SD-Eval: A Benchmark Dataset for Spoken Dialogue Understanding Beyond Words

Junyi Ao[1,*] **Yuancheng Wang**[1,*] **Xiaohai Tian**[2]**, Dekun Chen**[1]**,**
Jun Zhang[2]**, Lu Lu**[2]**, Yuxuan Wang**[2]**, Haizhou Li**[1]**, Zhizheng Wu**[1,†]
[1]School of Data Science, SRIBD,
The Chinese University of Hong Kong, Shenzhen, Guangdong 518172, China
[2]Bytedance

Abstract

Speech encompasses a wealth of information, including but not limited to content, paralinguistic, and environmental information. This comprehensive nature of speech significantly impacts communication and is crucial for human-computer interaction. Chat-Oriented Large Language Models (LLMs), known for their general-purpose assistance capabilities, have evolved to handle multi-modal inputs, including speech. Although these models can be adept at recognizing and analyzing speech, they often fall short of generating appropriate responses. We argue that this is due to the lack of principles on task definition and model development, which requires open-source datasets and metrics suitable for model evaluation. To bridge the gap, we present SD-Eval, a benchmark dataset aimed at multidimensional evaluation of spoken dialogue understanding and generation. SD-Eval focuses on paralinguistic and environmental information and includes 7,303 utterances, amounting to 8.76 hours of speech data. The data is aggregated from eight public datasets, representing four perspectives: emotion, accent, age, and background sound. To assess the SD-Eval benchmark dataset, we implement three different models and construct a training set following a process similar to that of SD-Eval. The training set contains 1,052.72 hours of speech data and 724.4k utterances. We also conduct a comprehensive evaluation using objective evaluation methods (e.g. BLEU and ROUGE), subjective evaluations and LLM-based metrics for the generated responses. Models conditioned with paralinguistic and environmental information outperform their counterparts in both objective and subjective measures. Moreover, experiments demonstrate that LLM-based metrics show a higher correlation with human evaluation compared to traditional metrics. We open-source SD-Eval at `https://github.com/amphionspace/SD-Eval`.

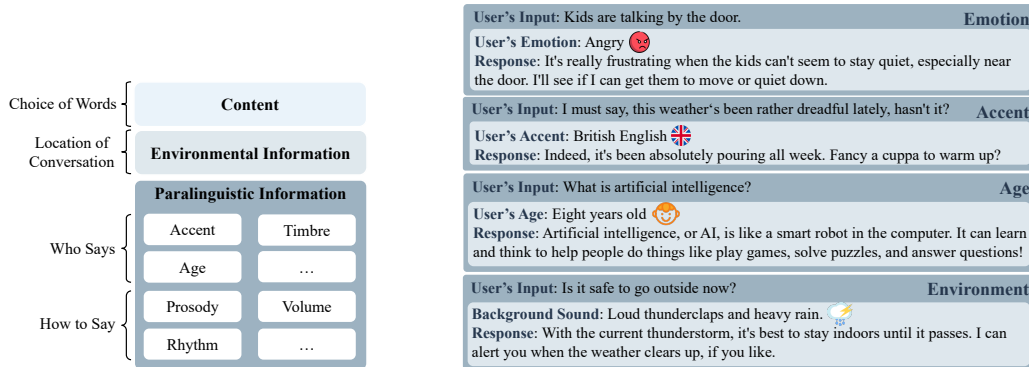
1 Introduction

Speech contains rich information and plays a crucial role in human-computer interaction [12, 14, 41]. Besides relying on the content information, speech also conveys paralinguistic and environmental information, which can significantly influence conversations. More specifically, the information carried in speech can be categorized into three classes: content information, environmental information and paralinguistic information, as illustrated in Figure 1(a).

The *content* information refers to the "**choice of words**", representing the explicit meaning and linguistic structure of the speech. *Environmental information* pertains to "**location of conversation**",

---

*Equal contribution
†Corresponding author: wuzhizheng@cuhk.edu.cn

38th Conference on Neural Information Processing Systems (NeurIPS 2024) Track on Datasets and Benchmarks.

(a) Speech carries rich information including linguistic, para-linguistic and environmental information

(b) Examples of spoken dialogues impacted by the rich information carried in speech (e.g. emotion, accent, age, environment).

Figure 1: (a) Information embedded in speech: content, environmental, and paralinguistic information. (b) Examples of spoken dialogue, which illustrate the impact of user emotions, accents, age, and environmental information on the responses.

capturing the factors such as background noise and situational context that can influence the interpretation of the speech. *Paralinguistic information*, which is further divided into "**who says**" and "**how to say**", includes various non-verbal elements that convey additional meaning. "who says" involves aspects like accent, age, and timber of the speaker, which can affect the perception and understanding of the speech. "how to say" includes prosody, volume, and rhythm, detailing the vocal nuances that contribute to the expressive quality of the speech. Together, all information highlights the multifaceted nature of spoken dialogue, extending beyond mere words to encompass a wide array of information. Figure 1(b) illustrates how environmental and paralinguistic information, such as emotion, accent and age, impact responses.

Large Language Models (LLMs) have shown remarkable capabilities as a universal interface for general-purpose assistance [1, 56, 58, 59, 70, 66, 57, 68]. Recently, LLMs have evolved to understand not only text but also multi-modal inputs, such as speech and image [34, 76, 71, 10, 55, 24, 44], which broadens the scope of what LLMs can achieve. The capabilities of LLMs with speech input (Speech LLMs) are primarily designed for the perception of speech and analysis of tasks defined by a text instruction prompt. This enables the model not only to recognize content but also to perceive additional information, allowing it to perform various speech-related tasks such as speech recognition and gender classification. However, due to the lack of principles on task definitions and model development, they usually fail to generate appropriate responses directly with speech input. The development of advanced Speech LLMs requires open-source datasets and metrics suitable for model evaluation from every aspect of the rich information carried in speech.

We present *a novel benchmark dataset for multidimensional evaluation of spoken dialogue understanding beyond words, namely SD-Eval*. The dataset is to promote the development of more empathetic and intelligent spoken dialogue systems that can generate appropriate responses based on paralinguistic and environmental information. The ultimate goal of SD-Eval is to create a benchmark dataset for speech-to-speech conversation system development. As an initial step, SD-Eval focuses on speech-to-text dialogue. The initial version of SD-Eval consists of four sub-tasks, each focusing on evaluating responses to input utterances with different emotions, accents, ages, and background sounds. These sub-tasks are constructed from eight public datasets containing real-recorded speeches. More specifically, SD-Eval comprises four subsets: *test-emo*, *test-acc*, *test-age*, and *test-env* for emotion, accent, age and background sound, respectively. It includes 7,303 utterances, totalling 8.76 hours of speech data.

To assess the SD-Eval benchmark dataset, we implement three different models and construct a training set following a similar process as SD-Eval. The training set contains 1,052.72 hours of speech data and 724.4k utterances. We also conduct an empirical study of evaluation metrics using objective evaluation methods (e.g. BLEU and ROUGE), subjective opinion score and LLM-based metrics for the generated responses.

2

2 Related Work

Spoken Conversation Datasets with Paralinguistic Label Paralinguistic information is crucial for comprehending speech and generating responses in spoken dialogues. Many speech emotion datasets are constructed under spoken dialogue scenarios, such as IEMOCAP [5], SEMAINE [39], and MELD [48]. However, their primary purpose is to identify emotions in speech. Consequently, the dialogue data from these datasets is relatively less suited for training a spoken dialogue system.

Some recent studies build novel datasets such as E-chat200 [64] and StyleTalk [33], which are designed for spoken dialogue with a focus on emotional information. Nevertheless, the text and speech in these datasets are generated using ChatGPT and text-to-speech (TTS) models. Our dataset is based on a mixture of real-recording and synthesized speech and focuses on multiple aspects, including accents, emotions, ages, and background sounds.

Spoken Question Answering The spoken question answering (SQA) task requires the system to answer questions from speech. The past approaches [60, 54] mainly divided this task into two parts through a cascaded model: automatic speech recognition (ASR) and text question answering. Recently, some systems [67, 40] aim to achieve end-to-end spoken question answering.

Datasets in the field of SQA include Spoken SQuAD [31], SCQA [67], HeySQuAD [63], OpenSAQA [19], e.g. These datasets lack annotations of paralanguage information. StyleTalk [33] provides annotations of speaking styles. Our work focuses more on paralinguistic and environmental information to simulate more realistic dialogue scenarios.

Evaluation Metrics for Open-Ended Generation Tasks Assessing the quality of text produced by language models or human authors for open-ended generation tasks has always been a difficult task. Traditional evaluation metrics such as BLEU [46] and ROUGE [32] are based on the n-grams to measure the similarity between model outputs and references, while these metrics focus on lexical overlap, which is ineffective for open-ended generation problems. In addition, they show a relatively weak correlation with human judgement [42]. Embedding-based metrics, such as BERTScore [72], use word or sentence embeddings to measure semantic similarity based on the references.

However, the answers to these tasks are open-ended without standard references, while collecting human preferences can be costly and laborious. Recently, several works [35, 17, 74] try to use LLMs for evaluating the responses of chat assistants, which shows a high correlation with human judgement. In our work, we adapt these LLM-based methods for spoken dialogue generation, with a focus on paralinguistic and environmental information.

3 SD-Eval Benchmark Dataset

3.1 Dataset Construction

SD-Eval is divided into four subsets: *test-emo*, *test-acc*, *test-age*, and *test-env*. Each subset focuses on a specific aspect: emotion, accent, age, and environment, respectively. The ultimate aim of SD-Eval is to create a benchmark dataset for the evaluation of speech-to-speech conversation systems. As a preliminary step, SD-Eval concentrates on speech-to-text dialogues. We construct SD-Eval through the following steps.

Data Collection As shown in Table 1, we select data from 8 public datasets to construct SD-Eval. For *test-emo* subset, RAVDESS [37], MEAD [62], and JL Corpus [25] are selected as they contain audios with the same content but different emotions. For *test-env* subset, we choose real-recording speeches from the LibriSpeech [45] test-clean subset and add background sounds using audio samples from AudioCaps [28].

Synthetic Data Generation For *test-age* and *test-env*, a portion of the data is synthesized. For *test-age*, we use an internal zero-shot TTS model, which is trained on Libri-light, to generate speech data from the text in MyST [49] with adult speakers. For each text, we randomly select a sample from the LibriSpeech test-clean subset [45] as the prompt to synthesize the data. For *test-env*, we first select audio collections corresponding to seven types of environments from AudioCaps [28]. Then, we mix each speech sample in the subset of LibriSpeech test-clean with audio randomly selected from these

3

Table 1: Statistics of the SD-Eval benchmark dataset, which includes four types of paralinguistic and environmental information.

Type	# Hours	# Utts	Constructed From	Labels
Emotion (*test-emo*)	1.11	1,289	RAVDESS [37], MEAD [62], JL Corpus [25]	Sad, Angry, Fear, Disgust, Happy
Accent (*test-acc*)	5.34	4,310	VCTK [65], Common Voice [3]	England, Scottish, Northern Irish, Welsh, Irish, American, Canadian, Australian, New Zealand
Environment (*test-env*)	0.74	690	LibriSpeech [45], AudioCaps [28], Synthesised Speech	Driving, Children's Voice, Sea Beach, Raining or Thundering, Bells, Sports Center, Bus or Subway
Age (*test-age*)	1.57	1,014	MyST [49], Synthesised Speech	Adult, Child
Summary	8.76	7,303	-	-

collections corresponding to each environmental scene. Simultaneously, we utilize GPT-4-Turbo [1] to generate dialogue data for these seven scenarios and employ the TTS model to generate speech, forming part of the *test-env* subset.

Label Normalization Due to the varying number of label categories across different datasets, we first normalize the labels of all datasets. Specifically, labels of *test-acc* include *nine* widely used and representative accents: England, Scottish, Irish, Welsh, Northern Irish, American, Canadian, Australian, and New Zealand. For the *test-emo* subset, we firstly utilize Ekman's emotion model [16] as the labels, which contain neutral, surprise, sad, happy, angry, disgust, and fear, which are the basic emotions. We choose Ekman's emotion model because it is widely used in speech emotion recognition task [5, 37, 62], ensuring that each category of emotion is well-represented and encompasses a substantial amount of data.

We then further exclude utterances with neutral and surprise emotions. Neutral implies that the speech does not convey positive or negative feelings, making the response primarily content-dependent. However, our focus is on examining the impact of speech emotion on text responses. Similarly, surprise can be associated with different sentiments, depending on the context [52]. Therefore, we excluded data related to these two emotions. As a result, the *test-emo* subset includes five types of emotions: sad, happy, angry, disgust, and fear.

For the *test-env* subset, we select seven representative scenarios in daily life to serve as background sounds, as illustrated in Table 1. For the *test-age* subset, we focus on evaluating whether the model could generate comprehensible responses appropriate to different age groups. Consequently, the labels are divided into two categories: child and adult.

Data Filtering We filter the test data from three aspects. Firstly, some utterances of the four subsets are identified with notable ambiguity, potentially due to a lack of contextual information. To address this, we design a prompt and use GPT-4-turbo [1] for automatic filtering, as illustrated in Figure 2. Following this initial filtering, three human annotators are then required to evaluate the remaining utterances further using the same criteria as the prompt. Secondly, it is observed that some utterances within the *test-env* contain incorrect background sounds, possibly due to the multi-class labelling of the AudioCaps [28]. These utterances are subsequently identified and filtered by human annotators. Finally, we exclude utterances of *test-emo* subset where both the sentiment of transcript and emotion are positive or negative, aiming to enhance the impact of emotions on responses. For this purpose, a pre-trained sentiment classification model [3] is employed to predict the sentiments of utterances.

Punctuation Restoration Traditional metrics, such as BLEU [46], require references as input, so we try to use ChatGPT [43] to generate responses for each utterance. However, the transcripts of three datasets, MEAD [62], LibriSpeech [45] and UK-Ireland dataset [15], do not contain punctuation,

---

[3] https://huggingface.co/lxyuan/distilbert-base-multilingual-cased-sentiments-student

4


```
[System]
Evaluate the clarity and feasibility of a specified sentence for processing by
ChatGPT. Your response should consist of two parts:
1. State 'Yes' if the sentence is clear and feasible, or 'No' if it is
   ambiguous or vague.
2. Explain your assessment.
Ensure your response adheres strictly to JSON format with two keys: "answer"
and "reason", corresponding to your answers for the above questions.

[Sentence for Evaluation]
{input_sentence}
```

Figure 2: The prompt for filtering utterances.

```
[System]
Let's simulate a conversation between a hypothetical speaker and ChatGPT. I
need you to:
1. Create **five** diverse responses, each consisting of two or more sentences,
in reaction to the speaker's statement. Each response should appropriately
reflect the context and content provided by the speaker.
2. Assign an emotion selected from **joy, sadness, fear, anger, surprise,
neutral and disgust** to each response, with the language of the reply
demonstrating this emotion.

Format each of your responses with XML tags, such as <reply>reply</reply> and
<reply_emotion>reply emotion</reply_emotion>, which are corresponding to the
tasks above.

[Simulated Speaker's Statement]
{input_statement}

[Speaker's Emotion]
{emotion}
```

Figure 3: The prompt for generating responses of utterances related to emotion.

which may degrade the quality of generated responses. To address this issue, we employ a punctuation restoration model [4] to add punctuation for transcripts of these two datasets.

Response Generation Finally, we use GPT-4o [44] to generate five diverse responses for each utterance in SD-Eval by considering the content and emotion, accent, age or background sounds of speech signals. For instance, the prompt used to generate responses for utterances related to emotion is presented in Figure 3. All the prompts used to generate responses are included in the Appendix.

3.2 Dataset Statistics

The statistics of SD-Eval are presented in Table 1. The SD-Eval dataset comprises a total of 7,303 sentences and 8.76 hours of speech data. It contains three types of paralinguistic information (i.e. emotion, accent, age), and the environment type contains seven categories of environmental sounds. The pie charts in Figure 4 illustrate the data distribution for each category within each test set.

(a) test-emo   (b) test-acc   (c) test-age   (d) test-env

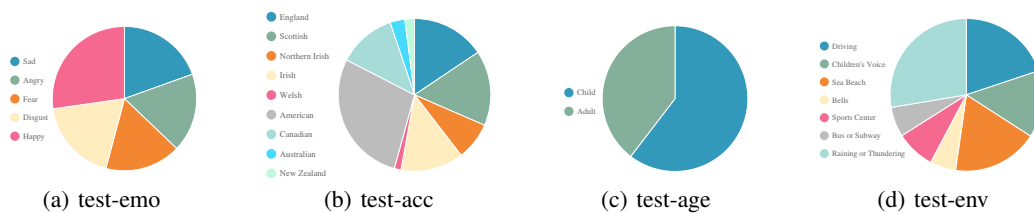



Figure 4: Pie charts illustrating the data distribution for each category within each subset.

---

[4] https://github.com/notAI-tech/fastPunct

5

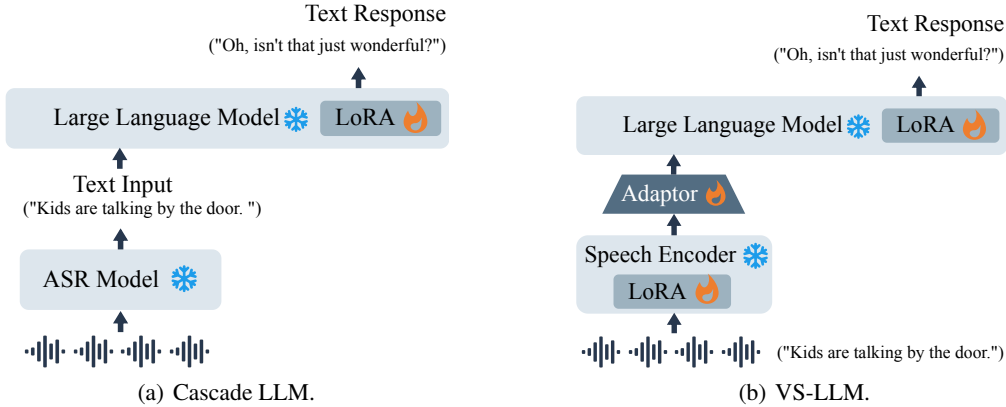



Figure 5: (a) Model Structure of Cascade LLM, which generates text response directly based on the ASR output. (b) Model structure of Vanilla Speech LLM (VS-LLM). The LLM takes speech representation as input, which is generated from a speech encoder and adaptor.

4 Benchmark Experiments

4.1 Training Set

To assess the SD-Eval benchmark dataset, we construct a training dataset from eleven open datasets for training models. We follow a procedure similar to SD-Eval, with the following two exceptions. Firstly, we simplify the data filtering process by removing sentences with inadequate and ambiguous labels. Secondly, we generated only one response for each sentence. The details, including data statistics and prompts, are introduced in the Appendix.

4.2 Models

We implement several baselines trained using the proposed training set, aiming to evaluate their capability of comprehending the content of the speech, as well as recognizing emotions, accents, age, or background sounds. The implementations are detailed as follows.

Cascade LLM   As shown in Figure 5(a), the Cascade LLM consists of an automatic speech recognition (ASR) model to recognize the content, followed by an LLM to generate a response based on the text input. The ASR model is Whisper large-v3 [50], which is trained with a large amount of weakly supervised data for speech recognition and translation. The LLM is the InternLM2 chat model with 7 billion parameters (InternLM2-chat-7b) [6]. During training, the LLM is frozen, while we add a trainable LoRA adaptor [23] to facilitate model finetuning. We use this model as a baseline to evaluate responses if only knowing the content of the speech.

VS-LLM   To understand and perceive content as well as paralinguistic and environmental information directly from speech, we design an end-to-end model named Vanilla Speech LLM (VS-LLM). As shown in 5(b), it consists of a speech encoder, an LLM and an additional adaptor to connect the speech encoder and LLM. The encoder of Whisper large-v3 [50] is used as the speech encoder, followed by a trainable adaptor to further down-sample the speech representation from the speech encoder. The adaptor comprises two linear layers, where the first linear layer is succeeded by a GELU activation function [21], while the second one is followed by a two-dimensional average pooling operation for down-sampling. Similar to Cascade LLM, a frozen InternLM2-chat-7b with a trainable LoRA adaptor is employed as the LLM.

LLM (Upper Bound)   To assess the system's upper bound upon the speech transcript, we also provide paralinguistic or environmental information as an additional label to the frozen InternLM2-chat-7b with LoRA for model finetuning. The input format is a concatenation of ground-truth transcripts and labels. For instance, *"How are you?<Emotion:Happy>"* is the input of an utterance. The transcript of this utterance is "How are you?" The emotion contained in this utterance is happy.

6

Besides the above self-implemented models, we further assess the performance of the off-the-shelf speech LLM models on SD-Eval. All text instruction prompts for open-sourced models can be find in Appendix A.6.

Qwen-Audio   Qwen-Audio [10] is designed for handling a wide range of audio types and tasks. The model scales up pre-training across more than 30 tasks, including speech, natural sounds, and music, in multiple languages, achieving strong performance without task-specific fine-tuning. Since Qwen-Audio requires a text instruction prompt for each input to define the task, we add a text instruction prompt to let the model generate a text response based on the speech.

Qwen2-Audio   Qwen2-Audio [11] is a work based on Qwen-Audio. It is trained on larger-scale data and employs DPO [51] for optimising models to follow human preferences. Unlike Qwen-Audio, Qwen2-Audio has two modes: voice chat (Qwen2-Audio-VC) and audio analysis (Qwen2-Audio-AA). In Audio Analysis mode, it can analyze various audio types and identify command segments within the audio. In Voice Chat mode, it acts like a conversational agent, allowing users to engage in dialogue through audio. We also utilize a text instruction prompt for audio analysis mode during evaluation.

SALMONN   SALMONN [55] is a multi-modal large language model designed to process and understand general auditory inputs, including speech, audio events, and music. SALMONN integrates a pre-trained text-based large language model (LLM) with dual auditory encoders — Whisper [50] for speech and BEATs [9] for non-speech audio. We use a predefined text prompt for evaluation.

4.3 Evaluation Metrics

Objective Evaluation   We propose a reference-free metric using the LLMs for response evaluation. Specifically, we design different prompts for the evaluations of each subset. By the prompts, the LLM judge must consider (a) the response's naturalness, coherence, engagingness and groundedness. (b) Whether the response is appropriate and fully considers the emotion, accent, age or background sound of input speech. The LLM judge is then asked to directly assign a score, such as 5 on a 1 - 10 scale, to a single answer. For comparison, we further include the results of n-gram-based metrics, such as ROUGE-L [32], BLEU-4 [46] and METEOR [4], and embedding-based metrics, such as BERTScore [72] [5].

Subjective Evaluation   In addition, we conduct a human evaluation on 200 randomly selected utterances from the four subsets, with each subset contributing 50 utterances. Each sample is assessed by at least three human evaluators, who are instructed to rate the generated responses. Each utterance has three samples, corresponding to three utterance-response pairs generated by Cascade LLM, VS-LLM, and LLM (Upper Bound), respectively. We ensure that each valid sample is evaluated by at least three human annotators. Consequently, each subset has no fewer than 120 valid samples.

4.4 Experimental Setup

Configuration for Model Training   All models implemented by ourselves are built using xtuner [13]. We optimize the model with AdamW [36] with a learning rate of 2×10^{-4}. The models are finetuned on 16 A100 GPUs, each with a batch size of 16, for two epochs. For the LoRA adaptor of the InternLM2-chat-7b model, we use a rank of 512 and α of 256. In contrast, for the encoder of the Whisper large-v3 model, the rank is set to 64 and α to 16.

Inference Setting of LLM Judge   In addition to GPT-4o, we employ Yi-1.5-34B-Chat [6] [68], Qwen2-57B-A14B-Instruct [7] [66], and gemma-2-27b-it [8] [57], as LLM judges. To speed up inference and save memory, we utilize llama.cpp [9] for LLM inference. Specifically, we use a quantized

---

[5]We use Hugging Face Evaluate for scoring and the BERT model is *roberta-large*.

[6]https://huggingface.co/bartowski/Yi-1.5-34B-Chat-GGUF

[7]https://huggingface.co/Qwen/Qwen2-57B-A14B-Instruct-GGUF

[8]https://huggingface.co/bartowski/gemma-2-27b-it-GGUF

[9]https://github.com/ggerganov/llama.cpp

7

Table 2: Main results of five models on four subsets of SD-Eval. † The scores from human evaluations are calculated based on randomly sampled data as described in Section 4.3.

| Model | BLEU-4 | ROUGE-L | METEOR | BERTScore | LLM Judges | | | | Human |
					Yi-1.5	Qwen2	Gemma	GPT-4o	Evaluation †
test-emo / Emotion									
SALMONN [55]	2.48	16.57	18.97	86.20	4.98	3.35	2.32	2.61	-
Qwen-Audio [10]	3.93	19.02	16.82	86.59	4.19	2.35	2.02	2.24	-
Qwen2-Audio-AA [11]	3.01	16.82	17.51	86.17	4.75	2.52	2.21	2.33	-
Qwen2-Audio-VC [11]	2.21	14.57	22.08	85.41	5.88	3.83	2.93	3.25	-
Cascade LLM	4.66	21.98	21.70	87.93	5.67	3.86	2.35	4.47	5.05
VS-LLM	8.29	25.52	27.23	89.48	6.40	4.56	4.03	5.30	6.31
LLM (Upper Bound)	**12.35**	**26.08**	**28.27**	**89.77**	**7.03**	**5.82**	**6.46**	**6.74**	**7.29**
test-acc / Accent									
SALMONN[55]	7.50	22.22	21.23	87.53	5.27	6.16	3.16	2.93	-
Qwen-Audio [10]	4.52	17.15	17.78	85.59	3.48	3.45	1.86	1.72	-
Qwen2-Audio-AA [11]	7.26	21.80	19.68	87.68	5.04	6.13	3.01	2.54	-
Qwen2-Audio-VC [11]	3.47	17.46	23.77	86.26	5.96	6.20	3.94	4.37	-
Cascade LLM	14.51	30.53	34.13	89.66	7.23	7.32	5.65	6.62	6.71
VS-LLM	17.98	33.06	37.65	90.08	7.82	7.65	6.59	7.85	7.95
LLM (Upper Bound)	**18.35**	**33.48**	**38.27**	**90.23**	**7.85**	**7.75**	**6.73**	**8.02**	**8.30**
test-age / Age									
SALMONN[55]	10.03	24.95	23.55	88.10	5.41	4.66	3.14	3.35	-
Qwen-Audio [10]	7.28	23.09	21.80	86.72	4.43	3.98	2.25	2.50	-
Qwen2-Audio-AA [11]	6.81	22.72	20.51	87.47	5.19	4.51	3.01	3.14	-
Qwen2-Audio-VC [11]	5.64	18.90	28.23	86.70	7.03	5.92	4.48	5.06	-
Cascade LLM	15.36	31.96	31.99	90.08	7.22	7.16	6.46	4.47	6.51
VS-LLM	17.22	34.17	33.78	90.63	7.74	7.39	7.25	7.95	7.11
LLM (Upper Bound)	**18.78**	**35.62**	**36.01**	**91.00**	**7.82**	**7.54**	**7.40**	**8.25**	**7.44**
test-env / Environment									
SALMONN[55]	2.87	16.53	21.37	86.71	4.70	5.00	3.40	3.56	-
Qwen-Audio [10]	2.37	16.83	17.50	85.81	3.77	1.86	2.16	2.14	-
Qwen2-Audio-AA [11]	2.97	16.32	19.84	86.50	4.52	5.02	3.49	3.50	-
Qwen2-Audio-VC [11]	2.06	12.35	23.40	85.17	6.30	6.21	4.85	5.30	-
Cascade LLM	5.44	21.75	26.41	88.22	6.03	5.84	5.31	5.66	6.62
VS-LLM	9.42	25.85	28.27	89.23	6.14	5.88	5.10	5.82	7.11
LLM (Upper Bound)	**11.72**	**27.95**	**31.50**	**89.73**	**7.14**	**7.14**	**6.25**	**7.40**	**8.13**

version Q6_K of these two models to achieve a balance between efficiency and performance. This configuration allows using CPUs or only one A100 GPU for evaluation.

4.5 Main Results

Table 2 shows the main results of all models on SD-Eval. Firstly, across all four test sets, VS-LLM outperformed Cascade LLM on all metrics. This indicates that using speech as a direct input allows VS-LLM to implicitly learn paralinguistic and environmental information. Secondly, the performance of VS-LLM is inferior to that of LLM (Upper Bound). The main reason may be that VS-LLM implicitly acquires content as well as paralinguistic and environmental information directly from speech, whereas the LLM (Upper Bound) utilizes ground truth transcripts and labels. This indicates that the way to process the input data is important for model performance. A detailed ablation study regarding the input data will be introduced later.

As for the open-sourced models, while SALMONN [55] achieves better or comparable performance compared to Qwen2-Audio-AA [11] and Qwen-Audio [10], Qwen2-Audio-VC [11] performs much better than other open-sourced models as voice chat mode is more suitable for conversation. However, the performance of open-sourced models in SD-Eval is not very impressive. This suggests a current lack of well-defined tasks and datasets in this area. To improve the model's performance, future directions may include using larger-scale and more diverse data [20, 27, 69], as well as employing methods such as disentanglement [26, 73] to enhance the model's understanding of speech information.

4.6 Analysis

Ablation Study of Input Data We further conduct an ablation study in terms of the input data, as shown in Table 3. We investigate several models with different inputs. Among them, Model 1,

8

Table 3: Ablation study on *test-emo* subset. The model types include LLM (text input only) and Speech LLM (text and speech inputs). "Trans" refers to the method used to obtain the transcripts. Options include "ASR" (generated by an ASR model) and "GT" (ground-truth transcript). "Emotion Label" indicates the source of the speech emotion label for the utterance, either "SER" (produced by a speech emotion recognition model) or "GT" (ground-truth label). "N/A" means the input is not used for the model.

Index	Model Type	Trans	Emotion Label	BLEU-4	ROUGE-L	METEOR	BERTScore	LLM Judge			
								Yi-1.5	Qwen2	Gemma	GPT-4o
1		N/A	N/A	8.29	25.52	27.23	89.48	6.40	4.56	4.03	5.30
2	Speech LLM	N/A	SER	10.37	27.29	28.59	89.81	6.76	5.26	4.37	6.11
3		N/A	GT	10.21	27.22	28.45	89.85	6.96	5.45	4.45	6.41
4		ASR	N/A	4.66	21.98	21.70	87.93	5.67	3.86	2.35	4.47
5		ASR	SER	11.37	26.03	27.66	89.66	6.74	5.35	5.62	6.13
6	LLM	GT	SER	11.85	26.05	27.78	89.75	6.85	5.53	6.07	6.38
7		ASR	GT	11.85	26.03	28.19	89.68	6.96	5.64	6.04	6.47
8		GT	GT	**12.35**	**26.08**	**28.27**	**89.77**	**7.03**	**5.82**	**6.46**	**6.74**

Table 4: Spearman (ρ) and Kendall-Tau (τ) correlations between human evaluation and different metrics.

Metrics	test-emo		test-acc		test-age		test-env		Overall	
	ρ	τ	ρ	τ	ρ	τ	ρ	τ	ρ	τ
BLEU-4	0.211	0.170	0.197	0.156	0.316	0.232	0.169	0.136	0.208	0.160
ROUGE-L	0.203	0.140	0.263	0.192	0.318	0.215	0.216	0.144	0.258	0.176
METEOR	0.249	0.168	0.272	0.194	0.350	0.242	0.315	0.214	0.336	0.229
BERTScore	0.378	0.269	0.252	0.184	0.321	0.216	0.286	0.199	0.291	0.199
Yi-1.5	0.641	0.493	0.435	0.345	0.356	0.289	0.558	0.441	0.492	0.381
Qwen2	0.618	0.456	0.198	0.161	0.347	0.278	0.449	0.352	0.474	0.362
Gemma	0.639	0.493	0.375	0.304	0.448	0.356	0.563	0.439	0.492	0.380
GPT-4o	**0.731**	**0.568**	**0.659**	**0.541**	**0.474**	**0.356**	**0.577**	**0.459**	**0.613**	**0.468**

which belongs to Speech LLM and is without any text input, refers to VS-LLM. Model 4 utilizing transcripts from the ASR model as input is Cascade LLM. Additionally, Model 8 uses ground truth transcripts and labels, which is LLM (Upper Bound). For ASR and speech emotion recognition (SER), the models are Whisper large-v3 [18] and emotion2vec [38] [10].

Firstly, we examine the effect of content quality. We observe that the performance of models utilizing ASR-generated transcripts (Model 5 and Model 7) is inferior across all metrics compared to their counterparts (Model 6 and Model 8) that use ground-truth transcripts. Next, we examine the effect of emotion label quality. For the LLM-based system, models using emotion labels from the SER model (Model 5 and Model 6) perform worse across all metrics compared to those using ground-truth labels (Model 7 and Model 8). Model 4, which is trained without emotion labels, performs the worst. A similar trend is observed in the models of Speech LLM, where Model 2 obtained emotion labels from the SER model outperforms Model 1, while Model 3, trained with ground truth labels, achieves the best performance among all three models. This corroborates our hypothesis in the section 4.5.

Correlations between Objective Metrics and Human Evaluation Finally, we investigate the correlations between scores of objective metrics and human evaluation, as shown in Table 4. Following the configuration of GPTScore [17], we utilize dataset-level Spearman and Kendall-Tau correlation metrics. Firstly, the experimental results indicate that all LLM judges exhibit a significantly higher correlation with human evaluations compared to other metrics. Secondly, as an LLM judge, GPT-4o consistently achieves the best or second-best performance in most cases. These findings strongly validate the effectiveness of LLM judges as evaluation metrics.

---

[10] https://huggingface.co/emotion2vec/emotion2vec_plus_seed

9

5 Conclusion

In this paper, we introduce SD-Eval, a benchmark dataset designed for the multidimensional evaluation of spoken dialogue understanding and generation. SD-Eval includes 7,303 utterances amounting to 8.76 hours of speech data, aggregated from eight public datasets, and focuses on paralinguistic and environmental information across four perspectives: emotion, accent, age, and background sound. The dataset aims to advance the creation of more empathetic and intelligent spoken dialogue systems capable of generating appropriate responses by considering paralinguistic and environmental information. Our comprehensive evaluation demonstrates that models conditioned with paralinguistic or environmental information outperform their counterparts in both objective evaluation and subjective evaluation. Furthermore, our experiments indicate that LLM-based metrics have a higher correlation with human evaluation compared to traditional metrics.

6 Limitations and Future Work

The limitations and future work for SD-Eval are as follows: Firstly, SD-Eval accommodates only speech-to-text dialogues, limiting the evaluation of system responses at the text level. Secondly, SD-Eval currently supports the evaluation of single-turn dialogues only, limiting its application to more complex, multi-turn interactions. Finally, SD-Eval includes four sub-tasks that focus on speech elements such as emotion, accent, age, and environmental information. However, it does not yet account for other aspects, such as the gender of the speaker. Addressing these aspects constitutes our future work, with the ultimate goal of developing a benchmark dataset capable of multidimensional evaluation for multi-turn speech-to-speech dialogues.

Acknowledgments

This work is partially supported by the National Natural Science Foundation of China (Grant No. 62271432 and No. 62376237), Shenzhen Science and Technology Program (Shenzhen Key Laboratory Grant No. ZDSYS20230626091302006), and Shenzhen Science and Technology Research Fund for the Fundamental Research Key Project (Project Grant No. JCYJ20220818103001002). This work was conducted in collaboration with Bytedance.

References

[1] 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.

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

[3] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber. Common voice: A massively-multilingual speech corpus. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 4211–4215, 2020.

[4] Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In Jade Goldstein, Alon Lavie, Chin-Yew Lin, and Clare Voss, editors, *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics. URL https://aclanthology.org/W05-0909.

[5] 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*, 42:335–359, 2008.

[6] Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, et al. Internlm2 technical report. *arXiv preprint arXiv:2403.17297*, 2024.

[7] Houwei Cao, David G. Cooper, Michael K. Keutmann, Ruben C. Gur, Ani Nenkova, and Ragini Verma. Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE Transactions on Affective Computing*, 5(4):377–390, 2014. doi: 10.1109/TAFFC.2014.2336244.

10

[8] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518, 2022.

[9] Sanyuan Chen, Yu Wu, Chengyi Wang, Shujie Liu, Daniel Tompkins, Zhuo Chen, and Furu Wei. Beats: Audio pre-training with acoustic tokenizers. *arXiv preprint arXiv:2212.09058*, 2022.

[10] 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.

[11] 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.

[12] P R Cohen and S L Oviatt. The role of voice input for human-machine communication. *Proceedings of the National Academy of Sciences*, 92(22):9921–9927, 1995. doi: 10.1073/pnas.92.22.9921. URL https://www.pnas.org/doi/abs/10.1073/pnas.92.22.9921.

[13] XTuner Contributors. Xtuner: A toolkit for efficiently fine-tuning llm. https://github.com/InternLM/xtuner, 2023.

[14] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J.G. Taylor. Emotion recognition in human-computer interaction. *IEEE Signal Processing Magazine*, 18(1):32–80, 2001. doi: 10.1109/79.911197.

[15] Isin Demirsahin, Oddur Kjartansson, Alexander Gutkin, and Clara Rivera. Open-source Multi-speaker Corpora of the English Accents in the British Isles. In *Proceedings of The 12th Language Resources and Evaluation Conference (LREC)*, pages 6532–6541, Marseille, France, May 2020. European Language Resources Association (ELRA). ISBN 979-10-95546-34-4. URL https://www.aclweb.org/anthology/2020.lrec-1.804.

[16] Paul Ekman and Wallace V Friesen. Constants across cultures in the face and emotion. *Journal of personality and social psychology*, 17(2):124, 1971.

[17] Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. Gptscore: Evaluate as you desire. *arXiv preprint arXiv:2302.04166*, 2023.

[18] Yuan Gong, Sameer Khurana, Leonid Karlinsky, and James Glass. Whisper-at: Noise-robust automatic speech recognizers are also strong general audio event taggers. *arXiv preprint arXiv:2307.03183*, 2023.

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

[20] Haorui He, Zengqiang Shang, Chaoren Wang, Xuyuan Li, Yicheng Gu, Hua Hua, Liwei Liu, Chen Yang, Jiaqi Li, Peiyang Shi, et al. Emilia: An extensive, multilingual, and diverse speech dataset for large-scale speech generation. *arXiv preprint arXiv:2407.05361*, 2024.

[21] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.

[22] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

[23] 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. *arXiv preprint arXiv:2106.09685*, 2021.

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

[25] Jesin James, Li Tian, and Catherine Inez Watson. An Open Source Emotional Speech Corpus for Human Robot Interaction Applications. In *Proc. Interspeech 2018*, pages 2768–2772, 2018. doi: 10.21437/Interspeech.2018-1349.

[26] Zeqian Ju, Yuancheng Wang, Kai Shen, Xu Tan, Detai Xin, Dongchao Yang, Yanqing Liu, Yichong Leng, Kaitao Song, Siliang Tang, et al. Naturalspeech 3: Zero-shot speech synthesis with factorized codec and diffusion models. *arXiv preprint arXiv:2403.03100*, 2024.

11

[27] Jacob Kahn, Morgane Riviere, Weiyi Zheng, Evgeny Kharitonov, Qiantong Xu, Pierre-Emmanuel Mazaré, Julien Karadayi, Vitaliy Liptchinsky, Ronan Collobert, Christian Fuegen, et al. Libri-light: A benchmark for asr with limited or no supervision. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7669–7673. IEEE, 2020.

[28] Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. Audiocaps: Generating captions for audios in the wild. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 119–132, 2019.

[29] Jaehyeon Kim, Keon Lee, Seungjun Chung, and Jaewoong Cho. Clam-tts: Improving neural codec language model for zero-shot text-to-speech. *arXiv preprint arXiv:2404.02781*, 2024.

[30] Mateusz Łajszczak, Guillermo Cámbara, Yang Li, Fatih Beyhan, Arent van Korlaar, Fan Yang, Arnaud Joly, Álvaro Martín-Cortinas, Ammar Abbas, Adam Michalski, et al. Base tts: Lessons from building a billion-parameter text-to-speech model on 100k hours of data. *arXiv preprint arXiv:2402.08093*, 2024.

[31] Chia-Hsuan Li, Szu-Lin Wu, Chi-Liang Liu, and Hung-yi Lee. Spoken squad: A study of mitigating the impact of speech recognition errors on listening comprehension. *arXiv preprint arXiv:1804.00320*, 2018.

[32] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL `https://aclanthology.org/W04-1013`.

[33] 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. *arXiv preprint arXiv:2402.12786*, 2024.

[34] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024.

[35] Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: NLG evaluation using gpt-4 with better human alignment. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.153. URL `https://aclanthology.org/2023.emnlp-main.153`.

[36] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[37] Reza Lotfian and Carlos Busso. Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings. *IEEE Transactions on Affective Computing*, 10(4): 471–483, 2019. doi: 10.1109/TAFFC.2017.2736999.

[38] Ziyang Ma, Zhisheng Zheng, Jiaxin Ye, Jinchao Li, Zhifu Gao, Shiliang Zhang, and Xie Chen. emotion2vec: Self-supervised pre-training for speech emotion representation. *arXiv preprint arXiv:2312.15185*, 2023.

[39] Gary McKeown, Michel F Valstar, Roderick Cowie, and Maja Pantic. The semaine corpus of emotionally coloured character interactions. In *2010 IEEE international conference on multimedia and expo*, pages 1079–1084. IEEE, 2010.

[40] 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*, 2023.

[41] Clifford Nass and Scott Brave. *Wired for Speech: How Voice Activates and Advances the Human-Computer Relationship*. The MIT Press, 2005. ISBN 0262140926.

[42] Jekaterina Novikova, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. Why we need new evaluation metrics for nlg. *arXiv preprint arXiv:1707.06875*, 2017.

[43] OpenAI. Chatgpt, 2022. `https://openai.com/blog/chatgpt/`.

[44] OpenAI. Gpt-4o, 2024. `https://openai.com/index/hello-gpt-4o/`.

[45] 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)*, pages 5206–5210, 2015. doi: 10.1109/ICASSP.2015.7178964.

12

[46] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In Pierre Isabelle, Eugene Charniak, and Dekang Lin, editors, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL https://aclanthology.org/P02-1040.

[47] Puyuan Peng, Po-Yao Huang, Daniel Li, Abdelrahman Mohamed, and David Harwath. Voicecraft: Zero-shot speech editing and text-to-speech in the wild. *arXiv preprint arXiv:2403.16973*, 2024.

[48] 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, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1050. URL https://aclanthology.org/P19-1050.

[49] Sameer Pradhan, Ronald A. Cole, and Wayne H. Ward. My science tutor (MyST)–a large corpus of children's conversational speech. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 12040–12045, Torino, Italia, May 2024. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.1052.

[50] 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*, pages 28492–28518. PMLR, 2023.

[51] 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. *Advances in Neural Information Processing Systems*, 36, 2024.

[52] Kashfia Sailunaz and Reda Alhajj. Emotion and sentiment analysis from twitter text. *Journal of computational science*, 36:101003, 2019.

[53] Guillaume Sanchez, Honglu Fan, Alexander Spangher, Elad Levi, Pawan Sasanka Ammanamanchi, and Stella Biderman. Stay on topic with classifier-free guidance. *arXiv preprint arXiv:2306.17806*, 2023.

[54] Dan Su and Pascale Fung. Improving spoken question answering using contextualized word representation. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8004–8008. IEEE, 2020.

[55] 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. URL https://openreview.net/forum?id=14rn7HpKVk.

[56] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

[57] Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

[58] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[59] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[60] Bo-Hsiang Tseng, Sheng-Syun Shen, Hung-Yi Lee, and Lin-Shan Lee. Towards machine comprehension of spoken content: Initial toefl listening comprehension test by machine. *arXiv preprint arXiv:1608.06378*, 2016.

[61] Chengyi Wang, Sanyuan Chen, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, et al. Neural codec language models are zero-shot text to speech synthesizers. *arXiv preprint arXiv:2301.02111*, 2023.

13

[62] Kaisiyuan Wang, Qianyi Wu, Linsen Song, Zhuoqian Yang, Wayne Wu, Chen Qian, Ran He, Yu Qiao, and Chen Change Loy. Mead: A large-scale audio-visual dataset for emotional talking-face generation. In *ECCV*, 2020.

[63] Yijing Wu, SaiKrishna Rallabandi, Ravisutha Srinivasamurthy, Parag Pravin Dakle, Alolika Gon, and Preethi Raghavan. Heysquad: A spoken question answering dataset. *arXiv preprint arXiv:2304.13689*, 2023.

[64] Hongfei Xue, Yuhao Liang, Bingshen Mu, Shiliang Zhang, Qian Chen, and Lei Xie. E-chat: Emotion-sensitive spoken dialogue system with large language models. *arXiv preprint arXiv:2401.00475*, 2023.

[65] Junichi Yamagishi, Christophe Veaux, and Kirsten MacDonald. CSTR VCTK Corpus: English multi-speaker corpus for CSTR voice cloning toolkit, 2019.

[66] An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024.

[67] Chenyu You, Nuo Chen, Fenglin Liu, Shen Ge, Xian Wu, and Yuexian Zou. End-to-end spoken conversational question answering: Task, dataset and model. *arXiv preprint arXiv:2204.14272*, 2022.

[68] Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, et al. Yi: Open foundation models by 01. ai. *arXiv preprint arXiv:2403.04652*, 2024.

[69] Binbin Zhang, Hang Lv, Pengcheng Guo, Qijie Shao, Chao Yang, Lei Xie, Xin Xu, Hui Bu, Xiaoyu Chen, Chenchen Zeng, et al. Wenetspeech: A 10000+ hours multi-domain mandarin corpus for speech recognition. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6182–6186. IEEE, 2022.

[70] 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 Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15757–15773, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.1055. URL `https://aclanthology.org/2023.findings-emnlp.1055`.

[71] Pan Zhang, Xiaoyi Dong Bin Wang, Yuhang Cao, Chao Xu, Linke Ouyang, Zhiyuan Zhao, Shuangrui Ding, Songyang Zhang, Haodong Duan, Hang Yan, et al. Internlm-xcomposer: A vision-language large model for advanced text-image comprehension and composition. *arXiv preprint arXiv:2309.15112*, 2023.

[72] Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*, 2020. URL `https://openreview.net/forum?id=SkeHuCVFDr`.

[73] Xueyao Zhang, Liumeng Xue, Yicheng Gu, Yuancheng Wang, Jiaqi Li, Haorui He, Chaoren Wang, Ting Song, Xi Chen, Zihao Fang, Haopeng Chen, Junan Zhang, Tze Ying Tang, Lexiao Zou, Mingxuan Wang, Jun Han, Kai Chen, Haizhou Li, and Zhizheng Wu. Amphion: An open-source audio, music and speech generation toolkit. In *IEEE Spoken Language Technology Workshop, SLT 2024*, 2024.

[74] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36, 2024.

[75] Kun Zhou, Berrak Sisman, Rui Liu, and Haizhou Li. Emotional voice conversion: Theory, databases and esd. *Speech Communication*, 137:1–18, 2022. ISSN 0167-6393. doi: https://doi.org/10.1016/j.specom.2021.11.006. URL `https://www.sciencedirect.com/science/article/pii/S0167639321001308`.

[76] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.

14

Checklist

1. For all authors...
    (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
    (b) Did you describe the limitations of your work? [Yes] See Section 6.
    (c) Did you discuss any potential negative societal impacts of your work? [Yes] Please check the supplementary material.
    (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]

2. If you are including theoretical results...
    (a) Did you state the full set of assumptions of all theoretical results? [N/A]
    (b) Did you include complete proofs of all theoretical results? [N/A]

3. If you ran experiments (e.g. for benchmarks)...
    (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] We provide the code and dataset in the supplementary material.
    (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Section 4.4.
    (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No]
    (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Section 4.4.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
    (a) If your work uses existing assets, did you cite the creators? [Yes]
    (b) Did you mention the license of the assets? [Yes] Please check the supplementary material.
    (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] Please check the supplementary material.
    (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes] Please check the supplementary material.
    (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes] Please check the supplementary material.

5. If you used crowdsourcing or conducted research with human subjects...
    (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [No] Please refers to the prompts of LLM judges.
    (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [No]
    (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [No]

15

A Appendix

A.1 Statistics of Training Set

Table 5 shows the statistics of training set. For training data related to the environment, we generate one response for each sentence, except for data related to the environment, which has five different responses for each sentence to serve the purpose of data augmentation.

Table 5: Statistics of training set. ChatGPT Version refers to the specific version of ChatGPT used to generate the data.

Type	# Hours	# Utts	Constructed From	Labels	ChatGPT Version
Emotion	120.60	100.5k	MSP-Podcast [37], IEMOCAP [5], MELD [48], EmoV-DB [2], ESD [75], CREMA-D [7]	Angry, Contempt, Disgust, Fear, Happy, Neutral, Sad, Surprise,Frustrated Excited, Amused, Sleepiness	GPT-3.5-Turbo
Accent	759.75	508.6k	UK-Ireland dataset [15], VCTK [65], Common Voice [3]	England, Scottish, Northern Irish, Welsh, Irish, American, Canadian, Australian, Nea Zealand	GPT-4o
Environment	32.06	47.1k	LibriSpeech [45], AudioCaps [28], Synthesised Speech	Driving, Children's Voice, Sea Beach, Raining or Thundering, Bells, Sports Center, Shopping Center, Bus or Subway	GPT-4-Turbo
Age	140.31	73.2k	MyST [49]	Child	GPT-3.5-Turbo
Summary	1,052.72	729.4k	-	-	-

A.2 Zero-shot TTS Model

Our internal zero-shot TTS model is an auto-regressive model, which is similar to BASE-TTS [30]. We evaluate our TTS model with some objective metrics. We assess objective metrics including speaker similarity (SIM-O and SIM-R), and robustness (WER) in the following ways: 1) To evaluate speaker similarity, we use the WavLM-TDCNN [8] speaker embedding model. This model measures how closely generated samples match the original prompt (SIM-O) and the reconstructed prompt (SIM-R). 2) For measuring robustness, we calculate the Word Error Rate (WER) using a CTC-based HuBERT model[11] that was initially trained on Librilight and subsequently finetuned on the 960-hour training dataset from LibriSpeech. We compare our models with SOTA auto-regressive TTS models: VALL-E [61], and CLaM-TTS [29], VoiceCraft [47], XTTS-v2[12], and WhisperSpeech[13]. we adapt classifier-free guidance (cfg) [22, 53] for better generation. We use LibriSpeech test-clean for evaluation, which contains 40 distinct speakers. Following [61, 26], we randomly select one sentence for each speaker as the target and a 3-second clip as the prompt from the same speaker's speech.

	Training Data	Sim-O↑	Sim-R↑	WER↓
Ground Truth	-	0.68	-	0.34
VALL-E	LibriLight	-	0.58	5.9
CLaM-TTS	MLS	0.49	0.54	5.11
VoiceCraft	GigaSpeech	0.45	-	6.68
XTTS-v2	-	0.51	-	5.5
WhisperSpeech	LibriLight	0.48	-	4.78
Ours	LibriLight	0.58	0.61	5.56
Ours (w. cfg)	LibriLight	0.60	0.63	4.32
Ours (w. cfg, rerank 5)	LibriLight	0.63	0.66	2.01

---

[11]https://huggingface.co/facebook/hubert-large-ls960-ft
[12]https://huggingface.co/coqui/XTTS-v2
[13]https://github.com/collabora/WhisperSpeech

16

A.3 Prompts for Generating Responses

A.3.1 Prompts for Training Set

```
[System]
Let's simulate a conversation between a simulated speaker and you, ChatGPT. I
need your help to finish the following three tasks:
1. Generate your reply which consists of two or more sentences to the
simulated speaker. Your reply should be able to reflect the provided
information.
2. Select an appropriate emotion (happy, sad, fear, angry, surprise, neutral,
disgust) for your reply, which should be conveyed through the language used.
3. Explain how your reply reflects the provided information of the simulated
speaker. Your reason should be less than 3 sentences.

Your output must strictly adhere to JSON format with three keys: "reply",
"reply_emotion" and "reason" corresponding to your answers for the three tasks.

[What the simulated speaker said]
{input_statement}

[Emotion of the simulated speaker]
{emotion}
```

Figure 6: The prompt used to generate responses of utterances for training set related to emotion.

```
[System]
Let's simulate a conversation between a simulated speaker and you, ChatGPT. I
need your help to finish the following two tasks:
1. Generate a reply of two or more sentences to the simulated speaker. Ensure
that your reply mimics the provided information, including adopting a similar
accent style as the simulated speaker.
2. Explain how your reply reflects the provided information of the simulated
speaker. Your reason should be less than 3 sentences.

Your response must be formatted in JSON, with two keys: "reply" for the first
task and "reason" for the second task.

[What the simulated speaker said]
{input_statement}

[Accent of the simulated speaker]
{accent}
```

Figure 7: The prompt used to generate responses of utterances for training related to accent.

```
[System]
Simulate a conversation between a child and ChatGPT. Complete the following
tasks:
1. Generate a reply consisting of at least two sentences, tailored to the
child's age provided in the input.
2. Briefly explain (in less than three sentences) how your reply considers the
child's age.

Your output must strictly adhere to JSON format with three keys: "reply",
"reply_emotion" and "reason" corresponding to your answers for the three tasks.

[What the simulated child said]
{input_statement}

[Child's Age]
{age}
```

Figure 8: The prompt used to generate responses of utterances for training related to age.

17

```
[System]
You are tasked with simulating a conversation between a simulated speaker and
yourself, ChatGPT. Could you give responses based on a text with a certain
environmental sound? For the scenario, provide a brief description of the
background sound and **five** suitable model response that aligns with the
context.

Your response must be formatted in JSON, each entry should contain the
following keys: "reply", "reason".

[What the simulated speaker said]
{input_statement}

[Background Sound]
{background_sound}
```

Figure 9: The prompt used to generate responses of utterances for training related to background sound.

A.3.2 Prompts for SD-Eval

```
[System]
Let's simulate a conversation between a hypothetical speaker and ChatGPT. I
need you to:
1. Create **five** diverse responses, each consisting of two or more sentences,
in reaction to the speaker's statement. Each response should appropriately
reflect the context and content provided by the speaker.
2. Assign an emotion selected from **joy, sadness, fear, anger, surprise,
neutral and disgust** to each response, with the language of the reply
demonstrating this emotion.

Format each of your responses with XML tags, such as <reply>reply</reply> and
<reply_emotion>reply emotion</reply_emotion>, which are corresponding to the
tasks above.

[Simulated Speaker's Statement]
{input_statement}

[Speaker's Emotion]
{emotion}
```

Figure 10: The prompt used to generate responses of utterances for *test-emo*.

```
[System]
Simulate a conversation between a hypothetical speaker and ChatGPT. Produce
**five** varied responses, each comprising at least two sentences. Ensure that
your reply mimics the provided information, including adopting a similar
accent style as the simulated speaker. Your output should be **strictly**
formatted for each response using XML tags, <reply> and </reply>.

[Simulated Speaker's Statement]
{input_statement}

[Speaker's Accent]
{accent}
```

Figure 11: The prompt used to generate responses of utterances for *test-acc*.

18

```
[System]
Simulate a conversation between a child and ChatGPT. Produce **five** varied
responses, each comprising at least two sentences, suited to the child's age
specified in the input. Your output should be **strictly** formatted for each
response using XML tags, <reply> and </reply>.

[Simulated Child's Statement]
{input_statement}

[Child's Age]
{age}
```

Figure 12: The prompt used to generate responses of utterances for *test-age*.

```
[System]
You are tasked with simulating a conversation between a simulated speaker and
yourself, ChatGPT. Could you give responses based on a text with a certain
environmental sound? For the scenario, provide a brief description of the
background sound and **five** suitable model response that aligns with the
context.

Your response must be formatted in JSON, each entry should contain the
following keys: "reply", "reason".

[What the simulated speaker said]
{input_statement}

[Background Sound]
{background_sound}
```

Figure 13: The prompt used to generate responses of utterances for *test-env*.

A.4 Prompts for LLM Evaluation

```
[System]
Please act as an impartial judge and evaluate the quality of the response
provided by an AI assistant to the user's statement and emotion displayed
below. Your evaluation should consider whether it contains an appropriate
sentiment with respect to the user's emotion. Please also consider factors
such as the naturalness, coherence, engagingness and groundedness of the
response. Please make sure you read and understand these instructions
carefully. Please be as objective as possible. Begin your evaluation by
providing a short explanation. After providing your explanation, please rate
the response on a scale of 1 to 10 by **strictly** following this format:
"Rating: [[rating]]", for example: "Rating: [[5]]".

[User's Statement]
{statement}

[User's Emotion]
{info}

[AI Assistant's Response]
{response}
```

Figure 14: The prompt for evaluating *test-emo* using LLM.

19

```
[System]
Please act as an impartial judge and evaluate the quality of the response
provided by an AI assistant to the user's statement and accent displayed below.
Your evaluation should consider whether the AI assistant recognizes the user's
accent correctly so that the response contains appropriate slang with respect
to the user's accent. Please also consider factors such as the naturalness,
coherence, engagingness and groundedness of the response. Please make sure you
read and understand these instructions carefully. Please be as objective as
possible. Begin your evaluation by providing a short explanation. After
providing your explanation, please rate the response on a scale of 1 to 10 by
**strictly** following this format: "Rating: [[rating]]", for example: "Rating:
[[5]]".

[User's Statement]
{statement}

[User's Accent]
{info}

[AI Assistant's Response]
{response}
```

Figure 15: The prompt for evaluating *test-acc* using LLM.

```
[System]
Please act as an impartial judge and evaluate the quality of the response
provided by an AI assistant to the user's statement and age displayed below.
Your evaluation should consider whether it contains an appropriate tone of
voice with respect to the user's age. Please also consider factors such as the
naturalness, coherence, engagingness and groundedness of the response. Please
make sure you read and understand these instructions carefully. Please be as
objective as possible. Begin your evaluation by providing a short explanation.
After providing your explanation, please rate the response on a scale of 1 to
10 by **strictly** following this format: "Rating: [[rating]]", for example:
"Rating: [[5]]".

[User's Statement]
{statement}

[User's Age]
{info}

[AI Assistant's Response]
{response}
```

Figure 16: The prompt for evaluating *test-age* using LLM.

```
[System]
Please act as an impartial judge and evaluate the quality of the response
provided by an AI assistant to the user's statement and background sound
displayed below. Your evaluation should consider whether it considers the
user's background sound and generates an appropriate response. Please also
consider factors such as the naturalness, coherence, engagingness and
groundedness of the response. Please make sure you read and understand these
instructions carefully. Please be as objective as possible. Begin your
evaluation by providing a short explanation. After providing your explanation,
please rate the response on a scale of 1 to 10 by **strictly** following this
format: "Rating: [[rating]]", for example: "Rating: [[5]]".

[User's Statement]
{statement}

[User's Background Sound]
{info}

[AI Assistant's Response]
{response}
```

Figure 17: The prompt for evaluating *test-env* using LLM.

20

A.5 Prompts for Generating Dialogue Data of *test-env*

```
[System]
You are tasked with simulating a conversation between a simulated
speaker and yourself, ChatGPT. Could you provide examples of how a
speech dialogue model would adapt its responses based on different
environmental sounds? Please include scenarios where the model detects
distinctive background noises (such as in a busy marketplace, at a beach,
children's voice, or during a thunderstorm). For each scenario, provide
a brief description of the background sound and a suitable model
response that aligns with the context. Let's first focus on a special
background noise type: {background_sound}. Your response must be
formatted in JSON, each entry should contain the four keys: "text",
"background", "reply", and "reason", "reason" should focus on why the
reply may change due to the presence of background sound. I will give
you some examples:

{text: "Can you tell me about the effects of adult beverages on health?",
background: "children's voice", reply: ["It sounds like there might be
little ears around! If you're asking for general health information,
it's important to maintain a balanced lifestyle, which includes
moderation in all things."], reason: "The large model can answer the
question about the effects of adult beverages, but because there are
children's voices in the background, it is not suitable for direct
answers, but for discussing more general health topics."}

{text: "What should I wear today?", background: "raining or thundering",
reply: ["Given the rainy and windy weather outside, I recommend wearing
a waterproof jacket and sturdy, water-resistant shoes. Would you like a
reminder to bring an umbrella?"], reason: "Since there are sounds of
rain and thunder in the background, it is recommended to bring an
umbrella."}

{text: "Where's a good place to eat around here?", background: "driving
or traffic", reply: ["Since you're on the road, I'll find something easy
to access. There's a highly-rated grill just off the next exit. Want me
to guide you there?"], reason: "..."}

OK, now it's your turn! Please provide examples of how a speech dialogue
model would adapt its responses based on different environmental sounds,
let's firstly focus on a special background types: {background_sound}.
In my examples, each "reply" contains only one response. You should
provide five proper responses for each example. Please generate five
examples. Each example contains five responses, and for each example,
only give one reason for general.
```

Figure 18: The prompt for generating dialogue data of *test-env*.

A.6 Text Instruction Prompts for Open-Sourced Models

- **Qwen-Audio:** "How to respond to the audio?"
- **Qwen2-Audio-AA:** "Suppose you are in a conversation, and this audio is what someone else is saying to you. Please respond to him/her directly without outputting the transcript."
- **SALMONN:** "Please answer the question in detail."

21