incorporating speech data into the retrieval process is an interesting direction of future research.

7 Conclusion

In this paper, we presented a selection of recent advancements in multilingual and multimodal speech translation. We zoom into individual stages of the lifecycle of building a system: from determining model coverage and architecture, training procedures, to evaluation, and eventually deployment. This work is not an exhaustive survey, but rather a snapshot of ongoing developments related to multilingual and multimodal speech translation. We welcome the community's feedback on any relevant omitted works in the current version.

Acknowledgement

We thank the anonymous reviewers for constructive and insightful feedback. We also thank Annika Sauer for helpful pointers. This paper has received funding from the European Union's Horizon research and innovation programme under grant agreement No 101135798, project Meetween (My Personal AI Mediator for Virtual MEETtings BetWEEN People). Part of this work was supported by funding from the pilot program Core-Informatics of the Helmholtz Association (HGF).

References

Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. 2020. ASR is all you need: Cross-modal distillation for lip reading. In *2020 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2020, Barcelona, Spain, May 4-8, 2020*, pages 2143–2147. IEEE.

Kei Akuzawa, Yusuke Iwasawa, and Yutaka Matsuo. 2018. Expressive Speech Synthesis via Modeling Expressions with Variational Autoencoder. In *Proc. Interspeech 2018*, pages 3067–3071.

Antonios Anastasopoulos, Loïc Barrault, Luisa Bentivogli, Marcely Zanon Boito, Ondřej Bojar, Roldano

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

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 Hernández Ábrego, Junwhan Ahn, Jacob Austin, Paul Barham, Jan A. 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, Vladimir Feinberg, Fangxiaoyu Feng, Vlad Fienber, Markus Freitag, Xavier Garcia, Sebastian Gehrmann, Lucas Gonzalez, and et al. 2023. Palm 2 technical report. *CoRR*, abs/2305.10403.

Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Roee Aharoni, Melvin Johnson, and Wolfgang Macherey. 2019a. The missing ingredient in zero-shot neural machine translation. *CoRR*, abs/1903.07091.

Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Dmitry Lepikhin, Melvin Johnson, Maxim Krikun, Mia Xu Chen, Yuan Cao, George F. Foster, Colin Cherry, Wolfgang Macherey, Zhifeng Chen, and Yonghui Wu. 2019b. Massively multilingual neural machine translation in the wild: Findings and challenges. *CoRR*, abs/1907.05019.

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

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

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 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Bowen Baker, Otkrist Gupta, Nikhil Naik, and Ramesh Raskar. 2017. Designing neural network architectures using reinforcement learning. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

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.

C.F. Barnes, S.A. Rizvi, and N.M. Nasrabadi. 1996. Advances in residual vector quantization: a review. *IEEE Transactions on Image Processing*, 5(2):226–262.

Dan Berrebbi, Jiatong Shi, Brian Yan, Osbel López-Francisco, Jonathan D. Amith, and Shinji Watanabe. 2022. Combining spectral and self-supervised features for low resource speech recognition and translation. In *Interspeech 2022, 23rd Annual Conference of the International Speech Communication Association, Incheon, Korea, 18-22 September 2022*, pages 3533–3537. ISCA.

Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matthew Sharifi, Dominik Roblek, Olivier Teboul, David Grangier, Marco Tagliasacchi, and Neil Zeghidour. 2023a. Audiolm: A language modeling approach to audio generation. *IEEE ACM Trans. Audio Speech Lang. Process.*, 31:2523–2533.

Zalán Borsos, Matthew Sharifi, Damien Vincent, Eugene Kharitonov, Neil Zeghidour, and Marco Tagliasacchi. 2023b. Soundstorm: Efficient parallel audio generation. *CoRR*, abs/2305.09636.

Deng Cai, Yan Wang, Huayang Li, Wai Lam, and Lemao Liu. 2021. Neural machine translation with monolingual translation memory. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7307–7318, Online. Association for Computational Linguistics.

Rich Caruana. 1997. Multitask learning. *Mach. Learn.*, 28(1):41–75.

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. *Comput. Speech Lang.*, 66:101155.

Mingda Chen, Paul-Ambroise Duquenne, Pierre Andrews, Justine Kao, Alexandre Mourachko, Holger Schwenk, and Marta R. Costa-jussà. 2023. BLASER: A text-free speech-to-speech translation evaluation metric. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9064–9079, Toronto, Canada. Association for Computational Linguistics.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei. 2022a. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE J. Sel. Top. Signal Process.*, 16(6):1505–1518.

Zhehuai Chen, Yu Zhang, Andrew Rosenberg, Bhuvana Ramabhadran, Pedro J. Moreno, Ankur Bapna, and Heiga Zen. 2022b. MAESTRO: Matched Speech Text Representations through Modality Matching. In *Proc. Interspeech 2022*, pages 4093–4097.

Alexandra Chronopoulou, Dario Stojanovski, and Alexander Fraser. 2023. Language-family adapters for low-resource multilingual neural machine translation. In *Proceedings of the The Sixth Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2023)*, pages 59–72, Dubrovnik, Croatia. Association for Computational Linguistics.

Yu-An Chung, Yu Zhang, Wei Han, Chung-Cheng Chiu, James Qin, Ruoming Pang, and Yonghui Wu. 2021a. w2v-bert: Combining contrastive learning and masked language modeling for self-supervised speech pre-training. In *IEEE Automatic Speech Recognition and Understanding Workshop, ASRU 2021, Cartagena, Colombia, December 13-17, 2021*, pages 244–250. IEEE.

Yu-An Chung, Yu Zhang, Wei Han, Chung-Cheng Chiu, James Qin, Ruoming Pang, and Yonghui Wu. 2021b. w2v-bert: Combining contrastive learning and masked language modeling for self-supervised speech pre-training. In *IEEE Automatic Speech Recognition and Understanding Workshop, ASRU 2021, Cartagena, Colombia, December 13-17, 2021*, pages 244–250. IEEE.

Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. 2022. FLEURS: few-shot learning evaluation of universal representations of speech. In *IEEE Spoken Language Technology Workshop, SLT 2022, Doha, Qatar, January 9-12, 2023*, pages 798–805. IEEE.

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

Mattia Antonino Di Gangi, Matteo Negri, and Marco Turchi. 2019. One-to-many multilingual end-to-end speech translation. In *IEEE Automatic Speech Recognition and Understanding Workshop, ASRU 2019, Singapore, December 14-18, 2019*, pages 585–592. IEEE.

Tu Anh Dinh, Danni Liu, and Jan Niehues. 2022. Tackling data scarcity in speech translation using zero-shot multilingual machine translation techniques. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2022, Virtual and Singapore, 23-27 May 2022*, pages 6222–6226. IEEE.

Daxiang Dong, Hua Wu, Wei He, Dianhai Yu, and Haifeng Wang. 2015. Multi-task learning for multiple language translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1723–1732, Beijing, China. Association for Computational Linguistics.

Qianqian Dong, Zhiying Huang, Qiao Tian, Chen Xu, Tom Ko, yunlong zhao, Siyuan Feng, Tang Li, Kexin Wang, Xuxin Cheng, Fengpeng Yue, Ye Bai, Xi Chen, Lu Lu, Zejun MA, Yuping Wang, Mingxuan Wang, and Yuxuan Wang. 2024. Polyvoice: Language models for speech to speech translation. In *The Twelfth International Conference on Learning Representations*.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Yichao Du, Weizhi Wang, Zhirui Zhang, Boxing Chen, Tong Xu, Jun Xie, and Enhong Chen. 2022. Non-parametric domain adaptation for end-to-end speech translation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 306–320, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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

Paul-Ambroise Duquenne, Hongyu Gong, Benoît Sagot, and Holger Schwenk. 2022. T-modules: Translation modules for zero-shot cross-modal machine translation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5794–5806, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Paul-Ambroise Duquenne, Hongyu Gong, and Holger Schwenk. 2021. Multimodal and multilingual embeddings for large-scale speech mining. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 15748–15761.

Paul-Ambroise Duquenne, Holger Schwenk, and Benoît Sagot. 2023b. SONAR: sentence-level multimodal and language-agnostic representations. *CoRR*, abs/2308.11466.

Abteen Ebrahimi, Manuel Mager, Adam Wiemerslage, Pavel Denisov, Arturo Oncevay, Danni Liu, Sai Koneru, Enes Yavuz Ugan, Zhaolin Li, Jan Niehues, Monica Romero, Ivan G Torre, Tanel Alumäe, Jiaming Kong, Sergey Polezhaev, Yury Belousov, Wei-Rui Chen, Peter Sullivan, Ife Adebara, Bashar Talafha, Alcides Alcoba Inciarte, Muhammad Abdul-Mageed, Luis Chiruzzo, Rolando Coto-Solano, Hilaria Cruz, Sofía Flores-Solórzano, Aldo Andrés Alvarez López, Iván V. Meza-Ruíz, John E. Ortega, Alexis Palmer, Rodolfo Zevallos, Kristine Stenzel, Thang Vu, and Katharina Kann. 2021. Findings of the second americasnlp competition on speech-to-text translation. In *NeurIPS 2022 Competition Track, November 28 - December 9, 2022, Online*, volume 220 of *Proceedings of Machine Learning Research*, pages 217–232. PMLR.

David Eigen, Marc'Aurelio Ranzato, and Ilya Sutskever. 2014. Learning factored representations in a deep mixture of experts. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Workshop Track Proceedings*.

Carlos Escolano, Marta R. Costa-jussà, José A. R. Fonollosa, and Carlos Segura. 2021. Enabling zero-shot multilingual spoken language translation with language-specific encoders and decoders. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 694–701.

Patrick Fernandes, Behrooz Ghorbani, Xavier Garcia, Markus Freitag, and Orhan Firat. 2023. Scaling laws for multilingual neural machine translation. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 10053–10071. PMLR.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins. 2022. Results of WMT22 metrics shared task: Stop using BLEU – neural metrics are better and more robust. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 46–68, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Marco Gaido, Mauro Cettolo, Matteo Negri, and Marco Turchi. 2021. CTC-based compression for direct speech translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 690–696, Online. Association for Computational Linguistics.

Marco Gaido, Sara Papi, Dennis Fucci, Giuseppe Fiameni, Matteo Negri, and Marco Turchi. 2022. Efficient yet competitive speech translation: FBK@IWSLT2022. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, pages 177–189, Dublin, Ireland (in-person and online). Association for Computational Linguistics.

Marco Gaido, Sara Papi, Matteo Negri, and Luisa Bentivogli. 2024. Speech translation with speech foundation models and large language models: What is there and what is missing? *CoRR*, abs/2402.12025.

Gerard I. Gállego, Ioannis Tsiamas, Carlos Escolano, José A. R. Fonollosa, and Marta R. Costa-jussà. 2021. End-to-end speech translation with pre-trained models and adapters: UPC at IWSLT 2021. In *Proceedings of the 18th International Conference on Spoken Language Translation (IWSLT 2021)*, pages 110–119, Bangkok, Thailand (online). Association for Computational Linguistics.

Neeraj Gaur, Brian Farris, Parisa Haghani, Isabel Leal, Pedro J. Moreno, Manasa Prasad, Bhuvana Ramabhadran, and Yun Zhu. 2021. Mixture of informed experts for multilingual speech recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2021, Toronto, ON, Canada, June 6-11, 2021*, pages 6234–6238. IEEE.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. 2023. Imagebind one embedding space to bind them all. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 15180–15190. IEEE.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Trans. Assoc. Comput. Linguistics*, 10:522–538.

Albert Gu and Tri Dao. 2023. Mamba: Linear-time sequence modeling with selective state spaces. *CoRR*, abs/2312.00752.

Albert Gu, Karan Goel, and Christopher Ré. 2022a. Efficiently modeling long sequences with structured state spaces. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Shuhao Gu, Bojie Hu, and Yang Feng. 2022b. Continual learning of neural machine translation within low forgetting risk regions. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1707–1718, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Hongkun Hao, Guoping Huang, Lemao Liu, Zhirui Zhang, Shuming Shi, and Rui Wang. 2023. Rethinking translation memory augmented neural machine translation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2589–2605, Toronto, Canada. Association for Computational Linguistics.

Dan He, Minh-Quang Pham, Thanh-Le Ha, and Marco Turchi. 2023. Gradient-based gradual pruning for language-specific multilingual neural machine translation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 654–670, Singapore. Association for Computational Linguistics.

Cuong Hoang, Devendra Sachan, Prashant Mathur, Brian Thompson, and Marcello Federico. 2023. Improving retrieval augmented neural machine translation by controlling source and fuzzy-match interactions. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 289–295, Dubrovnik, Croatia. Association for Computational Linguistics.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 2790–2799. PMLR.

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

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Ke Hu, Bo Li, Tara Sainath, Yu Zhang, and Françoise Beaufays. 2023. Mixture-of-Expert Conformer for Streaming Multilingual ASR. In *Proc. INTERSPEECH 2023*, pages 3327–3331.

Wuwei Huang, Mengge Liu, Xiang Li, Yanzhi Tian, Fengyu Yang, Wen Zhang, Jian Luan, Bin Wang, Yuhang Guo, and Jinsong Su. 2023. The xiaomi AI lab's speech translation systems for IWSLT 2023 offline task, simultaneous task and speech-to-speech task. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pages 411–419, Toronto, Canada (in-person and online). Association for Computational Linguistics.

Hirofumi Inaguma, Kevin Duh, Tatsuya Kawahara, and Shinji Watanabe. 2019. Multilingual end-to-end speech translation. In *IEEE Automatic Speech Recognition and Understanding Workshop, ASRU 2019, Singapore, December 14-18, 2019*, pages 570–577. IEEE.

Hirofumi Inaguma, Sravya Popuri, Ilia Kulikov, Peng-Jen Chen, Changhan Wang, Yu-An Chung, Yun Tang, Ann Lee, Shinji Watanabe, and Juan Pino. 2023. UnitY: Two-pass direct speech-to-speech translation with discrete units. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15655–15680, Toronto, Canada. Association for Computational Linguistics.

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 *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8229–8233.

Pedro Jeuris and Jan Niehues. 2022. LibriS2S: A German-English speech-to-speech translation corpus. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 928–935, Marseille, France. European Language Resources Association.

Ye Jia, Melvin Johnson, Wolfgang Macherey, Ron J. Weiss, Yuan Cao, Chung-Cheng Chiu, Naveen Ari, Stella Laurenzo, and Yonghui Wu. 2019a. Leveraging weakly supervised data to improve end-to-end speech-to-text translation. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7180–7184.

Ye Jia, Michelle Tadmor Ramanovich, Tal Remez, and Roi Pomerantz. 2022a. Translatotron 2: High-quality direct speech-to-speech translation with voice preservation. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 10120–10134. PMLR.

Ye Jia, Michelle Tadmor Ramanovich, Quan Wang, and Heiga Zen. 2022b. CVSS corpus and massively multilingual speech-to-speech translation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6691–6703, Marseille, France. European Language Resources Association.

Ye Jia, Ron J. Weiss, Fadi Biadsy, Wolfgang Macherey, Melvin Johnson, Zhifeng Chen, and Yonghui Wu. 2019b. Direct Speech-to-Speech Translation with a Sequence-to-Sequence Model. In *Proc. Interspeech 2019*, pages 1123–1127.

Urvashi Khandelwal, Angela Fan, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2021. Nearest neighbor machine translation. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Yuka Ko, Ryo Fukuda, Yuta Nishikawa, Yasumasa Kano, Katsuhito Sudoh, and Satoshi Nakamura. 2023. Tagged end-to-end simultaneous speech translation training using simultaneous interpretation data. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pages 363–375, Toronto, Canada (in-person and online). Association for Computational Linguistics.

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. 2023. 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*, pages 1–42, Singapore. Association for Computational Linguistics.

Yeskendir Koishekenov, Alexandre Berard, and Vassilina Nikoulina. 2023. Memory-efficient NLLB-200: Language-specific expert pruning of a massively multilingual machine translation model. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3567–3585, Toronto, Canada. Association for Computational Linguistics.

Yoohwan Kwon and Soo-Whan Chung. 2023. Mole : Mixture of language experts for multi-lingual automatic speech recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 2023, Rhodes Island, Greece, June 4-10, 2023*, pages 1–5. IEEE.

Chenyang Le, Yao Qian, Long Zhou, Shujie Liu, Yanmin Qian, Michael Zeng, and Xuedong Huang. 2023. Comsl: A composite speech-language model for end-to-end speech-to-text translation. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Hang Le, Juan Pino, Changhan Wang, Jiatao Gu, Didier Schwab, and Laurent Besacier. 2021. Lightweight adapter tuning for multilingual speech translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 817–824, Online. Association for Computational Linguistics.

Ann Lee, Peng-Jen Chen, Changhan Wang, Jiatao Gu, Sravya Popuri, Xutai Ma, Adam Polyak, Yossi Adi, Qing He, Yun Tang, Juan Pino, and Wei-Ning Hsu. 2022. Direct speech-to-speech translation with discrete units. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3327–3339, Dublin, Ireland. Association for Computational Linguistics.

Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. 2021. Gshard: Scaling giant models with conditional computation and automatic sharding. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Bo Li, Ruoming Pang, Yu Zhang, Tara N. Sainath, Trevor Strohman, Parisa Haghani, Yun Zhu, Brian Farris, Neeraj Gaur, and Manasa Prasad. 2022. Massively multilingual asr: A lifelong learning solution. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6397–6401.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 2023. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 19730–19742. PMLR.

Xian Li, Changhan Wang, Yun Tang, Chau Tran, Yuqing Tang, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. 2021. Multilingual speech translation from efficient finetuning of pretrained models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 827–838, Online. Association for Computational Linguistics.

Zehui Lin, Liwei Wu, Mingxuan Wang, and Lei Li. 2021. Learning language specific sub-network for multilingual machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 293–305, Online. Association for Computational Linguistics.

Danni Liu, Jan Niehues, James Cross, Francisco Guzmán, and Xian Li. 2021. Improving zero-shot translation by disentangling positional information. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1259–1273, Online. Association for Computational Linguistics.

Danni Liu, Gerasimos Spanakis, and Jan Niehues. 2020a. Low-Latency Sequence-to-Sequence Speech Recognition and Translation by Partial Hypothesis Selection. In *Proc. Interspeech 2020*, pages 3620–3624.

Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo P. Mandic, Wenwu Wang, and Mark D. Plumbley. 2023a. Audioldm: Text-to-audio generation with latent diffusion models. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 21450–21474. PMLR.

Junpeng Liu, Kaiyu Huang, Hao Yu, Jiuyi Li, Jinsong Su, and Degen Huang. 2023b. Continual learning for multilingual neural machine translation via dual importance-based model division. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12011–12027, Singapore. Association for Computational Linguistics.

Yuchen Liu, Junnan Zhu, Jiajun Zhang, and Chengqing Zong. 2020b. Bridging the modality gap for speech-to-text translation. *CoRR*, abs/2010.14920.

Yizhou Lu, Mingkun Huang, Xinghua Qu, Pengfei Wei, and Zejun Ma. 2022. Language adaptive cross-lingual speech representation learning with sparse sharing sub-networks. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2022, Virtual and Singapore, 23-27 May 2022*, pages 6882–6886. IEEE.

Jean Maillard, Cynthia Gao, Elahe Kalbassi, Kaushik Ram Sadagopan, Vedanuj Goswami, Philipp Koehn, Angela Fan, and Francisco Guzman. 2023. Small data, big impact: Leveraging minimal data for effective machine translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2740–2756, Toronto, Canada. Association for Computational Linguistics.

Alireza Mohammadshahi, Vassilina Nikoulina, Alexandre Berard, Caroline Brun, James Henderson, and Laurent Besacier. 2022. SMaLL-100: Introducing shallow multilingual machine translation model for low-resource languages. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8348–8359, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. Crosslingual generalization through multitask finetuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15991–16111, Toronto, Canada. Association for Computational Linguistics.

Mathias Müller, Malihe Alikhani, Eleftherios Avramidis, Richard Bowden, Annelies Braffort, Necati Cihan Camgöz, Sarah Ebling, Cristina España-Bonet, Anne Göhring, Roman Grundkiewicz, Mert Inan, Zifan Jiang, Oscar Koller, Amit Moryossef, Annette Rios, Dimitar Shterionov, Sandra Sidler-Miserez, Katja Tissi, and Davy Van Landuyt. 2023. Findings of the second WMT shared task on sign language translation (WMT-SLT23). In *Proceedings of the Eighth Conference on Machine Translation*, pages 68–94, Singapore. Association for Computational Linguistics.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Y. Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loïc Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. *CoRR*, abs/2207.04672.

Siqi Ouyang, Rong Ye, and Lei Li. 2022. On the impact of noises in crowd-sourced data for speech translation. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, pages 92–97, Dublin, Ireland (in-person and online). Association for Computational Linguistics.

Siqi Ouyang, Rong Ye, and Lei Li. 2023. WACO: Word-aligned contrastive learning for speech translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3891–3907, Toronto, Canada. Association for Computational Linguistics.

Sara Papi, Marco Gaido, Matteo Negri, and Luisa Bentivogli. 2024. Simulseamless: Fbk at iwslt 2024 simultaneous speech translation. *Preprint*, arXiv:2406.14177.

Sara Papi, Marco Gaido, Matteo Negri, and Marco Turchi. 2022. Does simultaneous speech translation need simultaneous models? In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 141–153, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Sara Papi, Matteo Negri, and Marco Turchi. 2023. Attention as a guide for simultaneous speech translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13340–13356, Toronto, Canada. Association for Computational Linguistics.

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

Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. 2021. AdapterFusion: Non-destructive task composition for transfer learning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 487–503, Online. Association for Computational Linguistics.

Ngoc-Quan Pham, Tuan-Nam Nguyen, Thai Binh Nguyen, Danni Liu, Carlos Mullov, Jan Niehues, and Alexander Waibel. 2022. Effective combination of pretrained models - kit@iwslt2022. In *Proceedings of the 19th International Conference on Spoken Language Translation, IWSLT@ACL 2022, Dublin, Ireland (in-person and online), May 26-27, 2022*, pages 190–197. Association for Computational Linguistics.

Ngoc-Quan Pham, Tuan-Nam Nguyen, Sebastian Stüker, and Alex Waibel. 2021. Efficient weight factorization for multilingual speech recognition. In *Interspeech 2021, 22nd Annual Conference of the International Speech Communication Association, Brno, Czechia, 30 August - 3 September 2021*, pages 2421–2425. ISCA.

Ngoc-Quan Pham, Jan Niehues, Thanh-Le Ha, and Alexander Waibel. 2019. Improving zero-shot translation with language-independent constraints. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 13–23, Florence, Italy. Association for Computational Linguistics.

Ngoc-Quan Pham, Jan Niehues, and Alex Waibel. 2023. Towards continually learning new languages. In *Proc. INTERSPEECH 2023*, pages 3262–3266.

Jerin Philip, Alexandre Berard, Matthias Gallé, and Laurent Besacier. 2020. Monolingual adapters for zero-shot neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4465–4470, Online. Association for Computational Linguistics.

Juan Pino, Qiantong Xu, Xutai Ma, Mohammad Javad Dousti, and Yun Tang. 2020. Self-Training for End-to-End Speech Translation. In *Proc. Interspeech 2020*, pages 1476–1480.

Telmo Pires, Robin Schmidt, Yi-Hsiu Liao, and Stephan Peitz. 2023. Learning language-specific layers for multilingual machine translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14767–14783, Toronto, Canada. Association for Computational Linguistics.

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

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR.

Sylvestre-Alvise Rebuffi, Hakan Bilen, and Andrea Vedaldi. 2017. Learning multiple visual domains with residual adapters. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 506–516.

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

Paul K. Rubenstein, Chulayuth Asawaroengchai, Duc Dung Nguyen, Ankur Bapna, Zalán Borsos, Félix de Chaumont Quitry, Peter Chen, Dalia El Badawy, Wei Han, Eugene Kharitonov, Hannah Muckenhirn, Dirk Padfield, James Qin, Danny Rozenberg, Tara N. Sainath, Johan Schalkwyk, Matthew Sharifi, Michelle Tadmor Ramanovich, Marco Tagliasacchi, Alexandru Tudor, Mihajlo Velimirovic, Damien Vincent, Jiahui Yu, Yongqiang Wang, Vicky Zayats, Neil Zeghidour, Yu Zhang, Zhishuai Zhang, Lukas Zilka, and Christian Havnø Frank. 2023. Audiopalm: A large language model that can speak and listen. *CoRR*, abs/2306.12925.

Yusuke Sakai, Mana Makinae, Hidetaka Kamigaito, and Taro Watanabe. 2024. Simultaneous interpretation corpus construction by large language models in distant language pair. *CoRR*, abs/2404.12299.

Elizabeth Salesky, Kareem Darwish, Mohamed Al-Badrashiny, Mona Diab, and Jan Niehues. 2023. Evaluating multilingual speech translation under realistic conditions with resegmentation and terminology. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pages 62–78, Toronto, Canada (in-person and online). Association for Computational Linguistics.

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

Holger Schwenk. 2018. Filtering and mining parallel data in a joint multilingual space. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 228–234, Melbourne, Australia. Association for Computational Linguistics.

Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, Armand Joulin, and Angela Fan. 2021. CCMatrix: Mining billions of high-quality parallel sentences on the web. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6490–6500, Online. Association for Computational Linguistics.

Seamless Communication, Loïc Barrault, Yu-An Chung, Mariano Coria Meglioli, David Dale, Ning Dong, Mark Duppenthaler, Paul-Ambroise Duquenne, Brian Ellis, Hady Elsahar, Justin Haaheim, John Hoffman, Min-Jae Hwang, Hirofumi Inaguma, Christopher Klaiber, Ilia Kulikov, Pengwei Li, Daniel Licht, Jean Maillard, Ruslan Mavlyutov, Alice Rakotoarison, Kaushik Ram Sadagopan, Abinesh Ramakrishnan, Tuan Tran, Guillaume Wenzek, Yilin Yang, Ethan Ye, Ivan Evtimov, Pierre Fernandez, Cynthia Gao, Prangthip Hansanti, Elahe Kalbassi, Amanda Kallet, Artyom Kozhevnikov, Gabriel Mejia Gonzalez, Robin San Roman, Christophe Touret, Corinne Wong, Carleigh Wood, Bokai Yu, Pierre Andrews, Can Balioglu, Peng-Jen Chen, Marta R. Costa-jussà, Maha Elbayad, Hongyu Gong, Francisco Guzmán, Kevin Heffernan, Somya Jain, Justine Kao, Ann Lee, Xutai Ma, Alexandre Mourachko, Benjamin Peloquin, Juan Pino, Sravya Popuri, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, Anna Y. Sun, Paden Tomasello, Changhan Wang, Jeff Wang, Skyler Wang, and Mary Williamson. 2023a. Seamless: Multilingual expressive and streaming speech translation. *CoRR*, abs/2312.05187.

Seamless Communication, Loïc Barrault, Yu-An Chung, Mariano Coria Meglioli, David Dale, Ning Dong, Paul-Ambroise Duquenne, Hady Elsahar, Hongyu Gong, Kevin Heffernan, John Hoffman, Christopher Klaiber, Pengwei Li, Daniel Licht, Jean Maillard, Alice Rakotoarison, Kaushik Ram Sadagopan, Guillaume Wenzek, Ethan Ye, Bapi Akula, Peng-Jen Chen, Naji El Hachem, Brian Ellis, Gabriel Mejia Gonzalez, Justin Haaheim, Prangthip Hansanti, Russ Howes, Bernie Huang, Min-Jae Hwang, Hirofumi Inaguma, Somya Jain, Elahe Kalbassi, Amanda Kallet, Ilia Kulikov, Janice Lam, Daniel Li, Xutai Ma, Ruslan Mavlyutov, Benjamin Peloquin, Mohamed Ramadan, Abinesh Ramakrishnan, Anna Y. Sun, Kevin Tran, Tuan Tran, Igor Tufanov, Vish Vogeti, Carleigh Wood, Yilin Yang, Bokai Yu, Pierre Andrews, Can Balioglu, Marta R. Costa-jussà, Onur Celebi, Maha Elbayad, Cynthia Gao, Francisco Guzmán, Justine Kao, Ann Lee, Alexandre Mourachko, Juan Pino, Sravya Popuri, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, Paden Tomasello, Changhan Wang, Jeff Wang, and Skyler Wang. 2023b. Seamlessm4t-massively multilingual & multimodal machine translation. *CoRR*, abs/2308.11596.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc V. Le, Geoffrey E. Hinton, and Jeff Dean. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Bowen Shi, Wei-Ning Hsu, Kushal Lakhotia, and Abdelrahman Mohamed. 2022. Learning audio-visual speech representation by masked multimodal cluster prediction. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Simeng Sun, Maha Elbayad, Anna Sun, and James Cross. 2023. Efficiently upgrading multilingual machine translation models to support more languages. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1513–1527, Dubrovnik, Croatia. Association for Computational Linguistics.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun MA, and Chao Zhang. 2024. SALMONN: Towards generic hearing abilities for large language models. In *The Twelfth International Conference on Learning Representations*.

Yun Tang, Juan Pino, Xian Li, Changhan Wang, and Dmitriy Genzel. 2021. Improving speech translation by understanding and learning from the auxiliary text translation task. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4252–4261, Online. Association for Computational Linguistics.

Zineng Tang, Ziyi Yang, Mahmoud Khademi, Yang Liu, Chenguang Zhu, and Mohit Bansal. 2023. Codi-2: In-context, interleaved, and interactive any-to-any generation. *CoRR*, abs/2311.18775.

Andros Tjandra, Sakriani Sakti, and Satoshi Nakamura. 2019. Speech-to-speech translation between untranscribed unknown languages. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 593–600.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models. *CoRR*, 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, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

Ioannis Tsiamas, Gerard I. Gállego, José A. R. Fonollosa, and Marta R. Costa-jussà. 2024. Pushing the limits of zero-shot end-to-end speech translation. *CoRR*, abs/2402.10422.

Changhan Wang, Morgane Riviere, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. 2021a. VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 993–1003, Online. Association for Computational Linguistics.

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

Chen Wang, Yuchen Liu, Boxing Chen, Jiajun Zhang, Wei Luo, Zhongqiang Huang, and Chengqing Zong. 2022. Discrete cross-modal alignment enables zero-shot speech translation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5291–5302, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jiayi Wang, David Ifeoluwa Adelani, Sweta Agrawal, Ricardo Rei, Eleftheria Briakou, Marine Carpuat, Marek Masiak, Xuanli He, Sofia Bourhim, Andiswa Bukula, Muhidin Mohamed, Temitayo Olatoye, Hamam Mokayed, Christine Mwase, Wangui Kimotho, Foutse Yuehgoh, Aremu Anuoluwapo, Jessica Ojo, Shamsuddeen Hassan Muhammad, Salomey Osei, Abdul-Hakeem Omotayo, Chiamaka Chukwuneke, Perez Ogayo, Oumaima Hourrane, Salma El Anigri, Lolwethu Ndolela, Thabiso Mangwana, Shafie Abdi Mohamed, Ayinde Hassan, Oluwabusayo Olufunke Awoyomi, Lama Alkhaled, Sana Sabah Al-Azzawi, Naome A. Etori, Millicent Ochieng, Clemencia Siro, Samuel Njoroge, Eric Muchiri, Wangari Kimotho, Lyse Naomi Wamba Momo, Daud Abolade, Simbiat Ajao, Tosin P. Adewumi, Iyanuoluwa Shode, Ricky Macharm, Ruqayya Nasir Iro, Saheed S. Abdullahi, Stephen E. Moore, Bernard Opoku, Zainab Akinjobi, Afolabi Abeeb, Nnaemeka C. Obiefuna, Onyekachi Raphael Ogbu, Sam Brian, Verrah Akinyi Otiende, Chinedu Emmanuel Mbonu, Sakayo Toadoum Sari, and Pontus Stenetorp. 2023a. Afrimte and africomet: Empowering COMET to embrace under-resourced african languages. *CoRR*, abs/2311.09828.

Wenxuan Wang, Guodong Ma, Yuke Li, and Binbin Du. 2023b. Language-Routing Mixture of Experts for Multilingual and Code-Switching Speech Recognition. In *Proc. INTERSPEECH 2023*, pages 1389–1393.

Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. 2023. Next-gpt: Any-to-any multimodal LLM. *CoRR*, abs/2309.05519.

Haoran Xu, Weiting Tan, Shuyue Li, Yunmo Chen, Benjamin Van Durme, Philipp Koehn, and Kenton Murray. 2023. Condensing multilingual knowledge with lightweight language-specific modules. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1575–1587, Singapore. Association for Computational Linguistics.

Jitao Xu, Josep Crego, and Jean Senellart. 2020. Boosting neural machine translation with similar translations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1580–1590, Online. Association for Computational Linguistics.

Jichen Yang, Kai Fan, Minpeng Liao, Boxing Chen, and Zhongqiang Huang. 2023a. Towards zero-shot learning for end-to-end cross-modal translation models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13078–13087, Singapore. Association for Computational Linguistics.

Mu Yang, Andros Tjandra, Chunxi Liu, David Zhang, Duc Le, and Ozlem Kalinli. 2023b. Learning ASR pathways: A sparse multilingual ASR model. In *IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 2023, Rhodes Island, Greece, June 4-10, 2023*, pages 1–5. IEEE.

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

Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi. 2022. Soundstream: An end-to-end neural audio codec. *IEEE ACM Trans. Audio Speech Lang. Process.*, 30:495–507.

Jun Zhan, Junqi Dai, Jiasheng Ye, Yunhua Zhou, Dong Zhang, Zhigeng Liu, Xin Zhang, Ruibin Yuan, Ge Zhang, Linyang Li, Hang Yan, Jie Fu, Tao Gui, Tianxiang Sun, Yugang Jiang, and Xipeng Qiu. 2024. Anygpt: Unified multimodal LLM with discrete sequence modeling. *CoRR*, abs/2402.12226.

Biao Zhang, Ankur Bapna, Rico Sennrich, and Orhan Firat. 2021a. Share or not? learning to schedule language-specific capacity for multilingual translation. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. 2020. Improving massively multilingual neural machine translation and zero-shot translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1628–1639, Online. Association for Computational Linguistics.

Chen Zhang, Xu Tan, Yi Ren, Tao Qin, Kejun Zhang, and Tie-Yan Liu. 2021b. Uwspeech: Speech to speech translation for unwritten languages. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 14319–14327. AAAI Press.

Jingyi Zhang, Masao Utiyama, Eiichro Sumita, Graham Neubig, and Satoshi Nakamura. 2018. Guiding neural machine translation with retrieved translation pieces. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1325–1335, New Orleans, Louisiana. Association for Computational Linguistics.

Xin Zhang, Dong Zhang, Shimin Li, Yaqian Zhou, and Xipeng Qiu. 2023a. Speechtokenizer: Unified speech tokenizer for speech large language models. *CoRR*, abs/2308.16692.

Yu Zhang, Wei Han, James Qin, Yongqiang Wang, Ankur Bapna, Zhehuai Chen, Nanxin Chen, Bo Li, Vera Axelrod, Gary Wang, Zhong Meng, Ke Hu, Andrew Rosenberg, Rohit Prabhavalkar, Daniel S. Park, Parisa Haghani, Jason Riesa, Ginger Perng, Hagen Soltau, Trevor Strohman, Bhuvana Ramabhadran, Tara N. Sainath, Pedro J. Moreno, Chung-Cheng Chiu, Johan Schalkwyk, Françoise Beaufays, and Yonghui Wu. 2023b. Google USM: scaling automatic speech recognition beyond 100 languages. *CoRR*, abs/2303.01037.

Jinming Zhao, Hao Yang, Gholamreza Haffari, and Ehsan Shareghi. 2022. M-Adapter: Modality Adaptation for End-to-End Speech-to-Text Translation. In *Proc. Interspeech 2022*, pages 111–115.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Lingpeng Kong, Jiajun Chen, Lei Li, and Shujian Huang. 2023. Multilingual machine translation with large language models: Empirical results and analysis. *CoRR*, abs/2304.04675.

Yun Zhu, Parisa Haghani, Anshuman Tripathi, Bhuvana Ramabhadran, Brian Farris, Hainan Xu, Han Lu, Hasim Sak, Isabel Leal, Neeraj Gaur, Pedro J. Moreno, and Qian Zhang. 2020. Multilingual Speech Recognition with Self-Attention Structured Parameterization. In *Proc. Interspeech 2020*, pages 4741–4745.