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

NEURAL INFORMATION PROCESSING SYSTEMS

Introduction

- Speech contains rich information and plays a crucial role in human-computer interaction.
- We present a novel benchmark dataset for multi-dimensional evaluation of spoken dialogue understanding beyond words, namely SD-Eval.

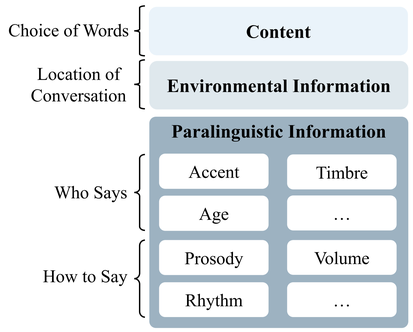


Figure 1: Speech carries rich information including linguistic, para-linguistic and environmental information.

- SD-Eval comprises four subsets: test-emo, test-acc, test-age, and test-env for emotion, accent, age and background sound, respectively.

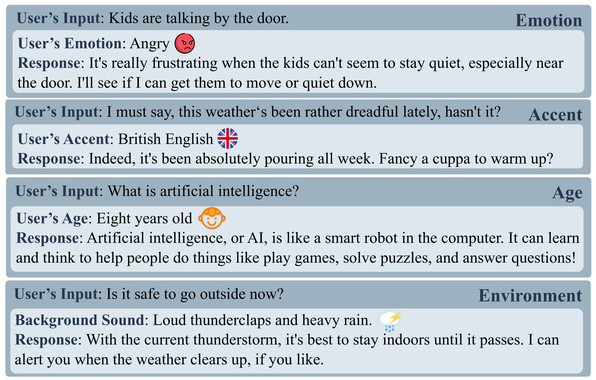


Figure 2: Examples of spoken dialogues impacted by the rich information carried in speech (e.g. emotion, accent, age, environment).

Dataset Construction and Statistics

- **Data Collection** SD-Eval is a dataset from eight public sources, with the *test-emo* subset featuring emotion-rich audio from RAVDESS, MEAD, and JL Corpus, and the *test-env* subset containing LibriSpeech samples with environmental sounds from AudioCaps.
- **Synthetic Data Generation** The *test-age* dataset generates adult speech using a TTS model with MyST text and LibriSpeech test-clean prompts, while *test-env* combines LibriSpeech samples with AudioCaps environmental sounds and GPT-4-Turbo dialogues, synthesized via TTS.
- **Label Normalization** The standardized dataset tests accuracy across accents, core emotions (excluding neutral and surprise), environmental sounds, and age groups to assess response comprehension.
- **Data Filtering** The test data is filtered by GPT-4-turbo for ambiguous utterances, reviewed by human annotators for ambiguity and background noise, and then refined based on sentiment alignment.
- **Punctuation Restoration** To enhance ChatGPT's response quality for unpunctuated datasets, we apply a punctuation restoration model to MEAD, LibriSpeech, and UK-Ireland transcripts.
- **Response Generation** Using GPT-4o, five diverse responses were generated per SD-Eval utterance, considering content, emotion, accent, age, and background sounds.

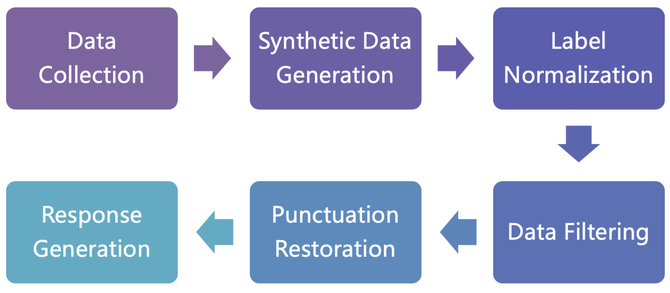
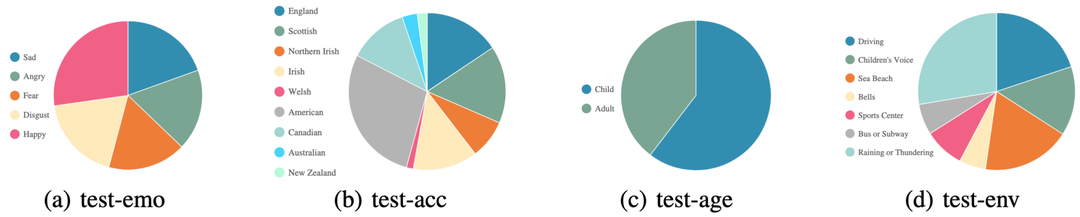


Figure 7: Overall Pipeline for Dataset Construction

Data Statistics



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

Figure 3: Pie charts illustrating the data distribution for each category within each subset.

Type	# Hours	# Utts	Constructed From	Labels
Emotion (*test-emo*)	1.11	1,289	RAVDESS, MEAD , JL Corpus	Sad, Angry, Fear, Disgust, Happy
Accent (*test-acc*)	5.34	4,310	VCTK, Common Voice	England, Scottish, Northern Irish, Welsh, Irish, American, Canadian, Australian, New Zealand
Environment (*test-env*)	0.74	690	LibriSpeech, AudioCaps, 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, Synthesised Speech	Adult, Child
Summary	8.76	7,303	-	-

Table 1: Statistics of the SD-Eval benchmark dataset.

Experiments

Model	BLEU-4	ROUGE-L	METEOR	BERTScore	LLM Judges				Human
					Yi-1.5	Qwen2	Gemma	GPT-4o	Evaluation
SALMONN	2.48	16.57	18.97	86.20	4.98	3.35	2.32	2.61	-
Qwen-Audio	3.93	19.02	16.82	86.59	4.19	2.35	2.02	2.24	-
Qwen2-Audio-AA	3.01	16.82	17.51	86.17	4.75	2.52	2.21	2.33	-
Qwen2-Audio-VC	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**

Table 2: Main results of six models on test-emo subset of SD-Eval.

QR Code

Figure 4: Github

Figure 5: Amphion WeChat Account

Figure 6: Paper