From TOWER to SPIRE: Adding the Speech Modality to a Translation-Specialist LLM

Kshitij Ambilduke[*†1], **Ben Peters**[*‡2], **Sonal Sannigrahi**[*3,4], **Anil Keshwani**[†5], **Tsz Kin Lam**[6], **Bruno Martins**[2,4], **André F.T. Martins**[¶§3,4,7,8], **Marcely Zanon Boito**[¶9]

[1]ENS Paris-Saclay [2]INESC-ID [3]Instituto de Telecomunicações
[4]Instituto Superior Técnico, Universidade de Lisboa [5]Sapienza University of Rome
[6]University of Edinburgh [7]ELLIS Unit Lisbon [8]TransPerfect [9]NAVER LABS Europe

Correspondence: benzurdopeters@gmail.com

Abstract

We introduce SPIRE, a speech-augmented language model (LM) capable of both translating and transcribing speech input from English into 10 other languages as well as translating text input in both language directions. SPIRE integrates the speech modality into an existing multilingual LM via speech discretization and continued pre-training using only 42.5K hours of speech. In particular, we adopt the pretraining framework of multilingual LMs and treat discretized speech input as an additional *translation language*. This approach not only equips the model with speech capabilities, but also preserves its strong text-based performance. We achieve this using significantly less data than existing speech LMs, demonstrating that discretized speech input integration as an additional language is feasible during LM adaptation. We make our code and models available to the community.

1 Introduction

Large language models (LLMs) have demonstrated remarkable success on various text-based natural language processing tasks (Achiam et al., 2023; Touvron et al., 2023; Yang et al., 2024; Alves et al., 2024; Martins et al., 2024), motivating research into extending them to other modalities. This has led to the development of multimodal LMs capable of processing speech, audio, images, and video (Gemini Team et al., 2023; Driess et al., 2023; Rubenstein et al., 2023; Liu et al., 2023; Tang et al., 2024; Défossez et al., 2024; Hu et al., 2024; Huang et al., 2024; Nguyen et al., 2025). However, the integration of new modalities often comes at the cost of existing capabilities (Zhai et al., 2024).

For speech-LLM integration, a simple approach is to link the output of an automatic speech recognition (ASR) system to a text-only LLM (Huang et al., 2024). This solution, however, is prone to error propagation and depends largely on individual model quality. More popular are solutions that investigate equipping LLMs natively with speech processing capabilities through modality projection (Shu et al., 2023; Radhakrishnan et al., 2023; Wu et al., 2023a; Tang et al., 2024; Xue et al., 2024; Hu et al., 2024). Typically, a speech foundation model generates speech representations that are mapped to the embedding space of the LLM, following which the model is then fine-tuned along with a projector on speech-to-text tasks to equip the LLM with speech processing capabilities. In this setting, key challenges include prompt overfitting and high training costs, as tuning these multimodal LLMs requires the adaptation of the speech projector module on vast amounts of raw speech data (Tang et al., 2024; Hu et al., 2024).

An alternative approach for integrating speech into a text-only LLM is to use *speech discretization*, where continuous speech features are transformed prior to training into sequences of "discrete speech units" (DSUs), which can be processed similarly to text (Chou et al., 2023a; Zhang et al., 2023; Rubenstein et al., 2023; Chang et al., 2024; Défossez et al., 2024; Trinh et al., 2024; Maiti et al., 2024; Nguyen et al., 2025). This approach simplifies training by eliminating the need for additional parameters beyond extended embedding matrices. Finally, while both projector-based and discretization-based solutions have shown promising results on text-to-speech and speech-to-text tasks, their development has prioritized speech-centric tasks at the expense of textual performance. Furthermore, limited research has focused on integrating speech while preserving the LLM's original capabilities in textual tasks (Chou et al., 2023b; Huang et al., 2024).

In this work we present SPIRE, a speech-

[*]Equal contribution. [¶]Equal contribution.
[†]Work begun during an internship at Instituto de Telecomunicações.
[‡]Work done while at Instituto de Telecomunicações.
[§]Work done while at Unbabel.

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 19658–19673
November 4-9, 2025

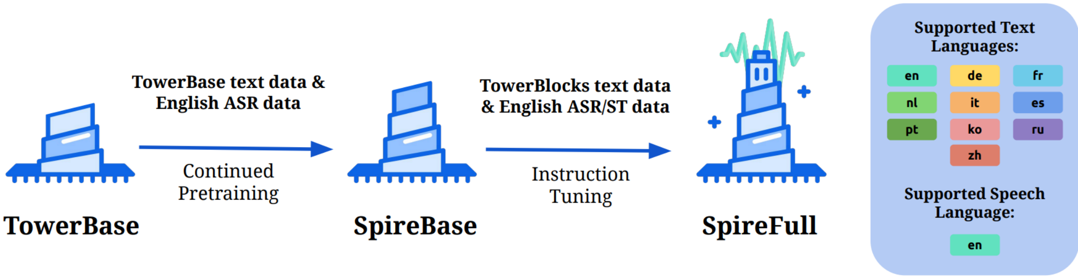

Figure 1: Illustration of the model training approach for SpireBase and SpireFull.

augmented LLM built from the open-weight multilingual model Tower (Alves et al., 2024). Spire can perform English ASR and from-English speech translation (ST) while maintaining Tower's strong performance on machine translation (MT) across all 10 languages[1] supported by Tower. Spire encodes speech via HuBERT-based (Hsu et al., 2021) k-means clustering, as in previous work (Zhang et al., 2023; Rubenstein et al., 2023; Chang et al., 2024). We perform training in two stages: Continued Pre-Training (CPT) and Instruction Tuning (IT). For the CPT stage, we use a mixture of ASR data and a small fraction of Tower's text CPT data. For IT, we leverage Tower's task-specific MT data, as well as additional English ASR and ST data. Spire is trained using only **42.5K** hours of speech, differing from the large scale of data used by existing models (Radford et al., 2023; Nguyen et al., 2025; Chu et al., 2024). Figure 1 illustrates our training process. We make the following contributions:

- We present a pipeline for integrating speech as an additional modality into an existing LLM, enabling it to transcribe and translate English speech while preserving its original text capabilities across 10 languages;
- We analyze speech integration at two stages, namely CPT and IT, demonstrating the necessity of both stages to achieve optimal performance across both modalities;
- We make our models, datasets, and scripts available to the community.[2]

2 Related Work

Speech-to-Text Models An increasing number of studies have explored integrating speech into LLMs (Zhang et al., 2023; Rubenstein et al., 2023; Hassid et al., 2024). For discrete speech input, Hassid et al. (2024) demonstrate the benefits of initializing a speech LLM from a text-based LLM. SpeechGPT (Zhang et al., 2023) applies IT on speech-to-text ASR, text-to-speech (TTS), and text-based question answering. AudioPALM (Rubenstein et al., 2023) is trained in a multi-task fashion, similarly to SpeechGPT, but on multilingual input. Recently, VoxtLM (Maiti et al., 2024) was trained jointly on DSUs and text data for ASR, TTS, and open-ended speech/text generation. Our work is most similar to SpiritLM (Nguyen et al., 2025), which adapts an LLM with an interleaved mixture of DSU and text data, which requires an expensive DSU-to-transcript step to create. In contrast, we adopt a more cost-effective input representation that can be extended to any language, regardless of the availability of a speech aligner. Our focus is on successfully incorporating speech input while preserving the original competence of the model, so that the resulting model can successfully perform both speech-to-text and text-only tasks. None of the aforementioned models are trained to preserve the original model's performance in text tasks.

Adapting LLMs Previous approaches involve training from scratch with task- and domain-specific data (Singhal et al., 2023; Lewkowycz et al., 2022), performing CPT with a diverse training data mix designed to broadly extend the model's knowledge (Wu et al., 2023b), or IT on use-case-specific data (Chen et al., 2023). Recent work has explored combining the latter two approaches (Xu et al., 2024a; Alves et al., 2024; Wei et al., 2021; Roziere et al., 2023). In our approach to integrating DSUs into Tower, we take inspiration from Alves et al. (2024) in adopting a two-step CPT+IT process. Our work differs in that we focus on adding the speech modality, whereas Alves et al. (2024) focused on increasing the multilingual

[1] en, de, fr, nl, it, es, pt, ko, ru, zh
[2] https://huggingface.co/collections/utter-project/spire-67d4253d6af8d6a0308527e0

capabilities of an LLM.

Continuous and Discrete Speech Representations Self-supervised speech representation models produce contextualized high-dimensional speech vectors directly from raw audio (Hsu et al., 2021; Baevski et al., 2020; Chen et al., 2022), largely outperforming statistical speech features on downstream tasks (Yang et al., 2021). These continuous representations can be used to derive DSUs that capture both linguistic content and prosody through clustering (Borsos et al., 2023; Kharitonov et al., 2022). DSUs provide better alignment with textual data, facilitating the transfer of successful training settings from the text domain (Cui et al., 2024). Building on Lakhotia et al. (2021), which demonstrated that HuBERT (Hsu et al., 2021) is a powerful feature extractor, several studies have adopted this approach, incorporating a k-means clustering step for discretization (Zhang et al., 2023; Rubenstein et al., 2023; Lam et al., 2024; Chang et al., 2024; Nguyen et al., 2025). Xu et al. (2024b) study the optimal settings to obtain DSUs in terms of cluster size and feature extraction layer. We use their findings to inform our initial choices.

3 SPIRE: A Speech-to-Text LLM

We introduce SPIRE, whose goal is to equip an LLM with speech capabilities while preserving its preexisting text capabilities. As our base LLM we choose TOWER (Alves et al., 2024), which was developed from Llama-2 (Touvron et al., 2023) with a two-step approach: CPT on a mixture of monolingual and parallel data (TOWERBASE), followed by IT on translation-related tasks (TOWERINSTRUCT). We use an approach similar to TOWER to extend the model to the speech modality. First, we perform CPT with a combination of text-only and aligned speech-to-text datasets, followed by IT using both text-only general-purpose and task-specific data curated in TOWERBLOCKS,[3] alongside task-specific speech-to-text datasets.

We choose TOWER in particular due to its competitive performance compared to other open alternatives. TOWER-based models were among the best participating systems in the WMT24 general translation task (Kocmi et al., 2024). TOWER's usage of open source data during the CPT phase along with the release of the TOWERBLOCKS dataset, used in the IT phase, further motivates our choice.

[3] `https://huggingface.co/datasets/Unbabel/TowerBlocks-v0.2`

3.1 Speech Discretization

To easily transfer the training setup of TOWER, we use DSUs as opposed to an auxiliary speech encoder. For all speech datasets that were used, we follow recent discretization methodology (Zhang et al., 2023; Rubenstein et al., 2023; Chang et al., 2024) to produce DSUs by first extracting continuous speech representations for our speech data from the 22nd layer of an HuBERT-large model, trained on 60K hours of English speech (Hsu et al., 2021), and then using k-means clustering ($K = 5000$) to produce centroids that are used to convert our continuous speech representation into a discrete sequence of cluster IDs.[4] We train our k-means model on a collection of 235K audio files (approximately 720 hours), drawn from three speech corpora: CoVoST-2 (Wang et al., 2021b), VoxPopuli (Wang et al., 2021a), and Multilingual LibriSpeech (MLS; Pratap et al., 2020). The CoVoST subset consists of 62K audio files from 10,049 speakers, with a maximum of 8 audio files per speaker. The VoxPopuli subset includes 65K audio files from 639 speakers, capped at 250 audio files per speaker. Finally, the MLS subset contains 107K audio files from 5,490 speakers.

3.2 SPIREBASE

The first CPT stage, yielding SPIREBASE, is trained from TOWERBASE-7B[5] using both text-only and aligned speech-to-text datasets. Following previous work, we include a fraction of TOWER's original training data to preserve its existing performance (Scialom et al., 2022; de Masson D'Autume et al., 2019).

3.2.1 Data

We use a mixture of monolingual and parallel text in Chinese (zh), Dutch (nl), English (en), French (fr), German (de), Italian (it), Korean (ko), Portuguese (pt), Russian (ru), and Spanish (es), that was sourced from the TOWER training data, as well as English ASR data sourced from popular open-source ASR datasets, as reported in Table 1. Both speech and text data are downsampled to create a 6B token data mixture (5B speech; 1B text), mea-

[4] Optimizing the layer selection for feature extraction is a complex research problem (Pasad et al., 2023; Mousavi et al., 2024). In this work we follow the insights from Gow-Smith et al. (2023) and Xu et al. (2024b).

[5] We used TOWER-7B models instead of the 13B or 70B versions due to its lower compute requirements.

sured by the model tokenizer.[6] Note that the 5B speech tokens include both DSUs (4.4B tokens) and their text transcriptions (0.6B tokens).

Text Data The monolingual text data split corresponds to data from mC4 (Raffel et al., 2019), a multilingual web-crawled corpus which we uniformly sample from across all languages. The parallel data split includes uniformly sampled instances to and from English (en↔xx) for the 10 languages, sourced from various public sources. Further details can be found in Alves et al. (2024).

Speech Data We collect 35K hours of speech data from SPGI Speech (O'Neill et al., 2021), GigaSpeech (Chen et al., 2021), MLS, and VoxPopuli. We normalize as described in Appendix A.1.

3.2.2 CPT Setup

We train SPIREBASE using MegatronLLM (Cano et al., 2023) on 8 A100-80GB GPUs for 6 days. We use the same hyperparameters as TOWER, except for the effective batch size, which in our case is 2304. To incorporate the DSUs in the CPT stage, we extend the model's original vocabulary by 5000 types, *e.g.*, `<extra_id_x>`. This allows us to have a vocabulary that can encode both text in subword units and speech in DSUs. For the extended vocabulary, we initialize new embeddings from a multivariate Gaussian distribution. The mean of this distribution is set to the average of the original embeddings, while the covariance is derived from the empirical covariance of the original embeddings, scaled by a factor of 1×10^{-5} (Hewitt, 2021).

3.3 SPIREFULL

SPIREFULL is obtained by instruction tuning SPIREBASE on task-specific text and speech data.

3.3.1 Data

We use a mixture of text and speech instructions for ASR, MT, and ST. The prompt formats used during training are shown in Appendix A.2.

Text Data We use TOWERBLOCKS (Alves et al., 2024), which includes high quality translation bitexts between English and the other languages supported by TOWER. It also includes instructions for the translation-related tasks of named entity recognition (NER) and automatic post-editing (APE).

[6]Preliminary experiments on the data mixture led to this particular choice.

Dataset	Task	Phase	# DSUs	# Hours
SPGI Speech	ASR	CPT	645M	5.1K
Gigaspeech	ASR	CPT	1.2B	9.9K
MLS	ASR	CPT	2.4B	19.2K
VoxPopuli	ASR	CPT	69M	0.5K
CV	ASR	IT	105M	0.8K
Europarl-ST	ST	IT	122M	1.0K
FLEURS	ST	IT	11M	0.09K
CoVoST-2	ST	IT	12M	0.09K
SPGI Speech	Pseudo-ST	IT	350M	2.8K
GigaSpeech	Pseudo-ST	IT	161M	1.3K
CV	Pseudo-ST	IT	212M	1.7K

Table 1: Statistics for speech training data. Hours are approximated from the number of deduplicated DSUs.

ASR Data We use 0.8K hours of ASR data from CommonVoice 18 (CV; Ardila et al., 2020), downsampling as described in Appendix A.1.

ST Data In our IT set, we use 842 hours of speech across three ST training sets: FLEURS (all nine language pairs; we filter out examples whose transcriptions overlap with the FLORES devtest set), Europarl-ST (Iranzo-Sánchez et al., 2020) (en → {de, es, fr, it, nl, pt}), and CoVoST-2 (en→zh). Since this amounts to far less data for ST than ASR, and since en→{ko, ru} have only examples from the tiny FLEURS set, we augment our speech collection with **pseudo-labeled** data, which has been effective for other ST systems (Barrault et al., 2023). We select 300k ASR examples each from CV, SPGI, and GigaSpeech and translate them to all nine target languages using `TowerInstruct-13B`.[7] We then filter examples whose transcript-translation combination has a COMET-QE[8] (Rei et al., 2022b) score under 85. Finally, for each language pair, we sample 60K examples to be used in direct ST prompts and another 60K to be used in multi-turn prompts. This results in 180K direct ST prompts and 180K multi-turn prompts for each language pair.[9] The prompt formats are shown in Appendix A.2.

3.3.2 IT Training Setup

We use the `chatml` template (OpenAI, 2023) to format our instructions in dialogue form. We train models using Axolotl[10] on 4 H100-80GB GPUs for 2.7 days. We use a learning rate of

[7]`https://huggingface.co/Unbabel/TowerInstruct-13B-v0.1`
[8]`https://huggingface.co/Unbabel/wmt22-cometkiwi-da`
[9]Due to our aggressive filtering, we were left with slightly fewer examples for en → zh.
[10]`https://github.com/axolotl-ai-cloud/axolotl`

7×10^{-6} and a cosine scheduler with 100 warm-up steps. We train for 4 epochs with an effective batch size of 576 and a weight decay of 0.01. We impose a maximum sequence length of 4096 and use the AdamW optimizer (Loshchilov and Hutter, 2019). Other hyperparameters are derived from TOWERINSTRUCT (Alves et al., 2024).

4 Experiments

We evaluate our models across three tasks: ASR, MT, and ST. First, we present our results for ASR (§4.1), confirming the new capabilities SPIRE has in the speech domain. We then present MT results (§4.2), demonstrating that the speech performance does not come at the expense of the original model's MT performance. Finally, we present results for ST (§4.3) to investigate model performance on a task that requires both ASR and MT capabilities.

Evaluation Setup Across models and tasks, we perform inference with greedy decoding with a maximum of 256 generated tokens. For the TOWER and SPIRE models, we decode with `vllm` (Kwon et al., 2023). However, since `vllm` does not support all of our baselines, we use alternative libraries (`transformers`; Wolf et al., 2019) where necessary. Unless specified otherwise, we use zero-shot prompts for all models and tasks.

4.1 ASR

Datasets and Metrics We evaluate ASR performance across multiple test sets, in order to cover a variety of recording styles: LibriSpeech (LS) test-clean and test-other (Panayotov et al., 2015), FLEURS (Conneau et al., 2023), and VoxPopuli.[11] We report the Word Error Rate (WER) between the hypotheses and gold transcripts, after Whisper normalization (Radford et al., 2023).

Baselines We include the following models:

- **Whisper** (Radford et al., 2023) is an encoder-decoder transformer trained on over 5 million hours of labeled data that performs multilingual ASR and to-English ST. We report results for Whisper-base (74M parameters) and Whisper-large-v3 (1.5B parameters).
- **SeamlessM4T** (Barrault et al., 2023) is an encoder-decoder transformer trained on 406K hours of speech that performs ASR, ST and MT across 100 languages. We report results for SeamlessM4T-large-v2 (2.3B parameters).
- **SALMONN** (Tang et al., 2024) integrates a pre-trained text LLM with separate speech and audio encoders into a single multimodal model.[12] SALMONN uses a LoRA adapter (Hu et al., 2022) to align the spaces.
- **Qwen2-Audio** (Chu et al., 2024) integrates audio into Qwen-7B (Bai et al., 2023) using a specialized encoder that is initialized from Whisper large-v3. The resulting model is pretrained on ~520K hours of data spanning speech, sound, and music.
- **SpiritLM** (Nguyen et al., 2025) is a decoder-only model, trained from Llama-2 on 307B tokens of text, 458K hours of unlabeled speech, and 111K hours of labeled speech. As in SPIRE, it uses HuBERT DSUs.
- **HuBERT-large+CTC** is a CTC-based ASR model trained using the same speech representation model we use for DSU generation, and using the same ASR data from the IT stage (Section 3.3.1).[13] Unlike SPIRE, this model has access to a very powerful speech representation backbone. However, it lacks strong language modeling capabilities.

	LibriSpeech Clean	LibriSpeech Other	FLEURS	VoxPopuli
Whisper-base	5.0	11.9	12.1	9.8
Whisper-large-v3	1.8	**3.7**	**5.8**	9.2
SeamlessM4T	2.6	4.9	8.1	7.5
SALMONN	2.4	5.3	9.3	8.9
Qwen2-Audio	**1.6**	3.9	6.6	**6.5**
SpiritLM	*6.0**	*11.0**	-	-
HuBERT-large+CTC	4.3	7.6	11.4	14.7
Our models				
SPIREBASE	28.9	56.3	11.0	13.7
SPIREFULL	4.2	7.1	10.7	15.8

*We were unable to reproduce SpiritLM's ASR performance; therefore, we report their self-reported LS results using ten-shot prompts.

Table 2: WER on various ASR test sets.

Results Our results are presented in Table 2. SPIREFULL's performance demonstrates that performing both the CPT and IT stages is an effective strategy to give speech capabilities to a text LLM.

[11] For CPT models, LS is an in-domain evaluation because its training set is part of MLS.

[12] SALMONN uses 4400 hours of speech/audio data in the IT phase but does not specify the large amount of pre-training ASR and audio captioning data used.

[13] The hyperparameters are described in Appendix B.

	en→xx		xx→en	
	C22	spB	C22	spB
SeamlessM4T	87.22	39.0	87.42	39.9
TOWERBASE-7B	87.38	37.8	88.02	41.7
TOWERINSTRUCT-7B	88.45	38.8	**88.27**	**42.0**
Our models				
SPIREBASE	87.41	37.4	87.97	41.4
SPIREFULL	**88.54**	**39.3**	88.21	41.8

Table 3: COMET-22 (C22) and spBLEU (spB) on the FLORES devtest set between English and the other languages supported by TOWER And SPIRE.

On the other hand, SPIREBASE does not consistently show reasonable speech performance; however, on FLEURS and VoxPopuli we obtain somewhat strong results in the zero-shot settings, which is surprising given that non-instruction-tuned models often struggle to work out-of-domain without in-context learning examples.[14]

Although SPIREFULL does not match the performance of SeamlessM4T, Whisper-large-v3, SALMONN, or Qwen2-Audio, these were trained on far more speech data than our models (around 10x for Qwen2-Audio and SeamlessM4T). Given this training data gap, it is notable that SPIREFULL *does* outperform Whisper-base on LS and FLEURS, and SpiritLM on all benchmarks SpiritLM reports, at a fraction of the speech data.

SPIREFULL also outperforms the HuBERT-large+CTC baseline on three out of four datasets. This is an impressive result given that the CTC model has access to continuous features, which SPIREFULL lacks. We believe this demonstrates that our compressed discrete representations capture the speech signal well enough to support speech-to-text tasks.

4.2 MT

Having demonstrated that our training approach works well to initially equip TOWER with speech processing capabilities, we now turn to MT to investigate whether SPIRE can maintain TOWER's strong performance on MT despite its speech-centric CPT.

Datasets and Metrics We evaluate on two datasets for MT: FLORES-200 (Team et al., 2024), which covers SPIRE's languages, and the WMT23 test set (Kocmi et al., 2023), which covers en↔{de, ru, zh}. We report COMET-22 (COMET; Rei et al., 2022a) and spBLEU[15] (Papineni et al., 2002) scores via the SacreBLEU toolkit (Post, 2018).

[14] We also tried prompting SPIREBASE with few-shot examples, but the results were much worse, possibly because the length of the DSU sequences led to in-context examples that were too long for the model to handle effectively.

	APE		NER
	en→xx	xx→en	Multilingual
TOWERINSTRUCT-7B	83.08	**80.29**	**71.56**
SPIREFULL	**83.13**	80.08	67.10

Table 4: Results on APE (COMET) and NER (seq. F1).

Baselines We compare the SPIRE models to the text-to-text translation performance of SeamlessM4T. Additionally, we report the performance of TOWERBASE-7B and TOWERINSTRUCT-7B.

Results Our results show that even after the speech-centric CPT and mixed speech and text IT stage, the SPIRE models retain the original text-only performance of TOWER on both FLORES (Table 3) and WMT23 (Table 5). This indicates that neither CPT nor IT on speech data negatively impacts the model's ability to perform MT. This is true for both SPIREBASE, which achieves performance comparable to TOWERBASE; and for IT models, where SPIREFULL slightly surpasses the performance of TOWERINSTRUCT on en→xx. Table 6 shows that between TOWERINSTRUCT and SPIREFULL, neither model consistently shows a significant improvement over the other. SPIREFULL also outperforms SeamlessM4T by both metrics on all WMT23 language pairs, and for both en→xx and xx→en on FLORES.

Translation-related Tasks We follow the evaluation set-up from TOWER (Alves et al., 2024) to additionally evaluate SPIRE on translation-related tasks. In Table 4 we report our results on APE for en↔{de, ru, zh} and NER for {de, en, es, fr, it, pt, zh}. SPIRE performs similarly to TOWERINSTRUCT across both tasks and all language directions, maintaining the original text-only capabilities even after training on speech data.

4.3 ST

As SPIRE has shown success at both ASR and MT, we now investigate its performance on ST.

Datasets For ST, we evaluate our models on FLEURS (Conneau et al., 2023), covering ST be-

[15] `nrefs:1|case:mixed|eff:no|tok:flores200|smooth:exp|version:2.5.1`

	en→de		en→ru		en→zh		de→en		ru→en		zh→en	
	C22	spB	C22	spB	C22	spB	C22	spB	C22	spB	C22	spB
SeamlessM4T	77.76	27.8	83.22	34.2	80.14	29.7	78.69	26.6	80.58	32.5	76.96	23.8
TOWERBASE-7B	79.96	36.1	83.08	34.2	83.49	33.3	83.56	41.1	80.06	32.7	78.48	23.5
TOWERINSTRUCT-7B	82.34	38.8	**84.66**	**34.9**	85.09	35.3	84.95	45.1	**82.94**	**36.7**	**80.14**	26.1
Our models												
SPIREBASE	79.88	34.7	83.04	33.7	83.85	32.4	83.19	40.5	80.20	32.4	78.65	23.1
SPIREFULL	**82.50**	**39.5**	84.60	**34.9**	**85.37**	**37.3**	**85.24**	**45.2**	82.58	36.4	79.92	**26.3**

Table 5: COMET-22 (C22) and spBLEU (spB) on the WMT23 test set.

Corpus	Metric	TI Better	SF Better	NS
FLORES	spBLEU	3	4	11
	COMET	1	1	16
WMT23	spBLEU	0	2	4
	COMET	2	2	2

Table 6: Counts of language pairs in which TOWERINSTRUCT significantly outperforms SPIREFULL (TI Better), SPIREFULL significantly outperforms TOWERINSTRUCT (SF Better), or differences are not significant (NS). Significance is reported at the $p < 0.5$ level. We used the paired bootstrap method for spBLEU.

	en→de		en→zh	
	C22	spB	C22	spB
Self-cascade				
SeamlessM4T	72.40	21.7	72.32	17.0
SPIREFULL	**78.05**	**31.8**	**79.50**	**28.1**
Direct				
SALMONN	74.98	22.7	80.92	27.8
Qwen2-Audio	82.29	34.5	**85.27**	**38.7**
SeamlessM4T	**85.95**	**42.3**	83.62	31.3
SPIREFULL	73.96	25.4	74.53	21.0

Table 7: ST results on CoVoST-2.

tween all en→xx pairs, and CoVoST-2 (Wang et al., 2021b) for en→{de, zh}.

ST approaches As well as direct ST, we report self-cascades, in which each model transcribes the audio before translating its own output to the target language (*i.e.*, ASR followed by MT).

Baselines We compare SPIRE to SeamlessM4T in both direct and cascaded settings. We also report the results of SALMONN and Qwen2-Audio, which are both 7B parameter models, like SPIRE. However, SALMONN and Qwen2-Audio do not support text-to-text translation, so we use them only for direct ST.[16] There are also coverage differences between the models: while SeamlessM4T can handle all of SPIRE's language pairs, neither SALMONN nor Qwen2-Audio supports en→ko; SALMONN also does not support en→ru.

Results Our FLEURS spBLEU ST results are reported in Table 8. For brevity, COMET-22 scores are reported in Appendix C. SeamlessM4T performs best at direct ST for all language pairs except en→zh. Among the 7B parameter models, SPIREFULL is the best direct model on average, notably beating SALMONN on all language pairs except en→zh. It also outperforms Qwen2-Audio on 6 out of 8 language pairs that Qwen2-Audio supports, and ties or beats it for all except en→zh and en→de.

Performance on CoVoST-2 (Table 7) tells a different story. Although SPIREFULL maintains its advantage over SeamlessM4T in self-cascaded translation, it attains the worst performance on en→zh, while performing similarly to SALMONN for en→de. This shows that the direct ST performance of SPIREFULL is dataset-dependent, which could be a consequence of its relatively small training data.

SPIREFULL achieves the best self-cascaded performance by a significant margin for both datasets, outperforming SeamlessM4T by a large margin in this setting. This demonstrates that SPIREFULL maintains greater robustness to its own outputs than SeamlessM4T, supporting the insight that LLM-based translation models can be very robust to perturbations (Peters and Martins, 2025).

5 Analysis

The key innovation of our approach is the application of the CPT followed by IT paradigm to dis-

[16] Although Whisper is frequently used for ST, we exclude it because it only supports to-English translation, whereas SPIRE is a from-English ST model. Therefore ST comparisons between the two models are impossible.

	de	es	fr	it	ko	nl	pt	ru	zh	avg_7	avg_{all}
Self-Cascade											
SeamlessM4T	24.2	21.5	37.7	18.9	12.5	16.9	28.2	27.1	14.6	23.1	22.4
SpireFull	**38.1**	**29.4**	**45.3**	**31.2**	**23.1**	**31.2**	**42.9**	**33.5**	**29.0**	**35.3**	**33.7**
Direct											
SeamlessM4T	**39.2**	**28.0**	**48.1**	**30.6**	**21.5**	**30.8**	**47.5**	**34.3**	23.2	**35.3**	**33.7**
SALMONN	25.5	20.8	34.3	16.7	0.1	20.5	32.6	3.1	21.9	24.6	19.5
Qwen2-Audio	31.8	23.5	31.3	23.5	5.4	22.3	36.1	23.7	**24.7**	27.6	24.7
SpireFull	31.1	23.5	37.9	25.5	15.4	25.7	37.3	26.9	21.0	28.9	27.1

Table 8: FLEURS ST ex→xx results with self-cascade and direct models in terms of spBLEU. The avg_7 column averages over the 7 language pairs that all models in the table support (excluding en→{ko, ru}).

Model	Base Model	CPT Speech	CPT Text	IT Speech	IT Pseudo	IT Text
TowerFull	TowerBase-7B	✗	✗	✓	✓	✓
SpireBase	SpireBase	✓	✓	✗	✗	✗
SpireFull	SpireBase	✓	✓	✓	✓	✓
Spire *Variants*						
SpireNoBlocks	SpireBase	✓	✓	✓	✓	✗
SpireNoPseudo	SpireBase	✓	✓	✓	✗	✓

Table 9: Ablations of our models. The CPT and IT columns indicate which data was seen during training.

	ASR	MT				ST	
		en→xx		xx→en		en→xx	
	WER	C22	spB	C22	spB	C22	spB
SpireFull	4.2	88.54	39.3	88.21	41.8	**81.33**	**27.1**
TowerFull	9.5	**88.57**	**39.4**	88.17	41.7	79.10	26.1
SpireNoBlocks	4.1	82.98	34.2	85.93	36.1	81.11	27.1
SpireNoPseudo	**3.9**	88.40	38.9	**88.22**	**42.0**	62.80	27.1

Table 10: Ablation models and SpireFull on LS Clean for ASR, FLORES devtest for MT, and Fleurs for ST reporting WER, COMET-22 (C22), and spBLEU (spB).

cretized speech allowing us to build upon existing text-only capabilities of our base model. Here, we analyze how the composition of these two training phases contributes overall to model performance across all tasks previously evaluated. To that end, we consider several variants of SpireBase and SpireFull which are described in Table 9 and whose results are reported in Table 10.

- *i*) No CPT was performed and IT was performed with the entire IT data mix (TowerFull);
- *ii*) CPT was performed and no data from TowerBlocks was seen during IT (SpireNoBlocks), and
- *iii*) CPT was performed and pseudo-labeled ST data and FLEURS were omitted from the IT data mix (SpireNoPseudo).

We report additional datasets in Appendix D.

Effectiveness of CPT and IT Our previous results demonstrated that using both CPT and IT was the most effective strategy. The performance gap between SpireFull and the TowerFull on ASR (5.3 points in LS test-clean) further shows that IT alone is also not as effective. However, for ST we observe that performing only IT leads to a strong model that is capable of performing speech translation. This contrasts from SpireBase, for which we also attempted direct ST but the model failed to produce output in the target language, even when given few-shot prompts. Despite the impressive results from TowerFull, we still observe the best performance by SpireFull showing that while the effect of CPT is not as drastic as in the case of ASR, we still observe gains with a speech-centric CPT phase.

Modality Interplay Our results show that text and speech modalities are orthogonal to each other. Specifically, the performances of TowerFull and SpireFull show that speech-centric CPT *does not* degrade the text performance of the base model. However, MT quality suffers when TowerBlocks is removed from the IT data, as is shown by SpireNoBlocks's much weaker performance than SpireFull. Simultaneously, SpireFull performs on par with SpireNoBlocks on both ASR and ST, indicating that adding text instructions also *does not* degrade performance on speech tasks. It is worth highlighting that a model strong at both MT and ASR (SpireNoPseudo) does not lead to a strong ST model, showing surprisingly that competence at MT is not very helpful for direct ST.

6 Conclusion

In this work we presented SPIRE, a simple and effective recipe for adapting a text-based, translation-specialist LLM to the speech modality while preserving the original performance on text-based tasks. We investigated the impact of speech integration on two stages of LLM adaptation, CPT and IT, finding that both contribute to the final model's performance on speech tasks. Our results demonstrate that we are able to successfully integrate a new modality without compromising the original model's capabilities. SPIRE achieves competitive performance on ASR, while its MT abilities remain on par with the original TOWER model. Finally, for the ST task, we find that the leveraging ASR and MT data does not directly transfer to ST performance. Nonetheless, the model achieves promising performance with both direct and self-cascaded ST. To benefit the community, we only used publicly available and licensed data to train our models, making our results reproducible. As future work, we intend to extend this recipe to multilingual settings by replacing our English HuBERT speech component by the multilingual mHuBERT-147 (Boito et al., 2024).

Limitations

The downstream tasks we evaluate on are restricted to MT and ASR/ST, which provides an idea of the model performance but do not give us the full picture. We plan to address this by utilizing the LM-harness evaluation (Gao et al., 2024) to evaluate on a suite of text-based benchmarks such as MMLU (multitask language understanding) (Hendrycks et al., 2021a,b), Arc (commonsense reasoning) (Clark et al., 2018), Belebele (reading comprehension) (Bandarkar et al., 2024), and HellaSwag (sentence completion) (Zellers et al., 2019). Lastly, our model handles speech and text on the input side but is currently limited to generating only text.

Acknowledgments

This work was supported by EU's Horizon Europe Research and Innovation Actions (UTTER, contract 101070631), by UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee (grant number 10039436: UTTER), by the project DECOLLAGE (ERC-2022-CoG 101088763), by the Portuguese Recovery and Resilience Plan through project C645008882-00000055 (Center for Responsible AI), by Fundação para a Ciência e Tecnologia (FCT) through the project with reference UIDB/50021/2020 (DOI:10.54499/UIDB/50021/2020), and by FCT/MECI through national funds and when applicable co-funded EU funds under UID/50008: Instituto de Telecomunicações. This work was performed using HPC resources from GENCI–IDRIS (Grant 2023-AD011014668R1). We thank Duarte Alves and Giuseppe Attanasio for their insightful comments.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Duarte Miguel Alves, José Pombal, Nuno M Guerreiro, Pedro Henrique Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, Pierre Colombo, José G. C. de Souza, and Andre Martins. 2024. Tower: An open multilingual large language model for translation-related tasks. In *First Conference on Language Modeling*.

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

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report. *Preprint*, arXiv:2309.16609.

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and

Madian Khabsa. 2024. The belebele benchmark: a parallel reading comprehension dataset in 122 language variants. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 749–775, Bangkok, Thailand. Association for Computational Linguistics.

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

Marcely Zanon Boito, Vivek Iyer, Nikolaos Lagos, Laurent Besacier, and Ioan Calapodescu. 2024. mHuBERT-147: A Compact Multilingual HuBERT Model. In *Interspeech 2024*.

Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matt Sharifi, Dominik Roblek, Olivier Teboul, David Grangier, Marco Tagliasacchi, et al. 2023. Audiolm: a language modeling approach to audio generation. *IEEE/ACM transactions on audio, speech, and language processing*, 31:2523–2533.

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. 2023. epfllm megatron-llm.

Xuankai Chang, Brian Yan, Kwanghee Choi, Jee-Weon Jung, Yichen Lu, Soumi Maiti, Roshan Sharma, Jiatong Shi, Jinchuan Tian, Shinji Watanabe, et al. 2024. Exploring speech recognition, translation, and understanding with discrete speech units: A comparative study. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 11481–11485. IEEE.

Guoguo Chen, Shuzhou Chai, Guan-Bo Wang, Jiayu Du, Wei-Qiang Zhang, Chao Weng, Dan Su, Daniel Povey, Jan Trmal, Junbo Zhang, Mingjie Jin, Sanjeev Khudanpur, Shinji Watanabe, Shuaijiang Zhao, Wei Zou, Xiangang Li, Xuchen Yao, Yongqing Wang, Zhao You, and Zhiyong Yan. 2021. GigaSpeech: An Evolving, Multi-Domain ASR Corpus with 10,000 Hours of Transcribed Audio. In *Proc. Interspeech 2021*, pages 3670–3674.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, and Xiong Xiao. 2022. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518.

Zeming Chen, Alejandro Hernández Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, et al. 2023. Meditron-70b: Scaling medical pretraining for large language models. *arXiv preprint arXiv:2311.16079*.

Ju-Chieh Chou, Chung-Ming Chien, Wei-Ning Hsu, Karen Livescu, Arun Babu, Alexis Conneau, Alexei Baevski, and Michael Auli. 2023a. Toward joint language modeling for speech units and text. *arXiv preprint arXiv:2310.08715*.

Ju-Chieh Chou, Chung-Ming Chien, Wei-Ning Hsu, Karen Livescu, Arun Babu, Alexis Conneau, Alexei Baevski, and Michael Auli. 2023b. Toward joint language modeling for speech units and text. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6582–6593, Singapore. Association for Computational Linguistics.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, Chang Zhou, and Jingren Zhou. 2024. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv:1803.05457v1*.

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

Wenqian Cui, Dianzhi Yu, Xiaoqi Jiao, Ziqiao Meng, Guangyan Zhang, Qichao Wang, Yiwen Guo, and Irwin King. 2024. Recent advances in speech language models: A survey. *arXiv preprint arXiv:2410.03751*.

Cyprien de Masson D'Autume, Sebastian Ruder, Lingpeng Kong, and Dani Yogatama. 2019. Episodic memory in lifelong language learning. *Advances in Neural Information Processing Systems*, 32.

Alexandre Défossez, Laurent Mazaré, Manu Orsini, Amélie Royer, Patrick Pérez, Hervé Jégou, Edouard Grave, and Neil Zeghidour. 2024. Moshi: a speech-text foundation model for real-time dialogue. *arXiv preprint arXiv:2410.00037*.

Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. 2023. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2024. A framework for few-shot language model evaluation.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Edward Gow-Smith, Alexandre Berard, Marcely Zanon Boito, and Ioan Calapodescu. 2023. NAVER LABS Europe's multilingual speech translation systems for the IWSLT 2023 low-resource track. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pages 144–158, Toronto, Canada (in-person and online). Association for Computational Linguistics.

Michael Hassid, Tal Remez, Tu Anh Nguyen, Itai Gat, Alexis Conneau, Felix Kreuk, Jade Copet, Alexandre Defossez, Gabriel Synnaeve, Emmanuel Dupoux, et al. 2024. Textually pretrained speech language models. *Advances in Neural Information Processing Systems*, 36.

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. 2021a. Aligning ai with shared human values. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021b. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.

John Hewitt. 2021. Initializing new word embeddings for pretrained language models. https:/nlp.stanford.edu/ johnhew//vocab-expansion.html.

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 transactions on audio, speech, and language processing*, 29:3451–3460.

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2022. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Shujie Hu, Long Zhou, Shujie Liu, Sanyuan Chen, Lingwei Meng, Hongkun Hao, Jing Pan, Xunying Liu, Jinyu Li, Sunit Sivasankaran, et al. 2024. Wavllm: Towards robust and adaptive speech large language model. *arXiv preprint arXiv:2404.00656*.

Rongjie Huang, Mingze Li, Dongchao Yang, Jiatong Shi, Xuankai Chang, Zhenhui Ye, Yuning Wu, Zhiqing Hong, Jiawei Huang, Jinglin Liu, et al. 2024. Audiogpt: Understanding and generating speech, music, sound, and talking head. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 23802–23804.

Javier Iranzo-Sánchez, Joan Albert Silvestre-Cerda, Javier Jorge, Nahuel Roselló, Adria 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. IEEE.

Eugene Kharitonov, Ann Lee, Adam Polyak, Yossi Adi, Jade Copet, Kushal Lakhotia, Tu Anh Nguyen, Morgane Riviere, Abdelrahman Mohamed, Emmanuel Dupoux, and Wei-Ning Hsu. 2022. Text-free prosody-aware generative spoken language modeling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8666–8681, Dublin, Ireland. 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, Marzena Karpinska, Philipp Koehn, Benjamin Marie, Christof Monz, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popović, Mariya Shmatova, Steinthór Steingrímsson, and Vilém Zouhar. 2024. Findings of the WMT24 general machine translation shared task: The LLM era is here but MT is not solved yet. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1–46, Miami, Florida, USA. 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.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Kushal Lakhotia, Eugene Kharitonov, Wei-Ning Hsu, Yossi Adi, Adam Polyak, Benjamin Bolte, Tu-Anh Nguyen, Jade Copet, Alexei Baevski, Abdelrahman Mohamed, and Emmanuel Dupoux. 2021. On generative spoken language modeling from raw audio. *Transactions of the Association for Computational Linguistics*, 9:1336–1354.

Tsz Kin Lam, Alexandra Birch, and Barry Haddow. 2024. Compact speech translation models via discrete speech units pretraining. *arXiv preprint arXiv:2402.19333*.

Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, et al. 2022. Solving quantitative reasoning problems with language models. *Advances in Neural Information Processing Systems*, 35:3843–3857.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual Instruction Tuning (LLaVA). *arXiv preprint*. ArXiv:2304.08485 [cs].

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Soumi Maiti, Yifan Peng, Shukjae Choi, Jee-weon Jung, Xuankai Chang, and Shinji Watanabe. 2024. Voxtlm: Unified decoder-only models for consolidating speech recognition, synthesis and speech, text continuation tasks. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 13326–13330. IEEE.

Pedro Henrique Martins, Patrick Fernandes, João Alves, Nuno M Guerreiro, Ricardo Rei, Duarte M Alves, José Pombal, Amin Farajian, Manuel Faysse, Mateusz Klimaszewski, et al. 2024. Eurollm: Multilingual language models for europe. *arXiv preprint arXiv:2409.16235*.

Pooneh Mousavi, Jarod Duret, Salah Zaiem, Luca Della Libera, Artem Ploujnikov, Cem Subakan, and Mirco Ravanelli. 2024. How should we extract discrete audio tokens from self-supervised models? *arXiv preprint arXiv:2406.10735*.

Tu Anh Nguyen, Benjamin Muller, Bokai Yu, Marta R Costa-Jussa, Maha Elbayad, Sravya Popuri, Christophe Ropers, Paul-Ambroise Duquenne, Robin Algayres, Ruslan Mavlyutov, et al. 2025. Spirit-lm: Interleaved spoken and written language model. *Transactions of the Association for Computational Linguistics*, 13:30–52.

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

OpenAI. 2023. URL `https://github.com/openai/openai-python/blob/release-v0.28.1/chatml.md`.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5206–5210. IEEE.

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.

Ankita Pasad, Bowen Shi, and Karen Livescu. 2023. Comparative layer-wise analysis of self-supervised speech models. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Ben Peters and André F. T. Martins. 2025. Did translation models get more robust without anyone even noticing? *Preprint*, arXiv:2403.03923.

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

Vineel Pratap, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve, and Ronan Collobert. 2020. MLS: A Large-Scale Multilingual Dataset for Speech Research. In *Proc. Interspeech 2020*, pages 2757–2761.

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

Srijith Radhakrishnan, Chao-Han Yang, Sumeer Khan, Rohit Kumar, Narsis Kiani, David Gomez-Cabrero, and Jesper Tegnér. 2023. Whispering LLaMA: A cross-modal generative error correction framework for speech recognition. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10007–10016, Singapore. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv e-prints*.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022a. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

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. 2022b. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine*

Translation (WMT), pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Baptiste Roziere, Jonas Gehring, Fabian Gloeckle, Sten Sootla, Itai Gat, Xiaoqing Ellen Tan, Yossi Adi, Jingyu Liu, Romain Sauvestre, Tal Remez, et al. 2023. Code llama: Open foundation models for code. *arXiv preprint arXiv:2308.12950*.

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, et al. 2023. Audiopalm: A large language model that can speak and listen. *arXiv preprint arXiv:2306.12925*.

Thomas Scialom, Tuhin Chakrabarty, and Smaranda Muresan. 2022. Fine-tuned language models are continual learners. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6107–6122, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yu Shu, Siwei Dong, Guangyao Chen, Wenhao Huang, Ruihua Zhang, Daochen Shi, Qiqi Xiang, and Yemin Shi. 2023. Llasm: Large language and speech model. *arXiv preprint arXiv:2308.15930*.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

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

NLLB Team et al. 2024. Scaling neural machine translation to 200 languages. *Nature*, 630(8018):841.

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. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Viet Anh Trinh, Rosy Southwell, Yiwen Guan, Xinlu He, Zhiyong Wang, and Jacob Whitehill. 2024. Discrete multimodal transformers with a pretrained large language model for mixed-supervision speech processing. *arXiv preprint arXiv:2406.06582*.

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. *Interspeech 2021*.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Jian Wu, Yashesh Gaur, Zhuo Chen, Long Zhou, Yimeng Zhu, Tianrui Wang, Jinyu Li, Shujie Liu, Bo Ren, Linquan Liu, et al. 2023a. On decoder-only architecture for speech-to-text and large language model integration. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–8. IEEE.

Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. 2023b. Bloomberggpt: A large language model for finance. *arXiv preprint arXiv:2303.17564*.

Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. 2024a. A paradigm shift in machine translation: Boosting translation performance of large language models. In *The Twelfth International Conference on Learning Representations*.

Yaoxun Xu, Shi-Xiong Zhang, Jianwei Yu, Zhiyong Wu, and Dong Yu. 2024b. Comparing discrete and continuous space llms for speech recognition. In *Proc. Interspeech 2024*.

Hongfei Xue, Wei Ren, Xuelong Geng, Kun Wei, Longhao Li, Qijie Shao, Linju Yang, Kai Diao, and Lei Xie. 2024. Ideal-llm: Integrating dual encoders and language-adapted llm for multilingual speech-to-text. *arXiv preprint arXiv:2409.11214*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

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. 2021. SUPERB: Speech Processing Universal PERformance Benchmark. In *Proc. Interspeech 2021*, pages 1194–1198.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.

Yuexiang Zhai, Shengbang Tong, Xiao Li, Mu Cai, Qing Qu, Yong Jae Lee, and Yi Ma. 2024. Investigating the catastrophic forgetting in multimodal large language model fine-tuning. In *Conference on Parsimony and Learning*, pages 202–227. PMLR.

Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. 2023. SpeechGPT: Empowering large language models with intrinsic cross-modal conversational abilities. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15757–15773, Singapore. Association for Computational Linguistics.

A Data

A.1 Speech Data Preprocessing

Normalization In order to make transcripts consistent across the different datasets, the following normalization is applied:

- **GigaSpeech (CPT):** we lower-case the text and replace punctuation tags: <COMMA>, <PERIOD>, QUESTIONMARK>, <EXCLAMATIONPOINT> with their appropriate punctuation.
- **MLS (CPT):** we apply a tail-end normalization step here which uniformly samples each speaker to have at maximum 13 transcriptions. This allows us to have a better distribution of speakers.
- **CV (IT):** we subsampled from CommonVoice to ensure a minimum duration of 3 seconds per sample. To enhance transcript diversity, we limit each transcript to 4 unique speakers.

Deduplication As in previous work (Zhang et al., 2023; Rubenstein et al., 2023; Chang et al., 2024), we merge consecutive repeated DSU tokens into a single token to reduce sequence length.

A.2 Prompt Format

Table 11 show the prompts used during both training stages.

ASR (CPT)
Speech:<extra_id_i>···<extra_id_j> English: {TRANSCRIPT}
MT (CPT)
Source_lang: Source-sentence Target_lang: {TRANSLATION}
ASR (IT)
Speech: <extra_id_i>···<extra_id_j> English: {TRANSCRIPT}
Direct ST (IT)
Speech: <extra_id_i>···<extra_id_j> TARGET_LANG: {TRANSLATION}
Multi-turn ST (IT)
Speech: <extra_id_i>···<extra_id_j> English:{TRANSCRIPT} TARGET_LANG: {TRANSLATION}

Table 11: Prompt formats for CPT and IT.

B CTC-based ASR model

We train a CTC-based ASR model using the HuggingFace `transformers` library (Wolf et al., 2019), leveraging the ASR data from the IT stage (CV, Table 1) as training data after whisper normalization. Our ASR model is made of the HuBERT-Large[17] speech representation model, followed by three hidden layers and a vocabulary projection layer. We train for 50 epochs with a dropout of 0.3 and a learning rate of 1e-4 with a warm-up ratio of 0.15. We perform step-based best checkpoint selection based on CER scores. Our best checkpoint was obtained at step 220K (at epoch 12.8).

C ST results

Table 12 report results of ST on FLEURS across baseline models and SpireFull. We report COMET-22. We observe the same trend in scores as reported by spBLEU where in SpireFull obtains the best self-cascaded performance while beating Qwen2-Audio and SALMONN on direct ST across most language pairs. SeamlessM4T obtains the overall best performance in direct ST.

D Ablation results

Table 13 reports results from all remaining evaluation datasets across ASR, MT, and ST. We report the same metrics as in Section 4. Here as well, we note that in MT, the inclusion of speech data did not degrade text-only performance (SpireFull vs. TowerFull). Similarly, the inclusion of task-specific text data also did not harm performance on ASR (SpireNoBlocks vs. SpireFull). Lastly, SpireFull has the best performing direct ST system, further showing that individual task competencies (in MT and ASR) do not contribute directly to a compositional task (ST) but rather the inclusion of task-specific data leads to the highest gains (SpireNoPseudo vs SpireFull).

[17] https://huggingface.co/facebook/hubert-large-ll60k

	de	es	fr	it	ko	nl	pt	ru	zh	avg_7	avg_{all}
Self-Cascade											
SeamlessM4T	72.69	76.97	78.06	76.03	75.33	72.58	78.25	79.38	69.76	74.91	75.45
SPIREFULL	**84.26**	**83.32**	**84.70**	**85.16**	**86.89**	**84.91**	**86.01**	**86.45**	**85.21**	**84.80**	**85.21**
Direct											
SeamlessM4T	**84.79**	**83.20**	**85.32**	**85.03**	**85.17**	**85.17**	**86.75**	**86.31**	79.90	**84.31**	**84.63**
SALMONN	77.41	77.99	79.95	74.47	61.07	77.18	80.94	53.05	**81.63**	78.51	73.74
Qwen2-Audio	79.82	80.43	79.44	81.28	69.33	78.75	83.41	77.90	80.71	80.55	79.01
SPIREFULL	80.16	79.82	80.68	81.63	82.62	81.93	83.18	82.19	79.76	81.02	81.33

Table 12: FLEURS ST ex→xx results with self-cascade and direct models in terms of COMET-22. avg_7 covers the 7 language pairs that all models in the table support (excluding en→{ko, ru}).

	ASR			MT				ST	
	WER			C22	spB	C22	spB	C22	spB
	LS Other	Fleurs	VoxPopuli	en→xx		xx→en		en→xx	
SPIREFULL	**7.1**	10.7	15.8	84.16	**37.2**	**82.58**	**41.8**	**81.33**	**27.1**
TOWERFULL	13.8	14.3	40.7	**84.19**	36.9	82.25	35.6	71.52	20.1
SPIRENOBLOCKS	7.4	**10.4**	15.8	73.12	26.9	74.78	25.1	74.02	23.2
SPIRENOPSEUDO	7.3	11.1	**14.3**	83.93	36.9	82.50	35.9	59.88	6.8

Table 13: Ablation models and SPIREFULL on LS Other, Fleur, VoxPopuli for ASR, WMT23 for MT, and CoVoST-2 for ST reporting WER, COMET-22 (C22), and spBLEU (spB).