S2S-Bench, Evaluating Instruction Following with Paralinguistic Information in Speech2Speech Scenarios

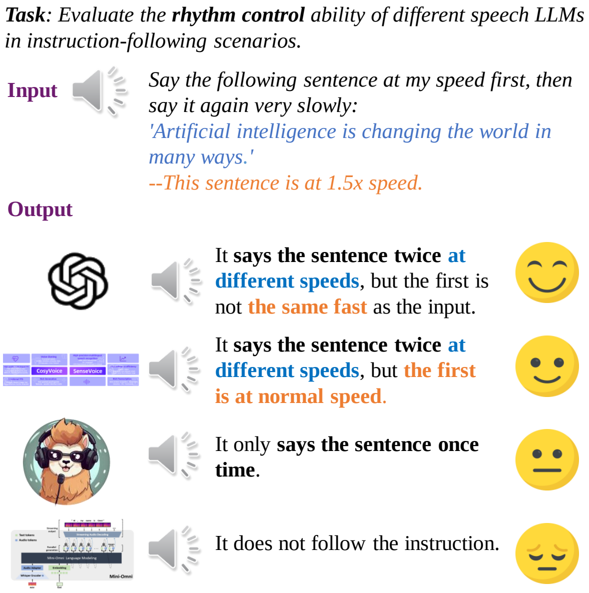
Anonymous ACL submission

Abstract

The rapid development of large language models (LLMs) has brought significant attention to speech large models (SLMs), particularly recent speech-to-speech (S2S) models supporting both speech input and output. However, current benchmarks for evaluating instruction-following abilities in these models exhibit notable gaps, primarily due to insufficient consideration of paralinguistic information in both input and output. Existing studies either focus on paralinguistic elements in input while neglecting output or prioritize semantic alignment at the expense of paralinguistic factors. Additionally, direct comparisons of speech output across models are scarce, limiting comprehensive assessment. To address these issues, we introduce a novel arena-style S2S benchmark that evaluates instruction-following capabilities across real-world tasks, covering four domains and 21 tasks with open-ended samples generated via a three-stage method. Preliminary experiments using an ELO rating system reveal performance comparisons among leading S2S models, indicating GPT-4o's overall strength in knowledge-intensive tasks yet significant challenges in expressive speech generation. Our findings provide key insights for S2S model development and offer a robust framework for evaluating model performance across semantic and paralinguistic dimensions.[1]

1 Introduction

Voice-based human-computer interaction is one of the most natural ways (Card et al., 1983; Allen et al., 2001). In such speech-to-speech (S2S) scenarios, the machine is expected to have higher abilities to interact with humans, not only understanding the voice commands issued by humans (Chu et al., 2023, 2024; Tang et al., 2023; Ghosh et al., 2024; Hu et al., 2024) but also generating corresponding voice replies and executing corresponding tasks (Wang et al., 2024b; Chen et al., 2024c; Liao et al., 2024).



***Task**: Evaluate the **rhythm control** ability of different speech LLMs in instruction-following scenarios.*

Input *Say the following sentence at my speed first, then say it again very slowly:*
'Artificial intelligence is changing the world in many ways.'
--This sentence is at 1.5x speed.

Output

It **says the sentence twice at different speeds**, but the first is not **the same fast** as the input.

It **says the sentence twice at different speeds**, but **the first is at normal speed**.

It only **says the sentence once time**.

It does not follow the instruction.

Figure 1: An Example of Evaluating Instruction Following with Rhythm Controlling in Speech-in and Speech-out for Speech2Speech Model.

Thanks to the excellent semantic understanding ability of large-scale language models (LLMs) (Dubey et al., 2024), the current work of LLM-based Speech2Speech models (Zhang et al., 2023; SpeechTeam, 2024a; Fang et al., 2024; Xie and Wu, 2024) focuses more on the paralinguistic information involved in speech, as shown in Figure 1. As an important part of communication, paralinguistic information (Trager, 1958) primarily includes biological characteristics (Schuller et al., 2010), emotional features (Batliner et al., 2011), speaking style (Nose et al., 2007), and social roles (Ipgrave, 2009). These features can be inferred by analyzing the acoustic characteristics of speech, such as pitch, tone, speech rate, and voice quality (Schuller et al., 2013).

[1] The source code and dataset for this project is available at GitHub and Hugging Face.

Model Name	Release Date	Input	Backbone	Output	Input/Output Form
SpeechGPT	2023-05	Speech Tokens	LLaMA 7B	Speech Tokens	Text or Speech / Text or Speech
AnyGPT	2024-02	Text w/ Special Tokens	LLaMA-2 7B	Text w/ Special Tokens	Text or Speech / Text or Speech
GPT-4o	2024-05	Unknown	Unknown	Unknown	Text or Speech / Text or Speech
FunAudioLLM	2024-07	Text w/ Special Token	Qwen-2 72B	Text w/ Special Tokens	Text or Speech / Text and Speech
LSLM	2024-08	Speech Tokens	Transformers	Speech Tokens	Speech / Speech
Mini-Omni	2024-09	Embeddings	Qwen-2 0.5B	Speech Tokens	Text or Speech / Text and Speech
LLaMA-Omni	2024-09	Embeddings	LLaMA-3.1 8B	Speech Tokens	Speech / Text and Speech
Moshi	2024-10	Embeddings	Transformers	Embeddings	Speech / Speech
Westlake-Omni	2024-10	Speech Tokens	Qwen-2	Speech Tokens	Text and Speech / Text and Speech
GLM-4-Voice	2024-10	Speech Tokens	GLM-4 9B	Speech Tokens	Text or Speech / Text and Speech
Freeze-Omni	2024-11	Embeddings	Qwen-2 7B	Speech Tokens	Speech / Speech

Table 1: The Comparison of Speech2Speech Models.

However, existing benchmarks for speech models have lagged behind the rapid iteration of these models, ignoring the requirements for considering the paralinguistic information both in input and output when communicating with people. Some of them (Huang et al., 2024; Wang et al., 2024a; Ao et al., 2024; Bu et al., 2024), similar to FLAN (Wei et al., 2022) in NLP, attempt to evaluate S2S model capabilities using foundation tasks. Others (Chen et al., 2024b) focus on assessing S2S models' dialogue abilities, but most only consider paralinguistic information in speech input, neglecting the importance of paralinguistic elements in speech out during instruction following (Ji et al., 2024). Additionally, although some evaluations produce results in speech form, they often convert outputs to text via Automatic Speech Recognition (ASR) for assessment, which introduces bias (Chen et al., 2024a; Ye et al., 2024) and loss of information embedded in speech output (Zhang et al., 2023).

To address these issues, we propose the S2S-Bench, a benchmark designed to evaluate the instruction-following capabilities of S2S models with paralinguistic information, and an example is shown in Figure 1. It not only requires the S2S model to understand the paralinguistic information (rhythm) in the speech input but also to be able to follow semantic instructions to control the generation of speech output with paralinguistic information (rhythm) attached.

To build this benchmark, we designed a three-stage construction process: task determination, instruction design, and sample recording (see Section 3 for details). It covers four practical domains with 21 tasks, includes 154 instructions of varying difficulty levels, and features a mix of samples from TTS synthesis, human recordings, and existing audio datasets. We then implemented an arena-style pairwise comparison to conduct head-to-head evaluations of popular S2S models(see Section 4 for details). After conducting 400 evaluations across 22 individual assessments, we obtained initial comparative results for current models. Additionally, we conduct an in-depth analysis of aspects such as semantic speech incongruity, language consistency, instruction-following failures, and evaluation agreement and bias (see Section 5 for details). We hope that these preliminary results can provide some insights for the development of the S2S model in the future. Our contributions are as follows:

1. We propose the S2S-Bench, a novel benchmark designed to evaluate the instruction following capabilities with paralinguistic information for Speech2Speech models.

2. We conduct an intuitive performance comparison of existing Speech2Speech models across 21 tasks in four distinct domains via our carefully designed benchmark and arena-style evaluation methods.

3. We analyze the advantages and disadvantages of existing Speech2Speech models from multiple perspectives and discuss potential ways for improving S2S modeling in the future.

2 Related Work

2.1 Speech2Speech Models (S2S Models)

Despite the proprietary nature of some commercial speech models, which did not public the technical details, publicly available speech models can generally be categorized into three types based on input and output processing: Speech Tokens, Text with Special Tokens, and Embeddings, as shown in Table 1.

Models like SpeechGPT (Zhang et al., 2023), LSLM (Ma et al., 2024), and Westlake-Omni[2] transform audio into speech tokens for input and feed these into an LLM for inference. These models also output speech tokens for the voice decoder,

[2] https://github.com/xinchen-ai/Westlake-Omni

preserving more paralinguistic information but sacrificing some semantic modeling.

In contrast, AnyGPT (Zhan et al., 2024) and FunAudioLLM (SpeechTeam, 2024b) embed paralinguistic information into text using special tokens that are then processed by an LLM. Their outputs are text with these tokens, capturing some paralinguistic information while maintaining semantic content.

Other models like Mini-Omni (Xie and Wu, 2024), LLaMA-Omni (Fang et al., 2024), and Moshi (Défossez et al., 2024) employ an encoder to convert speech input into embeddings. These embeddings are then integrated with the adaptor and fed into the LLM to obtain the speech output, which helps retain paralinguistic and semantic information, though at the cost of increased training complexity. Notably, Mini-Omni and LLaMA-Omni produce speech tokens as output, whereas Moshi outputs the embeddings.

In addition, most models support both text and speech for inputs and outputs, with some, like LSLM and Moshi, exclusively supporting speech.

2.2 Benchmarks for Speech Models

Table 2 compares benchmarks for speech models, which can be broadly classified into three categories: (1) foundational task completion (e.g., Dynamic-SUPERB (Huang et al., 2024), SGAI (Bu et al., 2024), AudioBench (Wang et al., 2024a), MMAU (Sakshi et al., 2024) and AV-Odyssey Bench (Gong et al., 2024)), (2) dialogue capabilities (e.g., VoiceBench (Chen et al., 2024b), SD-Eval (Ao et al., 2024)), and (3) comprehensive assessments, such as AIR-Bench (Yang et al., 2024). Most of these benchmarks primarily evaluate semantic and paralinguistic understanding in speech input, often neglecting the evaluation of paralinguistic elements in speech output.

Specifically, SGAI, AudioBench, MMAU and AV-Odyssey Bench focus on speech understanding tasks, with output presented in text. VoiceBench assesses instruction-following in a speech modality like a speech-based version of Alpaca Eval. SD-Eval considers additional paralinguistic features in input, such as speaker age and emotion, though outputs are still text-based. AIR-Bench includes both speech input and output but does not evaluate paralinguistic aspects in generated speech. Unlike existing benchmarks, MMAU emphasizes advanced perception and reasoning with domain-specific knowledge, challenging models to tackle tasks akin to those faced by experts. Although it considers the paralinguistic in the audio input, it still evaluates the text-based outputs.

Previous benchmarks primarily evaluate generated speech by converting it to text via ASR, then assessing it with advanced models like GPT-4. This method introduces biases (Chen et al., 2024a; Ye et al., 2024) and information loss (Zhang et al., 2023), especially for non-semantic features like emotion and intonation. Dynamic-SUPERB (Huang et al., 2024) directly evaluates speech output for some tasks but relies on objective metrics, which often fail to reflect human preferences (Streijl et al., 2016). Thus, direct human evaluation of speech output is essential.

3 S2S-Bench

To evaluate the current Speech2Speech models on Instruction following ability by considering Paralinguistic Information for both Speech-in and Speech-out, we build the S2S-bench. Unlike previous evaluations that focus primarily on paralinguistic information within the speech-in, S2S-Bench additionally examines the model's capability to generate paralinguistic information in the speech-out for instruction following ability in chat or foundational applications. Through careful task determination, instruction design, and sample recordings, we implement an arena-style evaluation framework, as shown in Figure 2.

3.1 Task Determination

We conduct a thorough investigation to identify key tasks across four domains (Education, Social Interaction, Entertainment, and Medical Consultation) that are currently in high demand. Each domain contains several popular tasks, such as pronunciation correction and rhythm control in education, implication understanding, and sarcasm detection in social interaction, as shown in Table 3.

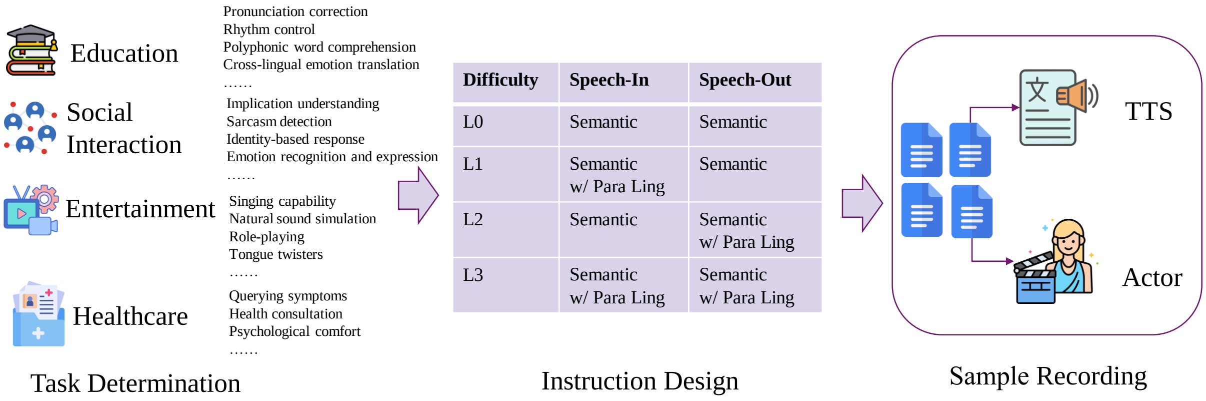
3.2 Instruction Design

For each task, we design the instruction with four difficulty levels to evaluate model performance under various settings, as shown in Figure 2.

The first level (L0) assesses only the model's ability to follow instructions without considering the paralinguistic information in speech-in and speech-out, which is similar to the VoiceBench. For example, in the "Querying symptoms" task, the model receives the instruction, "I have a headache,

Benchmarks	Types	Speech-In		Speech-Out		Evaluation	
		Semantic	ParaLing	Semantic	ParaLing	Modality	Evaluator
Dynamic-SUPERB	Foundation	✓	✓	✓		- *	Auto
SGAI	Foundation	✓	✓			Text	Auto
AudioBench	Foundation	✓	✓			Text	Auto
MMAU	Foundation	✓	✓			Text	Auto
AV-Odyssey Bench	Foundation	✓	✓			Text	Auto
VoiceBench	Chat	✓		✓		Text	Auto
SD-Eval	Chat	✓	✓			Text	Auto
AIR-Bench	Chat/Foundation	✓	✓	✓		Text	Auto
S2S-Bench (Ours)	Chat/Foundation	✓	✓	✓	✓	Speech	Human/Auto#

Table 2: Benchmark Comparison for Speech Models. The star* means that the evaluation modality of the Dynamic-Superb is decided by the tested task. We will provide the auto judge# in the next version.



Difficulty	Speech-In	Speech-Out
L0	Semantic	Semantic
L1	Semantic w/ Para Ling	Semantic
L2	Semantic	Semantic w/ Para Ling
L3	Semantic w/ Para Ling	Semantic w/ Para Ling

Figure 2: The Three-Stage Process of S2S-Bench Construction.

what could be the cause?" The Model is expected to provide possible causes of the headache that simply follows the instruction without any paralinguistic information.

The second level (L1) further evaluates the model's ability to produce corresponding speech output after comprehending paralinguistic information embedded in speech-in, which aligns with the evaluation priorities of Dynamic-SUPERB and AIR-Bench. For example, in the "Identity-based response" task, the model is given a spoken input from a child asking, "If it rains tomorrow, how should I plan my day?" The model is expected to discern the speaker's age using paralinguistic information and respond with suggestions suitable for children rather than adults.

The third level (L2) focuses on evaluating the model generating the speech with paralinguistic information while still adhering to the speech instruction-following requirements. It resembles mainstream TTS evaluation (such as TTS-Arena [3]) but uses speech input in the form of instructions without paralinguistic information. For example, one of the instructions in the "Tongue twisters" task is "Recite a tongue twister at three different speeds: fast, medium, and slow." The model's speech response should not only recite a tongue twister but also demonstrate each recitation at three different speeds.

The fourth level (L3) is the most comprehensive, assessing the model's ability to understand paralinguistic information in the speech input while simultaneously generating speech output that includes corresponding paralinguistic information. This level closely approximates real-world speech-to-speech scenarios. For example, in the "Cross-lingual emotional translation" task, one prompt with a happy emotion is "Help me tell him in Chinese that Mike is coming to my house tomorrow for a week." The model should fully recognize the expressed happiness and translate the message into Chinese with an equivalent emotional tone. Additionally, the translation should use third-person narration, not first-person.

3.3 Sample Recording

We craft several distinct prompts for each instruction difficulty level as scripts for recording sample

[3] https://huggingface.co/spaces/TTS-AGI/TTS-Arena

Domain	Task	Evaluation Target
Education	Pronunciation correction	Can the model correct inaccurate pronunciations?
Education	Emphasis control	Can the model understand stress emphasis and emphasize specific content with the right stress?
Education	Rhythm control	Can the model adjust the output pace, speaking faster or slower as required?
Education	Polyphonic word comprehension	Can the model accurately understand polyphonic word?
Education	Pause and segmentation	Can the model accurately pause and segment in ambiguous cases?
Education	Cross-lingual emotional translation	Can the model accurately convey emotions during translation?
Education	Language consistency	Does the model respond in the same language as the query when asked in different languages?
Social Companionship	Implication understanding	Can the model respond humorously, understanding implied meanings?
Social Companionship	Sarcasm detection	Can the model detect sarcasm in phrases like "You're amazing!"?
Social Companionship	Identity-based response	Can the model adapt responses based on the user's age (child, adult, elderly) and handle identity-based queries?
Social Companionship	Emotion recognition and expression	Can the model recognize emotions and provide appropriate responses based on different emotions?
Entertainment	Singing capability	Can the model sing a song upon request?
Entertainment	Natural sound simulation	Can the model simulate certain natural sounds?
Entertainment	Poetry recitation	Can the model recite poems?
Entertainment	Role-playing	Can the model simulate a character with specific age, gender, accent, and voice tone?
Entertainment	Storytelling	Can the model narrate a story with emotional depth?
Entertainment	Tongue twisters	Can the model correctly pronounce a given tongue twister?
Entertainment	Stand-up comedy/skit performance	Can the model perform a skit, playing both roles in a comedic dialogue?
Medical Consultation	Querying symptoms	Can the model answer questions related to symptoms?
Medical Consultation	Health consultation	Can the model provide general health advice?
Medical Consultation	Psychological comfort	Can the model provide comforting psychological support?

Table 3: Task Description Across Four Domains.

responses (See Appendix A for more details). We utilized Doubao TTS for synthetic speech generation in tasks that current TTS systems can adequately handle. We sample from the existing datasets, which are manually generated for existing tasks, such as emotional speech (RAVDESS (Livingstone and Russo, 2018)). For other tasks with more paralinguistic information, like "Singing capability" that TTS cannot handle, we use manual performance to create samples. To enhance the robustness of the benchmark, we designed varied prompt scripts for each task, using different vocal tones and adding 8 input noises, such as airport background sounds, to simulate diverse acoustic environments.

3.4 Data Quality

In the end, we collected 154 independent speech instruction samples in 21 tasks (See Appendix A for more details about the datasets.). For privacy and security reasons, we did not allow users to upload audio files as input themselves. However, our tasks and sample data are open-access, allowing anyone to submit test cases. These submissions are manually verified before being integrated into the testing process.

To validate the effectiveness of our data, we conducted a manual evaluation. Four native Mandarin speakers (two males and two females) with IELTS scores above 6.5 were recruited to assess data quality. If any of the participants identified an issue with a particular data entry, that entry was discarded.

4 Experiments

4.1 Experimental Settings

Metrics Inspired by the popular evaluation method (Zheng et al., 2023) for LLMs, S2S-Bench adopts Arena-style metrics where human evaluation assesses each model's speech output across various tasks directly and then applies ELO ranking (Elo and Sloan, 1978) to score performance quantitatively.

We chose the ELO ranking method as our metric for two main reasons. First, it enables relative comparisons. While Mean Opinion Score (MOS) (Saeki et al., 2022) is commonly used to assess speech generation quality, our open-ended tasks lack reference answers, making absolute scoring challenging; thus, relative comparisons offer a more practical evaluation approach. Second, ELO ranking allows for dynamic evaluation, accommodating the rapid development and frequent introduction of new models by enabling their seamless inclusion in ongoing assessments.

Specifically, In our evaluation framework, all models start with an initial ELO rating of 1000. Each comparison round is conducted in a no-tie format, with the winning model's ELO score updated based on its relative performance to the competing model. Specifically, we calculate the expected score E_A for model A against model B using the Eq. (1):

$$E_A = \frac{1}{1 + 10^{\frac{R_B - R_A}{400}}} \tag{1}$$

where R_A and R_B are the current ratings of models A and B, respectively. The updated rating R'_A for model A is then computed as Eq. (2):

$$R'_A = R_A + K \times (S_A - E_A) \tag{2}$$

where S_A represents the actual outcome for model A (1 for a win, 0 for a loss), and K is the adjustment factor, set to 32.

Human Evaluation For human evaluation, we adopt a pair-wise scoring method inspired by ChatArena (Chiang et al.) [4], performing reference-free evaluation. Specifically, given input audio, we invite human evaluators to rank the audio outputs from two different models, considering both speech and semantic quality.

To facilitate this, we developed a web-based evaluation tool for annotators. The UI is designed to streamline the evaluation process for human assessors, as shown in Figure 3. Once evaluators have selected a specific domain and task for evaluation, a random sample from the chosen task is presented to them. To mitigate potential misunderstandings regarding the task, the corresponding text transcript of the audio is also provided alongside the audio input and the description of the evaluation target for this task.

For the models under evaluation, we employ an Elo-based selection mechanism to choose the two audio outputs that offer the most meaningful comparison from a pool of outputs generated by models such as ChatGPT-4o, SpeechGPT, Mini-Omni, and FanAudioLLM. Evaluators are then required to select the model that performs best based on the speech and semantic quality. The evaluation interface can be seen in Figure 3.

4.2 Benchmarked models

We select the following representative models for testing.

GPT-4o-realtime[5]: To conduct batch testing efficiently, we utilize the speech-enabled API version of GPT-4o instead of the app version. It is important to note that performance differences exist between these versions, and we use GPT-4o-realtime for the short name of the API.

FunAudioLLM (4o) (SpeechTeam, 2024b): FunAudioLLM is another state-of-the-art model designed for seamless S2S interactions, optimized for handling diverse audio inputs while maintaining semantic consistency. During the replication process, we utilized the gpt-4o-2024-08-06 to replace the Qwen2 72B for the LLM module.

Mini-Omni (Xie and Wu, 2024): It takes the whisper as the audio encoder, followed by an audio adapter to Qwen2 0.5B, and uses the MusicGen as the streaming audio decoding.

SpeechGPT (Zhang et al., 2023): SpeechGPT is regarded as one of the first LLM-based speech models that can deal with speech input. It is established on the LLaMA and takes the traditional HuBERT as the discrete speech unit extractor and HiFi-GAN as the unit vocoder.

LLaMA-Omini (Fang et al., 2024). With the help of the most advanced LLaMA 3.1, LLaMA-Omni uses HuBERT as the encoder and leverages the unit-based vocoder to generate the speech in a two-stage training framework.

Cascade: We also construct a traditional cascade model comprising Automatic Speech Recognition (ASR) via Whisper, a Large Language Model (LLM) using GPT-4o, and Text-to-Speech (TTS) with cosyVoice[6].

We did not take into LSLM (Ma et al., 2024) and Moshi (Défossez et al., 2024) for fair comparison because they did not use the LLM as the backbone.

We standardize the samples to a 24,000 Hz sample rate to ensure fairness in testing. However, due to some models' limited support for certain input formats, we are required to use alternative formats. Specifically, for SpeechGPT, we convert the input audio to a 22,500 Hz sample rate.

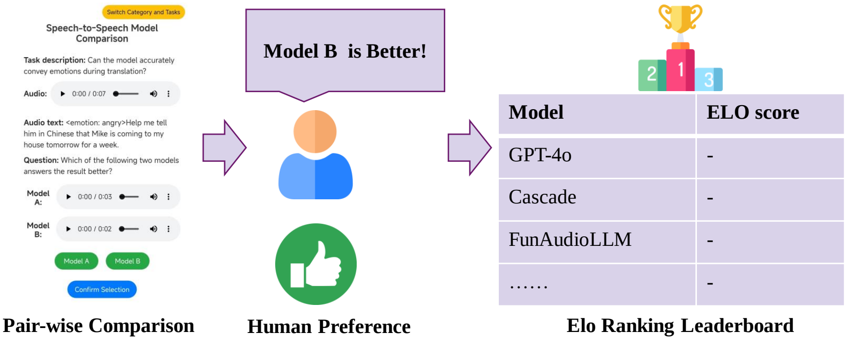
4.3 Results

We conduct a preliminary experimental investigation and receive about 400 pair-wise comparison results with over 22 individuals. We select 10% of the samples annotated by different annotators

[4] https://huggingface.co/spaces/lmsys/chatbot-arena-leaderboard

[5] gpt-4o-realtime-preview-2024-10-01.

[6] whisper-large-v3 for ASR, gpt-4o-2024-08-06 for LLM and CosyVoice-300M-Instruct for TTS.



Speech-to-Speech Model Comparison

Task description: Can the model accurately convey emotions during translation?

Audio: 0:00 / 0:07

Audio text: <emotion: angry>Help me tell him in Chinese that Mike is coming to my house tomorrow for a week.

Question: Which of the following two models answers the result better?

Model A: 0:00 / 0:03

Model B: 0:00 / 0:02

Model	ELO score
GPT-4o	-
Cascade	-
FunAudioLLM	-
……	-

Figure 3: The Evaluation Process of S2S-Bench.

Model	Overall	Education	Social Companionship	Entertainment	Medical Consultation
GPT-4o-realtime	**1365**	**1185**	1064	970	**1146**
Cascade	1207	1065	995	1069	1077
FunAudioLLM (4o)	1025	1105	**1077**	850	993
SpeechGPT	849	906	919	**1095**	929
Mini-Omni	841	857	1000	1041	943
LLaMA-Omni	714	882	945	975	911

Table 4: Elo Rank across Various S2S Models.

simultaneously, and the agreement between annotators is 83.7%.

4.3.1 Overall Model ELO Ranking

Table 4 presents the overall ELO rankings of each model based on their performance across all tasks in the four evaluation domains: Education, Social Companionship, Entertainment Dubbing, and Medical Consultation. It can be seen that GPT-4o is still the most effective model, accompanied by its significant advantages in the fields of education and healthcare. This is because it can better understand speech, especially for rhythm, pronunciation, and better language support.

Surprisingly, without considering other factors such as latency and full duplex, Cascade's model achieved the second-best performance due to its powerful LLM model (GPT-4o). However, other non-GPT-4o-based models exhibit significant performance degradation due to knowledge-intensive education and healthcare scenarios. But there is not much difference between entertainment and social scenes, even better than GPT-4o. We conduct a thorough analysis of this in section 5.4.

4.3.2 Fraction of Model A Wins for A vs. B Battles

We further analyze the win rate between the pairwise battle models, as shown in Figure 4. It can be seen that the first three models based on GPT-4o (GPT-4o real-time, Cascade and FunAudioLLM

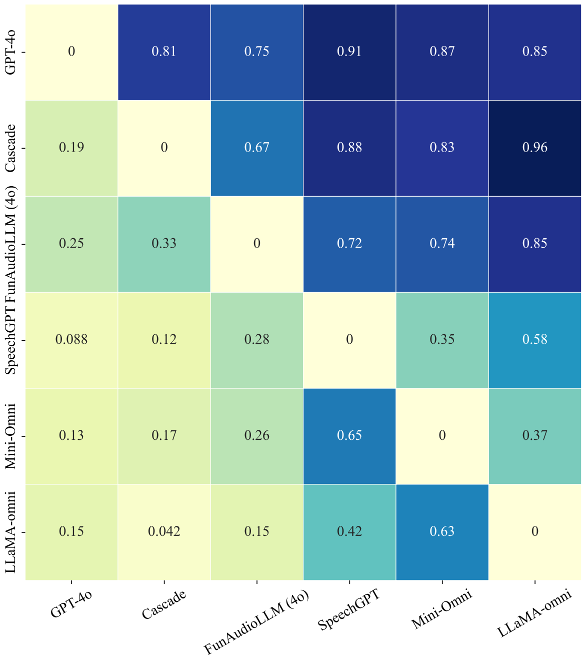


Figure 4: Fraction of Model A Wins for A vs. B Battles.

(40)) are significantly better than the last three models (SpeechGPT, Mini OMini, LLaMA omni). Moreover, although FunAudioLLM has the optimal threat power for GPT-4o real-time, it can be seen that Casecade outperforms the other three open-source models. And the last three models each have their own strengths, making it difficult to determine which one is more outstanding and significant.

5 Analysis

To investigate the challenges and limitations of current S2S models, this study focuses on four key research questions (RQs) that underscore both technical and behavioral complexities in speech-in and speech-out tasks:

RQ1: How do inconsistencies between paralinguistic information and semantic content impact model performance?

RQ2: How effectively do current models support multiple languages?

RQ3: To what extent are there positional biases in annotations?

RQ4: What are the underlying reasons for instruction-following failures?

5.1 Inconsistency between paralinguistic information and semantics in speech-in

We first examine whether S2S models can perceive and appropriately respond to paralinguistic information when there is a mismatch between speech and semantic content. This challenge is particularly relevant in tasks like sarcasm detection, where tone, pitch, and emotional intonation can convey a meaning opposite to the literal semantic content.

In a preliminary analysis, we evaluate the performance of several models (Cascade, GPT-4o, and FunAudioLLM) on the Chinese sarcasm detection task in our dataset. Across the samples, all models prioritized prosodic features, such as sarcastic tone, in 66.67% of cases while focusing on the literal semantic content in the remaining 33.3%. This indicates that the current speech models have a decent ability to understand speech and can, to some extent, detect semantic changes contained in the speech.

5.2 The Language Support

We further examine the language support capabilities of current Speech-to-Speech (S2S) models, focusing on the consistency between the instructional and expressive languages, as shown in Table 5.

Model	Speech-In	Speech-Out
GPT-4o-realtime	EN, CN, JP, TH	EN, CN, JP, TH
Cascade	EN, CN, JP	EN, CN, JP
FunAudioLLM (4o)	EN, CN, JP	EN, CN, JP
LLaMA-Omni	EN, CN	EN
Mini-Omni	EN	EN
SpeechGPT	EN	EN

Table 5: Language Support by Model for Input and Output in Four Languages: English (EN), Chinese (CN), Japanese (JP), and Thai (TH).

It can be seen from the table that GPT-4o real-time does indeed use speech encoding different from traditional cascading methods (such as Whisper in the Cascade model and Sensitivity in FuanAudioLLM 4o), making it better able to handle more diverse languages.

Interestingly, LLaMA Omni, although unable to express Chinese, can understand Chinese input and provide corresponding responses in English. We think this may be related to its unique HuBERT encoding method.

5.3 Position Bias

We further explored the role of positional biases, which usually occur in the comparison of LLMs. Notably, positional shifts influenced the outcomes in 5 out of 22 comparisons. Moreover, within the samples that displayed inconsistencies, 5 out of 7 instances were affected by positional changes, representing a significant proportion, suggesting that positional bias may substantially impact annotator agreement. (**Need more analysis**)

5.4 Instruction Failures in S2S Models

Then, we delve into the specific reasons behind the instruction failures of S2S models. This includes instances where the model is aware of its own limitations and appropriately refuses to respond, cases where it attempts to execute an instruction but fails to complete it, and scenarios where the model is unable to recognize or understand the given instruction. By identifying these distinct failure types, we aim to pinpoint areas for improvement in instruction-following accuracy and reliability within S2S models.

We categorize the causes of failure into three types: Case 1 – the model follows the instruction but performs worse compared to the other model; Case 2 – the model attempts to execute the instruction but fails to complete it; and Case 3 – the model is unable to recognize or understand the given instruction. Overall, the primary causes of failure are Case 1 and Case 3, accounting for 171/456 and 215/456 of instances, while Case 2 accounts for only 70/456.

Specifically, for the three models with higher ELO scores (GPT-4o, Cascade, and FunAudioLLM), the main failure reason is Case 1, with proportions of 17/25, 28/40, and 41/64, respectively. In contrast, for the three models with lower ELO scores (Mini-Omni, SpeechGPT, and LLaMa-Omni), the primary failure reason is Case 3, with proportions of 59/110, 85/128, and 39/89, respectively.

References

James F Allen, Donna K Byron, Myroslava Dzikovska, George Ferguson, Lucian Galescu, and Amanda Stent. 2001. Toward conversational human-computer interaction. *AI magazine*, 22(4):27–27.

Junyi Ao, Yuancheng Wang, Xiaohai Tian, Dekun Chen, Jun Zhang, Lu Lu, Yuxuan Wang, Haizhou Li, and Zhizheng Wu. 2024. Sd-eval: A benchmark dataset for spoken dialogue understanding beyond words. *arXiv preprint arXiv:2406.13340*.

Anton Batliner, Björn Schuller, Dino Seppi, Stefan Steidl, Laurence Devillers, Laurence Vidrascu, Thurid Vogt, Vered Aharonson, and Noam Amir. 2011. *The automatic recognition of emotions in speech*. Springer.

Fan Bu, Yuhao Zhang, Xidong Wang, Benyou Wang, Qun Liu, and Haizhou Li. 2024. Roadmap towards superhuman speech understanding using large language models. *arXiv preprint arXiv:2410.13268*.

Stuart K Card, Allen Newell, and Thomas P Moran. 1983. The psychology of human-computer interaction.

Guiming Hardy Chen, Shunian Chen, Ziche Liu, Feng Jiang, and Benyou Wang. 2024a. Humans or llms as the judge? a study on judgement biases. *arXiv preprint arXiv:2402.10669*.

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

Yushen Chen, Zhikang Niu, Ziyang Ma, Keqi Deng, Chunhui Wang, Jian Zhao, Kai Yu, and Xie Chen. 2024c. F5-tts: A fairytaler that fakes fluent and faithful speech with flow matching. *arXiv preprint arXiv:2410.06885*.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael Jordan, Joseph E Gonzalez, et al. Chatbot arena: An open platform for evaluating llms by human preference. In *Forty-first International Conference on Machine Learning*.

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

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

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

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

Arpad E Elo and Sam Sloan. 1978. The rating of chess-players: Past and present. *(No Title)*.

Qingkai Fang, Shoutao Guo, Yan Zhou, Zhengrui Ma, Shaolei Zhang, and Yang Feng. 2024. Llama-omni: Seamless speech interaction with large language models. *arXiv preprint arXiv:2409.06666*.

Sreyan Ghosh, Sonal Kumar, Ashish Seth, Chandra Kiran Reddy Evuru, Utkarsh Tyagi, Sakshi Singh, Oriol Nieto, Ramani Duraiswami, and Dinesh Manocha. 2024. GAMA: A large audio-language model with advanced audio understanding and complex reasoning abilities. *arXiv preprint arXiv:2406.11768*.

Kaixiong Gong, Kaituo Feng, Bohao Li, Yibing Wang, Mofan Cheng, Shijia Yang, Jiaming Han, Benyou Wang, Yutong Bai, Zhuoran Yang, et al. 2024. Av-odyssey bench: Can your multimodal llms really understand audio-visual information? *arXiv preprint arXiv:2412.02611*.

Shujie Hu, Long Zhou, Shujie Liu, Sanyuan Chen, Hongkun Hao, Jing Pan, Xunying Liu, Jinyu Li, Sunit Sivasankaran, Linquan Liu, et al. 2024. Wavllm: Towards robust and adaptive speech large language model. *arXiv preprint arXiv:2404.00656*.

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, Roshan S. Sharma, Shinji Watanabe, Bhiksha Ramakrishnan, Shady Shehata, and Hung-Yi Lee. 2024. Dynamic-superb: Towards a dynamic, collaborative, and comprehensive instruction-tuning benchmark for speech. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2024, Seoul, Republic of Korea, April 14-19, 2024*, pages 12136–12140. IEEE.

Julia Ipgrave. 2009. The language of friendship and identity: Children's communication choices in an interfaith exchange. *British Journal of Religious Education*, 31(3):213–225.

Shengpeng Ji, Yifu Chen, Minghui Fang, Jialong Zuo, Jingyu Lu, Hanting Wang, Ziyue Jiang, Long Zhou, Shujie Liu, Xize Cheng, et al. 2024. Wavchat: A survey of spoken dialogue models. *arXiv preprint arXiv:2411.13577*.

Shijia Liao, Yuxuan Wang, Tianyu Li, Yifan Cheng, Ruoyi Zhang, Rongzhi Zhou, and Yijin Xing. 2024. Fish-speech: Leveraging large language models for advanced multilingual text-to-speech synthesis. *arXiv preprint arXiv:2411.01156*.

Steven R Livingstone and Frank A Russo. 2018. The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. *PloS one*, 13(5):e0196391.

Ziyang Ma, Yakun Song, Chenpeng Du, Jian Cong, Zhuo Chen, Yuping Wang, Yuxuan Wang, and Xie Chen. 2024. Language model can listen while speaking. *arXiv preprint arXiv:2408.02622*.

Takashi Nose, Yoichi Kato, and Takao Kobayashi. 2007. Style estimation of speech based on multiple regression hidden semi-markov model. In *INTERSPEECH*, pages 2285–2288.

Takaaki Saeki, Detai Xin, Wataru Nakata, Tomoki Koriyama, Shinnosuke Takamichi, and Hiroshi Saruwatari. 2022. Utmos: Utokyo-sarulab system for voicemos challenge 2022. *Interspeech 2022*.

S Sakshi, Utkarsh Tyagi, Sonal Kumar, Ashish Seth, Ramaneswaran Selvakumar, Oriol Nieto, Ramani Duraiswami, Sreyan Ghosh, and Dinesh Manocha. 2024. Mmau: A massive multi-task audio understanding and reasoning benchmark. *arXiv preprint arXiv:2410.19168*.

Björn Schuller, Stefan Steidl, Anton Batliner, Felix Burkhardt, Laurence Devillers, Christian Müller, and Shrikanth Narayanan. 2010. The interspeech 2010 paralinguistic challenge. In *Proc. INTERSPEECH 2010, Makuhari, Japan*, pages 2794–2797.

Björn Schuller, Stefan Steidl, Anton Batliner, Felix Burkhardt, Laurence Devillers, Christian MüLler, and Shrikanth Narayanan. 2013. Paralinguistics in speech and language—state-of-the-art and the challenge. *Computer Speech & Language*, 27(1):4–39.

Tongyi SpeechTeam. 2024a. Funaudiollm: Voice understanding and generation foundation models for natural interaction between humans and llms. *arXiv preprint arXiv:2407.04051*.

Tongyi SpeechTeam. 2024b. Funaudiollm: Voice understanding and generation foundation models for natural interaction between humans and llms. *arXiv preprint arXiv:2407.04051*.

Robert C. Streijl, Stefan Winkler, and David S. Hands. 2016. Mean opinion score (MOS) revisited: methods and applications, limitations and alternatives. *Multim. Syst.*, 22(2):213–227.

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*.

George L Trager. 1958. Paralanguage: A first approximation. *Stud. Linguist.*, 13:1–12.

Bin Wang, Xunlong Zou, Geyu Lin, Shuo Sun, Zhuohan Liu, Wenyu Zhang, Zhengyuan Liu, AiTi Aw, and Nancy F Chen. 2024a. Audiobench: A universal benchmark for audio large language models. *arXiv preprint arXiv:2406.16020*.

Yuancheng Wang, Haoyue Zhan, Liwei Liu, Ruihong Zeng, Haotian Guo, Jiachen Zheng, Qiang Zhang, Xueyao Zhang, Shunsi Zhang, and Zhizheng Wu. 2024b. Maskgct: Zero-shot text-to-speech with masked generative codec transformer. *arXiv preprint arXiv:2409.00750*.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2022. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

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

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

Jiayi Ye, Yanbo Wang, Yue Huang, Dongping Chen, Qihui Zhang, Nuno Moniz, Tian Gao, Werner Geyer, Chao Huang, Pin-Yu Chen, et al. 2024. Justice or prejudice? quantifying biases in llm-as-a-judge. *arXiv preprint arXiv:2410.02736*.

Jun Zhan, Junqi Dai, Jiasheng Ye, Yunhua Zhou, Dong Zhang, Zhigeng Liu, Xin Zhang, Ruibin Yuan, Ge Zhang, Linyang Li, et al. 2024. Anygpt: Unified multimodal llm with discrete sequence modeling. *arXiv preprint arXiv:2402.12226*.

Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. 2023. Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities. In *Proceedings of EMNLP, 2023*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

A Details for Dateset