Do Spoken Language Models Hear Speech as They Read Text? Bridging Structural Gaps Between Speech and Text

Hyeonyu Kim[1], Hwayeon Kim[1], Youngwon Choi[1], Myeongkyun Cho[1,2], Huu-Kim Nguyen[3],
[1]Maum AI Inc., [2]KAIST, [3]Atmanity Inc.
hykim@maum.ai, khy0908@maum.ai, youngwonchoi@maum.ai,
mkcho@maum.ai, huukim136@gmail.com
Correspondence: hykim@maum.ai

Abstract

Spoken Language Models (SLMs) generate textual responses directly from speech, offering an alternative to cascaded systems. Despite recent advances, existing SLMs still exhibit weaker instruction-following behavior and limited generalization across diverse tasks compared to text-based language models. Our analysis shows that speech and text representations in current SLMs remain weakly aligned despite strong downstream performance, indicating that structural differences between continuous, temporally varying speech and discrete text remain insufficiently addressed. To address this, we propose a simple framework that decouples length mismatch from semantic alignment and encourages closer correspondence between speech and text representations. Experiments across multiple benchmarks demonstrate competitive performance against strong baselines, underscoring the importance of explicitly addressing structural differences between speech and text in SLM training. Our code is publicly available at https://github.com/jaykim9870/Do_SLMs_Hear_Speech_as_They_Read_Text.

1 Introduction

Spoken Language Models (SLMs) have attracted significant attention as a paradigm for enabling more general-purpose interaction with speech. SLMs broadly encompass pure speech LMs, speech+text LMs, and speech-aware text LMs, which differ in how speech and text are represented and modeled (Arora et al., 2025). In this work, we focus specifically on *speech-aware text LMs*, which combine a text LLM with a speech encoder to generate textual responses from speech and natural-language instructions, and refer to this class as SLMs throughout the paper. By processing speech directly rather than relying on an explicit ASR–LLM cascade, these models can mitigate error propagation and preserve speech-specific information such as acoustic and paralinguistic cues (Fathullah et al., 2023; Wang et al., 2024a).

Prior studies (Tang et al., 2023; Wang et al., 2023a; Yu et al., 2024; Pan et al., 2023) have shown that when SLMs are trained on automatic speech recognition (ASR) data to predict the transcription, they tend to focus only on speech content while ignoring textual instructions, a behavior referred to as *speech anchor bias* (Yu et al., 2024) or *task overfitting* (Tang et al., 2023). To mitigate this issue, prior works (Fathullah et al., 2023; Yu et al., 2024; Kang et al., 2024; Lu et al., 2025b) leverage diverse instructions by generating responses from text descriptions of speech and training SLMs to reproduce the same behaviors directly from speech. This *behavior alignment* strategy improves instruction-following while allowing SLMs to capture paralinguistic information, even without updating the LLM weights (Kang et al., 2024; Lu et al., 2025b).

However, these approaches primarily enforce alignment at the *behavioral* level by encouraging the same responses, while leaving the alignment of internal speech representations implicit. Recent studies have shown that current SLMs still exhibit notable limitations, including difficulties with instruction-following on Speech-IFEval (Lu et al., 2025c), limited generalization of diverse tasks on Dynamic-SUPERB (Huang et al., 2024b,a), and a consistent performance gap relative to text-only models on SpeechR (Yang et al., 2025). These observations raise a fundamental question: *Do SLMs Hear Speech as They Read Text?*

We are motivated by this question and analyze the internal representations of speech and text in current SLMs. As detailed in Section 3.1, we observe that even when SLMs perform well on downstream tasks, their internal representations of speech remain weakly aligned with text. This suggests that, despite the strong semantic correspondence between speech and its transcription, current SLMs map speech into representations that remain

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 47439–47454
October 24-29, 2026 ©2026 Association for Computational Linguistics

structurally different from text embeddings.

We argue that structural differences between speech and text are a key factor underlying this discrepancy. Unlike text, speech is a continuous and time-varying signal, which leads to longer and structurally distinct representations compared to text embeddings. Prior studies (Zhang et al., 2023; Wang et al., 2023b; Tang et al., 2023) have recognized these structural differences and have largely focused on reducing the length of mapped speech features to narrow this gap.

In parallel, other works have explored explicitly treating text embeddings as alignment targets, by measuring an L2 distance using only a subset of mapped speech tokens (Held et al., 2024) or employing the Wasserstein distance (Züfle and Niehues, 2024). Alternatively, another line of work (Wang et al., 2024b; Deng et al., 2024) explicitly aligns sequence lengths using a CIF mechanism and applies an internal alignment loss. However, this design requires the modality adapter to perform length matching and semantic alignment simultaneously.

In this work, we propose a simple framework that explicitly addresses structural differences between speech and text and encourages closer correspondence between the two representations. Specifically, during training, we dynamically match the length of mapped speech features to text embeddings, thereby decoupling length matching from semantic alignment. Building on this design, we further incorporate a *token-level internal alignment* alongside *behavior alignment* to encourage more consistent correspondence between speech and text representations. Experimental results across multiple benchmarks demonstrate that our approach achieves competitive performance against strong baselines, highlighting the importance of explicitly addressing structural differences between speech and text in SLM training.

In summary, our contributions are threefold:

- We analyze the internal representations of speech and text in current SLMs and show that they remain weakly aligned, even when models perform well on downstream tasks.
- We propose a simple framework that decouples length mismatch from semantic alignment and incorporates *token-level internal alignment*, encouraging closer correspondence between speech and text features.
- We demonstrate competitive performance across multiple benchmarks, including comparisons with strong closed-source models, highlighting the importance of explicitly addressing structural differences between speech and text in SLM training.

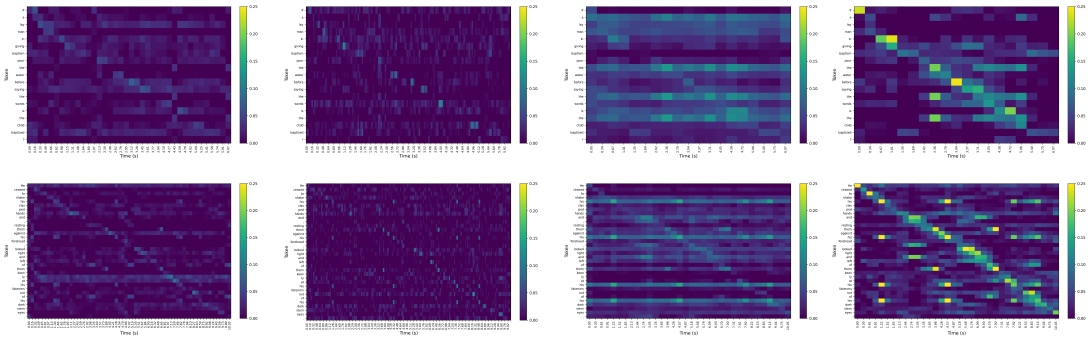
2 Related Work

Researchers have explored SLMs with a variety of input/output modality setups and training methods (Arora et al., 2025). Here, we concentrate on the line of research that generates textual responses from speech inputs.

Early SLM research often involve single-task training such as automatic speech recognition (ASR) or automatic speech translation (AST) (Lakhotia et al., 2021; Kharitonov et al., 2021; Chang et al., 2022; Wu et al., 2023; Zhang et al., 2023; Chen et al., 2024). More recent studies have shown that involving multiple speech tasks can enhance SLMs with a broader understanding of spoken language (Tang et al., 2023; Chu et al., 2024; Deshmukh et al., 2023). These architectures integrate a pre-trained speech encoder and a large language model with a modality adapter, and it is common to freeze the pre-trained model (Fathullah et al., 2023; Wang et al., 2023a; Kang et al., 2024) or to apply lightweight LoRA-based fine-tuning (Gong et al., 2023a; Wang et al., 2024a) to preserve the rich knowledge in each component.

A central challenge when stitching an LLM with a speech encoder is the mismatch between the encoded speech representation and the LLM's input space. This discordance arises not only from differences in semantic representations between speech and text, but also from the substantially longer sequence lengths (Wang et al., 2024c). To address this, many approaches introduce a modality adapter that downsamples speech features (Fathullah et al., 2023; Wang et al., 2023b; Ma et al., 2024; Li et al., 2024), although such temporal compression may not explicitly preserve higher-level linguistic structure. Some studies have proposed techniques to match the length of speech representations to that of the text transcription (Wu et al., 2023; Deng et al., 2024; Ma et al., 2025; Wang et al., 2024b).

In addition to temporal mismatches, ensuring robust instruction-following capabilities in SLM remains a critical challenge. Studies have shown that SLMs trained solely on ASR objectives overlook textual prompts and focus exclusively on speech inputs (Tang et al., 2023; Wang et al., 2023a; Yu et al.,



(a) BLSP-emo (b) Qwen2-Audio-Instruct (c) Ours(Cos) (d) Ours(InfoNCE)

Figure 1: Token-wise similarity maps between Z_s and Z_t on LibriTTS (Zen et al., 2019) test-clean for BLSP-emo (Wang et al., 2024a), Qwen2-Audio-Instruct (Chu et al., 2024), and our models. Existing SLMs show weak diagonal patterns, whereas our model exhibits a clear diagonal trend by addressing structural difference between speech and text. Non-diagonal activations correspond to identical text tokens appearing at different positions.

Model	CKA
BLSP-emo (Wang et al., 2024a)	0.3570
Qwen2-Audio-Instruct (Chu et al., 2024)	0.4113
DiVA (Held et al., 2024)	0.3992
DeSTA2 (Lu et al., 2025a)	0.4302
Ours	**0.6399**

Table 1: Centered Kernel Alignment (CKA) metric between speech and text in the LLM input space.

2024; Pan et al., 2023). To address this, *behavior alignment* frameworks have been introduced, which generate synthetic instruction–response pairs and then train the SLM on those examples (Fathullah et al., 2023; Yu et al., 2024; Kang et al., 2024; Lu et al., 2025b).

3 Method

3.1 Do SLMs Hear Speech as They Read Text?

Given speech input s, SLMs encode the input using a speech encoder $Enc(\cdot)$ and project the output into the LLM input space with a modality adapter $\psi(\cdot)$. Previous works have observed that when SLMs are trained on paired data from automatic speech recognition (ASR) (s, t) to predict transcriptions t from speech s, as in Equation 1, they focus only on the speech content while ignoring textual instructions, a phenomenon called *speech anchor bias* (Yu et al., 2024) or *task overfitting* (Tang et al., 2023).

$$\mathcal{L}_{\text{asr}} = -\log P\big(t \mid \psi(Enc(s))\big), \tag{1}$$

To mitigate this issue, several works (Fathullah et al., 2023; Yu et al., 2024; Kang et al., 2024; Lu et al., 2025b) incorporate a diverse instruction set I. Given text descriptions of speech $\tilde{t}$ and an instruction $I_i \in I$, the backbone LLM is used to generate a target response y_i, as formalized in Equation 2. The text descriptions $\tilde{t}$ may include not only transcriptions but also additional attributes such as emotion or intent (Kang et al., 2024; Lu et al., 2024).

$$y_i \sim P\big(y \mid \tilde{t}, I_i\big) \tag{2}$$

Then the SLM is trained to predict y_i directly from speech, as in Equation 3. This *behavior alignment* (Wang et al., 2023a) strategy has been shown to improve the instruction-following capability of SLMs, and demonstrates that models can capture paralinguistic information even without updating the LLM weights (Kang et al., 2024).

$$\mathcal{L}_{\text{behavior}} = -\log P\big(y_i \mid \psi(Enc(s)), I_i\big), \tag{3}$$

However, these behavior-alignment approaches mainly encourage the model to produce the same responses from speech inputs, while the alignment between speech and text representations inside the model remains implicit. Recent studies report that current SLMs still struggle with simple instruction-following (Lu et al., 2025c), and exhibit limited generalization across dynamic tasks (Huang et al., 2024b,a) compared to text-only LLMs. These observations motivate a fundamental question: *Do SLMs hear speech as they read text?*

To examine this, we compare mapped speech features $Z_s = \psi(Enc(s))$ in the LLM input space with text embeddings Z_t. We first measure their similarity using CKA (Kornblith et al., 2019; Raghu et al., 2021), which measures shared subspace structure. Existing SLMs exhibit low CKA scores (Table 1), which indicates limited structural similarity between speech and text representations.

To obtain a more fine-grained perspective, we visualize the token-wise similarity maps between Z_s and Z_t. As shown in Figure 1, existing SLMs exhibit weak or inconsistent diagonal patterns. This

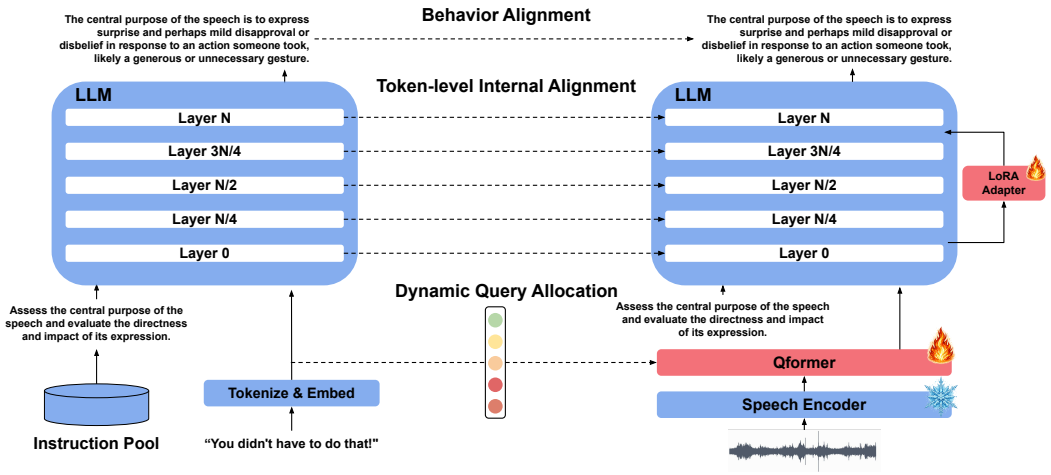


Figure 2: Model architecture. Given a speech input, a frozen speech encoder extracts frame-level features, which are mapped into the LLM input space with a windowed Q-former adapter. To explicitly resolve length mismatch, the adapter dynamically allocates the number of query tokens to match the target text length during training, decoupling length alignment problem from semantic alignment. The model is trained with behavior alignment, together with an token-level internal alignment which encourages fine-grained correspondence between speech and text.

observation implies that the learned cross-modal mapping in current SLMs may yield representations with a different structure from text, rather than naturally aligning with them at the token level.

3.2 Persistent Structural Differences in SLMs

In this section, we take a closer look at structural differences between speech and text that may contribute to this discrepancy. That is, speech consists of continuous and temporally varying signals and the encoded speech $Enc(s) \in \mathbb{R}^{L_s \times d_s}$ is much longer and structurally different from their textual counterparts, which are discrete and symbolic.

Prior approaches reduce the length of mapped speech features using convolutional layers (Zhang et al., 2023; Das et al., 2024), downsampling (Wang et al., 2023b; Gong et al., 2023b; Kang et al., 2024), or a windowed Q-former (Tang et al., 2023; Wang et al., 2025b; Mousavi et al., 2025). For example, Qwen2-Audio (Chu et al., 2024) produces 25 tokens per second of audio, while DeSTA2 (Held et al., 2024) yields 64 tokens regardless of input length. However, such compression can merge multiple subword or phonetic units into a single token, making the representation sensitive to temporal variations such as hesitations, stuttering, elongated vowels, or speaking rate changes. Recent methods further explore length reduction using CTC posteriors (Ma et al., 2025) or residual vector quantization (RVQ) (Tseng et al., 2025), but they still primarily focus on alleviating length mismatch as a way to address speech-text structural differences.

In parallel, several studies have attempted to improve cross-modal alignment by explicitly treating text embeddings as alignment targets. For instance, DiVA (Held et al., 2024) measures an L2 distance using only a subset of mapped speech features, while Züfle and Niehues (2024) employ the Wasserstein distance to compare representations with different lengths. Alternatively, methods that explicitly align sequence lengths using CIF mechanisms allow KL divergence (Wang et al., 2024b) or mean squared error (Deng et al., 2024) to be directly applied. However, these approaches rely on additional objectives to train the CIF module itself and require the adapter to perform length matching and semantic alignment simultaneously.

3.3 Our Approach

In the previous sections, we show that existing SLMs process speech differently from text (Section 3.1), and that structural differences between speech and text can persist (Section 3.2). In this work, we argue that explicitly addressing these structural differences can improve cross-modal alignment between speech and text. To this end, we introduce a simple framework that decouples length mismatch from semantic alignment and encourages closer correspondence between speech and text representations. Figure 2 illustrates the overall architecture.

Our modality adapter $\psi(\cdot)$ is based on a win-

Benchmark	Model			
		Chat-speech↑		
AIR-Bench	SALMONN	6.16		
	BLSP	6.17		
	DeSTA2	7.16		
	Qwen2-Audio	7.18		
	Phi-4-Multimodal	7.47		
	Gemini-1.5-pro*	6.97		
	Gemini-2.0-Flash*	**7.92**		
	Ours	<u>7.85</u>		
		Multi-Choice↑	**Generative-Procedural (FC↑, LR↑, CoH↑)**	**Generative-Normative (FC↑, LR↑)**
SpeechR	SALMONN	34.73	(12.50, 1.90, 1.33)	(34.75, 3.03)
	Qwen2-Audio	12.83	(9.52, 1.50, 1.00)	(31.25, 2.82)
	Qwen2-Audio-Instruct	33.90	(25.00, 3.50, 2.16)	(38.91, <u>3.46</u>)
	Gemini-1.5-Pro*	**67.68**	(**83.04**, <u>4.49</u>, <u>4.47</u>)	(**51.92**, **3.58**)
	Ours	<u>52.91</u>	(<u>73.51</u>, **4.58**, **4.79**)	(<u>40.83</u>, 3.12)
		Perception↑	**Reasoning↑**	**Average↑**
MMSU	BLSP	28.36	44.77	35.96
	DiVA	33.95	65.04	48.31
	Qwen2-Audio-Instruct	39.02	68.90	53.27
	Gemini-1.5-Pro*	**46.10**	**76.16**	**60.68**
	Gemini-2.0-Flash*	<u>40.83</u>	59.18	51.03
	Ours	38.39	<u>70.31</u>	<u>53.85</u>
		(CEQ↑, CW↑)	**CoT↑**	**Forgetting Rate↑**
Speech-IFEval	SALMONN	(37.41, 61.25)	12.00	-50.20
	BLSP-emo	(66.35, 63.75)	50.50	-17.92
	Qwen2-Audio-Instruct	(41.59, 67.75)	32.00	-
	DeSTA2	(<u>83.71</u>, **92.49**)	**91.50**	**-3.57**
	Ours	(**96.14**, <u>69.50</u>)	<u>67.50</u>	<u>-12.18</u>

Table 2: Performances on Air-bench chat, SpeechR, MMSU, and Speech-IFeval benchmarks. Best results are highlighted in **bold**, and second-best results are <u>underlined</u>. For SpeechR, FC, LR, and CoH denote final correctness, logical relevance, and coherence, respectively. For Speech-IFEval, CEQ, CW, and CoT indicate close-ended question, creative writing, and chain of thought, respectively. Models marked with * are closed-source models.

dowed Q-former (Tang et al., 2023), in which mapped speech features attend to speech representations via cross-attention and their output length is determined by the learned query. Unlike prior approaches that rely on a fixed rate of query allocation, we adopt a *dynamic query allocation* strategy to adjust the length of mapped speech features. During training, we leverage speech–transcription pairs (s, t) and allocate L_t queries to match the length of the target text embeddings $Z_t \in \mathbb{R}^{L_t \times d}$, explicitly controlling the length of mapped speech features. At inference time, we employ a lightweight speech rate predictor (Yeo et al., 2025) to estimate the target token length from speech and allocate queries accordingly. Details of the speech rate predictor are provided in Appendix A.1, and its performance is reported in Table 6. By decoupling length alignment from semantic alignment, this simple design facilitates learning meaningful semantic correspondences between speech and text.

We adopt *behavior alignment* and train our model to generate the same LLM response y produced from the text descriptions of speech $\tilde{t}$, as defined in Equation 3. Note that we randomly sample instruction-response pair (I_i, y_i) from instruction set I during training, as illustrated in Figure 2, but for brevity we omit the instruction index below.

$$\mathcal{L}_{\text{behavior}} = -\sum_{j=1}^{T} \log P\big(y_j \mid y_{<j}, \psi(Enc(s)), I\big) \quad (4)$$

In addition, we apply a fine-grained *token-level internal alignment* loss between speech and text

	Air-bench	SpeechR	MMSU	Speech-IFeval	Rel. Δ (%) ↑
Ours(w/o $\mathcal{L}_{\text{behavior}}$)	7.59	47.33	52.45	-22.94	-16.20
Ours(w/o $\mathcal{L}_{\text{internal}}$)	7.24	52.00	53.14	-17.68	-10.66
Ours(MSE)	7.66	48.46	53.63	-24.51	-15.59
Ours(InfoNCE)	7.48	48.75	53.67	-23.30	-15.39
Ours(Cformer)	5.07	36.13	39.97	-29.60	-48.71
Ours(FQA)	7.78	52.41	53.25	-18.26	-9.07
Ours(w/o SRP)	**7.89**	51.93	**54.24**	-14.64	-4.37
Ours(w/o SRP, GT)	7.85	53.05	53.69	-12.70	-1.07
Ours	7.85	**52.91**	53.85	**-12.18**	**0.00**

Table 3: Ablation study across four benchmarks. We report performance on each benchmark and relative changes (%) with respect to the full model, where higher values indicate better performance.

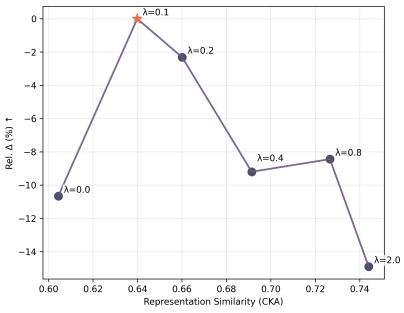


Figure 3: Trade-off between representation similarity and task performance across different values of λ.

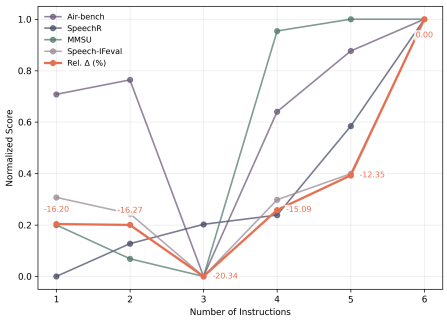


Figure 4: Effect of the number of instructions on model performance.

representations. Specifically, we use a cosine similarity loss, defined in Equation 5, to encourage alignment between the length-matched speech and text representations at selected token-level hidden states $h_s^{(\ell)}$ and $h_t^{(\ell)}$ of the LLM. Since dynamic query allocation matches the mapped speech sequence length to the target text length L_t during training, the loss is computed over corresponding token positions. We compute the token-level internal alignment loss at the input embedding layer, denoted as $\mathcal{L}_{\text{internal}}^{(0)}$, as well as at four evenly spaced hidden layers, $N/4$, $N/2$, $3N/4$, and N, where N denotes the number of layers in the LLM. The losses are obtained by averaging across layers.

$$\mathcal{L}_{\text{internal}}^{(\ell)} = \frac{1}{L_t} \sum_{j=1}^{L_t} \left(1 - \frac{\langle h_{s,j}^{(\ell)}, h_{t,j}^{(\ell)} \rangle}{\|h_{s,j}^{(\ell)}\|_2 \, \|h_{t,j}^{(\ell)}\|_2} \right). \quad (5)$$

The final training objective is defined as below:

$$\mathcal{L} = \mathcal{L}_{\text{behavior}} + \lambda \, \mathcal{L}_{\text{internal}}, \quad (6)$$

where λ controls the relative strength of the token-level internal alignment objective.

4 Experiments

4.1 Implementation Details

We use a total of approximately 69,000 hours of paired speech–text data, which includes diverse speech-specific attributes. The instruction set I consists of 18 instructions, constructed based on prior work (Yu et al., 2024). For the LLM backbone, we employ Qwen2.5-7B-Instruct (Team, 2024) and use Whisper-large-v3 (Radford et al., 2022) as the speech encoder. Further details on the datasets and training are provided in Appendix A.1.

4.2 Benchmarks

AIR-Bench (Chat-speech) (Yang et al., 2024) contains open-ended question–answer pairs designed to evaluate instruction-following and generative interaction capabilities from audio, and is presented as the first generative benchmark for SLMs.

SpeechR (Yang et al., 2025) is designed to evaluate speech-based reasoning capabilities of SLMs and is constructed using synthetic speech data. SpeechR evaluates three types of reasoning, factual retrieval, procedural inference, and normative judgment, and consists of three subsets of multiple-choice, generative, and acoustic-feature formats.

Model	AIR-Bench Foundation		AIR-Bench	SpeechR↑	MMSU↑	CKA
	Linguistic ↑	Speech-specific ↑	Chat-speech ↑			
Ours (text input)	**70.35**	44.11	**8.43**	**58.85**	34.43	-
Ours (w/o $\mathcal{L}_{\text{behavior}}$)	53.09	47.90	7.59	47.33	52.45	0.6436
Ours (InfoNCE)	52.94	47.89	7.48	48.75	53.67	0.7113
Ours	53.33	**50.63**	7.85	52.91	**53.85**	0.6399

Table 4: Analysis of linguistic and speech-specific performance. The text-input variant performs strongly on tasks primarily requiring linguistic content, but underperforms on tasks that rely more heavily on speech-specific information.

MMSU (Wang et al., 2025a) is a comprehensive benchmark that emphasizes the understanding of speech with diverse acoustic and paralinguistic signals, and is mostly built on real-world speech data. It contains 5,000 expert-annotated multiple-choice questions spanning 47 tasks that cover both perception and reasoning.

Speech-IFEval (Lu et al., 2025c) is designed to evaluate the instruction-following capability of SLMs. It disentangles instruction-following from speech perception and introduces instruction constraints that are independent of the speech content, enabling a focused assessment of whether models correctly follow textual instructions.

4.3 Main Results

Table 1 and Figure 1 show that our model exhibits stronger structural similarity between two representations, as evidenced by higher CKA scores and more consistent token-level similarity patterns.

Table 2 summarizes performance across four benchmarks. Our model achieves competitive results on most benchmarks, including comparisons with strong closed-source models.

SALMONN (Tang et al., 2023) employs a window-level Q-former as a modality adapter and introduces behavior alignment to mitigate task overfitting. Our model outperforms SALMONN on most benchmarks, with a particularly large improvement in forgetting rate on Speech-IFEval.

Qwen2-Audio-Instruct (Chu et al., 2024) is trained via a multi-stage pipeline with large-scale pre-training, supervised fine-tuning, and preference optimization. Despite using substantially less training data, our model achieves competitive performance and surpasses it on several subsets on SpeechR and Speech-IFEval.

DeSTA2 (Lu et al., 2025a) augments textual descriptions of speech using auxiliary models for better training, and incorporates ASR output along with the audio for inference. While DeSTA2 benefits from transcription access, our model achieves stronger performance on AIR-Bench and remains competitive on Speech-IFEval.

Finally, our model demonstrates competitive performance against strong closed-source baselines such as Gemini-1.5-Pro and Gemini-2.0-Flash across multiple benchmarks. Overall, these results suggest that explicitly addressing structural differences between speech and text can improve performance across diverse tasks.

4.4 Ablation Study

In this section, we conduct controlled ablations on key components of our model to analyze their contributions and training dynamics. The results are summarized in Table 3, where we report multiple-choice accuracy on SpeechR, average scores on MMSU, and the forgetting rate on Speech-IFEval.

4.4.1 Effects of Alignment Objectives and Instruction Diversity

We first train variants of our model by removing each term in Equation 6, and results are shown in the top rows of Table 3. Both settings lead to over 10% performance degradation, with a larger drop observed when excluding the behavior alignment.

We further analyze the effect of $\mathcal{L}_{\text{internal}}$ by varying λ. As shown in Figure 3, increasing λ consistently improves CKA, indicating stronger representational similarity between speech and text. However, excessively large λ values lead to performance degradation, suggesting that internal alignment is beneficial only at an appropriate strength, as overly strong alignment can compromise overall performance, potentially by reducing the preservation of speech-specific information.

Finally, we study the effect of instruction diversity by progressively expanding the instruction set. Among the 18 instructions spanning six categories, we add one category at a time, with three instructions per category. The categories are intro-

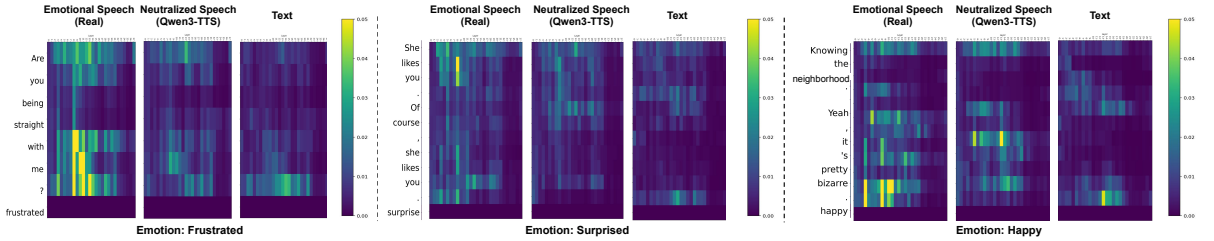


Figure 5: Attention analysis on IEMOCAP across real speech, neutralized cloned speech generated by Qwen3-TTS (Hu et al., 2026), and text. Brighter colors indicate higher attention weights from the input representations to the corresponding emotion label. Our model shows clear activations around regions associated with emotional expression in real speech.

duced in order of increasing semantic complexity: Speech Recognition, Content Repetition, Continuation, Keyword Extraction, Intent Recognition, and Sentiment Analysis. As shown in Figure 4, performance improves as instruction diversity increases, with noticeable gains when more complex instructions such as Keyword Extraction are introduced. These results suggest that adding instructions requiring more structured understanding can further improve performance, and that the best results in our setting are obtained by combining behavior alignment with token-level internal alignment. Detailed results are provided in Table 8.

4.4.2 Effect on Different Types of Token-level Internal Alignment Losses

We explore different types of token-level internal alignment losses and analyze their impact. In addition to Equation 5, we consider an MSE loss that enforces alignment in Euclidean space and an InfoNCE loss that introduces contrastive token-level alignment. Specifically, for each speech token, we treat the text token at the same index as a positive pair, while all other tokens serve as negative pairs, excluding identical text tokens at different positions from the negative set.

Table 3 (middle) summarizes the results. Both variants degrade performance by approximately 15%, despite higher CKA scores (MSE: 0.6679, InfoNCE: 0.7113). In particular, InfoNCE produces a clearer diagonal structure in the token-level similarity maps (Figure 1) due to its contrastive formulation. These results suggest that stronger alignment can improve representational similarity without necessarily improving performance, suggesting that overly constraining speech representations toward text may interfere with information that is not preserved in text. We examine this possibility in the following section.

4.4.3 Why Does Better Alignment Not Always Lead to Better Performance?

To understand why stronger speech–text alignment does not necessarily improve downstream performance, we analyze linguistic and speech-specific information separately. We use a text-input variant as a reference for linguistic information, providing ground-truth transcriptions when available and Whisper-large-v3 transcriptions otherwise. We additionally group the nine AIR-Bench Foundation tasks into linguistic tasks (Speech Grounding, Spoken Language Identification, Speech Entity Recognition, and Intent Classification) and speech-specific tasks based on whether they can primarily be solved from linguistic content. We exclude Speech-IFEval, whose forgetting-rate metric is defined relative to text-input performance.

First, linguistic understanding alone does not guarantee optimal downstream performance. As shown in Table 4, the text-input variant performs best on the linguistic subset of AIR-Bench Foundation and strongly on AIR-Bench Chat and SpeechR, but substantially worse on the speech-specific subset and MMSU. This indicates that some speech tasks additionally require acoustic and paralinguistic information absent from text.

Second, stronger internal alignment objectives do not necessarily lead to stable or effective training. Ours (InfoNCE) achieves higher CKA than our final model but underperforms it across both AIR-Bench Foundation subsets and other downstream benchmarks. Likewise, removing $\mathcal{L}_{\text{behavior}}$ and relying only on internal alignment causes substantial degradation, suggesting that overly strong alignment can over-constrain speech representations toward text.

Together with Figure 3 and Section 4.4.2, these findings suggest that internal alignment should complement behavior alignment rather than be maximized.

4.4.4 Effect on Different Query Allocation Strategies

We analyze the impact of dynamic query allocation (DQA) by comparing it with a fixed query allocation (FQA) strategy. Following prior work (Tang et al., 2023), FQA assigns a constant number of 3 tokens per second of audio, whereas our approach dynamically adjusts the number of queries based on the target token length during training and uses a speech-rate predictor (SRP) at inference. To isolate the effect of each component, we consider two variants. *Ours (FQA)* replaces DQA with FQA during both training and inference, while *Ours (w/o SRP)* retains DQA during training but uses FQA only at inference to simulate prediction errors in SRP. Additionally, we compare with a CIF-based variant, *Ours (Cformer)* (Wang et al., 2024a), where the modality adapter jointly learns to handle both length mismatch and semantic alignment.

Table 3 (bottom) summarizes the results. *Ours (Cformer)* shows a substantial performance drop, highlighting the benefit of separating length matching from representation alignment. *Ours (FQA)* also exhibits significant degradation across most benchmarks, whereas *Ours (w/o SRP)* results in only minor performance loss. These results suggest that the main benefit of DQA comes from length matching during training rather than from precise length prediction at inference.

4.5 How SLMs Understand Speech-Specific Information

A potential concern with token-level internal alignment is that close alignment with text may hinder the model's ability to capture speech-specific information. While MMSU results suggest that the model retains sensitivity to prosodic and paralinguistic cues, we further analyze how such information is reflected internally.

We analyze this on the IEMOCAP (Busso et al., 2008) dataset by prompting the model to classify emotions from three types of inputs: real emotional speech, neutralized cloned speech, and text. The neutralized speech is generated using Qwen3-TTS (Hu et al., 2026) by cloning the real speech, while instructing the model to remove emotional expression. To examine which parts of the input contribute to the prediction, we compute attention weights from the input to the emotion labels. Additional details are provided in Appendix A.2.

Figure 5 presents the results. For real speech, strong activations appear around regions associated with emotional expression. When the same utterances are converted into neutralized speech, these activations become noticeably weaker, suggesting that the highlighted regions reflect emotional expression rather than lexical content alone. In contrast, text inputs show the strongest activations on punctuation tokens. These findings suggest that, despite token-level internal alignment, the model still captures speech-specific cues absent from transcripts. Additional examples and audio samples are provided in the supplementary material.

4.6 Discussion

Our results show that speech-text alignment in SLMs should be treated as a balanced objective rather than a quantity to be maximized. While moderate token-level alignment improves performance, overly strong alignment can increase representational similarity without improving downstream results. This highlights the need to preserve speech-specific cues while encouraging linguistic correspondence with text. Although our analysis in Section 4.5 suggests that the model retains speech-specific cues, understanding how SLMs encode and balance linguistic content with such cues remains an important direction for future work.

SLMs have also been extended toward broader audio understanding beyond speech. In such settings, it becomes hard to define a strong correspondence analogous to speech–transcription pairs. Investigating how Large Audio Language Models (LALMs) can learn and align representations for non-speech audio represents another promising direction for future research.

5 Conclusion

In this work, we investigate how current Spoken Language Models process speech relative to text and find that their internal representations remain weakly aligned, suggesting persistent structural differences between the two modalities. We propose a simple framework that addresses this issue by decoupling length matching from semantic alignment and encouraging speech-text correspondence. Experiments across multiple benchmarks show that our approach improves representational alignment while achieving competitive performance against strong baselines. Our findings highlight the importance of addressing structural differences between speech and text for more effective SLM training.

6 Limitations

As discussed in Section 4.6, this work has several limitations. First, while our analysis shows that explicitly mitigating structural differences between speech and text and modeling fine-grained internal alignment can improve downstream performance, as shown in Section 4.4, the extent to which internal alignment should be encouraged may depend on the target task, model architecture, and training data. Thus, our findings should not be interpreted as suggesting that improving internal speech–text alignment alone will necessarily lead to better downstream performance.

Second, although our analysis in Section 4.5 suggests that SLMs can retain speech-specific cues, we do not fully characterize how linguistic content and speech-specific information are jointly encoded and balanced inside the model. A more detailed analysis of this interaction remains an important direction for future work.

Finally, this work focuses on speech–text correspondence, where paired speech and transcription data provide a natural basis for alignment. Extending the analysis to broader audio understanding settings is less straightforward, since non-speech audio often lacks a direct textual counterpart. Investigating how Large Audio Language Models can learn and align representations for non-speech audio remains an interesting direction for future research.

References

Adaeze Adigwe, Noé Tits, Kevin El Haddad, Sarah Ostadabbas, and Thierry Dutoit. 2018. The emotional voices database: Towards controlling the emotion dimension in voice generation systems. *arXiv preprint arXiv:1806.09514*.

Siddhant Arora, Kai-Wei Chang, Chung-Ming Chien, Yifan Peng, Haibin Wu, Yossi Adi, Emmanuel Dupoux, Hung-Yi Lee, Karen Livescu, and Shinji Watanabe. 2025. On the landscape of spoken language models: A comprehensive survey. *arXiv preprint arXiv:2504.08528*.

Emanuele Bastianelli, Andrea Vanzo, Pawel Swietojanski, and Verena Rieser. 2020. Slurp: A spoken language understanding resource package. *arXiv preprint arXiv:2011.13205*.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335–359.

Kai-Wei Chang, Wei-Cheng Tseng, Shang-Wen Li, and Hung-yi Lee. 2022. Speechprompt: An exploration of prompt tuning on generative spoken language model for speech processing tasks. *arXiv preprint arXiv:2203.16773*.

Guoguo Chen, Shuzhou Chai, Guanbo Wang, Jiayu Du, Wei-Qiang Zhang, Chao Weng, Dan Su, Daniel Povey, Jan Trmal, Junbo Zhang, and 1 others. 2021. Gigaspeech: An evolving, multi-domain asr corpus with 10,000 hours of transcribed audio. *arXiv preprint arXiv:2106.06909*.

Zhehuai Chen, He Huang, Andrei Andrusenko, Oleksii Hrinchuk, Krishna C Puvvada, Jason Li, Subhankar Ghosh, Jagadeesh Balam, and Boris Ginsburg. 2024. Salm: Speech-augmented language model with in-context learning for speech recognition and translation. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 13521–13525. IEEE.

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*.

CLAPv2. 2025. Jl-corpus. https://huggingface.co/datasets/CLAPv2/JL-Corpus. Hugging Face dataset.

Nilaksh Das, Saket Dingliwal, Srikanth Ronanki, Rohit Paturi, Zhaocheng Huang, Prashant Mathur, Jie Yuan, Dhanush Bekal, Xing Niu, Sai Muralidhar Jayanthi, and 1 others. 2024. Speechverse: A large-scale generalizable audio language model. *arXiv preprint arXiv:2405.08295*.

Keqi Deng, Guangzhi Sun, and Philip C Woodland. 2024. Wav2prompt: End-to-end speech prompt generation and tuning for llm in zero and few-shot learning. *arXiv preprint arXiv:2406.00522*.

Soham Deshmukh, Benjamin Elizalde, Rita Singh, and Huaming Wang. 2023. Pengi: An audio language model for audio tasks. *Advances in Neural Information Processing Systems*, 36:18090–18108.

Yassir Fathullah, Chunyang Wu, Egor Lakomkin, Ke Li, Junteng Jia, Yuan Shangguan, Jay Mahadeokar, Ozlem Kalinli, Christian Fuegen, and Mike Seltzer. 2023. Audiochatllama: Towards general-purpose speech abilities for llms. *arXiv preprint arXiv:2311.06753*.

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. *arXiv preprint arXiv:2111.09344*.

Yuan Gong, Alexander H Liu, Hongyin Luo, Leonid Karlinsky, and James Glass. 2023a. Joint audio and speech understanding. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–8. IEEE.

Yuan Gong, Hongyin Luo, Alexander H Liu, Leonid Karlinsky, and James Glass. 2023b. Listen, think, and understand. *arXiv preprint arXiv:2305.10790*.

William Held, Ella Li, Michael Ryan, Weiyan Shi, Yanzhe Zhang, and Diyi Yang. 2024. Distilling an end-to-end voice assistant without instruction training data. *arXiv preprint arXiv:2410.02678*.

Hangrui Hu, Xinfa Zhu, Ting He, Dake Guo, Bin Zhang, Xiong Wang, Zhifang Guo, Ziyue Jiang, Hongkun Hao, Zishan Guo, and 1 others. 2026. Qwen3-tts technical report. *arXiv preprint arXiv:2601.15621*.

Chien-yu Huang, Wei-Chih Chen, Shu-wen Yang, Andy T Liu, Chen-An Li, Yu-Xiang Lin, Wei-Cheng Tseng, Anuj Diwan, Yi-Jen Shih, Jiatong Shi, and 1 others. 2024a. Dynamic-superb phase-2: A collaboratively expanding benchmark for measuring the capabilities of spoken language models with 180 tasks. *arXiv preprint arXiv:2411.05361*.

Chien-yu Huang, Ke-Han Lu, Shih-Heng Wang, Chi-Yuan Hsiao, Chun-Yi Kuan, Haibin Wu, Siddhant Arora, Kai-Wei Chang, Jiatong Shi, Yifan Peng, and 1 others. 2024b. Dynamic-superb: Towards a dynamic, collaborative, and comprehensive instruction-tuning benchmark for speech. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 12136–12140. IEEE.

Keith Ito and Linda Johnson. 2017. The lj speech dataset. `https://keithito.com/LJ-Speech-Dataset/`.

Wei Kang, Xiaoyu Yang, Zengwei Yao, Fangjun Kuang, Yifan Yang, Liyong Guo, Long Lin, and Daniel Povey. 2023. Libriheavy: a 50,000 hours asr corpus with punctuation casing and context. *Preprint*, arXiv:2309.08105.

Wonjune Kang, Junteng Jia, Chunyang Wu, Wei Zhou, Egor Lakomkin, Yashesh Gaur, Leda Sari, Suyoun Kim, Ke Li, Jay Mahadeokar, and 1 others. 2024. Frozen large language models can perceive paralinguistic aspects of speech. *arXiv preprint arXiv:2410.01162*.

Eugene Kharitonov, Ann Lee, Adam Polyak, Yossi Adi, Jade Copet, Kushal Lakhotia, Tu-Anh Nguyen, Morgane Rivière, Abdelrahman Mohamed, Emmanuel Dupoux, and 1 others. 2021. Text-free prosody-aware generative spoken language modeling. *arXiv preprint arXiv:2109.03264*.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. 2019. Similarity of neural network representations revisited. In *International conference on machine learning*, pages 3519–3529. PMlR.

Kushal Lakhotia, Eugene Kharitonov, Wei-Ning Hsu, Yossi Adi, Adam Polyak, Benjamin Bolte, Tu-Anh Nguyen, Jade Copet, Alexei Baevski, Abdelrahman Mohamed, and 1 others. 2021. On generative spoken language modeling from raw audio. *Transactions of the Association for Computational Linguistics*, 9:1336–1354.

Keon Lee, Kyumin Park, and Daeyoung Kim. 2022. Dailytalk: Spoken dialogue dataset for conversational text-to-speech. *Preprint*, arXiv:2207.01063.

Mohan Li, Cong-Thanh Do, Simon Keizer, Youmna Farag, Svetlana Stoyanchev, and Rama Doddipatla. 2024. Whisma: A speech-llm to perform zero-shot spoken language understanding. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pages 1115–1122. IEEE.

Ke-Han Lu, Zhehuai Chen, Szu-Wei Fu, He Huang, Boris Ginsburg, Yu-Chiang Frank Wang, and Hung-yi Lee. 2024. Desta: Enhancing speech language models through descriptive speech-text alignment. *arXiv preprint arXiv:2406.18871*.

Ke-Han Lu, Zhehuai Chen, Szu-Wei Fu, Chao-Han Huck Yang, Jagadeesh Balam, Boris Ginsburg, Yu-Chiang Frank Wang, and Hung-yi Lee. 2025a. Developing instruction-following speech language model without speech instruction-tuning data. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Ke-Han Lu, Zhehuai Chen, Szu-Wei Fu, Chao-Han Huck Yang, Sung-Feng Huang, Chih-Kai Yang, Chee-En Yu, Chun-Wei Chen, Wei-Chih Chen, Chien-yu Huang, and 1 others. 2025b. Desta2. 5-audio: Toward general-purpose large audio language model with self-generated cross-modal alignment. *arXiv preprint arXiv:2507.02768*.

Ke-Han Lu, Chun-Yi Kuan, and Hung-yi Lee. 2025c. Speech-ifeval: Evaluating instruction-following and quantifying catastrophic forgetting in speech-aware language models. *arXiv preprint arXiv:2505.19037*.

Loren Lugosch, Mirco Ravanelli, Patrick Ignoto, Vikrant Singh Tomar, and Yoshua Bengio. 2019. Speech model pre-training for end-to-end spoken language understanding. *arXiv preprint arXiv:1904.03670*.

Rao Ma, Tongzhou Chen, Kartik Audhkhasi, and Bhuvana Ramabhadran. 2025. Legoslm: Connecting llm with speech encoder using ctc posteriors. *arXiv preprint arXiv:2505.11352*.

Ziyang Ma, Guanrou Yang, Yifan Yang, Zhifu Gao, Jiaming Wang, Zhihao Du, Fan Yu, Qian Chen, Siqi Zheng, Shiliang Zhang, and 1 others. 2024. An embarrassingly simple approach for llm with strong asr capacity. *arXiv preprint arXiv:2402.08846*.

Pooneh Mousavi, Shubham Gupta, Cem Subakan, and Mirco Ravanelli. 2025. Listen: Learning soft token

embeddings for neural audio llms. *arXiv preprint arXiv:2505.18517*.

Tu Anh Nguyen, Wei-Ning Hsu, Antony d'Avirro, Bowen Shi, Itai Gat, Maryam Fazel-Zarani, Tal Remez, Jade Copet, Gabriel Synnaeve, Michael Hassid, and 1 others. 2023. Expresso: A benchmark and analysis of discrete expressive speech resynthesis. *arXiv preprint arXiv:2308.05725*.

Kari Ali Noriy, Xiaosong Yang, and Jian Jun Zhang. 2023. Emns/imz/corpus: An emotive single-speaker dataset for narrative storytelling in games, television and graphic novels. *arXiv preprint arXiv:2305.13137*.

Jing Pan, Jian Wu, Yashesh Gaur, Sunit Sivasankaran, Zhuo Chen, Shujie Liu, and Jinyu Li. 2023. Cosmic: Data efficient instruction-tuning for speech in-context learning. *arXiv preprint arXiv:2311.02248*.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an asr corpus based on public domain audio books. In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, pages 5206–5210. IEEE.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. Meld: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 527–536.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *arXiv preprint*.

Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. 2021. Do vision transformers see like convolutional neural networks? *Advances in neural information processing systems*, 34:12116–12128.

ShoukanLabs. 2024. Anispeech: A dataset for anime-style speech. `https://huggingface.co/datasets/ShoukanLabs/AniSpeech`. Hugging Face dataset.

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*.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Paden Tomasello, Akshat Shrivastava, Daniel Lazar, Po-Chun Hsu, Duc Le, Adithya Sagar, Ali Elkahky, Jade Copet, Wei-Ning Hsu, Yossi Adi, and 1 others. 2023. Stop: A dataset for spoken task oriented semantic parsing. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 991–998. IEEE.

Liang-Hsuan Tseng, Yi-Chang Chen, Kuan-Yi Lee, Da-Shan Shiu, and Hung-yi Lee. 2025. Taste: Text-aligned speech tokenization and embedding for spoken language modeling. *arXiv preprint arXiv:2504.07053*.

Irina-Elena Veliche, Zhuangqun Huang, Vineeth Ayyat Kochaniyan, Fuchun Peng, Ozlem Kalinli, and Michael L Seltzer. 2024. Towards measuring fairness in speech recognition: Fair-speech dataset. *arXiv preprint arXiv:2408.12734*.

Chen Wang, Minpeng Liao, Zhongqiang Huang, Jinliang Lu, Junhong Wu, Yuchen Liu, Chengqing Zong, and Jiajun Zhang. 2023a. Blsp: Bootstrapping language-speech pre-training via behavior alignment of continuation writing. *arXiv preprint arXiv:2309.00916*.

Chen Wang, Minpeng Liao, Zhongqiang Huang, Junhong Wu, Chengqing Zong, and Jiajun Zhang. 2024a. Blsp-emo: Towards empathetic large speech-language models. *arXiv preprint arXiv:2406.03872*.

Chen Wang, Minpeng Liao, Zhongqiang Huang, and Jiajun Zhang. 2024b. Blsp-kd: Bootstrapping language-speech pre-training via knowledge distillation. *arXiv preprint arXiv:2405.19041*.

Dingdong Wang, Jincenzi Wu, Junan Li, Dongchao Yang, Xueyuan Chen, Tianhua Zhang, and Helen Meng. 2025a. Mmsu: A massive multi-task spoken language understanding and reasoning benchmark. *arXiv preprint arXiv:2506.04779*.

Hankun Wang, Haoran Wang, Yiwei Guo, Zhihan Li, Chenpeng Du, Xie Chen, and Kai Yu. 2024c. Why do speech language models fail to generate semantically coherent outputs? a modality evolving perspective. *arXiv preprint arXiv:2412.17048*.

Mingqiu Wang, Wei Han, Izhak Shafran, Zelin Wu, Chung-Cheng Chiu, Yuan Cao, Nanxin Chen, Yu Zhang, Hagen Soltau, Paul K Rubenstein, and 1 others. 2023b. Slm: Bridge the thin gap between speech and text foundation models. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–8. IEEE.

Wenbin Wang, Yang Song, and Sanjay Jha. 2024d. Globe: A high-quality english corpus with global accents for zero-shot speaker adaptive text-to-speech. *arXiv preprint arXiv:2406.14875*.

Ziqian Wang, Xianjun Xia, Xinfa Zhu, and Lei Xie. 2025b. U-sam: An audio language model for unified speech, audio, and music understanding. *arXiv preprint arXiv:2505.13880*.

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

Junichi Yamagishi, Christophe Veaux, and Kirsten MacDonald. 2019. Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit (version 0.92). *The Rainbow Passage which the speakers read out can be found in the International Dialects of English Archive:(http://web. ku. edu/˜ idea/readings/rainbow. htm)*.

Qian Yang, Jin Xu, Wenrui Liu, Yunfei Chu, Ziyue Jiang, Xiaohuan Zhou, Yichong Leng, Yuanjun Lv, Zhou Zhao, Chang Zhou, and 1 others. 2024. Air-bench: Benchmarking large audio-language models via generative comprehension. *arXiv preprint arXiv:2402.07729*.

Wanqi Yang, Yanda Li, Yunchao Wei, Meng Fang, and Ling Chen. 2025. Speechr: A benchmark for speech reasoning in large audio-language models. *arXiv preprint arXiv:2508.02018*.

Jeong Hun Yeo, Hyeongseop Rha, Se Jin Park, and Yong Man Ro. 2025. Mms-llama: Efficient llm-based audio-visual speech recognition with minimal multimodal speech tokens. *arXiv preprint arXiv:2503.11315*.

Tengfei Yu, Xuebo Liu, Zhiyi Hou, Liang Ding, Dacheng Tao, and Min Zhang. 2024. Self-powered llm modality expansion for large speech-text models. *arXiv preprint arXiv:2410.03798*.

Heiga Zen, Viet Dang, Rob Clark, Yu Zhang, Ron J Weiss, Ye Jia, Zhifeng Chen, and Yonghui Wu. 2019. Libritts: A corpus derived from librispeech for text-to-speech. *arXiv preprint arXiv:1904.02882*.

Hao Zhang, Nianwen Si, Yaqi Chen, Wenlin Zhang, Xukui Yang, Dan Qu, and Xiaolin Jiao. 2023. Tuning large language model for end-to-end speech translation. *arXiv preprint arXiv:2310.02050*.

Guanlong Zhao, Sinem Sonsaat, Alif Silpachai, Ivana Lucic, Evgeny Chukharev-Hudilainen, John Levis, and Ricardo Gutierrez-Osuna. 2018. L2-arctic: A non-native english speech corpus. In *Proc. Interspeech*, pages 2783–2787.

Kun Zhou, Berrak Sisman, Rui Liu, and Haizhou Li. 2021. Seen and unseen emotional style transfer for voice conversion with a new emotional speech dataset. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 920–924. IEEE.

Maike Züfle and Jan Niehues. 2024. Contrastive learning for task-independent speechllm-pretraining. *arXiv preprint arXiv:2412.15712*.

Score	Criterion
0	Useless: ignores the instruction or refuses to respond
1	Poor: barely related, incomplete, or unhelpful
2	Weak: partially relevant but lacking clarity or depth
3	Fair: mostly relevant and informative, but limited in quality
4	Good: relevant, coherent, and reasonably informative
5	Excellent: directly follows the instruction and provides clear, useful content

Table 5: LLM-as-a-judge evaluation scale.

A Appendix

A.1 Further Details on Dataset and Implementation

Table 9 summarizes the datasets used for training. In addition to speech transcriptions, the datasets include various speech-specific annotations such as emotion, intent, and gender. These attributes are jointly paired with the corresponding transcriptions and are utilized during response generation, as described in Equation 2.

Table 10 presents the instruction set I adopted in our training framework. The instruction set is largely based on prior work (Yu et al., 2024) but the Speech Translation category is excluded in our setting since it yields too many language pairs.

After response generation, we apply an additional filtering step to remove low-quality responses using an LLM-as-a-judge strategy. Specifically, we employ the same backbone model, Qwen2.5-7B-Instruct (Team, 2024), to score each instruction–response pair on a 5-point scale as shown in Table 5. Only samples with a score of 3 or higher are utilized for training.

Our modality adapter $\psi(\cdot)$ is implemented using a Q-former architecture with a maximum query length of 512. It adopts the same hidden dimensionality as the LLM and consists of 2 transformer layers with 4 attention heads. During training, we freeze the speech encoder, and apply LoRA to the LLM with conservative settings ($r = 2$, $\alpha = 2$), as we observed that larger LoRA configurations led to excessive deviation in the LLM behavior. Our model contains approximately 350M trainable parameters.

Training is conducted on 8 NVIDIA H100 GPUs, with a per-device batch size of 10 and 30 gradient accumulation steps to stabilize the token-wise internal alignment loss. Our final model is trained with $\lambda = 0.1$ for 8K training steps with a learning rate of 5×10^{-5}. Optimization is performed using AdamW with $\beta_1 = 0.9$, $\beta_2 = 0.99$, and $\epsilon = 1\text{e}{-08}$.

	LibriSpeech	LibriTTS	IEMOCAP
L1 distance	8.198	4.073	4.071
Pearson Correlation	0.971	0.972	0.920

Table 6: Performance of our speech rate predictor across datasets. We report L1 distance and Pearson correlation with respect to ground-truth token lengths.

A.2 Additional Details for Speech-Specific Information Analysis

We use the following instruction for the emotion classification analysis:

"Please select the most appropriate emotion expressed in the following speech. Choose only from <angry, happy, sad, neutral, frustrated, excited, fear, surprise, disgust>. Respond with exactly one label only."

We construct the input prompts using a chat template that includes both a system prompt and the above instruction. For the purpose of analysis, the ground-truth emotion label is appended to the input sequence, allowing us to examine attention patterns directed toward the label tokens. For each layer, we average the attention weights across grouped-query attention (GQA) heads and visualize the resulting attention maps for both speech inputs and text transcriptions.

In some emotion categories, the tokenizer splits the emotion label into multiple tokens. In such cases, we compute the average attention weights across the corresponding label tokens and visualize them as a single emotion label. In Figure 5, the attention weights at the emotion label positions are set to zero for clarity.

A.3 Artifact Licenses

We use publicly available datasets (Table 9), benchmarks, and pretrained models in accordance with their respective licenses and terms of use. The benchmarks used in our experiments, including Air-Bench, SpeechR, MMSU, and Speech-IFeval, are used solely for research evaluation. We do not redistribute the original datasets, benchmark data, or model checkpoints. For pretrained models and codebases, we follow the licenses and usage conditions specified by the original authors or providers.

	Air-bench	SpeechR	MMSU	Speech-IFeval	Rel. Δ (%) ↑	CKA
$\lambda = 0.0$	7.24	52.00	53.14	-17.68	-10.66%	0.6044
$\lambda = 0.1$	**7.85**	52.91	**53.85**	**-12.18**	**0.00%**	0.6399
$\lambda = 0.2$	7.78	**54.90**	53.77	-13.81	-2.31%	0.6601
$\lambda = 0.4$	7.57	52.17	53.12	-17.48	-9.20%	0.6914
$\lambda = 0.8$	7.64	53.15	53.62	-17.66	-8.44%	0.7266
$\lambda = 2.0$	7.62	50.77	51.81	-23.62	-14.90%	**0.744**

Table 7: Ablation results across different values of λ. We report performance on multiple benchmarks along with representation similarity (CKA).

	Air-bench	SpeechR	MMSU	Speech-IFeval	Rel. Δ (%) ↑
instructions=1	7.59	47.33	52.45	-22.94	-16.20%
instructions=2	7.64	48.04	52.22	-23.92	-16.27%
instructions=3	6.96	48.46	52.10	-27.71	-20.34%
instructions=4	7.53	48.66	53.77	-23.08	-15.09%
instructions=5	7.74	50.59	**53.85**	-21.51	-12.35%
instructions=6	**7.85**	**52.91**	**53.85**	**-12.18**	**0.00%**

Table 8: Ablation results with varying numbers of instructions. All metrics improve consistently as the number of instructions increases, with the largest gains observed when $instructions = 6$. Higher is better for all metrics, while Speech-IFeval is better when closer to zero.

Dataset	**Hours**	**Information**
GigaSpeech (Chen et al., 2021)	10,044.55	transcription, data source
DailyTalk (Lee et al., 2022)	21.67	transcription, emotion, action
LJSpeech (Ito and Johnson, 2017)	23.92	transcription
SLURP (Bastianelli et al., 2020)	26.27	transcription, intent, action, scenario
VCTK (Yamagishi et al., 2019)	43.89	transcription, age, gender, accent, region
Libriheavy (Kang et al., 2023)	51,024.12	transcription
LibriTTS (Zen et al., 2019)	585.83	transcription
Librispeech (Panayotov et al., 2015)	961.05	transcription
People's speech (Galvez et al., 2021)	6,246.09	transcription
AniSpeech (ShoukanLabs, 2024)	34.79	transcription
EMNS (Noriy et al., 2023)	1.91	transcription, emotion, gender, age
EmoV-DB (Adigwe et al., 2018)	9.49	transcription, emotion
ESD (Zhou et al., 2021)	13.41	transcription, emotion
EXPRESSO (Nguyen et al., 2023)	10.18	transcription, emotion
Fair-Speech (Veliche et al., 2024)	55.55	transcription, gender, age, first language, socioeconomic background, ethnicity
FSC (Lugosch et al., 2019)	14.72	transcription, action, object, location
GLOBE (Wang et al., 2024d)	611.99	transcription, accent, age, gender
IEMOCAP (Busso et al., 2008)	12.44	transcription, gender, speaking rate, pitch, relative dB, emotion, emotion intensity
JL-Corpus (CLAPv2, 2025)	1.41	transcription, emotion, country
L2-ARTIC (Zhao et al., 2018)	27.51	transcription
MELD (Poria et al., 2019)	8.72	transcription, emotion
STOP (Tomasello et al., 2023)	116.59	transcription, gender, speaker nativeness, intent

Table 9: Summary of datasets

Category	Instruction
Content Repetition	1. Repeat the provided speech, ensuring to maintain its original meaning and details. 2. Provide the speech exactly as given—do not alter wording, structure, or omit any content. 3. Echo the content of the speech, maintaining its exact purpose and details.
Keyword Extraction	1. Extract the most frequently occurring words or phrases in the speech, excluding common stopwords, to identify main topics. 2. Identify and list the most common words or phrases from the speech, omitting typical stopwords, to highlight central themes. 3. Extract significant words or phrases that appear often in the speech, exclude basic stopwords, to uncover the main subjects.
Intent Recognition	1. Determine the primary purpose of the speech and evaluate how clearly and effectively the message is conveyed. 2. Identify the main intent of the speech and assess the clarity and effectiveness of its delivery. 3. Assess the central purpose of the speech and evaluate the directness and impact of its expression.
Sentiment Analysis	1. Determine the sentiment of the speech and identify which sections contribute most to sentiment. 2. Evaluate the emotional tone of the speech and determine which segments primarily affect the sentiment. 3. Assess the sentiment expressed in the speech and highlight which areas contribute most to this feeling.
Continuation	1. Please write a coherent and engaging continuation of the given speech with less than 50 words. 2. Compose a logical and captivating follow-up to the provided speech within 50 words. 3. Write a fluent and engaging continuation of the speech, limited to 50 words.
Speech Recognition	1. Provide the transcription according to the speech. 2. Convert the spoken language into a written transcript. 3. Write down the speech as a text transcript.

Table 10: Instruction Set