Instituto de Telecomunicações at IWSLT 2025: Aligning Small-Scale Speech and Language Models for Speech-to-Text Learning

Giuseppe Attanasio[♣], **Sonal Sannigrahi**[♣♠], **Ben Peters**[♣], **André F.T. Martins**[♣♠♢]
[♣]Instituto de Telecomunicações, Lisbon, Portugal
[♠]Instituto Superior Técnico, Universidade de Lisboa, Portugal
[♢]Unbabel, Lisbon, Portugal
giuseppe.attanasio@lx.it.pt

Abstract

This paper presents Instituto de Telecomunicações's submission to the IWSLT 2025 Shared Task on Instruction Following Speech Processing. We submit results for the Short Track, i.e., speech recognition, translation, and spoken question answering. Our model is a unified speech-to-text model that integrates a pretrained continuous speech encoder and text decoder through a first phase of modality alignment and a second phase of instruction finetuning. Crucially, we focus on using small-scale language model backbones (< 2B) and restrict to high-quality, CC-BY data along with synthetic data generation to supplement existing resources.[1]

1 Introduction

This paper presents our submission to the IWSLT 2025 Instruction Following track for the tasks of Automatic Speech Recognition (ASR), Speech Translation (ST), and Spoken Question Answering (SQA) for English, Chinese, and German. Our work builds upon a long line of previous research equipping LMs with additional multimodal capabilities, aligning an LM's semantic spaces with that of a pretrained speech encoder (Tang et al., 2023; Huang et al., 2023; Hu et al., 2024; Chu et al., 2024; Grattafiori et al., 2024, *inter alia)*. Our contribution is particularly motivated by *efficiency*, i.e., the goal of achieving strong performance using small-scale (< 2B) models. Recent work has explored audio quantization techniques (Zhang et al., 2024; Défossez et al., 2023, *inter alia)*, quantized input mel spectrograms (Shen et al., 2024), or extreme compression of input data over the time dimension through convolutional kernels paired with strong small-scale LM backbones (Abouelenin et al., 2025).

We take stock of such advancements and propose a model for even smaller scales. We use established methods for speech integration in LMs (Gaido et al., 2024; Grattafiori et al., 2024) using pretrained base models of up to 1.5B learnable parameters, finding empirically that with highly filtered and synthetic data, we can enable similar results at a fraction of the cost of larger LMs. The main contributions of our system are as follows:

- **Adapting pretrained, small-scale LMs:** We experiment with Qwen 2.5 1.5B and 0.5B (Yang et al., 2024a) as our LM of choice and use w2v-BERT 2.0 (Barrault et al., 2023) as our speech encoder.
- **Two-stage Training Curriculum:** We use a *modality alignment* and *instruction fine tuning* (IFT) phase for training our models, where the first equips the model with general speech capabilities and the second enables multi-task capabilities.
- **Training on open-licensed data:** To guarantee reproducibility and facilitate future research, we train on established CC-BY data collections and synthetic data filtered for quality. We release training and modeling artefacts under a permissive license.

2 Related Work

Efficient LMs. Recent works on efficient, small-scale language models (SLMs) have shown impressive knowledge compression capabilities by maintaining similar performance to larger, more computationally-intensive models. Models such as Phi-4 Mini (Abouelenin et al., 2025) and Gemma 2 (Gemma Team et al., 2024) have reported strong performance relative to size with a focus on computational efficiency. Lu et al. (2024) have shown how scaling laws operate differently for SLMs and

[1]Code and data at https://github.com/deep-spin/it-iwslt-2025.

Proceedings of the 22nd International Conference on Spoken Language Translation (IWSLT 2025), pages 347–353

have further demonstrated the efficiency of such models in subsequent reasoning tasks.

Multimodal LM Extension. Equipping a text-based model with multimodal capabilities is often done using an auxiliary modality encoder that is then used to jointly learn a semantic mapping between speech and text. Early works in joint text-speech modeling include AudioPaLM (Rubenstein et al., 2023), VioLA (Wang et al., 2023), and VoxtLM (Maiti et al., 2024). Other approaches combine a pretrained continuous speech encoder with an LM by concatenating speech embeddings to the text context (Tang et al., 2023; Huang et al., 2023; Hu et al., 2024; Chu et al., 2024; Grattafiori et al., 2024, *inter alia)*. Such works rely on strong multilingual capabilities of the speech encoder and those of large-scale LMs (i.e., 7B or more) to learn how to use speech-related parts of the context (Grattafiori et al., 2024). Our system echoes this compositional approach to speech and language modeling but leverages recent language models in the scale 0.5-1.5B.

3 System Overview

3.1 Model Architecture

Our model follows a standard speech encoder, text decoder architecture (e.g., Tang et al., 2023; Grattafiori et al., 2024; Chu et al., 2024).

Speech stack. We extract 80-dimensional Mel-filterbank audio representations with a stride of 2 using w2v-BERT 2.0's standard processor. Then we compress the audio over the time dimension using three 1D convolutional layers with a kernel width of 3 and a stride of 2. This input is then processed by the pretrained w2v-BERT 2.0 model. The output representations are processed by a modality and length adapter composed of two Conformer-like (Gulati et al., 2020) layers that further compress the audio representations on the time dimension and project them into the embedding space of the language model.

Text stack. We prepend the audio representations computed from the audio stack to the text input embeddings extracted from the input embedding matrix of the language model. We use a bidirectional self-attention for the audio positions and a causal (autoregressive) one for the text part of the context. Following prior work (Chu et al., 2023; Radford et al., 2022), we constrain text generation

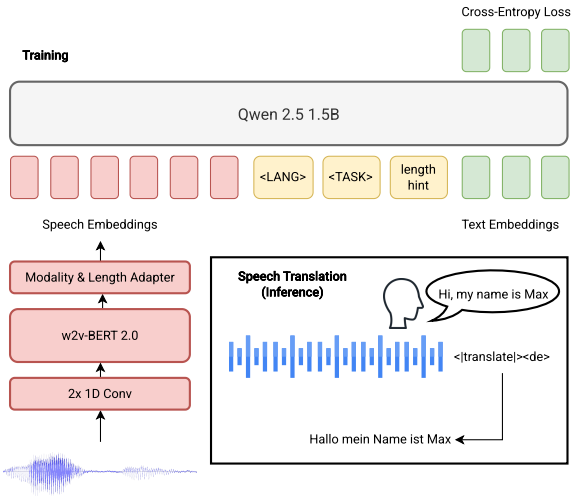

Figure 1: Illustration of our model. During training, the speech stack (red) generates speech representations which are prepended to task and language tags and an (optional) length hint (yellow), and text tokens (green). At inference time, we only provide the language and task tokens.

using a target language and task token. Figure 1 illustrates the model architecture.

3.2 Training Curriculum

We train our model in two stages: a modality alignment stage followed by an instruction fine-tuning (IFT) stage.

Modality Alignment. This first stage aligns the speech stack output representations to the language model's embedding space. We use the pretrained w2v-BERT 2.0 (v2)[2] as our speech encoder and randomly initialize the pre-encoder convolutional layers and the modality adapter. We choose Qwen 2.5 1.5B,[3] a multilingual LM, as our text decoder. Choosing a small (< 2B) model allows for the exploration of more efficient alternatives and is often overlooked in the literature.

In this phase, we train only the pre-encoder convolutional layers and the modality adapter with a learning rate of 3×10^{-3} for a single epoch. We train the model only on ASR data. The model is trained using standard cross-entropy loss on the reference transcript tokens. With a 95% chance, we prepend to the language and task tags a *length hint*, as suggested by Deitke et al. (2024), to let the model learn a length distribution. This stage leads to a model that can perform ASR but does not yet have other capabilities.

[2] https://huggingface.co/facebook/w2v-bert-2.0
[3] https://huggingface.co/Qwen/Qwen2.5-1.5B

Task	Data	License	Hours
	Modality Alignment (MA)		
ASR	LibriSpeech (LS)	CC-BY 4.0	1K
	Multilingual LS	CC-BY 4.0	2K
	FLEURS	CC-BY 4.0	24
	CommonVoice 16.1	CC-BY 4.0	4K
	Instruction Fine-Tuning (IFT)		
ASR	*All MA data*		
	VoxPopuli	CC-BY 4.0	1.8K
	Peoples Speech	CC-BY 4.0	12K
	CV 16.1 PL	-	30K
ST	CoVoST-2	CC-BY NC 4.0	3K
	CoVoST-2 PL	CC-BY NC 4.0	3K
SQA	SpokenSQuAD	-	-
	Generated Data	-	-

Table 1: Data statistics with licence, hours of speech data across all languages, and task splits.

Instruction Fine-Tuning. Following the modality alignment phase, we perform IFT using speech-to-text tasks included in the IWSLT campaign (AST, SQA) in English, German, and Chinese. During this stage, we train every component jointly end-to-end.[4]

Generation Parameters. For all tasks, we let the model generate up to 1024 tokens with beam search decoding (beam size of 3), a repetition penalty of 1.6, and nucleus sampling with temperature of 1.2.

3.3 Data

Where possible, we use CC-BY licensed data across all tasks. When sufficient data is not available, we generate synthetic corpora using the procedures described below. Table 1 provides an overview of the data sources used for each task and training phase.

Speech Recognition. We use CommonVoice 16.1 (Ardila et al., 2020), FLEURS (Conneau et al., 2023), MLS (Pratap et al., 2020), and LibriSpeech (Panayotov et al., 2015) for the modality alignment (MA) data mixture. For IFT, we reuse all MA data plus VoxPopuli (Wang et al., 2021) and The People's Speech (clean) (Galvez et al., 2021).

Speech Translation. We use CoVoST2 (Wang et al., 2020) as gold-standard AST data across English, Chinese, and German. We supplement this gold standard parallel data by **pseudolabeling** ASR transcriptions. This technique has proven effective for previous systems (Barrault et al., 2023; Ambilduke et al., 2025) and is simple to implement. Concretely, we translate all transcriptions from the English portion of CommonVoice 16.1 using four strong MT models: the 13B and Mistral-7B versions of TowerInstruct (Alves et al., 2024), EuroLLM-9B-Instruct (Martins et al., 2024), and NLLB-3.3B (NLLB Team et al., 2022). For each transcription, we use a COMETKiwi (Rei et al., 2022) oracle to select the best translation among the four systems. We then filter out examples for which the best translation records a score under 0.85. This process allows for fewer examples to be filtered than in conventional single-model pseudolabeling, as is shown in Table 2, and also increases diversity because the translations come from a mixture of several models.

Model	% Kept en→de	% Kept en→zh
NLLB-3B	58.1	34.3
TowerInstruct-Mistral-7B	60.0	51.5
TowerInstruct-13B	59.4	49.9
EuroLLM-9B-Instruct	62.5	52.3
Oracle	71.0	64.3

Table 2: Percentage of English transcriptions in CommonVoice 16.1 for which various models produce a translation with a COMETKiwi score of at least 0.85. The oracle keeps a much larger share of hypotheses than any individual model.

Spoken Question Answering. We use the Spoken SQuAD (Lee et al., 2018) dataset for English SQA. This dataset consists of several texts which are synthesized into speech and has a total of 37K questions and answers. Due to the limited availability of multilingual SQA datasets, we follow the same pseudolabeling process as for ST to create synthetic German and Chinese questions and answers for each example. The question and answer were translated separately using the same mixture of models as for ST. Question-answer pairs were kept if the best translated question had a COMETKiwi score of at least 0.80. The same system was used to translate the answer regardless of how it compared to the translated answers from other systems.[5] As the SQA task also includes questions where the answer cannot be inferred from the context, we additionally generate synthetic *unanswerable* questions for each context in English, German, and Chinese using Qwen2.5-

[4] MA and IFT runs required a total of three days using four H100 GPUs in an in-house infrastructure.

[5] As the answers were generally very short, we found that COMETKiwi performed unreliably for them.

Given a text passage and some questions about it, write 2 questions in [LANG_ID] as close to the style of the original questions as possible but that are not answerable. The questions must be of similar difficulty as the example questions, i.e., they have to mention aspects and topics of the passage, but the answer cannot be inferred from the text. Be creative. Provide one question per line.

Text passage: [CONTEXT]

Example questions: [QUESTION]
Unanswerable questions:

Figure 2: Prompt used to generate unanswerable questions from Qwen2.5-70B where **context** is the transcript used to synthesize speech in Spoken SQuAD, **question** is an answerable question from Spoken SQuAD, and **lang id** is the language in which we want to generate questions.

70B (Yang et al., 2024a). Following insights from (Sannigrahi et al., 2024), we provide the LM with context along with example questions to guide the style and quality of the generated answers. We find that without example questions based on the original dataset, the LM often produces i) questions not adhering to the topic of the context and ii) verbose questions. We also experiment with prompts that do not explicitly request the model to mention aspects/topics of the context provided and find this to be suboptimal. As the number of positive instances in the Spoken QA dataset is small, in order to maintain a balanced dataset, we limit unanswerable questions to two per context. We further experimented with using the audio directly as opposed to the text transcript for context, but found this approach to be more prone to errors, as the Spoken SQuAD dataset is *not* a native spoken dataset but rather a synthesized QA dataset, often leading to minor pronunciation errors. The final prompt used to obtain additional questions is shown in Figure 2.

Preprocessing. We restrict our model to process audio of up to 120 seconds, discarding all training input longer than that. We preprocess all the instances appending to the speech embeddings (and prepending to the text embeddings) the task and language tags following the input template in Figure 1. The task task tag can be either <|transcribe|>, <|translate|>, or <|reply|> for ASR, AST, and SQA, respectively, and the language tag is <|en|>, <|de|>, or <|zh|>.

en		en-de		en-zh	
ASR	SQA	ST	SQA	ST	SQA
0.15	0.14	0.34	0.22	0.34	0.21

Table 3: Official normalized ASR (WER (↓)), ST (COMET (↑)), and SQA (BertScore (↑)) scores.

4 Results

Our results for all three short-track tasks are in Table 3. For further details about the evaluation campaign as well as the metrics, we refer readers to Abdulmumin et al. (2025).

Our model obtains **reasonably good ASR scores for English**. This result is particularly relevant, considering that the test data originate from the technical domain, exhibit high speaker variability, and consist primarily of spontaneous speech. However, while the model successfully performs ASR, **SQA and ST prove to be more complex**. Through manual inspection, we observed poor quality outputs for ST. At times, the model repeats the same word or ignores the task tag and transcribes the audio segment rather than translating it. This finding aligns with prior work that has found ASR data dominates the multitask capabilities of models (Tang et al., 2023). Moreover, it emphasizes the importance of a more carefully designed training curriculum, where SQA, ST, and ASR data are more evenly distributed. Lastly, due to the audio length cutoff—set to 120 seconds due to technical limitations—we were unable to use all of the available SQA data. At test time, when prompted to perform the SQA task, the model sometimes generates the question itself, rather than the answer. We believe that by utilizing a combination of more data, enhanced base models with stronger multilingual capabilities, and extended context support, we will be able to improve upon these results substantially.

5 Conclusions

We have presented our submission for the IWSLT 2025 Instruction Following Short Track. We explored the usage of a small-scale LM in modality adaptation through a continuous speech encoder. In particular, we equip an existing text model, Qwen 2.5 1.5B, with the speech modality for a joint multilingual and multitask model.

We used standard modality alignment approaches, including building on pretrained speech encoders and autoregressive text decoder mod-

els, and a two-stage curriculum learning. In future work, we plan to support longer contexts, better filtered data, and further push small-scale LMs to be fully multimodal. We will incorporate more high-quality multilingual data to enhance the model's language identification capabilities. Additionally, we will extend the evaluation beyond standard performance-oriented benchmarks, e.g., by accounting for safety (Yang et al., 2024b) and fairness (Koudounas et al., 2024; Attanasio et al., 2024).

Acknowledgements

We thank Duarte Alves and Patrick Fernandes for their feedback and insightful discussions in the earlier versions of the paper. This work was supported by the Portuguese Recovery and Resilience Plan through project C645008882-00000055 (Center for Responsible AI), by EU's Horizon Europe Research and Innovation Actions (UTTER, contract 101070631), by the project DECOLLAGE (ERC-2022-CoG 101088763), and by FCT/MECI through national funds and when applicable co-funded EU funds under UID/50008: Instituto de Telecomunicações.

Limitations

Currently, our model supports audio up to 2 minutes in length. Ideally, we would also like to support longer audio contexts while maintaining computationally inexpensive training. With the current length filters, we do not see much of the SQA data, which hinders the model's multi-task capabilities. Additionally, we have worked with a small LM (1.5B) in our model, which did not have the best language modeling capabilities. We plan to run additional experiments within the 3B scale. Lastly, there is limited research on the filtering of synthetically generated data for the QA domain. For future work, we plan to further refine the pipeline to generate synthetic QA data from spoken contexts.

References

Idris Abdulmumin, Victor Agostinelli, Tanel Alumäe, Antonios Anastasopoulos, Ashwin, Luisa Bentivogli, Ondřej Bojar, Claudia Borg, Fethi Bougares, Roldano Cattoni, Mauro Cettolo, Lizhong Chen, William Chen, Raj Dabre, Yannick Estève, Marcello Federico, Marco Gaido, Dávid Javorský, Marek Kasztelnik, and 30 others. 2025. Findings of the iwslt 2025 evaluation campaign. In *Proceedings of the 22nd International Conference on Spoken Language Translation (IWSLT 2025)*, Vienna, Austria (in-person and online). Association for Computational Linguistics. To appear.

Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, and 1 others. 2025. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras. *arXiv preprint arXiv:2503.01743*.

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*.

Kshitij Ambilduke, Ben Peters, Sonal Sannigrahi, Anil Keshwani, Tsz Kin Lam, Bruno Martins, Marcely Zanon Boito, and André FT Martins. 2025. From tower to spire: Adding the speech modality to a text-only llm. *arXiv preprint arXiv:2503.10620*.

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

Giuseppe Attanasio, Beatrice Savoldi, Dennis Fucci, and Dirk Hovy. 2024. Twists, humps, and pebbles: Multilingual speech recognition models exhibit gender performance gaps. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21318–21340, Miami, Florida, USA. Association for Computational Linguistics.

Loïc Barrault, Yu-An Chung, Mariano Coria Meglioli, David Dale, Ning Dong, Mark Duppenthaler, Paul-Ambroise Duquenne, Brian Ellis, Hady Elsahar, Justin Haaheim, and 1 others. 2023. Seamless: Multilingual expressive and streaming speech translation. *arXiv preprint arXiv:2312.05187*.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, and 1 others. 2024. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*.

Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. 2023. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models. *arXiv preprint arXiv:2311.07919*.

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.

Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. 2023. High fidelity neural audio compression. *Transactions on Machine Learning Research*. Featured Certification, Reproducibility Certification.

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, and 1 others. 2024. Molmo and pixmo: Open weights and open data for state-of-the-art multimodal models. *arXiv preprint arXiv:2409.17146*.

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? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14760–14778, Bangkok, Thailand. Association for Computational Linguistics.

Daniel Galvez, Greg Diamos, Juan Ciro, Juan Felipe Cerón, Keith Achorn, Anjali Gopi, David Kanter, Maximilian Lam, Mark Mazumder, and Vijay Janapa Reddi. 2021. The people's speech: A large-scale diverse english speech recognition dataset for commercial usage. *Preprint*, arXiv:2111.09344.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, and 1 others. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. 2020. Conformer: Convolution-augmented transformer for speech recognition. *Preprint*, arXiv:2005.08100.

Shujie Hu, Long Zhou, Shujie Liu, Sanyuan Chen, Lingwei Meng, Hongkun Hao, Jing Pan, Xunying Liu, Jinyu Li, Sunit Sivasankaran, Linquan Liu, and Furu Wei. 2024. WavLLM: Towards robust and adaptive speech large language model. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4552–4572, Miami, Florida, USA. Association for Computational Linguistics.

Zhichao Huang, Rong Ye, Tom Ko, Qianqian Dong, Shanbo Cheng, Mingxuan Wang, and Hang Li. 2023. Speech translation with large language models: An industrial practice. *arXiv preprint arXiv:2312.13585*.

Alkis Koudounas, Eliana Pastor, Giuseppe Attanasio, Vittorio Mazzia, Manuel Giollo, Thomas Gueudre, Elisa Reale, Luca Cagliero, Sandro Cumani, Luca de Alfaro, and 1 others. 2024. Towards comprehensive subgroup performance analysis in speech models. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:1468–1480.

Chia-Hsuan Lee, Szu-Lin Wu, Chi-Liang Liu, and Hung-yi Lee. 2018. Spoken squad: A study of mitigating the impact of speech recognition errors on listening comprehension. *Proc. Interspeech 2018*, pages 3459–3463.

Zhenyan Lu, Xiang Li, Dongqi Cai, Rongjie Yi, Fangming Liu, Xiwen Zhang, Nicholas D Lane, and Mengwei Xu. 2024. Small language models: Survey, measurements, and insights. *CoRR*.

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, Pierre Colombo, Barry Haddow, José G. C. de Souza, Alexandra Birch, and André F. T. Martins. 2024. Eurollm: Multilingual language models for europe. *Preprint*, arXiv:2409.16235.

NLLB Team, Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia-Gonzalez, Prangthip Hansanti, and 20 others. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

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.

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. 2022. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*.

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. 2022. 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.

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, and 1 others. 2023. Audiopalm: A large language model that can speak and listen. *arXiv preprint arXiv:2306.12925*.

Sonal Sannigrahi, Thiago Fraga-Silva, Youssef Oualil, and Christophe Van Gysel. 2024. Synthetic query generation using large language models for virtual assistants. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2837–2841.

Kai Shen, Zeqian Ju, Xu Tan, Eric Liu, Yichong Leng, Lei He, Tao Qin, Sheng Zhao, and Jiang Bian. 2024. Naturalspeech 2: Latent diffusion models are natural and zero-shot speech and singing synthesizers. In *The Twelfth International Conference on Learning Representations*.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. 2023. Salmonn: Towards generic hearing abilities for large language models. *arXiv preprint arXiv:2310.13289*.

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

Changhan Wang, Anne Wu, and Juan Pino. 2020. Covost 2 and massively multilingual speech-to-text translation. *arXiv preprint arXiv:2007.10310*.

Tianrui Wang, Long Zhou, Ziqiang Zhang, Yu Wu, Shujie Liu, Yashesh Gaur, Zhuo Chen, Jinyu Li, and Furu Wei. 2023. Viola: Unified codec language models for speech recognition, synthesis, and translation. *arXiv e-prints*, pages arXiv–2305.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 22 others. 2024a. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Hao Yang, Lizhen Qu, Ehsan Shareghi, and Reza Haf. 2024b. Towards probing speech-specific risks in large multimodal models: A taxonomy, benchmark, and insights. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10957–10973, Miami, Florida, USA. Association for Computational Linguistics.

Xin Zhang, Dong Zhang, Shimin Li, Yaqian Zhou, and Xipeng Qiu. 2024. Speechtokenizer: Unified speech tokenizer for speech language models. In *ICLR*.