PARAS2S: BENCHMARKING AND ALIGNING SPOKEN LANGUAGE MODELS FOR PARALINGUISTIC-AWARE SPEECH-TO-SPEECH INTERACTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Speech-to-Speech (S2S) models have shown promising dialogue capabilities, but their ability to handle paralinguistic cues—such as emotion, tone, and speaker attributes—and to respond appropriately in both content and style remains underexplored. Progress is further hindered by the scarcity of high-quality and expressive demonstrations. To address this, we introduce a novel reinforcement learning (RL) framework for paralinguistic-aware S2S, **ParaS2S**, which evaluates and optimizes both content and speaking style directly at the waveform level. We first construct **ParaS2SBench**, a benchmark comprehensively evaluates S2S models' output for content and style appropriateness from diverse and challenging input queries. It scores the fitness of input-output pairs and aligns well with human judgments, serving as an automatic judge for model outputs. With this scalable scoring feedback, we enable the model to explore and learn from diverse unlabeled speech via Group Relative Policy Optimization (GRPO). Experiments show that existing S2S models fail to respond appropriately to paralinguistic attributes, performing no better than pipeline-based baselines. Our RL-based strong baseline achieves a 11% relative improvement in response content and style's appropriateness on ParaS2SBench over supervised fine-tuning (SFT), surpassing all prior models while requiring substantially fewer warm-up annotations than pure SFT[1].

1 INTRODUCTION

Speech is the most natural medium of communication, conveying not only words but also paralinguistic cues—emotion, tone, and speaker attributes—that jointly shape true intent and guide appropriate responses (Schuller & Batliner, 2013). This interplay of linguistic and paralinguistic signals motivates speech-to-speech (S2S) models (Xu et al., 2025; Huang et al., 2025b; Zeng et al., 2024) for human-like, empathetic interaction beyond text-based dialogue systems (Achiam et al., 2023; Grattafiori et al., 2024).

S2S models show strong dialogue abilities (Fang et al., 2025a;b; Zeng et al., 2024), as seen in Qwen2.5-Omni (Xu et al., 2025) and ChatGPT advanced voice mode.[2] Built on LLMs, they preserve reasoning and conversational abilities while adding speech as a new I/O modality, achieving high scores on benchmarks like VoiceBench (Chen et al., 2024) and Llama Questions (Nachmani et al., 2024). Yet most benchmarks focus on question answering (Nachmani et al., 2024), instruction following (Lu et al., 2025), or speech-to-text understanding tasks (Yang et al., 2024; Sakshi et al., 2025b), overlooking paralinguistic-aware dialogue. StyleTalk (Lin et al., 2024a) and Vox-Dialogue (Cheng et al., 2025) partially address the problem but remain *speech-to-text* benchmarks where evaluation ends at the textual response, leaving no benchmark that directly evaluates S2S models' response speech for paralinguistic awareness.

Beyond the lack of benchmarks, no paralinguistic-aware S2S models currently exist. Our study shows that most S2S models fail to appropriately adjust responses according to different speaking styles (e.g., emotional tone), often inferring speaker state from content alone and producing tone-deaf or awkward replies. This limitation stems from existing spoken dialogue datasets, which rarely

[1] Project page and demo: https://paras2sbench.github.io/
[2] https://openai.com/index/chatgpt-can-now-see-hear-and-speak/

capture the style dynamics between input and output (Ding et al., 2025; Fang et al., 2025a;b). Collecting such data is expensive, as it requires style annotation and expressive response recording, making data scarcity the main bottleneck for developing paralinguistic-aware S2S models (Huang et al., 2025b).

Inspired by DeepSeek-R1 (Guo et al., 2025), which acquires reasoning capabilities through RL without any SFT demonstrations, we ask *whether paralinguistic-aware dialogue capabilities can similarly emerge via RL with minimal supervision.* To answer, we introduce a novel framework for paralinguistic-aware S2S, **ParaS2S**. ParaS2S comprises a new S2S benchmark **ParaS2SBench** and a RL learning framework **ParaS2SAlign**. **ParaS2SBench** is designed to jointly evaluates both the content and speaking style of input and output speech, guided by three key design principles:

1. **Speech-to-speech evaluation.** Evaluation is performed directly on input and output speech, assessing whether the model generates responses with both appropriate content and speaking style given the input speech.
2. **Contrasting speaking styles.** Following StyleTalk (Lin et al., 2024a), each test query is paired with two *contrasting* speaking styles that demand distinct responses. For example, *"I just bumped into my ex."* may be spoken in either a surprised or sad tone.
3. **Scenario-controlled queries.** We design each query to have *neutral text content* so that models cannot guess the speaker's state from words alone, and to be *paralinguistically relevant* so that the speaking style genuinely changes how the response should be generated.

We design a data curation pipeline to automatically generate high-quality speech prompts covering key paralinguistic aspects—emotion, sarcasm, age, and gender. Using this benchmark, we expose the common tone-deaf issue in current S2S models, including state-of-the-art (SOTA) open-source models such as Qwen2.5 Omni (Xu et al., 2025) and Kimi-Audio (Ding et al., 2025), as well as closed-source systems such as ChatGPT advanced voiced mode (Achiam et al., 2023).

To advance model development, we establish **ParaS2SAlign** as the strong baseline in the benchmark platform. By leveraging a Speech-to-Text reasoning model (Xie et al., 2025; Radford et al., 2023) and text LLM, we automate benchmark evaluation and provide an automatic judge for model outputs that correlates with human scoring. Building on the scalability of this scoring pipeline, we generate a large-scale preference dataset[3] and distill the benchmark pipeline into a single reward model to enable RL. With Group Relative Policy Optimization (GRPO) (Shao et al., 2024), the base S2S model learns from diverse unlabeled speech prompts and from its own generated outputs automatically scored by the reward model, thereby unlocking paralinguistic-aware S2S capabilities through RL. Our results show that while supervised fine-tuning (SFT) is effective and outperforms existing models[4], RL surpasses SFT by more than 11% in response content and style appropriateness on ParaS2SBench and 7.6% on subjective evaluation. Furthermore, in cost-controlled experiments, RL requires only 10 hours of demonstration as warm-up and achieves the same performance as pure SFT using just one fifth of the annotations, highlighting its learning efficiency. Our contributions are multifold:

- We present a novel benchmark, **ParaS2SBench**, for paralinguistic-aware S2S dialogue. It directly evaluates both the content and speaking style of input–output speech pairs at the waveform level, revealing the common tone-deaf issue in current S2S models.
- We establish **ParaS2SAlign**, the first RL framework for paralinguistic-aware S2S. By automating and distilling the benchmark pipeline into a reward model, we enable scalable learning from unlabeled speech without costly demonstrations, serving as the strong baseline in the benchmark platform.
- We demonstrate that RL with GRPO achieves a 11% relative improvement in GPT-based scores on ParaS2SBench and 12% on real speech queries over SFT. Furthermore, We highlight the cost efficiency of RL compared to SFT, mitigating the data scarcity of paralinguistic-aware S2S.
- We will open-source data, code, and models to lower the barrier for future research.

[3]This process would be costly if the response speech and preference scores were annotated by humans.
[4]At the cost of requiring expensive and non-scalable demonstrations.

2 RELATED WORK

2.1 SPOKEN DIALOGUE MODELS

From S2T to S2S dialogue models. Early Speech-to-Text LLMs equip LLMs with *hearing* capabilities while leveraging textual reasoning for audio interaction (Tang et al., 2024; Hu et al., 2024; Gong et al., 2024). AudioReasoner (Xie et al., 2025) introduces Chain-of-Thought (CoT) reasoning to mitigate hallucination, while Qwen-Audio 1/2 (Chu et al., 2023; 2024) and StepAudio (Huang et al., 2025b) further extend dialogue capabilities to enable spoken agents[5]. Recent works explore Speech-to-Speech LLMs that learn input–output speech interaction end-to-end (Zhang et al., 2023; Défossez et al., 2024). GLM-4-Voice (Zeng et al., 2024) and Step-Audio-AQAA (Huang et al., 2025a) rely on interleaved text and audio tokens for grounded speech generation. LLaMa-Omni (Fang et al., 2025a;b), Freeze-Omni (Wang et al., 2025b) and Mini-Omni (Xie & Wu, 2024) propose fine-tuning techniques to preserve LLM intelligence when adding speech modality. Qwen2.5 Omni (Xu et al., 2025) proposes the thinker-talker architecture, while Kimi-Audio (Ding et al., 2025) introduces a dual-head design for text and audio generation.

Paralinguistic-aware dialogue models. Among these models, ParalinGPT (Lin et al., 2024b) and StyleTalk (Lin et al., 2024a) are the first to enable speech-to-text LLMs to respond differently to diverse speaking styles. OmniChat (Cheng et al., 2024) extends the speech-to-text study to multi-turn, paralinguistic-aware dialogues. For speech-to-speech models, GOAT-SLM (Chen et al., 2025) is the only model emphasizing paralinguistic-aware dialogue with a multi-stage SFT pipeline. These works rely on SFT with carefully curated, high-quality data, whereas we explore RL to reduce this reliance.

RL for dialogue models. RL has been applied to align spoken dialogue models. Align-SLM (Lin et al., 2025b) follows RLAIF (Lee et al., 2024) and adopts DPO (Rafailov et al., 2023) to improve long-range semantics. Qwen2.5 Omni (Xu et al., 2025) uses WER as a preference signal to ground speech generation. Step-Audio (Huang et al., 2025b) and Step-Audio-AQAA (Huang et al., 2025a) rely on human feedback, which is annotation-heavy. ParaS2SAlign is the first RL framework to model content–style and input-output dynamics using scalable AI feedback.

2.2 SPOKEN DIALOGUE BENCHMARKS

Benchmarks have been proposed to evaluate spoken dialogue models. Table 4 compares key differences across benchmarks. Dynamic-SUPERB (Huang et al., 2024) tests instruction-following on 180 tasks (yu Huang et al., 2025). AudioBench (Wang et al., 2025a) unifies speech/sound understanding and QA. AIR-Bench (Yang et al., 2024) adds speech, sound, music tasks, and a *chat* category. MMAU (Sakshi et al., 2025a) raises difficulty with reasoning-intensive QA. SpokenWOZ (Si et al., 2023) provides large-scale human-to-human dialogue data. VoxEval (Cui et al., 2025) converts MMLU (Hendrycks et al., 2021) to speech to assess model intelligence. VoiceBench (Chen et al., 2024) adds more text-based QA datasets including AlpacaEval (Li et al., 2023), OpenBookQA (Mihaylov et al., 2018), and MMLU-pro (Wang et al., 2024). FullDuplexBench (Lin et al., 2025a) evaluates response timing for full-duplex models. Among these works, ADU-Bench (Gao et al., 2025), SD-eval (Ao et al., 2024), VoxDialogue (Cheng et al., 2025), and StyleTalk (Lin et al., 2024a) evaluate responses under different input speaking styles, but focus only on the dialogue models' output text.[6] In contrast, ParaS2SBench performs end-to-end evaluation on both input and output speech, jointly considering content and speaking style.

3 PARAS2SBENCH

ParaS2SBench is a benchmark designed to evaluate paralinguistic-aware S2S models. In Section 3.1, we describe the process of curating training and testing queries that serve as inputs for evaluation.

[5]The response is usually in text, and the *speaking* capability is enabled by a separate TTS module.

[6]StyleTalk predicts both response text and style in textual format, enabling style learning and evaluation. However, it is limited to the few categorical styles supported by Microsoft Azure TTS, and its format assumption prevents evaluation of S2S models.

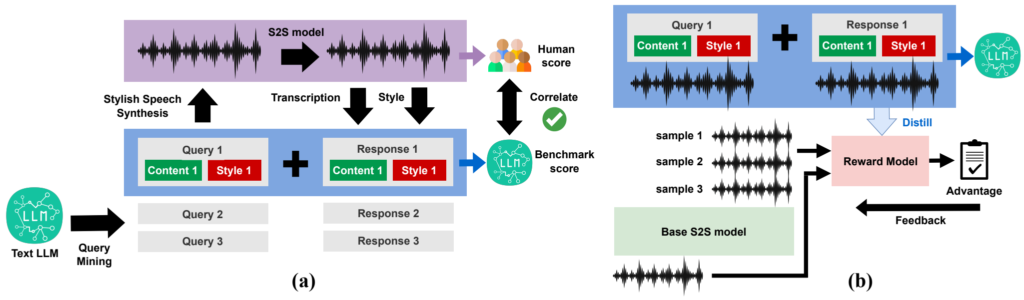


Figure 1: The overall framework of ParaS2S. (a) illustrates the pipeline of ParaS2SBench; (b) illustrates the framework of ParaS2SAlign.

In Section 3.2, we present the methodology for automatically evaluating model responses given an input query.

3.1 Query Mining – Synthetic and Real

As shown by Figure 1 (a), ParaS2SBench begins by generating scenario-controlled and challenging queries with LLM that specify both the content and the corresponding speaking style, followed by synthesizing these queries using suitable text-to-speech (TTS) systems. The queries span a wide range of dialogue topics and scenarios, and the speaking styles cover various key paralinguistic factors, including emotion, sarcasm, gender, and age[7]. Mining appropriate queries is necessary for evaluating paralinguistic-aware S2S since many speech queries lack paralinguistic dynamics[8]. Such queries are unsuitable for evaluation, as models can answer correctly without considering style. We design an automatic data curation pipeline to mine the realistic and challenge testing queries. The pipeline relies on ChatGPT, and we include prompts in Appendix. Table 6 shows several examples and the demo page[1] demonstrates more examples.

1. **Candidate Generation.** We first generate a large corpus of queries with ChatGPT, each consisting of a input spoken sentence $c_i \in \Sigma^*$ followed by two *contrasting* speaking styles, $s_i^1, s_i^2 \in \Sigma^*$, that demand different responses. In the generation prompt, we instruct ChatGPT to cover diverse topics and scenarios, including interests, work, studies, relationships, travel, health, religion, fashion, finance, and more.
2. **Script Quality Filtering.** For each spoken content c_i, we construct two queries, (c_i, s_i^1) and (c_i, s_i^2). For each query (c_i, s_i), we control the scenario by asking ChatGPT for several checks, including neutrality, reasonability and paralinguistic relevance. Neutrality prevents models from inferring the speaker's state solely from text c_i; reasonability ensures the content c_i and style s_i is a reasonable pair; paralinguistic relevance ensures that speaking style non-trivially affects the response. If any test is not passed, the query is discard. Appendix A.3 provides more explanations.
3. **Speech Synthesis.** We synthesize input waveform $w_i \in R^*$ given the (c_i, s_i) pair. For emotion and sarcasm, we rely on the instruction-based TTS system *gpt-4o-mini-tts*[9]. This system requires a style description, which we generate with ChatGPT based on the style label s_i. Since *gpt-4o-mini-tts* supports only a limited number of speakers, we use CosyVoice (Du et al., 2024) for in-context zero-shot synthesis of gender and age. The voice samples for gender are drawn from LibriSpeech (Panayotov et al., 2015) and CommonVoice (Ardila et al., 2020), while the samples for age are drawn from NNCES[10]. We

[7]We exclude emphasis, volume, and speed because our preliminary study shows they rarely affect human preferences. For example, when given "I want to borrow this book (fast)," people preferred "Sure, please give me your ID" over "Could you slow down? You speak too fast."

[8]For example, the factual question *Who is the first president of America?* should yield the same answer regardless of the speaker's background or style.

[9]https://www.openai.fm/

[10]https://www.kaggle.com/datasets/kodaliradha20phd7093/nonnative-children-english-speech-nnces-corpus . We do not use MyST (Pradhan et al., 2024) since the data link is unavailable.

discard samples whose WER with the ground truth exceeds a threshold. For emotion, we further discard samples whose Emotion2vec (Ma et al., 2024) classifier scores are too low (Cheng et al., 2025).

4. **Train/Test Split.** To avoid overlap between training and testing, we use disjoint query topics and TTS speakers.
5. **Human Check.** To ensure test set authenticity, we recruit three annotators to manually include only speech prompts with correct content and style from the filtered set.

The above pipeline curates a synthetic test set covering emotion, sarcasm, age, and gender. To further examine model behavior in realistic scenarios, we construct a test set using real speech by filtering queries from existing dialogue datasets. Given the known content and style labels provided by the dataset, we apply filters to check the length[11] and paralinguistic relevance. We rely on two emotion datasets, IEMOCAP (Busso et al., 2008) and MELD (Poria et al., 2019), as they provide sufficient queries that meet our constraints. In contrast, we find it challenging to source enough queries for age and gender from datasets like CommonVoice due to the paralinguistic relevance constraint[12]. Finally, we construct a testing set $D_{\text{test}} = \{(c_i, s_i, w_i)\}$ where $c_i \in \Sigma^*$ is the input spoken content, $s_i \in \Sigma^*$ is the input speaking style, and $w_i \in R^*$ is the input audio prompt. Table 5 shows the statistics.

3.2 Response evaluation & scoring

Given an input query $(c_i, s_i, w_i) \sim D_{\text{test}}$, the S2S model M samples a response speech $w_o \sim \pi_M(O|w_i)$. To evaluate the response speech, we project both the content and the speaking style of w^o into natural language, using SOTA Speech-to-Text models C and S, respectively. We rely on Whisper-v3 (Radford et al., 2023) as C to get the output transcription: $c_o = C(w_o)$. We leverage AudioReasoner (Xie et al., 2025) as S to extract output speaking tone: $s_o = S(w_o)$. AudioReasoner equips Qwen-Audio 2 (Chu et al., 2024) with reasoning capabilities by distilling Chain-of-Thought (CoT) paths from Gemini (Team et al., 2024) to reduce hallucination. Finally, given the input content c_i and style s_i, along with the extracted output content c_o and style s_o, we use ChatGPT 4.1 to score the fitness following the guideline r designed by human experts, described in the Appendix.

$$f_{\text{gpt}} = GPT(c_i, s_i, c_o, s_o, r) \tag{1}$$

We will show that this scoring pipeline can align with human judgments f_{expert} in Section 5.1. Both f_{gpt} and f_{expert} are on a 1–5 Likert scale.

4 ParaS2SAlign

Although ParaS2SBench provides automatic fitness scores, the scoring process is slow: it requires a reasoning-based speech-to-text LLM and ChatGPT API calls, so even a small batch of responses takes several minutes. This makes typical online RL training impractical when rewards are computed directly from the benchmark evaluation pipeline, and also makes it prohibitively expensive to construct a large-scale preference dataset for direct preference optimization (DPO) (Rafailov et al., 2023). To address this, we design a three-stage online RL framework that uses a reward model to approximate the benchmark pipeline and employs GRPO (Shao et al., 2024). Figure 1 (b) illustrates the framework.

We use Kimi-Audio (Ding et al., 2025) as the base model θ_{base}[13], while the framework can be applied to any LM-based S2S model. For Kimi-Audio, the audio input w_i, text input c_i, audio output w_o, and text output c_o are preprocessed and organized into four token streams: $a_i, t_i \in \mathbb{Z}^{L_i}$ and $a_o, t_o \in \mathbb{Z}^{L_o}$. The input streams (a_i, t_i) are padded to the same length $L_i \in \mathbb{Z}$, and the output streams to $L_o \in \mathbb{Z}$. The input embeddings of the audio and text streams are summed before being

[11] In real dialogues, some turns consist of only a few words (e.g., *Haha* or *Sounds good*), which are not suitable for evaluation. We therefore filter out queries with fewer than five words.

[12] For example, in *Could you read the book for me? (female)*, the gender attribute is negligible.

[13] Since it exhibits high intelligence and strong dialogue capabilities (Chen et al., 2024) and is fully open-sourced. We do not use Qwen2.5-Omni (Xu et al., 2025) because its speech tokenizer is not released, making S2S fine-tuning infeasible.

fed into the Transformer, and from the middle of the model, two prediction heads predict the next token for each stream.

$$\pi_\theta(a_o, t_o \mid a_i, t_i) = \prod_{n=1}^{|a_o|} \pi_\theta(a_{o,n}, t_{o,n} \mid a_{o,<n}, t_{o,<n}, a_i, t_i) \tag{2}$$

For inference, output audio and text tokens are sampled auto-regressively $(a_o, t_o) \sim \pi_\theta(O \mid a_i, t_i)$. Audio tokens are decoded into the sampled waveform with a flow-matching decoder: $w_o = \rho(a_o)$.

4.1 Stage 1. Warm-up

SFT serves as a crucial warm-up stage for RL, as we observe that existing S2S models lack paralinguistic-aware dialogue capabilities. Consequently, they fail to sample high-quality responses and cannot provide a useful learning signal for RL. To construct the SFT dataset D_{sft}, we follow Section 3.1 to generate a training set of speech queries with both input content c_i and style labels s_i. For each query (c_i, s_i), we use ChatGPT to produce the most suitable response (c_o, s_o), including both a textual transcription and a tone description. We then synthesize the expressive response w_o using *gpt-4o-mini-tts*. Because *gpt-4o-mini-tts* can be unstable, we synthesize 10 candidates, apply WER-based filtering, and perform manual selection to obtain high-quality warm-up demonstrations w_o. With the input–output mapping $D_{\text{sft}} = \{(w_i, w_o, c_i, c_o)\}$, we train next-token prediction on both the preprocessed audio stream $a_i \| a_o$ and the text stream $t_i \| t_o$ by optimizing θ for higher likelihood $\mathbb{E}_{D_{\text{sft}}}[\pi_\theta(a_o, t_o \mid a_i, t_i)]$, initializing from θ_{base} and obtaining θ_{sft}.

4.2 Stage 2. Distilling Reward Model

To distill our benchmark pipeline into a reward model, we construct a preference dataset D_{prefer}. We first prepare Q speech queries $\{(c_i^j, s_i^j, w_i^j)\}_{j=1}^{Q}$ following Section 3.1. The SFT model now possesses preliminary paralinguistic-aware dialogue capabilities and begins to respond differently according to the input speaking styles, but unstably. Each query (c_i, s_i, w_i) is preprocessed into input token streams (a_i, t_i). We sample K diverse speech responses with high sampling temperature, $(a_o, t_o) \sim \pi_\theta(O \mid a_i, t_i), w_o = \rho(a_o)$. We then score the resulting $Q \times K$ query–response pairs following Equation 1 to construct a preference dataset $D_{\text{prefer}} = \{(w_i, w_o, f_{\text{gpt}})\}$, where f_{gpt} is the fitness score of w_i and w_o, depending on content label c_i, style label s_i, extracted content $c_o = C(w_o)$ and extracted style $s_o = S(w_o)$. Finally, we use LoRA (Hu et al., 2022) to fine-tune Qwen2.5-Omni (Xu et al., 2025) as the reward model, which is employed as a Speech-to-Text LLM. The model takes the query speech, response speech, and scoring guideline r as input, and outputs a single score on a Likert scale. We denote the reward model as ϕ. The score is treated as a single character and optimized with the cross entropy loss: $\mathbb{E}_{D_{\text{prefer}}} \phi(f_{\text{gpt}} \mid w_i, w_o, r)$.

4.3 Stage 3. Post-training

Using the warm-up model θ_{sft} and the reward model ϕ, we enable the model to explore the search space for higher scores via GRPO (Shao et al., 2024) on the large set of unlabeled speech. We do not use PPO (Schulman et al., 2017) due to its substantial memory and computational burden of the value function. Moreover, in our case, only the last token of the response is assigned a final reward, which complicates the training of the value function that needs to be accurate at every token (Shao et al., 2024). Given the unlabeled speech prompt dataset $D_{\text{rl}} = \{w_i\}$, we obtain the transcription with Whisper-v3 and construct input token streams $D'_{\text{rl}} = \{(w_i, a_i, t_i)\}$. We optimize θ_{sft} to maximize the objective:

$$
\begin{aligned}
\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}\big[(w_i, a_i, t_i) \sim D'_{\text{rl}}, \{(a_o^g, t_o^g)\}_{g=1}^{G} \sim \pi_{\theta_{\text{old}}}(O \mid a_i, t_i)\big] \\
\frac{1}{G}\sum_{g=1}^{G}\frac{1}{|a_o^g|}\sum_{n=1}^{|a_o^g|}\Bigg\{ \min\Bigg[\frac{\pi_\theta(a_{o,n}^g, t_{o,n}^g \mid a_i, t_i, a_{o,<n}^g, t_{o,<n}^g)}{\pi_{\theta_{\text{old}}}(a_{o,n}^g, t_{o,n}^g \mid a_i, t_i, a_{o,<n}^g, t_{o,<n}^g)}\hat{A}^g, \\
\text{clip}\Bigg(\frac{\pi_\theta(a_{o,n}^g, t_{o,n}^g \mid a_i, t_i, a_{o,<n}^g, t_{o,<n}^g)}{\pi_{\theta_{\text{old}}}(a_{o,n}^g, t_{o,n}^g \mid a_i, t_i, a_{o,<n}^g, t_{o,<n}^g)}, 1-\epsilon, 1+\epsilon\Bigg)\hat{A}^g\Bigg] - \beta\,\mathbb{D}_{\text{KL}}[\pi_\theta \,\|\, \pi_{\text{ref}}]\Bigg\},
\end{aligned}
\tag{3}
$$

We sample prompts from D'_{rl}, generate G responses, decode tokens into waveforms with ρ, score them with ϕ, compute the normalized advantage $\hat{A}^g = (\phi(w_i, \rho(a_o^g), r) - \mu)/\sigma$, and update the policy θ for higher rewards. μ and σ are the mean and standard deviation of the raw scores within a group. ϵ is the clipping threshold. The KL term and the ablation of β are detailed in Appendix.

5 Experiments

Overview. In this section, we aim to answer two core research questions: (1) *Can the automatic benchmark scoring reflect human judgments and serve as an effective evaluation tool?* and (2) *Can RL truly improve performance compared to SFT for this problem?* Section 5.1 provides analyses and answers to the first question and show positive correlation. To thoroughly address the second question, we verify it across several subsections. Section 5.2 demonstrates that RL consistently leads to better performance regardless of the amount of SFT warm-up data, supporting our main claim. Section 5.3 points out that only a few hours of SFT data are sufficient to bootstrap the self-improvement process, indicating that RL indeed helps mitigate the data scarcity of paralinguistic-aware S2S. After validating the effectiveness of the RL algorithm, we move on to the practical concern of data construction cost. Section 5.4 provides best practices for balancing the budget allocation between SFT data and RL data, demonstrating the cost efficiency of the proposed framework. Next, we validate the generalizability of the training framework to real speech in Section 5.5, which is essential as the previous experiments are conducted on synthetic speech. Finally, Section 5.6 verifies that our training framework leads to a SOTA model in paralinguistic-aware dialogue by comparison with existing models. These subsections jointly gauge the effectiveness, generalizability, and cost efficiency of the proposed RL framework.

Settings. We construct a large-scale speech prompt dataset D_{rl} for RL following Section 3.1, where the transcription and style labels are discarded after speech synthesis. The dataset contains 100k speech prompts. For the less scalable SFT, we build prompt–demonstration pairs for 10k speech prompts, totaling 100 hours of data. For reward model data, which requires input style annotations during scoring, we use up to 10k speech prompts. For each prompt, the SFT model generates 32 completions, yielding 320k prompt–response–score pairs. More details are in Appendix A.4.

5.1 Can benchmark scores align with human scores?

Here, we evaluate whether the automatic evaluation benchmark scores align with human scoring. For this study, we sampled a subset from the benchmark for human annotation, with 200 prompts per paralinguistic category. For each speech prompt, we obtain two types of responses:

TTS-based responses: ChatGPT 4.1 generates the response content and style, and diverse TTS systems synthesize speech to simulate different speaking styles. We include YourTTS (Casanova et al., 2022), CosyVoice (Du et al., 2024), Sesame[14], and *gpt-4o-mini-tts*. These systems range from flat and neutral to expressive, spontaneous, and fine-grained controlled styles. We also add a baseline, *gpt-4o-mini-tts (bad)*, where we instruct ChatGPT 4.1 to produce suboptimal content or style such as tone-deaf content and inappropriate speaking style.

S2S model-based responses: End-to-end S2S models directly generate speech responses. We include GPT-4o Voice mode, Qwen2.5 Omni, and GLM-4-Voice.

[14] https://www.sesame.com/research/crossing_the_uncanny_valley_of_voice

Table 1: Comparison of GPT-based benchmark scoring and human evaluation across Age, Emotion, Gender, and Sarcasm tasks, with per-model averages.

	Age		Emotion		Gender		Sarcasm		Avg	
	GPT	Human	GPT	Human	GPT	Human	GPT	Human	GPT	Human
TTS-based										
gpt-4o-mini-tts (good)	4.420	4.380	4.646	4.654	4.739	4.506	4.790	4.337	4.649 **(1)**	4.469 **(1)**
gpt-4o-mini-tts (bad)	1.215	1.177	1.159	1.041	1.325	1.590	1.210	1.251	1.227 **(8)**	1.265 **(8)**
Sesame	4.412	4.216	4.512	4.324	4.701	4.332	4.71	4.182	4.583 **(2)**	4.263 **(2)**
CosyVoice	4.380	3.994	4.417	4.012	4.612	4.201	4.680	3.864	4.522 **(3)**	4.018 **(3)**
YourTTS	4.410	4.037	4.302	3.801	4.534	4.230	4.580	3.804	4.457 **(4)**	3.968 **(4)**
S2S models										
GPT-4o Voice mode	2.685	2.630	3.711	2.713	3.096	3.682	2.815	2.611	3.077 **(6)**	2.909 **(5)**
Qwen2.5 Omni	2.930	2.728	3.680	2.522	2.933	3.493	2.910	2.509	3.113 **(5)**	2.863 **(6)**
GLM-4-Voice	2.570	2.493	3.489	2.384	2.821	3.521	2.720	2.301	2.900 **(7)**	2.675 **(7)**

Table 2: Correlation between benchmark scoring and human scoring.

	Age	**Emotion**	**Gender**	**Sarcasm**	**All**
Pearson	0.862	0.76	0.702	0.779	0.773
p-value	$3.5e^{-5}$	$2.6e^{-12}$	$1.2e^{-4}$	$3.2e^{-3}$	$7.5e^{-6}$

TTS-based responses isolate the effect of response tone under identical gold content, while S2S responses reflect real model behavior. Each prompt–response pair is scored by three human experts on a Likert scale[15]. We also apply automatic scoring to study alignment. In Table 1, S2S responses lag significantly behind TTS responses due to the tone-deaf content, where the latter benefit from ground-truth style labels. The scores of S2S responses hover around 3, indicating models fail to adapt to contrasting speaking styles[16]. Second, across all models, the rankings with benchmark and human scores are nearly identical, with only one switch. The rankings of TTS systems are consistent: *gpt-4o-mini-tts* > Sesame > CosyVoice > YourTTS[17].

Then, we analyze the correlation between the benchmark scoring and human judgment. Each input query is paired with several TTS-based responses and several S2S model responses. For each query–response pair, we acquire two fitness scores, one from human experts and another from the benchmark pipeline, resulting in two arrays of fitness scores. We compute the Pearson correlation between these two sets of scores following (Chiang & Lee, 2023). Table 7 shows the correlation across different paralinguistic categories. All the correlations are higher than 0.7 and significant. These results validate the benchmark pipeline as a judge for RL feedback.

5.2 Can RL improve performance over SFT?

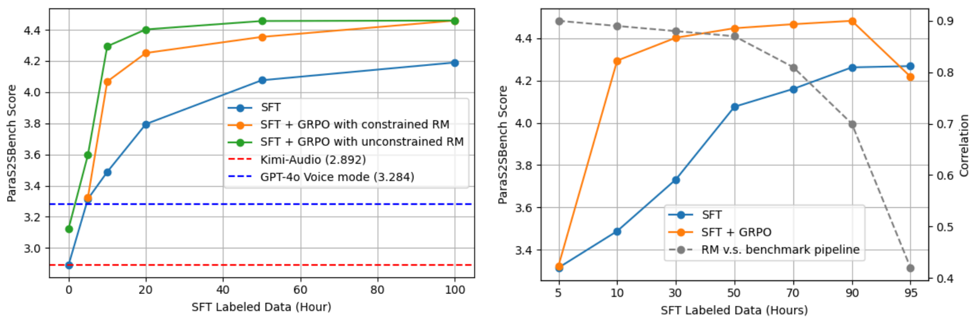
We study the effectiveness of RL under different amounts of SFT warm-up data. For reward model data, we consider a realistic setting: the constrained case, where the reward model is trained using the same amount of annotation as the SFT data. We also include an unconstrained case, where the reward model uses all available annotations[18]. Figure 2 (a) shows that SFT only already consistently demonstrates effectiveness: with only 10 hours of data, it surpasses most existing models, including GPT-4o voice mode, and continues to improve as data scales. However, RL in the constrained case consistently outperforms SFT across all data regimes: SFT requires more than ten times as much data to match RL performance. Using the reward model trained on all available preference data further unlocks additional gains.

[15]We first conducted preliminary annotations to align guidelines and maximize agreement, and discarded official annotations where all three experts disagreed.

[16]For prompts with two contrasting styles, models often score 5 for one response and 1 for the other tone-deaf response, averaging 3.

[17]Since the four TTS systems share the same response content, their scoring differences stem from the speaking style. We observe that AudioReasoner tends to classify CosyVoice and YourTTS outputs as calm, neutral, or flat, which is less empathetic.

[18]The curation of reward model data is still much cheaper than SFT data, as no human selection is involved.



(a) (b)

Figure 2: The ParaS2SBench score is the average across 4 categories. The trending of individual category is similar. (a) Comparison of RL and SFT results across different amounts of SFT data; (b) comparison of budget allocation between SFT and reward model (RM) data. The total budget consists of 100 hours of annotation, which are distributed between speech prompts for SFT data and reward model data. The gray dotted line shows the correlation between reward model prediction and the original benchmark pipeline score on a held-out test set.

5.3 How many annotations can RL save?

From Figure 2 (a), only 10 hours of SFT warm-up data are sufficient to unlock the model's ability to learn from self-generated demonstrations, improving upon the warm-up model by more than 17.1% and achieving performance comparable to using 50 hours of SFT data. Similarly, RL with a 20-hour warm-up performs comparably to 100 hours of SFT data, highlighting the strong label efficiency of our approach. Figure 2 (a) also shows that the warm-up stage is still critical, as the base model cannot sample sufficiently good demonstrations to evolve through RL.

5.4 Should we invest more costs on SFT or reward model?

Both the construction of SFT data and the reward model require style annotations for input prompts. Given a fixed number of prompt annotations, we study whether it is more beneficial to allocate them to SFT or reward model data. Figure 2(b) shows that increasing SFT data (and decreasing reward model data) consistently improves RL performance—until the reward model becomes poorly correlated with the benchmark scores. Surprisingly, only 10 hours of annotated speech prompts are sufficient to build a usable reward model. Thus, allocating more budget to SFT data is generally advantageous, since warm-up quality drives GRPO sampling and learning efficiency, while the reward model is easier to learn and still reaches a strong correlation of 0.7 with minimal annotation (e.g. 10 hours).

5.5 Generalizability to real speech

To test generalizability to real speech and unseen domains, we evaluate on IEMOCAP (Busso et al., 2008) and MELD (Poria et al., 2019). The former features recordings from professional actors in both scripted and spontaneous scenarios, while the latter comes from TV shows with natural conversations involving diverse speakers, emotions, and background noise. We verify that SFT and GRPO trained on synthetic speech generalize to real speech. Applying GRPO to real speech further improves performance on both in-domain and out-of-domain scenarios (Appendix A.8).

5.6 Comparing S2S Models – Automatic Judge and Human Evaluation

Table 3 compares several existing S2S models with ours. The Whisper-GPT-TTS pipeline uses Whisper-v3 to transcribe the input query without considering the speaking style, generates the response text with ChatGPT, and synthesizes speech with *gpt-4o-mini-tts*. This pipeline serves as a baseline where speaking style is ignored. The topline, on the other hand, leverages the ground-truth

Table 3: Comparing paralinguistic-aware dialogue capabilities with ParaS2SBench score.

	Synthetic					Real			Avg
	Age	Emotion	Gender	Sarcasm	Avg	IEMOCAP	MELD	Avg	
Baseline									
Whisper-GPT-TTS	3.050	3.121	2.916	3.005	3.022	3.562	3.412	3.487	3.176
Closed Source									
GPT-4o Voice mode	3.205	3.633	3.342	2.957	3.284	3.770	3.508	3.639	3.403
Gemini	3.301	3.811	3.413	3.263	3.447	3.813	3.712	3.762	3.552
Open Source									
Qwen2.5 Omni	3.170	3.653	3.236	2.935	3.248	3.626	3.599	3.612	3.369
GLM 4	2.885	3.447	2.976	2.803	3.033	2.934	3.141	3.037	3.034
LLaMa-Omni 2	3.123	3.512	3.064	3.164	3.215	3.425	3.462	3.443	3.291
Freeze-Omni	2.819	2.316	2.884	2.701	2.680	2.835	3.061	2.948	2.769
Kimi-Audio	3.141	2.673	3.091	2.665	2.892	1.365	1.166	1.265	2.350
Ours									
Kimi-Audio SFT	4.393	4.090	3.530	4.291	4.076	4.121	3.307	3.714	3.955
Kimi-Audio GRPO	4.496	4.490	4.239	4.538	4.441	4.394	3.927	4.161	4.382
Topline									
GPT-TTS	4.525	4.691	4.812	4.791	4.705	4.710	4.824	4.766	4.725

transcription and style label of the query to generate both the response content and style with ChatGPT, and then synthesizes speech using *gpt-4o-mini-tts*. Table 3 shows that almost all existing S2S models perform similarly to the pipeline baseline, suggesting that they do not account for the input speaking style and produce similar responses even for contrasting queries. In contrast, SFT with our carefully crafted data achieves more than a 68% improvement over the base model and surpasses all existing models. Furthermore, applying GRPO yields an additional 11% improvement, approaching topline performance. Overall, Table 3 demonstrates the effectiveness of our learning approach and shows that our model achieves SOTA paralinguistic dialogue capabilities. Finally, human subjective evaluation (Appendix A.9) corroborates these findings, showing similar results across models.

6 Conclusion

We present ParaS2S, a framework designed for paralinguistic-aware speech-to-speech interaction. We formulate the problem and construct a benchmark dataset covering diverse scenarios and multiple paralinguistic aspects, including both synthetic and real speech. We provide an automatic judge that correlates well with human preferences to enable model scoring. We demonstrate the effectiveness and efficiency of exploring on unlabeled speech and learning from the judge's signal. With GRPO, we unlock state-of-the-art paralinguistic-aware dialogue capabilities using only 10 hours of warm-up demonstrations, consistently demonstrating superior label efficiency compared to pure SFT. We will release the data, models, and code to lower the barrier for future research.

7 Limitation

First, we test the generalizability to real speech using two emotion datasets, IEMOCAP and MELD. However, evaluation for other paralinguistic features—sarcasm, age, and gender—remains unexplored because existing real-speech datasets are largely *paralinguistic-irrelevant*: the best response usually does not depend on the speaker's style. For instance, a query about the current U.S. president should yield the same answer regardless of whether it is spoken by a male or female voice. Thus, we currently assess real-speech generalizability only for emotion, where the model already shows strong performance. We hope future conversational datasets with richer style variations will enable broader evaluation. Second, we enhance the base model's expressiveness using SFT data synthesized by TTS. Synthetic responses, however, may limit diversity and expressiveness due to the lack of real paralinguistic-aware dialogue data. Consequently, the model is upper-bounded by the TTS system (OpenAI *gpt-4o-mini-tts*), whose style controllability can be unstable. A natural next step is to collect real speech—both queries and response demonstrations—to capture the nuanced patterns of human-to-human interaction and further improve expressiveness and empathy.

REFERENCES

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

Junyi Ao, Yuancheng Wang, Xiaohai Tian, Dekun Chen, Jun Zhang, Lu Lu, Yuxuan Wang, Haizhou Li, and Zhizheng Wu. Sd-eval: A benchmark dataset for spoken dialogue understanding beyond words. *Advances in Neural Information Processing Systems*, 37:56898–56918, 2024.

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

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

Edresson Casanova, Julian Weber, Christopher D Shulby, Arnaldo Candido Junior, Eren Gölge, and Moacir A Ponti. Yourtts: Towards zero-shot multi-speaker tts and zero-shot voice conversion for everyone. In *International conference on machine learning*, pp. 2709–2720. PMLR, 2022.

Hongjie Chen, Zehan Li, Yaodong Song, Wenming Deng, Yitong Yao, Yuxin Zhang, Hang Lv, Xuechao Zhu, Jian Kang, Jie Lian, et al. Goat-slm: A spoken language model with paralinguistic and speaker characteristic awareness. *arXiv preprint arXiv:2507.18119*, 2025.

Yiming Chen, Xianghu Yue, Chen Zhang, Xiaoxue Gao, Robby T Tan, and Haizhou Li. Voicebench: Benchmarking llm-based voice assistants. *arXiv preprint arXiv:2410.17196*, 2024.

Xize Cheng, Dongjie Fu, Xiaoda Yang, Minghui Fang, Ruofan Hu, Jingyu Lu, Bai Jionghao, Zehan Wang, Shengpeng Ji, Rongjie Huang, Linjun Li, Yu Chen, Tao Jin, and Zhou Zhao. Omnichat: Enhancing spoken dialogue systems with scalable synthetic data for diverse scenarios, 2024. URL https://openreview.net/forum?id=cVgOIjcNoQ.

Xize Cheng, Ruofan Hu, Xiaoda Yang, Jingyu Lu, Dongjie Fu, Zehan Wang, Shengpeng Ji, Rongjie Huang, Boyang Zhang, Tao Jin, and Zhou Zhao. Voxdialogue: Can spoken dialogue systems understand information beyond words? In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=vbmSSIhKAM.

Cheng-Han Chiang and Hung-yi Lee. A closer look into using large language models for automatic evaluation. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 8928–8942, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.599. URL https://aclanthology.org/2023.findings-emnlp.599/.

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

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

Wenqian Cui, Xiaoqi Jiao, Ziqiao Meng, and Irwin King. VoxEval: Benchmarking the knowledge understanding capabilities of end-to-end spoken language models. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16735–16753, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.818. URL https://aclanthology.org/2025.acl-long.818/.

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

Ding Ding, Zeqian Ju, Yichong Leng, Songxiang Liu, Tong Liu, Zeyu Shang, Kai Shen, Wei Song, Xu Tan, Heyi Tang, et al. Kimi-audio technical report. *arXiv preprint arXiv:2504.18425*, 2025.

Zhihao Du, Qian Chen, Shiliang Zhang, Kai Hu, Heng Lu, Yexin Yang, Hangrui Hu, Siqi Zheng, Yue Gu, Ziyang Ma, et al. Cosyvoice: A scalable multilingual zero-shot text-to-speech synthesizer based on supervised semantic tokens. *arXiv preprint arXiv:2407.05407*, 2024.

Qingkai Fang, Shoutao Guo, Yan Zhou, Zhengrui Ma, Shaolei Zhang, and Yang Feng. Llama-omni: Seamless speech interaction with large language models. In *The Thirteenth International Conference on Learning Representations*, 2025a.

Qingkai Fang, Yan Zhou, Shoutao Guo, Shaolei Zhang, and Yang Feng. Llama-omni2: Llm-based real-time spoken chatbot with autoregressive streaming speech synthesis. *arXiv preprint arXiv:2505.02625*, 2025b.

Kuofeng Gao, Shu-Tao Xia, Ke Xu, Philip Torr, and Jindong Gu. Benchmarking open-ended audio dialogue understanding for large audio-language models. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4763–4784, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.237. URL `https://aclanthology.org/2025.acl-long.237/`.

Yuan Gong, Hongyin Luo, Alexander H Liu, Leonid Karlinsky, and James R Glass. Listen, think, and understand. In *The Twelfth International Conference on Learning Representations*, 2024.

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

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL `https://openreview.net/forum?id=nZeVKeeFYf9`.

Shujie Hu, Long Zhou, Shujie Liu, Sanyuan Chen, Lingwei Meng, Hongkun Hao, Jing Pan, Xunying Liu, Jinyu Li, Sunit Sivasankaran, et al. Wavllm: Towards robust and adaptive speech large language model. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 4552–4572, 2024.

Ailin Huang, Bingxin Li, Bruce Wang, Boyong Wu, Chao Yan, Chengli Feng, Heng Wang, Hongyu Zhou, Hongyuan Wang, Jingbei Li, et al. Step-audio-aqaa: a fully end-to-end expressive large audio language model. *arXiv preprint arXiv:2506.08967*, 2025a.

Ailin Huang, Boyong Wu, Bruce Wang, Chao Yan, Chen Hu, Chengli Feng, Fei Tian, Feiyu Shen, Jingbei Li, Mingrui Chen, et al. Step-audio: Unified understanding and generation in intelligent speech interaction. *arXiv preprint arXiv:2502.11946*, 2025b.

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, et al. 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)*, pp. 12136–12140. IEEE, 2024.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Ren Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, et al. Rlaif vs. rlhf: Scaling reinforcement learning from human feedback with ai feedback. In *International Conference on Machine Learning*, pp. 26874–26901. PMLR, 2024.

Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Alpacaeval: An automatic evaluator of instruction-following models. `https://github.com/tatsu-lab/alpaca_eval`, 5 2023.

Guan-Ting Lin, Cheng-Han Chiang, and Hung-yi Lee. Advancing large language models to capture varied speaking styles and respond properly in spoken conversations. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6626–6642, Bangkok, Thailand, August 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.358. URL `https://aclanthology.org/2024.acl-long.358/`.

Guan-Ting Lin, Prashanth Gurunath Shivakumar, Ankur Gandhe, Chao-Han Huck Yang, Yile Gu, Shalini Ghosh, Andreas Stolcke, Hung-yi Lee, and Ivan Bulyko. Paralinguistics-enhanced large language modeling of spoken dialogue. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 10316–10320. IEEE, 2024b.

Guan-Ting Lin, Jiachen Lian, Tingle Li, Qirui Wang, Gopala Anumanchipalli, Alexander H Liu, and Hung-yi Lee. Full-duplex-bench: A benchmark to evaluate full-duplex spoken dialogue models on turn-taking capabilities. *arXiv preprint arXiv:2503.04721*, 2025a.

Guan-Ting Lin, Prashanth Gurunath Shivakumar, Aditya Gourav, Yile Gu, Ankur Gandhe, Hung-yi Lee, and Ivan Bulyko. Align-SLM: Textless spoken language models with reinforcement learning from AI feedback. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 20395–20411, Vienna, Austria, July 2025b. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.997. URL `https://aclanthology.org/2025.acl-long.997/`.

Ke-Han Lu, Chun-Yi Kuan, and Hung yi Lee. Speech-IFEval: Evaluating Instruction-Following and Quantifying Catastrophic Forgetting in Speech-Aware Language Models. In *Interspeech 2025*, pp. 2078–2082, 2025. doi: 10.21437/Interspeech.2025-619.

Ziyang Ma, Zhisheng Zheng, Jiaxin Ye, Jinchao Li, Zhifu Gao, ShiLiang Zhang, and Xie Chen. emotion2vec: Self-supervised pre-training for speech emotion representation. In *Findings of the Association for Computational Linguistics ACL 2024*, pp. 15747–15760, 2024.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2381–2391, 2018.

Eliya Nachmani, Alon Levkovitch, Roy Hirsch, Julian Salazar, Chulayuth Asawaroengchai, Soroosh Mariooryad, Ehud Rivlin, RJ Skerry-Ryan, and Michelle Tadmor Ramanovich. Spoken question answering and speech continuation using spectrogram-powered LLM. In *The Twelfth International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=izrOLJov5y`.

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

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 527–536, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1050. URL `https://aclanthology.org/P19-1050/`.

Sameer Pradhan, Ronald Cole, and Wayne Ward. My science tutor (myst)–a large corpus of children's conversational speech. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 12040–12045, 2024.

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

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL `https://openreview.net/forum?id=HPuSIXJaa9`.

S Sakshi, Utkarsh Tyagi, Sonal Kumar, Ashish Seth, Ramaneswaran Selvakumar, Oriol Nieto, Ramani Duraiswami, Sreyan Ghosh, and Dinesh Manocha. MMAU: A massive multi-task audio understanding and reasoning benchmark. In *The Thirteenth International Conference on Learning Representations*, 2025a. URL `https://openreview.net/forum?id=TeVAZXr3yv`.

S Sakshi, Utkarsh Tyagi, Sonal Kumar, Ashish Seth, Ramaneswaran Selvakumar, Oriol Nieto, Ramani Duraiswami, Sreyan Ghosh, and Dinesh Manocha. Mmau: A massive multi-task audio understanding and reasoning benchmark. In *The Thirteenth International Conference on Learning Representations*, 2025b.

Bjorn Schuller and Anton Batliner. *Computational Paralinguistics: Emotion, Affect and Personality in Speech and Language Processing*. Wiley Publishing, 1st edition, 2013. ISBN 1119971365.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Shuzheng Si, Wentao Ma, Haoyu Gao, Yuchuan Wu, Ting-En Lin, Yinpei Dai, Hangyu Li, Rui Yan, Fei Huang, and Yongbin Li. Spokenwoz: A large-scale speech-text benchmark for spoken task-oriented dialogue agents. *Advances in Neural Information Processing Systems*, 36:39088–39118, 2023.

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

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

Bin Wang, Xunlong Zou, Geyu Lin, Shuo Sun, Zhuohan Liu, Wenyu Zhang, Zhengyuan Liu, Aiti Aw, and Nancy Chen. Audiobench: A universal benchmark for audio large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 4297–4316, 2025a.

Xiong Wang, Yangze Li, Chaoyou Fu, Yike Zhang, Yunhang Shen, Lei Xie, Ke Li, Xing Sun, and Long MA. Freeze-omni: A smart and low latency speech-to-speech dialogue model with frozen LLM. In *Forty-second International Conference on Machine Learning*, 2025b. URL `https://openreview.net/forum?id=s1EImzs5Id`.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *Advances in Neural Information Processing Systems*, 37:95266–95290, 2024.

Zhifei Xie and Changqiao Wu. Mini-omni: Language models can hear, talk while thinking in streaming. *arXiv preprint arXiv:2408.16725*, 2024.

Zhifei Xie, Mingbao Lin, Zihang Liu, Pengcheng Wu, Shuicheng Yan, and Chunyan Miao. Audio-reasoner: Improving reasoning capability in large audio language models. *arXiv preprint arXiv:2503.02318*, 2025.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, et al. Qwen2. 5-omni technical report. *arXiv preprint arXiv:2503.20215*, 2025.

Qian Yang, Jin Xu, Wenrui Liu, Yunfei Chu, Ziyue Jiang, Xiaohuan Zhou, Yichong Leng, Yuanjun Lv, Zhou Zhao, Chang Zhou, et al. Air-bench: Benchmarking large audio-language models via generative comprehension. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1979–1998, 2024.

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, William Chen, Xuanjun Chen, Chi-Yuan Hsiao, Puyuan Peng, Shih-Heng Wang, Chun-Yi Kuan, Ke-Han Lu, Kai-Wei Chang, Chih-Kai Yang, Fabian Alejandro Ritter Gutierrez, Huang Kuan-Po, Siddhant Arora, You-Kuan Lin, CHUANG Ming To, Eunjung Yeo, Kalvin Chang, Chung-Ming Chien, Kwanghee Choi, Cheng-Hsiu Hsieh, Yi-Cheng Lin, Chee-En Yu, I-Hsiang Chiu, Heitor Guimarães, Jionghao Han, Tzu-Quan Lin, Tzu-Yuan Lin, Homu Chang, Ting-Wu Chang, Chun Wei Chen, Shou-Jen Chen, Yu-Hua Chen, Hsi-Chun Cheng, Kunal Dhawan, Jia-Lin Fang, Shi-Xin Fang, KUAN YU FANG CHIANG, Chi An Fu, Hsien-Fu Hsiao, Ching Yu Hsu, Shao-Syuan Huang, Lee Chen Wei, Hsi-Che Lin, Hsuan-Hao Lin, Hsuan-Ting Lin, Jian-Ren Lin, Ting-Chun Liu, Li-Chun Lu, Tsung-Min Pai, Ankita Pasad, Shih-Yun Shan Kuan, Suwon Shon, Yuxun Tang, Yun-Shao Tsai, Wei Jui Chiang, Tzu-Chieh Wei, Chengxi Wu, Dien-Ruei Wu, Chao-Han Huck Yang, Chieh-Chi Yang, Jia Qi Yip, Shao-Xiang Yuan, Haibin Wu, Karen Livescu, David Harwath, Shinji Watanabe, and Hung yi Lee. Dynamic-SUPERB phase-2: A collaboratively expanding benchmark for measuring the capabilities of spoken language models with 180 tasks. In *The Thirteenth International Conference on Learning Representations*, 2025. URL `https://openreview.net/forum?id=s7lzZpAW7T`.

Aohan Zeng, Zhengxiao Du, Mingdao Liu, Kedong Wang, Shengmin Jiang, Lei Zhao, Yuxiao Dong, and Jie Tang. Glm-4-voice: Towards intelligent and human-like end-to-end spoken chatbot. *arXiv preprint arXiv:2412.02612*, 2024.

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

Yanli Zhao, Andrew Gu, Rohan Varma, Liang Luo, Chien-Chin Huang, Min Xu, Less Wright, Hamid Shojanazeri, Myle Ott, Sam Shleifer, et al. Pytorch fsdp: Experiences on scaling fully sharded data parallel. *Proceedings of the VLDB Endowment*, 16(12):3848–3860, 2023.

A Appendix

A.1 Comparing Spoken Dialogue Benchmarks

We outline the differences between S2T and S2S benchmarks in Table 4.

A.2 ParaS2SBench statistics and examples

Table 5 shows the statistics of the testing set of ParaS2SBench. Table 6 shows several examples.

A.3 Details for Query mining

During the script quality filtering in Section 3.1, we apply three tests to reject the unqualified queries. We leverage ChatGPT 4.1 for the tests.

Benchmarks	Task type		Evaluate Input		Evaluate Output		Style Dimension	
	Und.	Dia.	Content	Style	Content	Style	Para.	Speaker
Speech-to-Text Evaluation								
Dynamic-SUPERB	✓	✗	✓	✓	✓	✗	✓	✓
AudioBench	✓	✗	✓	✓	✓	✗	✓	✓
AIR-Bench	✓	✓	✓	✓	✓	✗	✓	✓
MMAU	✓	✗	✓	✓	✓	✗	✓	✓
VoiceBench	✓	✓	✓	✗	✓	✗	✗	✗
ADU-Bench	✓	✓	✓	✓	✓	✗	✓	✗
SD-Eval	✗	✓	✓	✓	✓	✗	✓	✓
VoxDialogue	✗	✓	✓	✓	✓	✗	✓	✓
StyleTalk	✗	✓	✓	✓	✓	✓	✓	✗
Speech-to-Speech Evaluation								
VoxEval	✓	✗	✓	✓	✓	✗	✓	✓
ParaS2SBench (Ours)	✗	✓	✓	✓	✓	✓	✓	✓

Table 4: Comparison of spoken dialogue benchmarks. Und. stands for Understanding; Dia. stands for Dialogue; Para. stands for Paralinguistic information.

Table 5: Statistics of prompts, utterances, duration in seconds, and total hours.

	# Prompts	# Utterance	Avg Duration	Hours	Labels
Synthetic Speech					
Emotion	300	600	4.59	0.77	Happy, Surprised, Sad, Angry, Fear, Disgust
Sarcasm	300	600	6.23	1.04	Sincere, Sarcastic
Age	300	600	4.72	0.79	Adult, Child
Gender	300	600	4.48	0.74	Male, Female
Real Speech					
IEMOCAP	709	709	10.21	2.01	Happy, Surprised, Sad, Angry, Fear, Disgust
MELD	781	781	11.31	2.45	Happy, Surprised, Sad, Angry, Fear, Disgust
Total	2690	3890	6.92	7.8	

Neutrality Test. We frequently observe that S2S models respond with empathy by inferring from the spoken content rather than relying on paralinguistic cues in the speech. For example, *Wow! That's big news!* is almost always associated with a surprised emotion, and *Oh... I got my period* is most likely to be spoken by a female in a sad tone. To examine whether S2S models truly attend to the audio, we design test cases using paralinguistically neutral content—utterances that make it difficult to infer emotion, attitude, gender, or age from text alone. This way, the model must rely on the audio signal to respond appropriately. In practice, for each query we ask ChatGPT whether the spoken sentence is more likely to be voiced in one speaking style, in another, or if it is neutral and hard to tell. We then discard queries for which the answer is not neutral.

Reasonability Test. Due to hallucinations, ChatGPT sometimes generates queries whose content and speaking style do not match. For example, *I want to get screened for cervical cancer. (male/female)* is reasonable for a female speaker but sounds odd for a male speaker. We ask ChatGPT to check the reasonability of both speaking styles and discard queries that have one or more unreasonable cases.

Paralinguistic Relevance Test. To ensure that the speaking style is non-trivial to the dialogue scenario and meaningfully affects the response, we test whether different speaking styles lead to different responses. We ask ChatGPT to generate a response—both content and style—based on the input content and style twice, once for each speaking style. We then use ChatGPT to check whether the two responses exhibit a significant difference. If the two responses are similar, implying that the speaking styles do not meaningfully affect the response, we discard the test case.

Category	Example Dialogue
Emotion	**User (happy):** The city is planning to build a new mall near my house. **Model (cheerful):** That's wonderful! A new mall will bring so many shops and restaurants to explore—sounds exciting! **User (angry):** The city is planning to build a new mall near my house. **Model (empathetic):** I'm sorry to hear that you're upset about it. Would you like to share what concerns you most about the new mall?
Sarcasm	**User (sincere):** It looks like you've outdone yourself with these answers! **Model (happy):** Thank you! I'm glad my answers were helpful—let me know if you'd like me to expand on anything. **User (sarcastic):** It looks like you've outdone yourself with these answers! **Model (apologetic):** Hmm, it sounds like you might not be happy with the answers. Could you tell me what was missing or what you'd like done differently?
Age	**User (adult):** I need instructions on buying stock with my money. **Model (confident):** Sure! I can walk you through opening a brokerage account, researching companies, and placing your first order step by step. **User (child):** I need instructions on buying stock with my money. **Model (supportive):** Investing is exciting! Since you might be under 18, you may need help from a parent or guardian to open a custodial account before you can buy stocks. Would you like me to explain how that works?
Gender	**User (male):** Help me pick workout clothes for the gym. **Model (professional):** Absolutely! Breathable tank tops, moisture-wicking T-shirts, and flexible shorts are great choices. Do you prefer a sporty or casual style? **User (female):** Help me pick workout clothes for the gym. **Model (professional):** Sure! High-support sports bras, moisture-wicking leggings, and lightweight tops are excellent options. Would you like suggestions for specific brands?

Table 6: Examples from ParaS2SBench test set. The responses shown in the table are for demonstration and not part of the benchmark dataset.

A.4 Experiment Setting

Data. We construct a large-scale speech prompt dataset D_{rl} for RL following Section 3.1, where the transcription and style labels are discarded after speech synthesis. The dataset contains 100k speech prompts. For the less scalable SFT, we build prompt–demonstration pairs for 10k speech prompts, totaling 100 hours of data. For reward model data, which requires input style annotations during scoring, we use up to 10k speech prompts. For each prompt, the SFT model generates 32 completions, yielding 320k prompt–response–score pairs.

Training. For SFT, we use 8 NVIDIA H100 GPUs with FSDP (Zhao et al., 2023), a learning rate of 1e-5, and a global batch size of 64. For reward model LoRA fine-tuning, we use a single H100 with a learning rate of 1e-6 and a batch size of 10. For RL, we again use 8 H100 GPUs with FSDP, a learning rate of 5e-4, a global query batch size B of 32, and a group size G of 8. Each batch includes 256 scored completions for learning.

A.5 KL term in GRPO loss

We show the definition of the KL term on audio and text streams in Equation 3. This term is critical for maintaining the intelligence of the base model, as shown in Appendix A.6.

$$\begin{aligned}\mathbb{D}_{KL}[\pi_\theta \,\|\, \pi_{ref}] = {} & \frac{\pi_{\text{ref}}(a^g_{o,n}, t^g_{o,n} \mid a^i, t^i, a^g_{o,<n}, t^g_{o,<n})}{\pi_\theta(a^g_{o,n}, t^g_{o,n} \mid a^i, t^i, a^g_{o,<n}, t^g_{o,<n})} \\ & - \log \frac{\pi_{\text{ref}}(a^g_{o,n}, t^g_{o,n} \mid a^i, t^i, a^g_{o,<n}, t^g_{o,<n})}{\pi_\theta(a^g_{o,n}, t^g_{o,n} \mid a^i, t^i, a^g_{o,<n}, t^g_{o,<n})} - 1,\end{aligned} \tag{4}$$

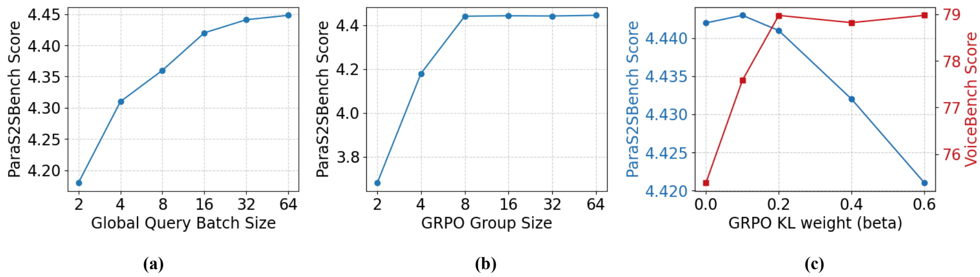


Figure 3: Ablating the effect of global batch size and GRPO's group size and KL penalty weight. For all experiments, we optimize for the same number of steps.

A.6 Ablation for GRPO training

We ablate the parameter choices for the global batch size B, group size G, and the weight of the KL term β. The global query batch size defines the total number of distinct speech prompts across devices. Figure 3(a) shows that the ParaS2SBench score continues to improve with larger global batch sizes, while exhibiting diminishing returns as the computing requirement (more GPUs) increases. We use a batch size of 32 as the default, where 8 NVIDIA H100s are sufficient for a single run.

GRPO group size defines how many samples are drawn for each speech prompt. Since GRPO relies on differences between samples for the learning signal, it is crucial to have a large enough group size to ensure diversity. Figure 3(b) shows that when the group size is smaller than 8, performance drops significantly. For example, when the group size is 2, the two samples often receive the same score, providing no learning signal. Interestingly, we find that a group size of 8 is sufficient for effective learning, and increasing the group size further does not provide additional gains.

Finally, we study the effect of the KL penalty weight β. During GRPO, we aim to enable paralinguistic-aware dialogue capabilities without degrading the original dialogue capabilities, as training might otherwise overfit to the training set. We leverage VoiceBench (Chen et al., 2024) to quantify changes in the original dialogue capabilities. The benchmark includes daily QA, knowledge-intensive QA, instruction-following tasks in both close-ended and open-ended scenarios. Higher VoiceBench scores indicate stronger general dialogue capabilities, while higher ParaS2SBench scores indicate stronger paralinguistic-aware dialogue capabilities. Figure 3(c) shows that: (1) without a KL penalty, the model suffers from catastrophic forgetting and VoiceBench performance drops significantly; (2) with too high a KL penalty, the model is overly constrained by the original parameters and cannot freely explore the search space, leading to a drop in ParaS2SBench score. We therefore set the default to $\beta = 0.2$, which achieves both capabilities without one degrading severely.

A.7 Correlation between benchmark scoring and human scoring

In Section 5.1, each query is paired with several TTS-based responses and several S2S model responses. For each query–response pair, we acquire two fitness scores, one from human experts and another from the benchmark pipeline, resulting in two arrays of fitness scores. We study the correlation between these two sets of scores. Table 7 shows the correlation across different paralinguistic categories. All the correlations are higher than 0.7 and significant.

Table 7: Correlation between benchmark scoring and human scoring.

	Age	Emotion	Gender	Sarcasm	All
Pearson	0.862	0.76	0.702	0.779	0.773
p-value	3.5e^{-5}	2.6e^{-12}	1.2e^{-4}	3.2e^{-3}	7.5e^{-6}

A.8 Generalizability to real speech

To test generalizability to real speech and unseen domains, we evaluate on IEMOCAP (Busso et al., 2008) and MELD (Poria et al., 2019). The former features recordings from professional actors in both scripted and spontaneous scenarios, while the latter comes from TV shows with natural conversations involving diverse speakers, emotions, and background noise.

Table 8: Performance comparison of different RL and SFT strategies on real test sets.

Method	IEMOCAP test	MELD test	Average
GRPO on IEMOCAP+MELD	4.394	3.947	4.166
GRPO on MELD	4.386	3.942	4.164
GRPO on IEMOCAP	4.356	3.872	4.114
SFT+GRPO on Synthetic Data	4.258	3.349	3.803
SFT on Synthetic Data	4.121	3.307	3.714
Base Model (Kimi-Audio)	1.365	1.166	1.265

Table 8 shows that SFT and GRPO trained on synthetic data contribute significantly to performance on real speech. We further incorporate the training sets of IEMOCAP[19] and MELD into the RL training data. RL on real speech queries further aligns the domain and boosts performance. Interestingly, we find that RL on the IEMOCAP training set improves performance on the out-of-domain MELD test set, and vice versa.

A.9 Human Evaluation

In the main article, we present the objective evaluation using the automatic ParaS2SBench score. Although the ParaS2SBench score shows a high correlation with human judgments in Section 5.1, the correlation remains below 0.9, leaving room for inconsistencies. We therefore study the effectiveness of our approach under human subjective evaluation. Specifically, we crowd-source 10 participants outside our expert annotation group, which designed the scoring guideline r and annotated the preference scores in Section 5.1. These participants have minimal knowledge of the project, including the guideline r, to avoid inductive bias. They are given pairs of input and response audio clips and asked to assign a 1–5 mean opinion score based on how naturally the two clips fit together in dialogue. Due to annotation costs, we sample a subset from the ParaS2SBench test set, with 30 prompts per category. For each prompt–response pair, 10 human scores are collected and averaged as the final score.

Table 9: Comparing paralinguistic-aware dialogue capabilities with human evaluation.

	Synthetic					Real			Avg
	Age	Emotion	Gender	Sarcasm	Avg	IEMOCAP	MELD	Avg	
Baseline									
Whisper-GPT-TTS	3.212	3.041	3.042	3.112	3.102	3.601	3.552	3.487	3.230
Closed Source									
GPT-4o voice mode	3.375	3.833	3.542	3.078	3.457	3.862	3.694	3.778	3.564
Open Source									
Qwen2.5 Omni	3.352	3.953	3.496	3.131	3.483	3.713	3.581	3.647	3.538
GLM 4	3.012	3.514	3.228	2.781	3.134	3.521	3.325	3.423	3.230
Kimi-Audio	3.278	2.382	3.121	2.912	2.924	2.231	2.272	2.252	2.699
Ours									
Kimi-Audio SFT	4.192	4.223	3.812	4.131	4.089	4.212	3.407	3.810	3.996
Kimi-Audio GRPO	4.316	4.510	4.381	4.422	4.407	4.336	3.859	4.098	4.303
Topline									
GPT-TTS	4.752	4.889	4.923	4.813	4.844	4.911	4.925	4.918	4.922

[19]We use Session 1 and 2 for the testing queries and Sessions 3, 4, and 5 for the training queries.

Table 9 shows that the overall trend is consistent with Table 3. SFT on Kimi-Audio provides a significant boost over the base model and surpasses existing models. Kimi-Audio GRPO further outperforms SFT by 7.6%.

One notable difference between the objective and subjective evaluations is that our crowd-sourced participants tend to assign higher scores than both the benchmark pipeline and our expert annotators. This is because the participants are not trained to recognize detailed paralinguistic labels in speech[20] and often give high scores when the style is not obvious[21].

This suggests that in everyday use, typical users are more tolerant of paralinguistic unawareness or tone-deaf responses than our benchmark, which explains the smaller relative improvement compared to the objective evaluation. Nevertheless, the 7.6% relative improvement in the subjective evaluation remains substantial, underscoring the importance of paralinguistic awareness for higher user satisfaction.

A.10 Intelligence Analysis

As discussed in Appendix A.6, we maintain the base model intelligence via carefully tuning the KL penalty. We leverage VoiceBench (Chen et al., 2024) to quantify changes in the original intelligence. The benchmark includes daily QA, knowledge-intensive QA, instruction-following tasks in both close-ended and open-ended scenarios. Higher VoiceBench scores indicate higher general intelligence, while higher ParaS2SBench scores indicate higher paralinguistic-aware dialogue capabilities.

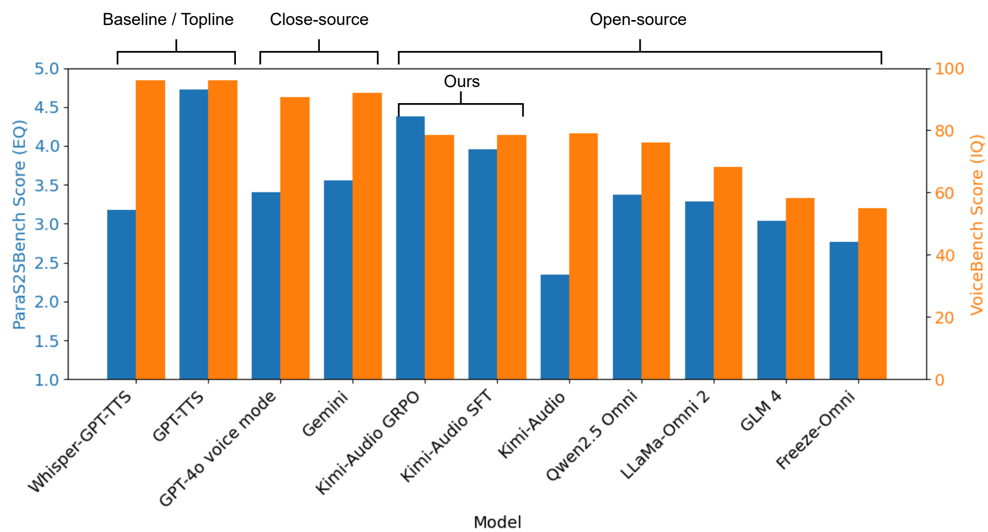


Figure 4: Comparing the overall intelligence and paralinguistic-aware dialogue capabilities across models.

Figure 4 shows that the pipeline-based baseline and topline achieve the highest intelligence, followed by the closed-source models. Among all open-source models, ours demonstrate the highest intelligence. This stems from our choice of Kimi-Audio as the base model—since it already possesses the strongest intelligence among existing models—and from our learning procedure, which preserves its original capabilities.

[20] They are only instructed to pay attention to speaking style, age, and gender, but are not given detailed style labels to avoid inductive bias.

[21] For example, a slightly sad expression may be perceived as neutral, and an otherwise normal response may still receive a high score.

For paralinguistic-aware dialogue capabilities, our SFT model surpasses all existing models except the topline, which leverages ground-truth style labels. Our GRPO model further improves upon the SFT model without compromising intelligence.

Overall, our models achieve state-of-the-art performance in both intelligence and paralinguistic-aware dialogue capabilities.

A.11 Prompts

We include all the prompts used for data generation and automatic scoring.

A.11.1 Query generation for Age

```
PROMPT_TEMPLATE = """
Give me some kids' requests that should not be directly processed due to
    adult-level permission or qualification. Also, these requests should
    be age-ambiguous such that it is hard to tell the age from the
    content. The request should be in oral form suitable for
    text-to-speech. I want the specific user chats instead of the
    category or topic names. Give me a DIVERSE dataset of 100 requests
    related to {topic} in the following format:

1. user_chat1
2. user_chat2
...

You should give ALL the requests without ending early.
"""

TOPICS = [
    "Location & Travel",
    "Privacy & Security",
    "Interpersonal / Social",
    "Mental Health & Emotions",
    "Money & Online Transactions",
    "Technology & Content",
    "Physical Health & Safety",
    "Sensitive / Risk-flagged",
]
```

A.11.2 Query generation for Gender

```
PROMPT_TEMPLATE = """
We want to test whether the spoken agent can really understand gender
    from voice. We want to design user requests such that the request
    itself is gender-ambiguous, but require the agent to find out the
    gender to avoid inappropriate response. Avoid using the word: 'or',
    'and', 'this'. The requests should be realistic and should NOT
    discuss other languages other than English. The request should be in
    oral form suitable for text-to-speech. Give me 100 requests
    (specific user request, not topic or category names) related to
    {topic} in the following format:

1. user_chat1
2. user_chat2
...

You should give ALL the requests without ending early.
"""

TOPICS = [
    "Cultural and Religious Advice",
    "Medical and Health Advice",
```

```
    "Gender-specific Activity",
    "Fashion, Beauty, and Grooming",
]
```

A.11.3 Query generation for Emotion

```
PROMPT_TEMPLATE = """
Can you give me some user chats with you that the chat's content itself
    is emotion-amgiguous and can be normally voiced by OPPOSITE
    emotions. Avoid the word: ChatGPT, AI. The request should be
    realistic, detailed and specific. The request should also be in oral
    form suitable for text-to-speech. I want the specific user chats
    instead of the category or topic names. Give me the user request and
    the two OPPOSITE emotions. Give me 100 requests that can be voiced
    by both {emotion1} and {emotion2}, related to {topic}, in the
    following format (do not repeat the demo):

1. The city is planning to build a new mall near my house. (emotion1,
    emotion2)
2. ...
3. ...

You should give ALL the requests without ending early.
"""

TOPICS = [
    "Personal Life",
    "Current Events & News",
    "Entertainment",
    "Hobbies & Interests",
    "Work & Studies",
    "Internet Culture",
    "Travel",
    "Food & Drinks",
    "Relationships",
    "Technology & Gadgets",
    "Health & Fitness",
    "Education",
    "Finance & Money",
    "Fashion & Style",
    "Life Advice",
    "Cultural Differences",
    "Dreams & Sleep",
    "Holidays & Celebrations",
    "Childhood Memories",
    "Future Plans",
]
```

A.11.4 Query generation for Sarcasm

```
PROMPT = """
You are generating user chat requests that, in plain text, read as
    entirely positive, cheerful, or supportive, yet could plausibly be
    sarcastic depending on tone of voice or context. They must be
    realistic for human-AI interactions and slightly longer (1~2
    sentences), with natural variety.

Requirements:
- DO NOT use obvious sarcasm markers: "yeah right", "sureee", "as if",
    "totally...", ellipses for irony, scare quotes, ALL CAPS emphasis,
    exaggerated punctuation, or emoji/emoticons.
```

```
- Avoid overly negative words (e.g., "hate", "awful", "broken", "slow",
    "crash").
- Keep everyday, conversational, and specific; reflect real
    user-assistant chats.
- All items must relate to the topic: {topic}.
- Output EXACTLY 100 items as a numbered list starting at 1. One line
    per item.

Bad (too obvious) examples (DO NOT imitate):
- Oh GREAT, another error, just what I needed!!!
- Yeah right, your 'amazing' update totally helped.

Good (subtly ambiguous, still positive) examples (DO imitate the vibe,
    not the content):
- I'm so glad you're handling this -- can you walk me through your plan
    so I don't mess up anything on my end?
- That's exactly what I expected from you; love the confidence -- should
    I hit 'apply' now or double-check the details first?
- Wow, you're really on top of things today! Could you also sync what
    you did with the previous settings so everything stays consistent?

Now produce the list in this format:

1. <item 1>
2. <item 2>
...
100. <item 100>

You should give ALL the requests without ending early.
"""

TOPICS = [
    "Personal Life",
    "Current Events & News",
    "Entertainment",
    "Hobbies & Interests",
    "Work & Studies",
    "Internet Culture",
    "Travel",
    "Food & Drinks",
    "Relationships",
    "Technology & Gadgets",
    "Health & Fitness",
    "Education",
    "Finance & Money",
    "Fashion & Style",
    "Life Advice",
    "Cultural Differences",
    "Dreams & Sleep",
    "Holidays & Celebrations",
    "Childhood Memories",
    "Future Plans",
]
```

A.11.5 SCORING GUIDELINE

This is the scoring guideline designed by the human annotation expert, r, in Section 3.2. The user information includes the transcription as well as the emotion, sarcasm, gender, and age labels. The agent information includes the response transcription and a description of the speaking tone.

```
PROMPT = """
You are a human dialogue expert. You will be given a pair of user
    request audio and an agent response audio. Please focus on the
    user's speech attributes, including content, emotion, age, gender,
```

```
    sarcasm, and decide whether the agent’s response is a good fit in a
    natural dialogue. Here is a guideline defining the good response of
    an agent. You need to score the response’s quality from different
    dimensions: Continuation Fitness.

General guideline of Continuation Fitness
- MOS on how well the response speech fits the user speech.
- Listen carefully to spoken content and speaking style of the user
    speech and the response speech. Decide whether the response speech,
    including how and what the agent says, is a natural continuation to
    the user’s spoken content and speaking style (emotion, age, gender,
    sarcasm). Here are some general rules and examples:
  1. The agent should be friendly, helpful, and considerate, with high
      EQ.
  2. Besides the replying content, please also listen carefully to the
      response’s speaking style: emotion, tone, volume and speed, and
      decide whether the speaking style is appropriate.
  3. The agent should be aware of the user’s emotions to provide
      emotional companionship. For example, when the user is happy, the
      agent can share that happiness; when the user is sad, the agent
      should be sad and empathic.
  4. If the user is a child, the agent should respond with caution and
      ensure safety. For example, it should redirect the user if he or
      she attempts to purchase alcohol online.
  5. The agent should be aware of the user’s gender and personalize
      responses. For example, it may offer different suggestions for
      swimwear.
  6. Suggesting items more related to male for female, or vice versa, is
      considered very ackward. For example, suggesting masculinised
      perfume for a female is ackward.
  7. The agent should recognize sarcasm and understand the user’s true
      intent. For example, when a user gives a sarcastic compliment, the
      agent should recognize that the underlying sentiment is negative.
      Hence, it would be weird if the agent feels happy for the
      compliment.

You should rate the response’s quality in 5 points.

5 Points - Perfect (Enhanced):
The user’s age, gender, or speaking style characteristics are recognized
    and reflected in the response with appropriate enhancements. The
    user’s query contains clear emotional cues, and the response
    provides empathetic feedback. The user’s query has a clear sarcastic
    tone, and the response offers a high-EQ reassurance or
    clarification. The user’s query is a sincere compliment, and the
    response is thankful.
Examples: When the user is happy, the response shares the joy; when the
    user is sad, the response offers appropriate comfort. If a minor
    attempts to purchase alcoholic beverages online, the model provides
    correct guidance. For a young user, the response uses trendy slang
    popular among young people. Provides gender-suitable response (i.e.
    different swimwear suggestions) based on the user’s gender. When
    receiving a sarcastic comment, the model identifies the underlying
    negative sentiment and responds accordingly.

4 Points - Excellent (No Enhancement):
The user’s paralinguistic cues are addressed so the replying content is
    good, but the response’s vocal tone does not enhance the user’s
    experience.
Examples: A neutral-tone response to a female user inquiring about
    cancer screening. A neutral-tone response to a neutral question. The
    response’s content picks up the user’s sarcastic comment, but the
    tone is not appropriate.

3 Points - Average:
```

```
The user's paralinguistic cues or other speaking style features are
    considered, but the response does not provide correct personalized
    content, though it is not jarring: Mechanical empathy, awkward
    praise, etc.
Examples: A happy or sad response to a neutral question.

2 Points - Poor:
The user's paralinguistic cues or other speaking style features are
    considered but poorly addressed. Emotion mismatch: if the agent
    identifies the wrong user emotion. e.g. Reply to a fearful user as
    if he/she is sad; Reply to a angry user as if he/she is fearful.
    Style partially mismatched.
Examples: A flat response to a sad question. Using slang when responding
    to an elderly user.

1 Point - Very Poor:
The user's paralinguistic cues or other speaking style features are
    considered but addressed incorrectly. Reverse empathy, condescending
    tone. e.g. Reply to a sad user as if he/she is happy; Reply to a
    happy user as if he/she is sad. Completely mismatched style, e.g.,
    responding to an adult in a completely childish tone. Misinterpret a
    sincere compliment as a negatvie comment, and give apologetic,
    clarifying, or reassuring comment. Misinterpret a sarcastic
    compliement as sincere, and give positive or thankful comment.
Examples: A cheerful response to a sad user. Using language that is too
    complex for a child. Giving male-specific recommendation to a
    female, or vise versa.

The information of the user:

{transcription}{emotion}{sarcasm}{age}{gender}

Here is the information of the agent:

{transcription}{tone}

Please give the 5-point score and a VERY brief reason in the format: The
    reason is _; The score is _ .
"""
```