

Text4Speech: Text-Driven Training for Massively Multilingual Speech Understanding

Hannah Sterz¹ Anna Korhonen¹ Ivan Vulić¹ Goran Glavaš²

¹University of Cambridge ²University of Würzburg

Abstract

Speech is a primary communication interface for humans, and thus a crucial modality for large language models (LLMs). Despite this, the vast majority of open speech-capable LLMs are developed for and evaluated on a small set of < 20 high-resource languages, adding to the already massive inequity of modern language technology. Massively multilingual speech understanding is primarily hindered by the limited amounts of speech training data for most languages and tasks. For the lion’s share of the world’s languages, speech datasets for tasks other than automatic speech recognition (ASR) simply do not exist. In this work, we bypass the sparsity of task-specific speech data by leveraging text—much more available for many languages—to enable an LLM to understand speech input. Our approach, dubbed Text4Speech, leverages state-of-the-art multilingual pre-aligned text and speech encoders: we first align the multilingual text encoder to an LLM using abundant text data and then further align the speech encoder to the LLM using a limited amount of speech. In both steps, we train the alignment parameters via self-distillation on the LLM. Our evaluation on three multilingual speech understanding benchmarks that encompass 90 languages renders Text4Speech highly effective, yielding average improvements of up to 6% on Fleurs-SIB over a very strong ASR-based cascade baseline. Crucially, we show that our text-based alignment unlocks a highly sample-efficient alignment of the speech encoder to the LLM.

1 Introduction

As a common and innate vehicle of human communication, speech is arguably the more intuitive of the two language modalities for interacting with LLMs. Accordingly, recent years have seen a surge of interest in speech-capable LLMs, boasting the ability to directly process spoken input, e.g., Qwen-Audio (Chu et al., 2024), Voxtral (Liu et al., 2025),

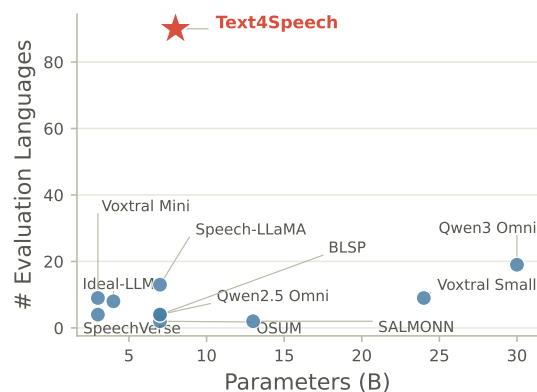


Figure 1: The landscape of open-weight speech-capable LLMs: plotting model size against the number of languages on which the model was empirically evaluated.

and Gemini (Comanici et al., 2025). These models unlock a range of applications, from voice assistants and spoken question answering to real-time speech translation (to speech or to text).

While some of these models (claim to) support a wide variety of languages, open-weight models in particular (i) target only ASR and translation tasks (Barrault et al., 2023; Omnilingual ASR Team et al., 2025) and/or (ii) are evaluated for a small set—typically less than 20—of high-resource languages (Dauvet et al., 2025; Liu et al., 2025). We illustrate the current landscape of open-weight speech-capable LLMs in Figure 1, contrasting the models’ size against the number of languages on which they were evaluated. The utility and accessibility of speech-capable LLMs are hindered by this limited set of languages that they (are empirically shown to) support. Their ability to support a diverse range of languages for a broad set of speech understanding tasks is constrained by the availability of (task-specific) speech data. Whereas high-quality text data is available across a wide range of languages and tasks, annotated speech data remains comparatively scarce, especially for generally mid-

resource and low-resource languages (Imam et al., 2025; Nagano et al., 2025).

In this work, we propose to circumvent this sparsity of multilingual multi-task speech data by leveraging the readily available pre-aligned massively multilingual speech-and-text encoders such as Seamlessm4T (Barrault et al., 2023). Illustrated in Figure 2, our proposed approach Text4Speech first aligns the text encoder of these massively multilingual models to an LLM, using the abundant text data. In the second step, we perform a very lightweight alignment of the speech encoder to the LLM, using the limited speech data: merely 2000 speech samples from only 10 languages. In both steps, we train the alignment by means of self-distillation on the LLM. We draw inspiration from the body of work that leverages pre-aligned audio-and-text encoders (Elizalde et al., 2023) (i.e., the so-called CLAP models) to enable LLM-based audio captioning using (only or predominantly) text data for alignment training (Kouzelis and Katsouros, 2023; Deshmukh et al., 2024; Wang et al., 2025). The key difference, however, is that the CLAP models—like the vision-and-language CLIP encoders (Radford et al., 2021) that inspired them—are sequence-level encoders that output a single representation (i.e., embedding) for the whole audio (or text) input. In contrast, speech encoders typically produce a sequence of speech embeddings (i.e., “speech tokens”) rather than a single representation, aligning with the autoregressive nature of the LLMs, operating on a tokenized (preceding) context. This makes aligning speech encoders to LLMs more challenging.

We thoroughly evaluate Text4Speech: (i) with two different LLM backbones, (ii) on several multilingual benchmarks covering translation and speech understanding—question answering and classification, (iii) encompassing a set of 90 diverse languages. This makes our Text4Speech variants the most multilingually evaluated speech-capable LLMs, by a wide margin (see Figure 1). The experimental results show that Text4Speech substantially outperforms both (1) open speech-capable LLMs of similar size and (2) the intuitive and competitive cascade approach that resorts to ASR to provide a text-only input to the (same) LLM. Crucially, we demonstrate that our text-based alignment unlocks a highly sample-efficient speech alignment: Text4Speech largely outperforms the baselines after being exposed to as few as 2000 speech samples from merely 10 different languages.

2 Background and Related Work

Speech-Capable LLMs. The great success of instruction-tuned text-only LLMs expectedly triggered efforts on extending them to be able to process input in other modalities, including speech (Tang et al., 2024; Chu et al., 2024; Hu et al., 2024; Chen et al., 2024; Ghosh et al., 2025; Goel et al., 2025; Zhang et al., 2025b). Most speech-capable LLMs, i.e., LLMs that can process speech input, heavily rely on transfer learning and share a common architecture: they align the representations of an existing, pre-trained speech encoder to an existing, pre-trained LLM. The parameters of an alignment adapter, a feed-forward network that learns to map the speech embeddings into the input embedding space of the LLM, are trained on speech-text aligned data; the LLM itself is typically also updated in a parameter-efficient manner, using LoRA adapters (Hu et al., 2022). To this end, many of the works resort to the Whisper’s massively multilingual speech encoder (Tang et al., 2024; Chen et al., 2024; Chu et al., 2024; Goel et al., 2025). These open speech-capable LLMs are evaluated on a small number (< 20) of high-resource languages (see Figure 1), which is somewhat surprising, considering (a) Whisper’s support of 97 languages and (b) the existence of massively multilingual speech understanding benchmarks (Keller and Glavaš, 2025; Schmidt et al., 2025).

Text Training for Audio and Speech Models

The emergence of CLAP (Elizalde et al., 2023), a model consisting of an audio and a text encoder with aligned sequence-level representations, triggered the efforts that leverage text data in order to align the audio encoder to an LLM, for the purposes of general audio captioning and understanding. Kouzelis and Katsouros (2023) performs automated audio captioning (ACC) with text-only training. During training, representations are generated by the CLAP’s text encoder; at inference, the encoders are swapped, and the model uses the representations from the CLAP’s audio encoder. Despite the CLAP’s pre-alignment of the two encoders on large corpus of captioned audio, the audio and text embeddings are still located in different regions of the embedding space, a phenomenon known as the modality gap. To mitigate the modality gap, Kouzelis and Katsouros (2023) couple an affine shifting of the (mean of the) audio embeddings towards the (mean of) text embeddings with a specialized decoding strategy. Other text-only AAC ap-

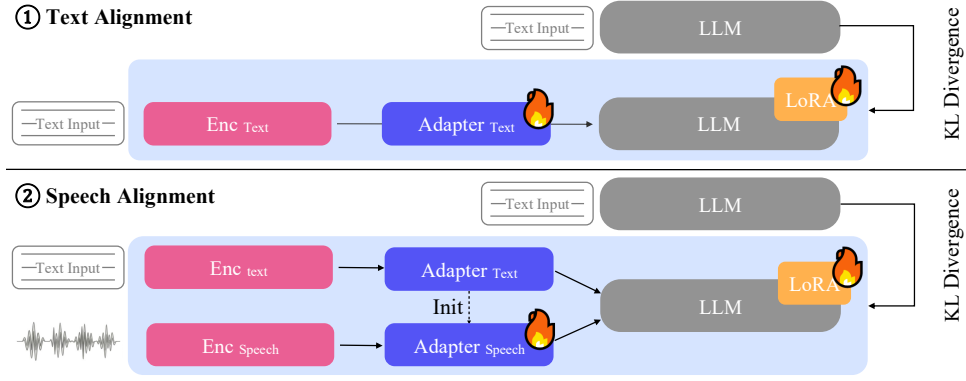


Figure 2: Illustration of Text4Speech. ① Aligning the text encoder of SeamlessM4T to the backbone LLM using a multilingual text-based instruction-tuning mix; the pre-alignment of the speech and text encoder of SeamlessM4T allows us to treat its text embeddings as pseudo-speech. ② Aligning the SeamlessM4T speech encoder to the backbone LLM by fine-tuning the alignment adapter learned in the previous step on limited-size multilingual ASR data. In both steps, we train via self-distillation, with the frozen backbone LLM as the teacher.

proaches utilize Gaussian noise and a linear adapter to bridge the gap between text and audio embeddings (Deshmukh et al., 2024), projection-based decoding in combination with a retrieval system that provides similar captions (Li et al., 2025), or rely on soft and hard prompts (Zhang et al., 2025a). Wang et al. (2025) extend this from audio captioning to an audio-language LLM capable of a variety of tasks. All of these approaches rely on CLAP’s pre-aligned text and audio encoders, which produce sequence-level embeddings, i.e., a single embedding for the entire input. The aggregated sequence-level representations make them apt for general audio, but inherently unfit for aligning speech to LLMs, where encoding of fine-grained semantic meaning of what was said is required.

TASU (Peng et al., 2025) uses an audio/speech encoder that outputs a sequence of audio embeddings. They use the corresponding tokenizer as the text encoder and modify the text embeddings via insertions, deletions, and noise to imitate the Label Synchronous Decoding posterior distribution of the audio encoder. While they report competitive ASR performance for English, they do not attempt to train a general-purpose speech-capable LLM. Speechless (Dao et al., 2025) introduces a text-only training strategy to align a speech encoder to an LLM, however, it requires extensive (>100h) training of an encoder that mimics the speech representations on ASR data. Similarly, LegoSLM (Ma et al., 2025) aligns a speech encoder to the LLM by mapping the encoder output to the token distribution using CTC. This mapping requires extensive training on ASR data. TESU-LLM (Kim and Ko,

2025) aligns Seamless encoder and decoder to an LLM. The decoder autoregressively decodes the sequence; thus, essentially much more similar to running ASR followed by an LLM on the text output. Fang et al. (2025) address a different scenario. They assume that the speech encoder is already aligned to the LLM and show that text data for new domains transfers to speech as well, meaning that training on text-only data from a new domain also improves the speech capabilities in that domain. AlignFormer (Fan et al., 2025) adapts the mapping from speech encoder to LLM to reduce the length mismatch between text and speech sequences.

3 Text4Speech: Text-Based Training for Speech Understanding

The central idea of Text4Speech, as illustrated in Figure 2, is as follows: we exploit the representational pre-alignment of the text and speech encoder of the massively multilingual translation-and-ASR models, such as SeamlessM4T (Barrault et al., 2023), to align the speech encoder to LLM primarily using abundant multilingual task-mix data, with the text encoder as the mediator of that alignment. If the two encoders were “ideally” (pre-)aligned, one could simply learn to align the text encoder to an LLM, using only text data, and then swap it with the speech encoder at inference time. In reality, the pre-trained encoders still exhibit a significant modality gap (see Figure 8), preventing such zero-shot swapping. We thus carry out the alignment in two stages:

① We align the SeamlessM4T’s multilingual *text* encoder to a backbone LLM, using a multilingual

instruction-tuning mix; the text-only data required for this training step is abundant.

② We then mitigate the modality gap by aligning the SeamlessM4T’s multilingual *speech* encoder to the LLM, using small-sized ASR data.

The alignment of the speech encoder in the second step is highly sample-efficient due to: (a) the pre-alignment of the speech and text encoder, and, crucially, (b) our initialization of the alignment adapter for the speech encoder with the parameters of the alignment adapter of the text encoder (see the dashed arrow in Figure 2), obtained in the first step. The second training step thus merely fine-tunes the alignment adapter learned on text data only, to account for the modality gap between the SeamlessM4T’s encoders. Notably (see §4), Text4Speech yields substantial empirical gains for many low-resource languages only seen in ①, for which no speech data was used in ②, validating the intended modality-gap-mitigating nature of ②.

3.1 Aligning the Text Encoder

We make use of large-scale, massively multilingual, and openly available text-only datasets to align the SeamlessM4T’s text encoder (TE) to a backbone LLM. This alignment step might seem redundant at first glance, since the LLM can already process text input. Our goal, however, is not to align the TE to the LLM. Instead, the TE here serves as a stand-in for its corresponding speech encoder (SE); we aim to capitalize on the fact that their representation spaces have been pre-aligned via massively multilingual ASR and translation training (Barrault et al., 2023). Thus, we term the text passed through the SeamlessM4T’s TE *pseudo-speech*, to distinguish it from text processed by the LLM’s tokenizer.

We can align the TE to an LLM using the abundant multilingual text data, whereas directly aligning the SE to the LLM would require a large, massively multilingual speech-text task-mix, which is much more difficult to obtain. We align the TE to the LLM via an *alignment adapter*, a trainable multi-layer feed-forward net. The task of the alignment adapter is to learn to transform the output embeddings from the TE into the input embedding space of the LLM. Conveniently, the backbone LLM processes text input too and can thus directly serve as a teacher in a self-distillation training procedure. We train the student model to produce the labels generated by the frozen teacher. Let x and x' be the sequences of tokens produced from the given text instance by the LLM’s and TE’s

tokenizers, respectively, let t be LLM-tokenized sequence of tokens of the task description, and $y = \{y_1, y_2, \dots, y_n\}$ the reference response text consisting of n tokens. Our self-distillation training loss (for one example), is then given as follows:

$$\mathcal{L}_{\text{text}} = \sum_{i=1}^n \text{KL} \left(P_{\text{teacher}}(y_i | x, t, \{y\}_0^{i-1}) \parallel P_{\text{student}}(y_i | x', t, \{y\}_0^{i-1}) \right)$$

where KL denotes the Kullback-Leibler divergence, the teacher is the frozen backbone LLM, i.e., $P_{\text{teacher}} = \text{LLM}(x, t, \{y\}_0^{i-1})$; and the student is the same LLM (*with LoRA adapters), but with the token embeddings produced by the composition of the SeamlessM4T’s TE (with its own tokenizer) and the corresponding alignment adapter MLP^T : $P_{\text{student}} = \text{LLM}^*(\text{MLP}^T(\text{TE}(x')), t, \{y\}_0^{i-1})$.

Training Tasks. We train the text alignment adapter and the LLMs’ LoRAs on three groups of training tasks: (i) standard massively multilingual prompt-response instruction-tuning, (ii) machine translation, and (iii) repetition and paraphrasing of the input text. The repetition is meant to mimic ASR, i.e., transcription of speech into text, which is commonly used for training speech-capable LLMs. In our case, the student LLM receives pseudo-speech, that is, text input processed with SeamlessM4T’s text encoder, which—due to SeamlessM4T’s pre-alignment—is meant to mimic the output of its speech encoder, but with text as the input. Forcing the LLM to repeat the text input to the SeamlessM4T’s text encoder thus trains the LLM for fine-grained semantic understanding of the pseudo-speech tokens, i.e., embeddings output by the TE. This, assuming sufficient alignment between the pseudo-speech embeddings (i.e., output of the TE) and real speech embeddings (i.e., output of the SeamlessM4T’s speech encoder, SE), will increase the LLMs’ ability to understand true speech tokens (output of SE). The shortcoming of repetition is that it is too easy a task for the teacher model, leading to low entropy of P_{teacher} and thus limited information gain for the student. To remedy this, we task the LLM with paraphrasing the text that was input to the TE instead of repeating it; p_{para} is the probability with which we replace repetition with paraphrasing in this training phase.

3.2 Speech Alignment

In the second step, we explicitly align the SeamlessM4T’s SE to the LLM. The goal is to miti-

gate the modality gap between the SE and TE so that we can benefit from prior large-scale multilingual TE alignment: this way, we hope to remove large-scale massively multilingual speech data as a precondition for obtaining performant multilingual speech-capable LLMs. To benefit from our multilingual TE alignment, we initialize the alignment adapter for the SE with the parameters of the previously learned adapter of the TE. We then use a small-scale ASR dataset to fine-tune the SE’s adapter and LLM’s LoRAs. This adjusts the expensively trained text alignment, carried out on pseudo-speech representations, for actual speech representations, which, due to the modality gap, are initially located in a different subspace of SeamlessM4T’s representation space.

Analogous to the TE alignment phase, we train on a mixture of exact speech transcription and paraphrasing (ratio determined with the probability p_{para}), again via a self-distillation objective. Self-distillation has an advantage over, for instance supervised fine tuning, when the goal is sample-efficiency, since matching the teacher’s full output distribution represents a richer training signal than enables more sample-efficient training. This time, however, we have two students: (1) the pseudo-speech student from the ① alignment and (2) the speech student, which is the LLM aligned to the SE. We include (1) to ensure the alignment of the encoders to the LLM stays compatible with pseudo-speech input so that the capabilities learned on pseudo-speech will transfer to speech. This loss is reducing the risk of catastrophic forgetting and overfitting to the ASR task. Let s be the sequence of speech tokens, output of the SE, and x and x' be the token sequences of the corresponding text (e.g., transcription of the speech), obtained with the LLM’s and TE’s tokenizer, respectively (as before); let t be task instruction tokens. The loss is then $\text{KL}(P_t \| P_{\text{speech}}) + \text{KL}(P_t \| P_{\text{text}})$, with:

$$\begin{aligned} P_t &= \text{LLM}(x, t, \{y\}_0^{i-1}), \\ P_{\text{text}} &= \text{LLM}^*(\text{MLP}_T(\text{TE}(x', t)), \{y\}_0^{i-1}), \\ P_{\text{speech}} &= \text{LLM}^*(\text{MLP}_S(\text{SE}(s)), t, \{y\}_0^{i-1}); \end{aligned}$$

where $*$ indicates the LLM with LoRA adapters, and MLP_S the alignment adapter of the SE. The LoRA adapters (shared between the students) and the parameters of MLP_S (initialized from MLP_T) are the only trainable parameters in this phase. The MLP and LoRA are initialized from the text stage to use the pre-aligned representations and increase the sample-efficiency of the speech training stage.

4 Evaluation

4.1 Experimental Setup

Training Data. To align the TE, we train on three types of tasks (see §3.1). For (i) **repetition and paraphrasing**, we source Wikipedia text across 78 languages that are supported by Seamless. For (ii) **translation**, we use FineTranslations corpora (Penedo et al., 2026), sampling instances that cover 92 languages. For each language X , we train by translating 20% of instances from English to X and the remaining 80% from X to English. Finally, we use (iii) the general-purpose massively multilingual **instruction tuning** dataset Aya (Singh et al., 2024), spanning 65 languages. We provide a detailed language breakdown of our training corpora in Table 8 in the Appendix. For SE alignment, we exclusively train on **ASR** data (see §3.2): unless specified otherwise, we sample 10K instances from the FLEURS dataset (Conneau et al., 2022) across 10 languages: English, German, French, Chinese, Korean, Vietnamese, Japanese, Russian, Turkish, and Swahili. In both alignment steps, we randomly sample whether the instance is to be repeated (for SE alignment, transcribed) or paraphrased, with p_{para} as the probability of choosing paraphrasing.

Backbones and Training Details. We align the TE and SE of the Seamless M4T v2 large (Barraut et al., 2023) to two instruction-tuned LLM backbones: Qwen-3 4B (Yang et al., 2025) and Gemma-3 4B (Team et al., 2025). In both alignment steps, we train for 1 epoch on the respective training mix. The alignment adapters are three-layer MLPs with Swish activation, and the LLM is efficiently fine-tuned with LoRA adapters of rank $r = 8$. Based on results from preliminary experiments, we set p_{para} to 0.6 in TE alignment, and to 0.8 in SE alignment. We provide further details on hyperparameters in the Appendix B.1.

Baselines. 1) Cascades. Our primary baselines are the cascades of ASR and (text-only) LLMs, which use exactly the same models as our Text4Speech variants: (1) the full Seamless-M4T (i.e., speech encoder *and* text decoder) to transcribe speech into text, then forwarded to (2) Qwen or Gemma to perform the task: we denote these with Qwen/Gemma-ASR. This cascade approach is common in practice; for languages with strong ASR performance, it represents a very competitive baseline.¹ To estimate

¹The cascade is significantly more computationally expensive than Text4Speech: next to speech encoding, it additionally

Model	#P	M	Fleurs-Belebele		Fleurs-SIB		SpeechTaxi		AVG all n = 90
			eng n = 1	other n = 83	eng n = 1	other n = 89	eng n = 1	other n = 17	
Qwen 2.5 7B Omni	7B	S	91.7	49.9	87.0	58.0	46.9	31.0	46.9
Voxtral Mini 3B	3B	S	81.7	51.5	84.7	69.2	45.7	46.0	55.7
Qwen3 4B - Oracle	4B	T	91.8	69.2 $\pm$ 2.0	87.0	76.9 $\pm$ 1.6	63.0	42.6 $\pm$ 3.3	63.4
Qwen3 4B - ASR	6B	S	91.6	65.1 $\pm$ 2.1	88.1	76.4 $\pm$ 1.6	58.0	38.8 $\pm$ 2.8	60.6
Text4Speech-2K	5B	S	91.4	64.2 $\pm$ 2.1	89.3	79.1 $\pm$ 2.4	60.5	48.8 $\pm$ 2.7	64.4
Text4Speech-10K	5B	S	91.6	65.1 $\pm$ 2.1	88.7	82.4 $\pm$ 1.1	61.7	47.3 $\pm$ 2.7	65.3
Gemma3 4B - Oracle	4B	T	87.3	66.5 $\pm$ 1.5	85.9	78.9 $\pm$ 1.0	61.7	44.0 $\pm$ 2.8	63.6
Gemma3 4B - ASR	6B	S	85.9	62.9 $\pm$ 1.6	85.9	77.5 $\pm$ 1.2	46.9	36.9 $\pm$ 2.6	59.4
Text4Speech-2K	5B	S	83.6	59.2 $\pm$ 1.5	87.0	79.8 $\pm$ 2.3	60.5	47.0 $\pm$ 2.7	62.4
Text4Speech-10K	5B	S	83.5	59.2 $\pm$ 2.4	87.6	79.5 $\pm$ 1.1	58.0	45.0 $\pm$ 2.9	61.6

Table 1: Performance of Text4Speech on massively multilingual SLU benchmarks for two LLM backbones, Qwen3 and Gemma3. “-ASR” denotes the respective cascade (ASR+LLM) baseline and “-Oracle” the LLM’s performance on ground-truth speech transcriptions. We report scores for smaller (2K speech samples) and larger (10K speech samples) SE alignment training in step ② of Text4Speech. T=Text and S=Speech indicate the modality (M) of the input. The subscript marks the standard error across languages and the best results on speech input in each column are marked in **bold**.

the contribution of ASR errors on the performance, we additionally evaluate Qwen and Gemma on ground truth transcriptions (Qwen/Gemma-Oracle), essentially defining the performance ceiling for a given LLM. **2) Open speech-capable LLMs.** We also evaluate two open speech-capable LLMs of similar size (i.e., the “small” variants, with 3-7B parameters): Qwen Omni 2.5 (Xu et al., 2025) and Voxtral (Liu et al., 2025), to see how Text4Speech compares against models that directly align a speech encoder to an LLM using orders of magnitude more speech data. **3) Seamless.** For translation, we compare Text4Speech, against the full SeamlessM4T model, predominantly trained for translation, with the aim of determining the extent to which Text4Speech—as a general-purpose speech-understanding model—retains the translation capabilities of SeamlessM4T.

Benchmarks. We measure the performance on two translation benchmarks and three multilingual speech understanding benchmarks.

Translation. We resort to Fleurs (Conneau et al., 2022) and Covost2 (Wang et al., 2020) as multilingual benchmarks widely used to evaluate speech-to-text translation (S2TT), where the speech in the source language is to be translated into target language text. We evaluate (BLEU) both EN→X and X→EN translation.

Speech Understanding. Fleurs-SLU (Schmidt et al., 2025) is a massively multilingual spoken language

understanding benchmark with two distinct tasks: multiple choice question answering based on a provided speech sample (Belebele-Fleurs), and topical classification of speech (SIB-Fleurs). SpeechTaxi (Keller and Glavaš, 2025) is a multilingual benchmark for topical classification of spoken Bible verses. For all three datasets, we measure accuracy for languages supported by SeamlessM4T.

4.2 Results

Speech Understanding. Table 1 summarizes the results on SLU benchmarks and Figure 3 shows Fleurs-SLU scores aggregated per-resource level. Overall, Text4Speech exhibits comparable performance to the cascade (ASR) baseline for English, while outperforming it for other languages. On the classification tasks,² Text4Speech beats the ASR cascade by wide margins: Qwen: +6.0% on Fleurs-SIB, +10.0% on SpeechTaxi; Gemma: +10.1% on SpeechTaxi. This indicates that Text4Speech is much more robust in coarse-level speech understanding (i.e., topic identification) than the ASR cascade. We believe this is due to the importance of lexical cues for the task: imperfect ASR likely distorts the key lexical cues in the cascade approach, Text4Speech avoids this problem by operating in the representation space. To our surprise, some scores on Fleurs-SIB surpass Oracle scores. We attribute this to SeamlessM4T’s representations, which provide sufficient content information to classify the content even though Qwen is not able to. When we evaluate the oracle setup with the

requires text decoding; Text4Speech, in contrast, requires just one forward pass through the SE.

²We provide per-language results on SIB in Figure 7, 6.

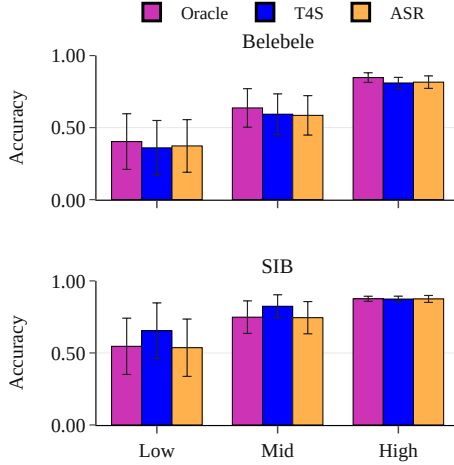


Figure 3: Performance of Qwen variants (Oracle, ASR, and Text4Speech) on SIB-Belebele and SIB-Fleurs aggregated by resource level and the standard deviation across languages.

trained LoRA we observe that the performance of Qwen 3 increases to 78.8 also improving over the oracle setup. This indicates that the LoRA improves the performance at the task.

On SIB (topic classification), the trends vary across languages, see Figure 7 for Qwen. We identify three main language groups. **1)** for high-resource languages (e.g., French, Chinese, Arabic), for which the Qwen-Oracle performance is above 80%, Text4Speech performs close, but slightly below the cascade (ASR) baseline; because of high-quality ASR output for these languages, the cascade performance approaches the ceiling performance of; for high-resource languages, the cascade (Qwen-ASR) is thus a very hard baseline to beat. **2)** languages where the oracle performance is $< 80\%$ but substantially above 40% benefit the most from Text4Speech, compared to the cascade (Qwen-ASR), with huge gains for some languages (e.g., +32 points for Nyanja and Somali, +30 points in Yoruba). These are mostly the languages for which the LLM is (at least somewhat) competent, but lower-quality ASR introduces significant noise. **3)** For languages where the LLM itself fails, both the cascade and Text4Speech perform around 40%.

Across all three benchmarks, Text4Speech consistently outperforms open speech-capable LLMs, Qwen 2.5 Omni 7B and Voxtral Mini 3B, most often with very wide margins for languages other than English (e.g., with Qwen as LLM, +14pp on Belebele, +13pp on SIB). We find this result par-

Model	Fleurs		CoVost2	
	x→en (84)	en→x (48)	x→en (21)	en→x (15)
Seamless	24.1	25.2	36.6	31.7
Qwen - Oracle	22.1	12.3	33.9	12.6
Qwen - ASR	20.4	10.4	30.3	11.5
Text4Speech-2K	23.4	11.5	30.7	11.3
Text4Speech-10K	23.6	11.3	30.8	11.3

Table 2: Translation results on Fleurs and CoVost2.³

ticularly encouraging, as the baseline models have been trained on orders of magnitude more speech.

Translation. Text4Speech is inherently limited by both the SeamlessM4T’s encoding abilities and the backbone LLMs’ ability to reason and generate in the desired language, especially because the LLM itself is the teacher in our self-distillation training. We thus do not expect Text4Speech’s performance to go beyond the LLMs’ performance on oracle transcription. To get a better understanding of how Text4Speech (with Qwen backbone) compares to the performance of Seamless and Qwen, we compare their translation performance in Table 2.

On the $X \rightarrow en$ translations, the translation performance of Text4Speech lies between the text-based translation performance of Qwen and the S2TT performance of SeamlessM4T for most languages. For high-resource languages (e.g., German, French, Spanish), Qwen-Oracle outperforms and, thus, so does Text4Speech, but it stays below the Qwen-Oracle performance. For low-resource languages, the massively multilingual translation-specific pre-training of SeamlessM4T delivers much stronger performance than Qwen-Oracle, so Qwen as the backbone LLM and teacher in self-distillation becomes the limiting factor for Text4Speech’s performance. This shows that Text4Speech retains near-maximal capabilities of its individual components, SeamlessM4T’s speech encoder and Qwen as the backbone LLM.

In $en \rightarrow X$ translation, Qwen itself compares very poorly, so we limit the analysis to languages X for which we observe $BLEU > 0$. Text4Speech overall performs very similarly to Qwen, which is what we hoped to observe since our training extends Qwen’s

³Note that Text4Speech extends Qwen’s understanding capabilities to speech. No part of our training procedure is designed to improve beyond Qwen’s (translation) performance on text. Because SeamlessM4T is a dedicated translation model, with additional translation-specialized decoders, we cannot expect Text4Speech, with Qwen as the backbone, to outperform the full SeamlessM4T on translation. We can only hope to approach the Qwen’s translation performance on ground-truth speech transcription (i.e., Qwen-Oracle).

Model	#P	Belebele		SIB	
		eng $n = 1$	other $n = 83$	eng $n = 1$	other $n = 89$
Gemma 4B - ASR	6B	85.9	62.9	85.9	77.5
Text4Speech-2K	5B	83.6	59.2	87.0	79.8
+ 12B Teacher	5B	81.8	58.8	86.4	77.3
Gemma 12B - ASR	14B	94.0	77.5	90.4	84.4
Text4Speech-2K	13B	93.6	75.9	90.4	83.8

Table 3: Text4Speech performance on Fleurs-SLU for different sizes of the backbone/teacher LLM (Gemma).

input modality to speech, but is in no way designed to improve Qwen’s generation capabilities. Thus the multilingual generation capabilities stay similar to the original LLM.

4.3 Further Analyses and Ablations

Speech-only training. To illustrate that the text training is a crucial step of our setup, we train the model without the text data. We run the speech training phase on 10K samples, the highest number of speech samples in the experiments, training for 3 epochs. After this, the model does not understand the speech input at all. On the SIB task, for instance, the results range from 5% to 26% with an average of 17.9%, which is close to random performance. When we increase the number of epochs to 10 to increase the number of training steps in this setup, we observe that the alignment overfit to the training set, and the SIB scores stay in a similar range. This showcases the importance of the text training data to learn a meaningful alignment such that the speech training data only needs to bridge the modality gap.

Text4Speech with Larger LLMs. We explore how Text4Speech scales with LLM size, using Gemma-3 as the LLM family. We evaluate two more configurations: (1) a larger model, Gemma 12B, as both the backbone for Text4Speech and self-distillation teacher, (2) a small model, Gemma 4B, as the Text4Speech backbone with a larger model, Gemma 12B, as the teacher. Table 3 summarizes the results. Text4Speech with a 4B backbone, but a 12B teacher stays below the performance of a self-distilled variant with the 4B Gemma as teacher and student. This aligns with other work that renders self-distillation highly effective for self-supervised learning (Mansourian et al., 2025). When self-distilled from the larger Gemma 12B as both the teacher and student, Text4Speech improves notably over the 4B variant. However, Gemma 12B is also more robust to ASR noise, so Text4Speech stays slightly below the cascade baseline.

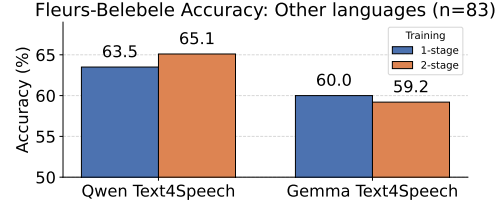


Figure 4: Comparison of Text4Speech with 1-stage vs. 2-stage alignment training on Belebele (average over 83 languages, without English).

Sequential vs. Simultaneous Training. We opt for a 2-step alignment procedure primarily to decouple the expensive, large-scale text-based training from the lightweight speech-based training. In this setup, any modification to the speech training data (e.g., further specialization for a particular language or task) requires only the repetition of the inexpensive SE alignment training. We now, investigate the performance of a single-step alignment procedure, one that *simultaneously* aligns ① the TE to the LLM using large-scale multilingual text data and ② SE to LLM using small-scale ASR data. To mitigate the imbalance in the number of samples between the two modalities, in this 1-stage training, we oversample the ASR training instances. Figure 4 compares the performance of the two alignment procedures (2-stage vs. 1-stage) on Belebele. We observe marginal differences between the two training procedures for Gemma and a small drop for Qwen. Compared to our 2-stage procedure, the 1-stage setup comes with the downside of expensive, large-scale retraining for any modification to the (small-scale) speech data.

Sample-Efficiency of Speech Alignment. We next analyze how much speech/ASR data is really necessary for the second training stage to bridge the modality gap. To this end, we train the SE alignment on 500, 1K, 2K, and 10K samples. We evaluate the S2TT performance on Fleurs and QA performance on Belebele for 20 languages.

Figure 5 summarizes the results. Without any speech-based training (zero-shot modality transfer), the model is not able to produce meaningful output. But with as few as 500 ASR examples used for SE alignment, Text4Speech already performs comparably to the Qwen-ASR baseline; with 1K ASR training samples, Text4Speech matches the translation (S4TT) performance of Qwen-Oracle. Further increase of the ASR training data brings diminishing gains, suggesting that our large-scale multilingual training on pseudo-speech (i.e., text em-

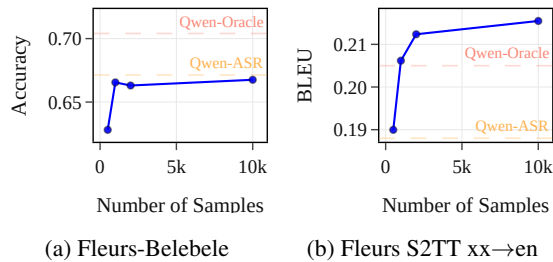


Figure 5: Analysis of speech sample efficiency: Text4Speech performance on Belebele (listening comprehension QA) and Fleurs (S4TT) with different sizes of ASR data used in SE alignment training (step ②).

beddings produced by SeamlessM4T’s TE) brings about impressive sample-efficiency for aligning actual speech representations to the LLM. This also strongly suggests that our second training step only serves to mitigate the modality gap, rather than to establish fine-grained semantic alignment between the output representations of SE and the input of the LLM: the first training stage on large(r)-scale massively multilingual text data achieves that.

5 Conclusion

Massively multilingual speech understanding is hindered by limited amounts of speech training data across languages and tasks, with open speech-capable LLMs evaluated only for small sets (i.e., <20) of high-resource languages. In this work, we introduce Text4Speech, a text-driven training approach to induce massively multilingual speech-capable LLMs from speech data that is very limited in size as well as language coverage. Leveraging the cross-modally aligned speech and text encoders of a massively multilingual speech-to-text translation model, we are able to train much of the alignment to an LLM on text data; a small ASR dataset, used to mitigate the remaining modality gap between the two encoders, then suffices to induce effective massively multilingual speech-capable LLMs. On several speech understanding benchmarks, Text4Speech outperforms the competitive cascade baseline that couples ASR and (text-based) LLM as well as open speech-capable LLMs of similar size, trained on orders of magnitude more speech data. Crucially, we show that our text-based alignment unlocks a highly sample-efficient alignment of the speech encoder to the LLM, with Text4Speech matching the performance of the cascade baseline with as few as 500 speech samples covering only 10 languages.

Limitations

Text4Speech is an approach for aligning a massively multilingual speech encoder to a (multilingual) LLM using very limited amounts of (multilingual) speech data. The method fundamentally relies on pre-aligned text and speech encoders that share a representation space. As such, the approach is inherently limited by the capabilities of the respective speech encoder (in our case SeamlessM4T’s speech encoders) and LLM. We bypass the need for large-scale multilingual speech data by aligning the corresponding SeamlessM4T’s text encoder to the LLM using large massively multilingual text training: this, however, means that the effectiveness of Text4Speech is limited by the quality of the representations coming both out of the SeamlessM4T’s (1) speech as encoder *and* its (2) text encoder. On the one hand, the quality of representations from the encoders defines the extent of useful information passed to the LLM; on the other hand, the capability of the LLM limits the performance additionally, because of its use as the teacher in our distillation training.

Further, using text as “pseudo-speech”, Text4Speech only supports languages for which significant amounts of text data exist. While this is fulfilled for a wide range of languages across all resource levels, Text4Speech cannot be used to support thousands of spoken-only languages.

Acknowledgment

Hannah Sterz thanks the Cambridge Trust for their support via the International Scholarship. The authors acknowledge the use of resources provided by the Isambard-AI National AI Research Resource (AIRR). Isambard-AI is operated by the University of Bristol and is funded by the UK Government’s Department for Science, Innovation and Technology (DSIT) via UK Research and Innovation; and the Science and Technology Facilities Council AIRR.

References

- Loïc Barrault, Yu-An Chung, Mariano Cora Meglioli, David Dale, Ning Dong, Paul-Ambroise Duquenne, Hady Elsahar, Hongyu Gong, Kevin Heffernan, John Hoffman, and 1 others. 2023. Seamlessm4t: Massively multilingual & multimodal machine translation. *arXiv preprint arXiv:2308.11596*.
- Xi Chen, Songyang Zhang, Qibing Bai, Kai Chen, and Satoshi Nakamura. 2024. [LLaST: Improved end-](#)

- to-end speech translation system leveraged by large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6976–6987, Bangkok, Thailand. Association for Computational Linguistics.
- 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*.
- Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Naveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.
- Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. 2022. [Fleurs: Few-shot learning evaluation of universal representations of speech](#). *arXiv preprint arXiv:2205.12446*.
- Alan Dao, Dinh Bach Vu, Huy Hoang Ha, Tuan Le Duc Anh, Shreyas Gopal, Yue Heng Yeo, Warren Keng Hoong Low, Eng Siong Chng, and Jia Qi Yip. 2025. [Speechless: Speech instruction training without speech for low resource languages](#). In *26th Annual Conference of the International Speech Communication Association, Interspeech 2025, Rotterdam, The Netherlands, 17-21 August 2025*. ISCA.
- Jonah Dauvet, Min Ma, Jessica Ojo, and David Ifeoluwa Adelani. 2025. [Reassessing speech translation for low-resource languages: Do LLMs redefine the state-of-the-art against cascaded models?](#) In *Proceedings of the 5th Workshop on Multilingual Representation Learning (MRL 2025)*, pages 149–160, Suzhou, China. Association for Computational Linguistics.
- Soham Deshmukh, Benjamin Elizalde, Dimitra Emmanouilidou, Bhiksha Raj, Rita Singh, and Huaming Wang. 2024. [Training audio captioning models without audio](#). In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2024, Seoul, Republic of Korea, April 14-19, 2024*, pages 371–375. IEEE.
- Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. 2023. [CLAP learning audio concepts from natural language supervision](#). In *IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 2023, Rhodes Island, Greece, June 4-10, 2023*, pages 1–5. IEEE.
- Ruchao Fan, Bo Ren, Yuxuan Hu, Rui Zhao, Shujie Liu, and Jinyu Li. 2025. [Alignformer: Modality matching can achieve better zero-shot instruction-following speech-llm](#). *IEEE J. Sel. Top. Signal Process.*, 19(7):1329–1337.
- Yangui Fang, Jing Peng, Xu Li, Yu Xi, Chengwei Zhang, Guohui Zhong, and Kai Yu. 2025. [Low-resource domain adaptation for speech llms via text-only fine-tuning](#). In *IEEE Automatic Speech Recognition and Understanding Workshop, ASRU 2025, Honolulu, HI, USA, December 6-10, 2025*, pages 1–7. IEEE.
- Sreyan Ghosh, Zhifeng Kong, Sonal Kumar, S Sakshi, Jaehyeon Kim, Wei Ping, Rafael Valle, Dinesh Manocha, and Bryan Catanzaro. 2025. Audio flamingo 2: An audio-language model with long-audio understanding and expert reasoning abilities. In *Forty-second International Conference on Machine Learning*.
- Arushi Goel, Sreyan Ghosh, Jaehyeon Kim, Sonal Kumar, Zhifeng Kong, Sang-gil Lee, Chao-Han Huck Yang, Ramani Duraiswami, Dinesh Manocha, Rafael Valle, and 1 others. 2025. Audio flamingo 3: Advancing audio intelligence with fully open large audio language models. *arXiv preprint arXiv:2507.08128*.
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. In *Proceedings of the 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, 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.
- Sukairaj Hafiz Imam, Babangida Sani, Dawit Ketema Gete, Bedru Yimam Ahmed, Ibrahim Sa’id Ahmad, Idris Abdulmumin, Seid Muhie Yimam, Muhammad Yahuza Bello, and Shamsuddeen Hassan Muhammad. 2025. Automatic speech recognition for african low-resource languages: Challenges and future directions. In *Proceedings of the Sixth Workshop on African Natural Language Processing (AfricaNLP 2025)*, pages 89–94.
- Lennart Keller and Goran Glavaš. 2025. [Speechtaxi: On multilingual semantic speech classification](#). In *2025 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2025, Hyderabad, India, April 6-11, 2025*, pages 1–5. IEEE.
- Taesoo Kim and Jong Hwan Ko. 2025. Tesu-llm: Training speech-llms without speech via unified encoder alignment. *arXiv preprint arXiv:2506.06343*.
- Theodoros Kouzelis and Vassilis Katsouras. 2023. Weakly-supervised automated audio captioning via text only training. *arXiv preprint arXiv:2309.12242*.
- Xiquan Li, Wenxi Chen, Ziyang Ma, Xuenan Xu, Yuzhe Liang, Zhisheng Zheng, Qiuqiang Kong, and Xie Chen. 2025. [Drcap: Decoding CLAP latents with retrieval-augmented generation for zero-shot audio captioning](#). In *2025 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*

- 2025, Hyderabad, India, April 6-11, 2025, pages 1–5. IEEE.
- Alexander H Liu, Andy Ehrenberg, Andy Lo, Clément Denoix, Corentin Barreau, Guillaume Lample, Jean-Malo Delignon, Khyathi Raghavi Chandu, Patrick von Platen, Pavankumar Reddy Muddireddy, and 1 others. 2025. Voxtral. *arXiv preprint arXiv:2507.13264*.
- Rao Ma, Tongzhou Chen, Kartik Audhkhasi, and Bhuvana Ramabhadran. 2025. [LegoSLM: Connecting LLM with speech encoder using CTC posteriors](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 18171–18186, Suzhou, China. Association for Computational Linguistics.
- Amir M. Mansourian, Rozhan Ahmadi, Masoud Ghafouri, Amir Mohammad Babaei, Elaheh Badali Golezani, Zeynab Yasamani Ghamchi, Vida Ramezani, Alireza Taherian, Kimia Dinashi, Amirali Miri, and Shohreh Kasaei. 2025. [A comprehensive survey on knowledge distillation](#). *Trans. Mach. Learn. Res.*, 2025.
- Simon McIntosh-Smith, Sadaf Alam, and Christopher Woods. 2024. Isambard-ai: a leadership-class supercomputer optimised specifically for artificial intelligence. In *Proceedings of the Cray User Group*, pages 44–54.
- Tohru Nagano, Gakuto Kurata, Samuel Thomas, Hong-Kwang Jeff Kuo, Daniel Bolaños, Hyun Jung, and George Saon. 2025. [LLM based text generation for improved low-resource speech recognition models](#). In *2025 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2025, Hyderabad, India, April 6-11, 2025*, pages 1–5. IEEE.
- Omnilingual ASR Team, Gil Keren, Artyom Kozhevnikov, Yen Meng, Christophe Ropers, Matthew Setzler, Skyler Wang, Ife Adebara, Michael Auli, Can Balioglu, Kevin Chan, Chierh Cheng, Joe Chuang, Caley Droof, Mark Duppenhaler, Paul-Ambroise Duquenne, Alexander Erben, Cynthia Gao, Gabriel Mejia Gonzalez, and 14 others. 2025. [Omnilingual asr: Open-source multilingual speech recognition for 1600+ languages](#). *Preprint*, arXiv:2511.09690.
- Guilherme Penedo, Hynek Kydlíček, Amir Hossein Kargaran, and Leandro von Werra. 2026. Finetranslations. <https://huggingface.co/datasets/HuggingFaceFW/finetranslations>.
- Jing Peng, Yi Yang, Xu Li, Yu Xi, Quanwei Tang, Yangui Fang, Junjie Li, and Kai Yu. 2025. Tasu: Text-only alignment for speech understanding. *arXiv preprint arXiv:2511.03310*.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. [Learning transferable visual models from natural language supervision](#). In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.
- Fabian David Schmidt, Ivan Vulić, Goran Glavaš, and David Ifeoluwa Adelani. 2025. [Fleurs-SLU: A massively multilingual benchmark for spoken language understanding](#). In *Second Conference on Language Modeling*.
- Shivalika Singh, Freddie Vargus, Daniel Dsouza, Börje F. Karlsson, Abinaya Mahendiran, Wei-Yin Ko, Herumb Shandilya, Jay Patel, Deividas Matciunas, Laura OMahony, Mike Zhang, Ramith Hettiarachchi, Joseph Wilson, Marina Machado, Luisa Souza Moura, Dominik Krzemiński, Hakimeh Fadaei, Irem Ergün, Ifeoma Okoh, and 14 others. 2024. [Aya dataset: An open-access collection for multilingual instruction tuning](#). *Preprint*, arXiv:2402.06619.
- Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. 2024. [SALMONN: towards generic hearing abilities for large language models](#). In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.
- Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 technical report.
- Changhan Wang, Anne Wu, and Juan Pino. 2020. [Covost 2: A massively multilingual speech-to-text translation corpus](#). *Preprint*, arXiv:2007.10310.
- Wen Wang, RuiBing Hou, Hong Chang, Shiguang Shan, and Xilin Chen. 2025. [MATS: An audio language model under text-only supervision](#). In *Forty-second International Conference on Machine Learning*.
- Bajian Xiang, Shuaijiang Zhao, Tingwei Guo, and Wei Zou. 2025. [Understanding the modality gap: An empirical study on the speech-text alignment mechanism of large speech language models](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 5187–5202, Suzhou, China. Association for Computational Linguistics.
- Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. 2025. [Qwen2.5-omni technical report](#). *Preprint*, arXiv:2503.20215.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Yiming Zhang, Xuenan Xu, Ruoyi Du, Haohe Liu, Yuan Dong, Zheng-Hua Tan, Wenwu Wang, and Zhanyu Ma. 2025a. [Zero-shot audio captioning using soft and hard prompts](#). *IEEE Transactions on Audio, Speech and Language Processing*, 33:2045–2058.

Yuhao Zhang, Zhiheng Liu, Fan Bu, Ruiyu Zhang, Benyou Wang, and Haizhou Li. 2025b. [Soundwave: Less is more for speech-text alignment in LLMs](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18718–18738, Vienna, Austria. Association for Computational Linguistics.

A Detailed Results

The detailed per-language results for Fleurs-Belebele are shown in Figure 6 and for Fleurs-SIB in Figure 7. We observe similar trends to the Fleurs-SIB. On languages where the Oracle and the ASR baseline perform similarly, Text4Speech performs similarly or slightly below the ASR baseline. Languages where Qwen performs close to random (25%) tend to have random performance across the baselines and Text4Speech. The results on languages with a notable gap between the Oracle and the ASR baseline, while scores on Nepali increase by 25 points and by 19 points on Icelandic, other languages have smaller or no gains.

The language breakdown for SpeechTaxi is shown in Table 4. Except for Text4Speech-Qwen on Telugu, Text4Speech surpasses the ASR baselines consistently. Surprisingly, even on high-resource languages such as French or English, we observe a gap between the Oracle and the ASR baselines.

Text4Speech vs. Cascading Setup. A common approach to dealing with speech is simply transcribing it to text and using established text-based methods to process the transcripts. This cascading setup (termed Qwen-ASR and Gemma-ASR in the previous sections) can use existing components without further training. This warrants the question: Why use Text4Speech instead of this cascading setup? While we report improvements over the cascading setup, another argument for using Text4Speech is shorter inference times. The transcription process uses auto regressive generation, requiring multiple passes through the Seamless encoder and decoder. Text4Speech, on the other hand, requires a single pass through the encoder to represent the speech.

We also hypothesized that the encoded speech sequences are could contain more information than the generated transcripts and thus using the embeddings could be more robust to errors. However,

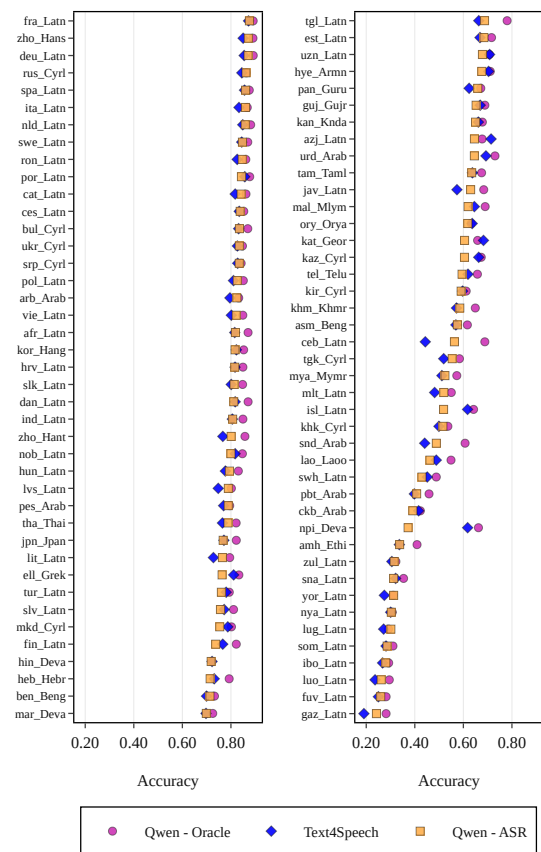


Figure 6: Per language performance of Qwen variants (Oracle, ASR, and Text4Speech) on Belebele-Fleurs.

Model	asm	ckb	eng	fra	guj	ibo	kan	luo	mal	mar	pan	rus	tam	tel	ukr	urd	vie	yor
Qwen 2.5 Omni	19.0	24.0	46.9	47.2	37.0	19.2	24.0	15.0	23.0	32.3	42.0	52.3	20.0	23.5	32.0	52.0	48.6	16.0
Voxtral Mini	40.0	40.6	45.7	49.1	51.0	35.4	46.0	29.0	48.0	49.5	55.0	43.9	48.0	37.0	52.0	57.0	55.0	27.0
Qwen - Oracle	39.0	30.2	63.0	50.0	51.0	15.2	51.0	18.0	50.0	57.6	48.0	53.3	50.0	52.0	53.0	35.0	50.5	20.0
Qwen - ASR	40.0	32.3	58.0	50.0	40.0	18.2	50.0	13.0	42.0	43.4	40.0	47.7	39.0	48.0	45.0	46.0	47.7	18.0
Text4Speech-2K	46.0	51.0	60.5	59.3	53.0	22.2	56.0	21.0	49.0	49.5	57.0	55.1	49.0	17.0	56.0	47.0	56.8	51.0
Text4Speech-10K	44.0	44.8	61.7	55.6	52.0	22.2	54.0	17.0	51.0	47.5	54.0	53.3	52.0	18.0	57.0	43.0	56.8	51.0
Gemma-Oracle	46.0	40.6	61.7	47.2	55.0	31.3	57.0	22.0	51.0	59.6	55.0	47.7	53.0	52.0	53.0	35.0	62.2	27.0
Gemma-ASR	35.0	25.0	46.9	44.4	49.0	17.2	43.0	14.0	38.0	40.4	40.0	38.3	41.0	41.8	41.0	49.0	47.7	22.0
Text4Speech-2K	46.0	50.0	60.5	55.6	53.0	19.2	51.0	17.0	48.0	46.5	55.0	47.7	51.0	55.1	48.0	49.0	52.3	54.0
Text4Speech-10K	45.0	53.1	58.0	55.6	52.0	20.2	55.0	16.0	50.0	49.5	57.0	49.5	48.0	48.0	48.0	56.0	50.5	61.0

Table 4: Accuracy per language on the SpeechTaxi benchmark

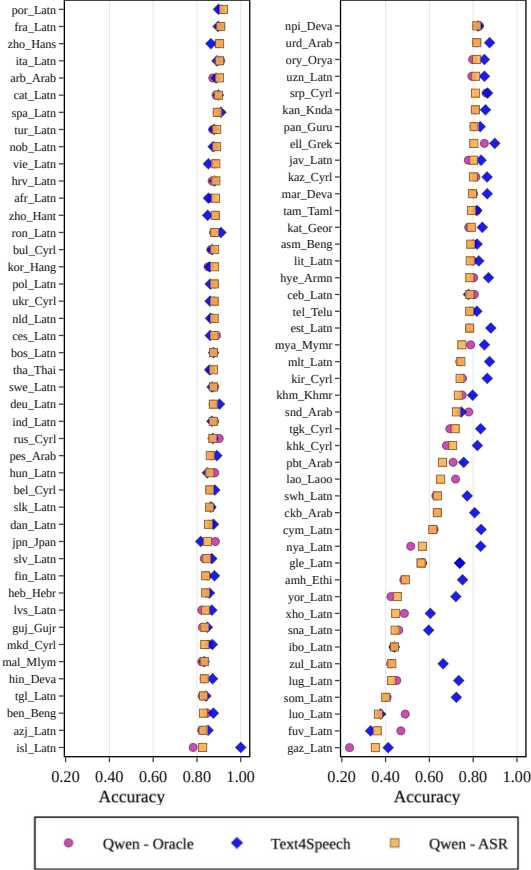


Figure 7: Per language performance of Qwen variants (Oracle, ASR, and Text4Speech) on SIB-Fleurs.

we could not find clear evidence that supports this hypothesis. While some languages notably benefit from using the embeddings, others perform very similarly to the cascading setup.

Speech Representations. Text4Speech exploits the alignment between the Seamless text and speech encoder. However, even though the two encoders are aligned, they still exhibit the modality gap. We want to explore the modality gap in parallel text and speech sequences with Text4Speech. Figure 8 shows representations of sequences in the Fleurs-SIB dataset before and after applying

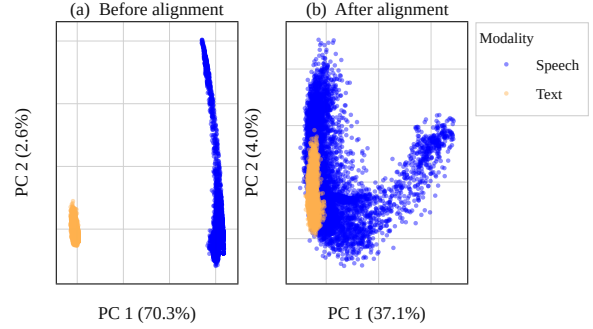


Figure 8: PCA visualization of the modality gap of (a) Seamless representations, (b) representations after the text and speech adapters.

the text and speech adapters across 10 languages. While both modalities clearly follow different distributions, the plot suggests that the gap is smaller after the adapters.

To obtain a clearer picture, we consider the average path score (APS) (Xiang et al., 2025), defined as the average similarity of the most similar token pairs for each text token.

$$APS = \frac{1}{N} \sum_{j=1}^N \max_{1 \leq i \leq M} (sim(t_i, t_j))$$

where N is the number of text tokens, M the number of speech tokens, and sim the similarity measure between two token representations. We use cosine similarity. Table 5 reports the average APS per language in the seamless embeddings and the representations after the adapters. The APS increases substantially after the adapters. The similarity also increases between unaligned sequences (0.0053 before adapters and 0.1711 after), but at a smaller scale. These numbers suggest that Text4Speech reduces the modality gap but does not close it. However, perfect alignment (APS=1.0) would lead to loss of information as speech is more information-rich than text.

Language	APS - Seamless	APS - Align	Δ
amh	0.1056	0.4832	+0.3776
arb	0.1169	0.5036	+0.3867
ben	0.1102	0.5078	+0.3976
deu	0.1161	0.4914	+0.3754
eng	0.1999	0.6363	+0.4364
fra	0.1241	0.4817	+0.3576
kor	0.0992	0.4749	+0.3758
rus	0.1274	0.5266	+0.3993
spa	0.1305	0.5040	+0.3736
tel	0.0981	0.4832	+0.3851
AVG	0.1228	0.5093	+0.3865

Table 5: APS between parallel speech and text of representations of the Seamless encoder (Seamless) and after the adapter (Align) and the Δ between the APS.

B Training Details

B.1 Hyperparameters

For training the two stages, we use the following hyperparameters:

Parameter	Text	Speech
LR	$1e-4$	$1e-4$
Batch Size	4	4
p_{para}	0.6	0.8
Num Layers	3	3
Hidden Dim	4092	4092

Table 6: Hyperparameters during the text alignment and speech alignment training.

The model checkpoints we use are: facebook/seamless-m4t-v2-large for the Seamless encoders, Qwen/Qwen3-4B-Instruct-2507 for the Qwen LLM, and google/gemma-3-4b-it (and google/gemma-3-12b-it for the larger models) as the Gemma LLM backbone. We run the experiments on the Isambard Cluster (McIntosh-Smith et al., 2024).

B.2 Dataset Details

We break down the large scale massively multilingual text-only dataset by language and task in Table 8. The dataset consists of samples from Aya (Apache License v2.0), FineTranslations (ODC-By v1.0), and Wikidata (CC BY-SA 3.0). We limit the dataset to languages supported by Seamless.

As speech samples, we use a subset of the Fleurs dataset (CC BY 4.0). The data is sampled uniformly across the 10 languages: English, German,

French, Chinese, Korean, Vietnamese, Japanese, Russian, Turkish, and Swahili.

C Use of AI Tools

The only AI tool we use is Claude for code generation. The generated code was thoroughly reviewed and tested to ensure integrity and correctness.

Code	Language	Script	AYA	R&P	TRANSLATION	Total
afr	Afrikaans	Latin		1,538	1,428	2,966
amh	Amharic	Ethiopic		1,538	1,428	2,966
ara	Arabic	Arabic		1,538		1,538
arb	Standard Arabic	Arabic			1,428	1,428
arb	Standard Arabic	Latin			1,428	1,428
ary	Moroccan Arabic	Arabic			1,428	1,428
arz	Egyptian Arabic	Arabic			1,428	1,428
asm	Assamese	Bengali		1,538	1,428	2,966
asm	Assamese	Latin			1,428	1,428
azj	North Azerbaijani	Latin			1,428	1,428
bel	Belarusian	Cyrillic		1,538	1,428	2,966
ben	Bangla	Bengali		1,538	1,428	2,966
ben	Bangla	Latin			1,428	1,428
bos	Bosnian	Latin		1,538	1,428	2,966
bul	Bulgarian	Cyrillic		1,538	1,428	2,966
cat	Catalan	Latin		1,538	1,428	2,966
ceb	Cebuano	Latin		1,538	1,428	2,966
ces	Czech	Latin		1,538	1,428	2,966
ckb	Central Kurdish	Arabic		1,538	1,428	2,966
cmn	Mandarin Chinese	Han			1,428	1,428
cym	Welsh	Latin		1,538	1,428	2,966
dan	Danish	Latin		1,538	1,428	2,966
deu	German	Latin	221	1,538	1,428	3,187
ell	Greek	Greek		1,538	1,428	2,966
eng	English	Latin	3,857	1,538		5,395
est	Estonian	Latin		1,538		1,538
eus	Basque	Latin			1,428	1,428
fin	Finnish	Latin		1,538	1,428	2,966
fra	French	Latin	1,390	1,538	1,428	4,356
fuv	Nigerian Fulfulde	Latin			1,428	1,428
gaz	West Central Oromo	Latin			1,428	1,428
gle	Irish	Latin		1,538	1,428	2,966
glg	Galician	Latin		1,538	1,428	2,966
guj	Gujarati	Gujarati		1,538	1,428	2,966
guj	Gujarati	Latin			1,428	1,428
heb	Hebrew	Hebrew		1,538	1,428	2,966
hin	Hindi	Devanagari	1,105	1,538	1,428	4,071
hin	Hindi	Latin			1,428	1,428
hrv	Croatian	Latin		1,538	1,428	2,966
hun	Hungarian	Latin		1,538	1,428	2,966
hye	Armenian	Armenian		1,538	1,428	2,966
ibo	Igbo	Latin		1,538	1,428	2,966
ind	Indonesian	Latin		1,538	1,428	2,966
isl	Icelandic	Latin		1,538	1,428	2,966
ita	Italian	Latin	693	1,538	1,428	3,659
jav	Javanese	Latin		1,538	1,428	2,966
jpn	Japanese	Japanese	6,255	1,538	1,428	9,221
kan	Kannada	Kannada		1,538	1,428	2,966
kan	Kannada	Latin			1,428	1,428
kat	Georgian	Georgian		1,538	1,428	2,966
kaz	Kazakh	Cyrillic		1,538	1,428	2,966
khk	Halh Mongolian	Cyrillic			1,428	1,428
khm	Khmer	Khmer		1,538	1,428	2,966
kir	Kyrgyz	Cyrillic		1,538	1,428	2,966
kor	Korean	Hangul	353	1,538	1,428	3, 319
lao	Lao	Lao		1,538	1,428	2,966
lit	Lithuanian	Latin		1,538	1,428	2,966
lug	Ganda	Latin		1,538	1,428	2,966
luo	Luo (Kenya and Tanzania)	Latin			1,428	1,428
lvs	Standard Latvian	Latin			1,428	1,428
lzh	Literary Chinese	Han (Traditional)		1,538		1,538
mai	Maithili	Devanagari			1,428	1,428
mal	Malayalam	Latin			1,428	1,428
mal	Malayalam	Malayalam		1,538	1,428	2,966
mar	Marathi	Devanagari		1,538	1,428	2,966

Table 7: Training data distribution across languages and datasets. R&P denotes the repeat and paraphrase task.

Code	Language	Script	AYA	R&P	TRANSLATION	Total
mar	Marathi	Latin			1,428	1,428
mkd	Macedonian	Cyrillic		1,538	1,428	2,966
mlt	Maltese	Latin		1,538	1,428	2,966
mni	Manipuri	Bengali			1,428	1,428
mni	Manipuri	Latin			1,428	1,428
mya	Burmese	Myanmar		1,538	1,428	2,966
nep	Nepali	Devanagari		1,538		1,538
nld	Dutch	Latin		1,538	1,428	2,966
nno	Norwegian Nynorsk	Latin			1,428	1,428
nob	Norwegian Bokmål	Latin			1,428	1,428
nor	Norwegian	Latin		1,538		1,538
npi	Nepali	Devanagari			1,428	1,428
npi	Nepali	Latin			1,428	1,428
nya	Nyanja	Latin		1,538	1,428	2,966
ory	Odia	Latin			1,428	1,428
ory	Odia	Odia			1,428	1,428
pan	Punjabi	Gurmukhi		1,538	1,428	2,966
pan	Punjabi	Latin			1,428	1,428
pbt	Southern Pashto	Arabic			1,428	1,428
pol	Polish	Latin		1,538	1,428	2,966
por	Portuguese	Latin	8,709	1,538	1,428	11,675
ron	Romanian	Cyrillic			1,428	1,428
ron	Romanian	Latin		1,538	1,428	2,966
rus	Russian	Cyrillic	400	1,538	1,428	3,366
sat	Santali	Latin			1,428	1,428
slk	Slovak	Latin		1,538	1,428	2,966
slv	Slovenian	Latin		1,538	1,428	2,966
sna	Shona	Latin		1,538	1,428	2,966
snd	Sindhi	Arabic		1,538	1,428	2,966
snd	Sindhi	Devanagari			1,428	1,428
snd	Sindhi	Latin			1,428	1,428
som	Somali	Latin		1,538	1,428	2,966
spa	Spanish	Latin	3,496	1,538	1,428	6,462
srp	Serbian	Cyrillic		1,538	1,428	2,966
srp	Serbian	Latin			1,428	1,428
swa	Swahili	Latin		1,538	1,428	2,966
swe	Swedish	Latin		1,538	1,428	2,966
tam	Tamil	Latin			1,428	1,428
tam	Tamil	Tamil		1,538	1,428	2,966
tel	Telugu	Latin			1,428	1,428
tel	Telugu	Telugu		1,538	1,428	2,966
tgk	Tajik	Cyrillic		1,538	1,428	2,966
tha	Thai	Thai		1,538	1,428	2,966
tur	Turkish	Latin	3,882	1,538	1,428	6,848
ukr	Ukrainian	Cyrillic		1,538	1,428	2,966
urd	Urdu	Arabic		1,538	1,428	2,966
urd	Urdu	Latin			1,428	1,428
uzn	Northern Uzbek	Cyrillic			1,428	1,428
uzn	Northern Uzbek	Latin			1,428	1,428
vie	Vietnamese	Latin	6,091	1,538	1,428	9,057
yor	Yoruba	Latin		1,538	1,428	2,966
yue	Cantonese	Han		1,538	1,428	2,966
zul	Zulu	Latin		1,538	1,428	2,966
Total			36,452	119,964	159,936	316,352

Table 8: Training data distribution across languages and datasets. R&P denotes the repeat and paraphrase task.