SpeechLM for Automatic Speech Recognition in Low-resource Languages

Md Abdur Razzaq Riyadh[1], **Eneko Agirre**[1], **Eva Navas**[1], **Claudia Borg**[2]
[1]HiTZ Center, University of the Basque Country, Spain
[2]Dept. of Artificial Intelligence, University of Malta, Malta
{mdabdurrazzaq.riyadh,e.agirre,eva.navas}@ehu.eus, claudia.borg@um.edu.mt

Abstract

Multi-modal Speech Language Models (SpeechLMs) are a recent advancement in natural language processing. These SpeechLMs are instruction-tuned and optimized for general tasks. Their usefulness for Automatic Speech Recognition (ASR), particularly in relatively low-resource scenarios, remains largely understudied. This work developed SpeechLM for ASR in Basque and Maltese and studied the impact of language-adapted Large Language Model (LLM) and speech encoder within the SpeechLM for ASR. Using supervised learning, we fine-tuned LLaMA-Omni, a SpeechLM, for ASR. We have conducted comprehensive hyperparameter tuning and experimented with language-adapted SpeechLM components to improve performance and evaluated our best models on in-distribution datasets for both languages and an out-of-distribution dataset for Basque. LLaMA-Omni achieved 8.09% WER in Basque and 25.65% WER for Maltese on average across multiple test splits. The in-distribution results show that SpeechLM outperforms a fine-tuned ASR system under specific constraints, whereas it underperforms the baseline model on out-of-distribution Basque, indicating weaker overall robustness. We also find that a language-adapted LLM within SpeechLM improves in out-of-distribution settings when compared to the off-the-shelf LLM within SpeechLM.

Keywords: SpeechLM, ASR, LLM, Low-resource, Basque, Maltese

1. Introduction

ASR is the computational process of converting spoken language into written text, enabling machines to interpret and respond to human speech. It is an important task in natural language processing as speech is a natural and prevalent way of communication, while text is the more common modality of information processing. To bridge these modalities, SpeechLMs as a research field focus on seamless interaction between speech and language models, where the language model has an intrinsic capability to understand and generate speech. To directly understand speech, SpeechLMs encode raw audio signals or waveforms and convert them into discrete tokens or continuous representations. In this study, we adapt a SpeechLM to leverage its speech recognition capabilities, focusing primarily on the development and evaluation of ASR systems with a particular emphasis on the Basque and Maltese languages. The goal of this work is to analyze whether a general-purpose SpeechLM can be effectively adapted for supervised ASR in under-resourced languages, and to understand which components have a greater impact in different data regimes. Specifically, we have worked with LLaMA-Omni (Fang et al., 2025), a multi-modal speech language model that supports speech and text modalities as input and can produce output in both speech and text. We fine-tune LLaMA-Omni with supervised learning for ASR in different experimental setups to understand the impact of language-adapted components and evaluate them in terms of accuracy and robustness, where out-of-distribution data was available.

Our work demonstrates that LLaMA-Omni can be adapted for ASR, yielding better results than the fine-tuned speech encoder in some cases. On Basque in-distribution test sets, LLaMA-Omni outperforms the fine-tuned baseline model, `whisper-large-v3` (the speech encoder used within LLaMA-Omni) by 22%, demonstrating the effectiveness of SpeechLM architecture. However, the fine-tuned baseline outperforms LLaMA-Omni significantly on out-of-distribution evaluation. In Maltese, the performance decreases by 50% on the in-distribution test set due to a significantly smaller training dataset, highlighting a limitation in the SpeechLM architecture when handling data scarcity.

2. Related Work

Recent progress in SpeechLM has enabled end-to-end speech understanding and generation without relying on traditional cascaded ASR and text-to-speech pipeline. Among them, LLaMA-Omni (Fang et al., 2025) proposed a novel architecture toward low-latency, high-quality speech understanding and generation by directly mapping speech inputs to both textual and spoken outputs. However, their evaluation focuses exclusively on speech-to-text instruction-following and

Proceedings of SPEAKABLE @ LREC 2026, pages 125–131

speech-to-speech instruction-following tasks in English. This trend is common across SpeechLM research: evaluating on instruction-following generative tasks and downstream tasks such as ASR, text-to-speech, speech translation, etc., in high-resource languages. Although Soundwave (Zhang et al., 2025b) focuses on efficient training with limited resources and employs the same whisper-LLaMA 3.1-8B-Instruct architecture as LLaMA-Omni, its evaluation remains restricted to English speech tasks. EchoX (Zhang et al., 2025a) builds on Soundwave to emphasize acoustic robustness for knowledge-based question answering, again focusing on English. Despite its multilingual capabilities, GLM-4 Voice (Zeng et al., 2024) is restricted to high-resource training and evaluation regimes.

Although interest in SpeechLMs as general-purpose generative models is increasing, their performance on low-resource downstream tasks such as ASR remains largely unexplored. In this paper, we investigate whether recent SpeechLM architectures can be efficiently adapted to ASR for two low-resource languages: Basque and Maltese. We also isolate the effect of language adaptation at different components of the system. For the LLM, we incorporate Latxa Instruct (Etxaniz et al.), an instruction-tuned Basque LLM derived from the LLaMA 3.1-8B-Instruct (Grattafiori et al., 2024). Similarly, we analyze the impact of adapting the speech encoder, whisper. Our results provide the first systematic evaluation of SpeechLM-based ASR in low-resource language settings.

3. Methodology

We aim to investigate the adaptability of SpeechLM for low-resource ASR and further examine the impact of component-specific language-adaptation on ASR performance. Specifically, we address the following research questions: i) Can a general-purpose SpeechLM be adapted for supervised ASR in Basque and Maltese? ii) Does language adaptation of the LLM improve ASR performance in Basque? iii) Does adapting the speech encoder provide additional performance gains? While the first and third research questions focus on both Basque and Maltese, the second one only focuses on Basque, as a language-adapted LLM was not available for Maltese.

In this empirical study, we designed three controlled experiments using LLaMA-Omni to answer each of our research questions. The first experiment aimed to adapt LLaMA-Omni for ASR in Basque and Maltese, separately. Toward that goal, we fine-tuned the already trained English LLaMA-Omni[1] with supervised learning on respective datasets for each language. The goal of this experiment was to examine the effectiveness of adapting a SpeechLM for speech recognition in low-resource languages. LLaMA-Omni has four primary components: speech encoder, speech adaptor, LLM, and speech decoder, as presented in fig. 1. The second and third experiments allowed us to understand the impact of language-adapted LLM and speech encoder within the SpeechLM architecture on its performance in ASR. In the second experiment, we kept the same setting as in the first experiment, but replaced the LLaMA 3.1-8B-Instruct with Latxa Instruct (Etxaniz et al.), a continually pre-trained Basque-adapted LLaMA 3.1-8B-Instruct. We refer to this newly constructed SpeechLM architecture Latxa-Omni. We hypothesize that because Latxa Instruct has a better understanding of Basque, it should perform better at handling Basque-specific transcriptions, thereby improving ASR performance. The final experiment replicated the setup of the first experiment, with the exception that we swap the speech encoder with a version fine-tuned on the same training dataset for Basque and Maltese while keeping the remaining components fixed.

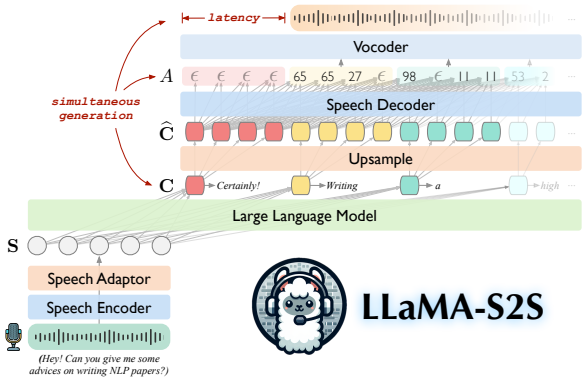

Figure 1: LLaMA-Omni's modular architecture, from Fang et al. (2025)

To study how training data size affects SpeechLM-based ASR, we conducted a control experiment by fine-tuning LLaMA-Omni on a 35-hour subset of Basque speech, matching the amount of training data available for Maltese. By keeping all other aspects identical to the first experiment, we could isolate the impact of extreme data constraints while controlling for language-specific factors.

The SpeechLM LLaMA-Omni combines speech and text into a unified input for the LLM. While the input audio for ASR is provided as speech to the LLM, the text prompt must be predetermined.

[1] https://huggingface.co/ICTNLP/Llama-3.1-8B-Omni/

Our preliminary experiments showed that the specific prompt used was not very important, as long as it was used consistently. For all our experiments, we fixed the prompt to be `<speech> Please directly repeat the sentence in <language>`, where `<speech>` gets replaced by the sequence of speech representation vectors from the speech adaptor and `<language>` is either "Basque" or "Maltese".

Following Fang et al. (2025), only the speech adaptor and the LLM are fine-tuned in all experiments, whereas the speech encoder is kept frozen, including in the third experiment in which it was replaced. The speech decoder was frozen throughout, as it is only relevant for speech generation and is not used in our text-only ASR setup. We conducted a hyperparameter search, focusing on learning rate and batch size. We explored learning rates of $(1 \times 10^{-4},\ 5 \times 10^{-5},\ 2 \times 10^{-5},\ 1 \times 10^{-5},\ 5 \times 10^{-6})$ and batch sizes of $(32, 64, 128, 256)$. All experiments on Basque run for 6 epochs and on Maltese for 4 epochs, a value determined through preliminary trials. Each training run took approximately 30 hours. We used the cross-entropy loss (Goodfellow et al., 2016) optimized with AdamW (Loshchilov and Hutter, 2019) optimizer in combination with a cosine learning rate scheduler and a 5% warm-up phase. During inference, nucleus sampling was used with the temperature set to 0.6 and a probability of p = 0.9.

We used `whisper-large-v3` (Radford et al., 2022) as the baseline for both languages. As `whisper-large-v3` is the underlying speech encoder in LLaMA-Omni, this setup isolates the impact of the SpeechLM architecture on ASR performance. For Basque, the baseline was fine-tuned by Aholab[2], and for Maltese, we fine-tune it ourselves.

4. Data

In this work, we have used multiple datasets, and their summaries are presented in tables 1 and 2 for Basque and Maltese, respectively. For the training, validation, and in-distribution evaluation of Basque speech recognition systems, an aggregated dataset titled Composite Corpus EU v2.1[3] was utilized. This publicly accessible resource compiles subsets from several other publicly available datasets, specifically Common Voice 18 (Ardila et al., 2020), Basque Parliament Speech Corpus 1.0 (Varona et al., 2024), and OpenSLR 76 (Kjartansson et al., 2020). To evaluate the robustness of our system in Basque, we utilized Faktoria dataset, from EJIE[4] for out-of-distribution evaluation. Faktoria is a radio talk show and it differs substantially from our training data both in domain and acoustic characteristics. The recordings include spontaneous conversational speech, background music, and noticeable dialectal variation, making the dataset particularly challenging. The dataset is divided into two subsets: 'Faktoria (EUS)', containing standard Basque, and 'Faktoria (Other)', consisting of samples from various Basque dialects. Collectively, the Basque corpus comprises a total of 676 hours of training data and 63 hours of testing data.

Dataset	Set	Hour	# Samples
CV 18	Train	300.0	198498
	Test	24.0	14312
	Validation	1.0	620
Basque Parl.	Train	370.0	185699
	Test	3.0	1521
	Validation	1.0	550
OpenSLR	Train	6.0	3229
	Test	1.0	526
	Validation	1.0	521
Faktoria	Test (EUS)	27.5	19142
	Test (other)	7.5	4863

Table 1: Summary of the Basque datasets used in this study, including Common Voice 18, Basque Parliament Speech Corpus 1.0, OpenSLR, and Faktoria test sets.

Dataset	Set	Hour	# Samples
Headset v2	Train	6.50	4979
Farfield	Train	9.50	4476
Booths	Train	2.50	1256
MEP	Train	1.25	656
TUBE	Train	13.25	8956
	Test	1.00	663
	Validation	1.00	638
CV 22	Train	2.37	1910
	Test	2.36	1661
	Validation	2.09	1625

Table 2: Maltese datasets used in this study. Datasets Headset v2, Farfield, Booths, MEP, and TUBE subsets are from the MASRI corpus.

The primary training dataset for Maltese was a speech-text parallel corpus, titled 'MASRI' (Hernandez Mena et al., 2020). It is a publicly accessible resource[5] developed to advance research in Maltese ASR. Along with MASRI, we also used Maltese data from Common Voice 22. There was no out-of-distribution dataset readily available for Maltese. In total, the Maltese datasets provide

[2] https://huggingface.co/HiTZ/whisper-large-v3-eu

[3] https://huggingface.co/datasets/asierhv/composite_corpus_eu_v2.1

[4] https://www.ejie.euskadi.eus

[5] https://www.um.edu.mt/projects/masri

35.37 hours of training material and 3.36 hours for testing.

5. Results

The goal of this work is to investigate the adaptability of SpeechLM for ASR in Basque and Maltese. To that end, we use Word Error Rate (WER) for evaluation, where lower values indicate better ASR performance. The baselines for both languages were fine-tuned on the same training set for respective languages, referred in section 4. We also evaluate them on the same test sets from tables 1 and 2, reported in tables 3 and 4, and their performance fluctuates by language, yielding an WER of 10.43% for Basque and 16.74% for Maltese. For all model comparisons, we report the average over multiple splits.

5.1. Quantitative Evaluation

Experiment 1: Off-the-shelf Adaptation. The result from the first experiment is presented as 'LLaMA-Omni' in tables 3 and 4. We examined the efficacy of adapting a standard SpeechLM through supervised fine-tuning. In Basque, LLaMA-Omni outperformed the baseline by approximately 22%. Conversely, the Maltese model exhibited a performance degradation of roughly 53%, suggesting that the architectural adaptation is highly sensitive to the target language or available data volume.

Experiment 2: Language-adapted LLMs. The second question explores whether replacing the LLM in the SpeechLM with a Basque-specific, instruction-tuned LLM leads to improved ASR performance. As shown in table 3, Latxa-Omni achieved the best overall performance with a WER of 7.98%, maintaining a significant lead over the baseline, though showing only marginal gains over the standard LLaMA-Omni in in-distribution settings.

Experiment 3: Adapted Speech Encoders. The third experiment explores the impact of a language-adapted speech encoder within the SpeechLM architecture. This experiment's result is presented as 'LLaMA-Omni (+ adapted SE)' in tables 3 and 4 which replaces the speech encoder adapted for each language. Integrating a language-adapted speech encoder yielded mixed results. For Basque, the model performed 21% better than the baseline, mirroring previous experiments. And in Maltese, this variant remained approximately 50% worse than the baseline. This suggests that while speech encoder adaptation can be beneficial, it cannot fully compensate for other bottlenecks in low-resource settings.

Data Scarcity Analysis. To investigate the Maltese performance gap, we conducted a controlled experiment in the Basque using a reduced corpus (LLaMA-Omni (35H) in table 3) which was sampled randomly. This model exhibited a marked performance degradation, with WER increasing by 76% relative to the baseline and 127% compared to LLaMA-Omni, which was fine-tuned on the entire corpus. These results support the hypothesis that data scarcity is the primary factor limiting the adaptation of SpeechLM architectures for Maltese.

5.2. Out-of-distribution Evaluation

To evaluate the robustness of our systems, we evaluate our models on an out-of-distribution dataset. We could only do this for Basque, utilizing the Faktoria corpus described in section 4. No out-of-distribution datasets were available for Maltese. Initial observations on the results presented in table 5 reveal a notable degradation in all models' performance on Faktoria. Interestingly, the baseline outperforms all the LLaMA-Omni variants from experiments 1-3 on both subsets, which contrasts our results in the in-distribution setting, reported in table 3. Moreover, the substantially higher WER observed on the Faktoria (other) subset by all models, including the baseline, compared to Faktoria (EUS only), clearly demonstrates the models' lack of robustness to dialectal variation. However, models with language-adapted components outperform LLaMA-Omni, while Latxa-Omni has the highest margin of 13% compared to the LLaMA-Omni from the first experiment. These results suggest that LLaMA-Omni negatively impacts generalizability ASR when compared to `whisper-large-v3`. Nevertheless, using language-adapted components generalize better than LLaMA-Omni, demonstrating better robustness. This pattern did not emerge in the in-distribution evaluation for Basque and Maltese.

5.3. Qualitative Analysis

Basque. To better understand the quality of the transcriptions produced across experiments in Basque, we analyzed both the best-performing and worst-performing transcription cases from the test sets, along with the most commonly mistaken words. In cases where a model achieved perfect transcriptions (WER = 0%), we observed that the corresponding audio featured clear pronunciation and minimal background noise. These recordings were typically well-articulated, contributing to the model's ability to recognize and transcribe the speech accurately. On the other hand, in poorly transcribed samples, particularly those with high WER, many samples contained speech that was phonetically ambiguous or acoustically similar to

Model/Corpus	CV 18	Basque Parl.	OpenSLR	Average
whisper-large-v3	5.08%	13.72%	**12.48%**	10.43%
LLaMA-Omni (35H)	16.62%	11.04%	27.45%	18.37%
LLaMA-Omni	4.80%	5.00%	14.47%	8.09%
Latxa-Omni	4.79%	**4.91%**	14.23%	**7.98%**
LLaMA-Omni (+ adapted SE)	4.86%	4.91%	14.78%	8.18%

Table 3: Basque ASR evaluation results for the baseline and LLaMA-Omni models. The first row shows the baseline system; the second shows LLaMA-Omni trained on only 35 hours of Basque data. The remaining rows correspond to Experiments 1-3: LLaMA-Omni fine-tuned on the full corpus, LLaMA-Omni with the LLM substituted, and LLaMA-Omni with an adapted speech encoder. Columns report WER on CV18, Basque Parliament Speech Corpus 1.0, and OpenSLR, along with their average.

Model/Corpus	CV 22	MASRI	Average
whisper-large-v3	**11.44%**	**22.04%**	**16.74%**
LLaMA-Omni	21.08%	30.22%	25.65%
LLaMA-Omni (+ adapted SE)	20.93%	29.33%	25.13%

Table 4: Maltese ASR results for the baseline and LLaMA-Omni experiments. The first row shows the baseline system. The next two rows correspond to Experiment 1 (LLaMA-Omni fine-tuning) and Experiment 3 (LLaMA-Omni with an adapted speech encoder). Columns report WER on CV22, MASRI, and their average.

other words, making it difficult for the model to differentiate. This suggests that the language model has difficulty handling words that sound alike, especially when there is little context to help disambiguate them. The transcription quality of Latxa-Omni and LLaMA-Omni (+ adapted SE) model is highly comparable to our fine-tuned 'LLaMA-Omni (2e-5, 256)'. Both models tend to struggle with a similar set of samples, primarily those containing phonetically close or noisy audio. In several samples, Latxa-Omni even performed worse than LLaMA-Omni. Interestingly, many of the word substitutions made by Latxa-Omni overlap with those observed in LLaMA-Omni's outputs, particularly for phonetically similar words.

Maltese. To understand the transcription quality in Maltese, we compare LLaMA-Omni and LLaMA-Omni (+ adapted SE) and find that the transcription quality is quite similar. Both model struggle with code-switching and transcribe phonetically instead of using the original spelling. The original transcriptions in the MASRI dataset contain filler words, but our fine-tuned models generally ignore them. The most common errors involve confusion between phonetically similar words. Another source of errors involves phonetic mis-segmentation, producing unintelligible tokens. These errors often take the shape of Italian-like morphological forms, reflecting the influence of contact language on the model's output. For example, "Ħalli niddistingwu" is transcribed as "Hallini d-distingu" by LLaMA-Omni. This example illustrates that while the speech encoder is able to phonetically capture Maltese sounds, the limited size of the dataset constrains LLM's ability to acquire and generalize Maltese grammatical rules.

6. Conclusions

This work investigated the adaptability of SpeechLM for ASR in the low-resource languages of Basque and Maltese, specifically examining how architectural components and data scale influence model performance. Our findings reveal that while LLaMA-Omni is highly effective and achieves comparable performance to traditional baselines when sufficient data is available, it is significantly more sensitive to data scarcity. The performance degradation observed in Maltese and the reduced-corpus Basque experiments suggests that the large-scale architecture of LLaMA-Omni requires a higher data size threshold for effective supervised fine-tuning than specialized ASR models.

The study further demonstrated a marked improvement in generalization to out-of-distribution samples and dialectal variations for Basque over off-the-shelf fine-tuning. This indicates that, in low-resource and linguistically diverse scenarios, the LLM plays a more critical role in handling distribution shifts than acoustic features alone. Despite these advancements, the reliance on substantial labeled datasets remains a limitation for truly low-resource scenarios. Future work should focus on optimizing adapter designs and exploring cross-lingual transfer mechanisms to reduce the data burden for SpeechLM adaptation. Finally, this research underscores that advancing ASR for under-resourced languages necessitates a shift toward

Model/Corpus	Faktoria (EUS only)	Faktoria (others)
whisper-large-v3	**18.56**%	**29.42**%
LLaMA-Omni	25.23%	54.81%
Latxa-Omni	21.87%	31.54%
LLaMA-Omni (+ adapted SE)	22.88%	45.14%

Table 5: Out-of-distribution (OOD) evaluation result for Basque. The first row is the baseline model's OOD evaluation. The remaining are models from the first, second, and third experiment respectively.

a more holistic integration of language-specific linguistic knowledge within the SpeechLM framework.

7. Acknowledgments

This work was partially supported by the Erasmus Mundus Master's Programme in Language and Communication Technologies (LCT), the Basque Government (IKER-GAITU project), the Spanish Government (Project AIA2025-163322-C61 funded by MICIU/AEI/10.13039/501100011033).

8. Bibliographical References

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.

Julen Etxaniz, Oscar Sainz, Naiara Miguel, Itziar Aldabe, German Rigau, Eneko Agirre, Aitor Ormazabal, Mikel Artetxe, and Aitor Soroa. Latxa: An open language model and evaluation suite for basque. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14952–14972. Association for Computational Linguistics.

Qingkai Fang, Shoutao Guo, Yan Zhou, Zhengrui Ma, Shaolei Zhang, and Yang Feng. 2025. Llama-omni: Seamless speech interaction with large language models. In *International Conference on Representation Learning*, volume 2025, pages 57607–57624.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning*. MIT Press. http://www.deeplearningbook.org.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, et al. 2024. The Llama 3 Herd of Models.

Carlos Daniel Hernandez Mena, Albert Gatt, Andrea DeMarco, Claudia Borg, Lonneke van der Plas, Amanda Muscat, and Ian Padovani. 2020. MASRI-HEADSET: A Maltese corpus for speech recognition. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6381–6388, Marseille, France. European Language Resources Association.

Oddur Kjartansson, Alexander Gutkin, Alena Butryna, Isin Demirsahin, and Clara Rivera. 2020. Open-Source High Quality Speech Datasets for Basque, Catalan and Galician. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced Languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 21–27, Marseille, France. European Language Resources association.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust Speech Recognition via Large-Scale Weak Supervision.

Amparo Varona, Mikel Penagarikano, Germán Bordel, and Luis Javier Rodriguez-Fuentes. 2024. A Bilingual Basque–Spanish Dataset of Parliamentary Sessions for the Development and Evaluation of Speech Technology. *Applied Sciences*, 14(5):1951.

Aohan Zeng, Zhengxiao Du, Mingdao Liu, Kedong Wang, Shengmin Jiang, Lei Zhao, Yuxiao Dong, and Jie Tang. 2024. GLM-4-Voice: Towards Intelligent and Human-Like End-to-End Spoken Chatbot.

Yuhao Zhang, Yuhao Du, Zhanchen Dai, Xiangnan Ma, Kaiqi Kou, Benyou Wang, and Haizhou

Li. 2025a. Echox: Towards mitigating acoustic-semantic gap via echo training for speech-to-speech llms.

Yuhao Zhang, Zhiheng Liu, Fan Bu, Ruiyu Zhang, Benyou Wang, and Haizhou Li. 2025b. Soundwave: Less is More for Speech-Text Alignment in LLMs.