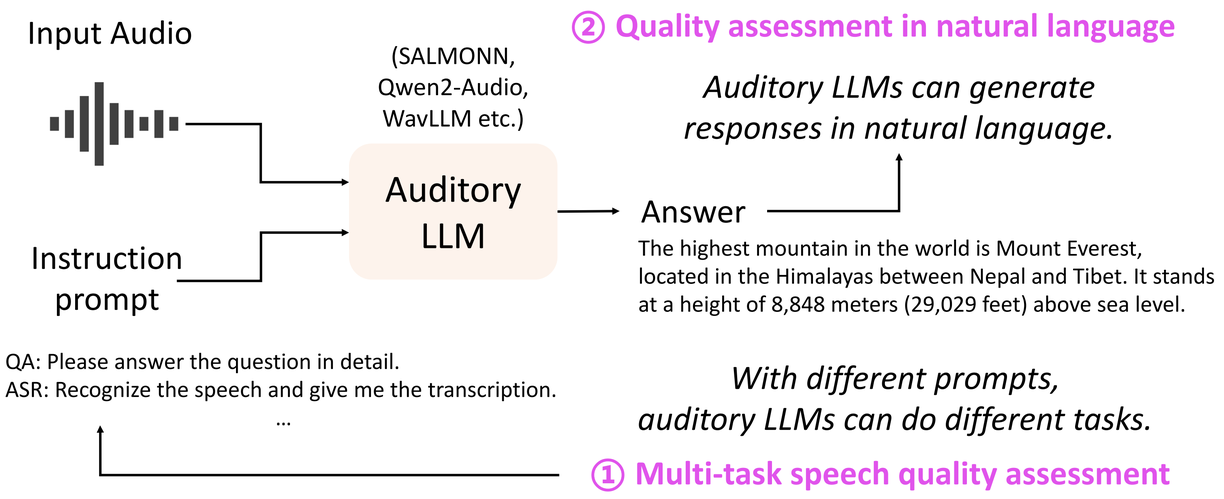
QualiSpeech: A Speech Quality Assessment Dataset with Natural Language Reasoning and Descriptions

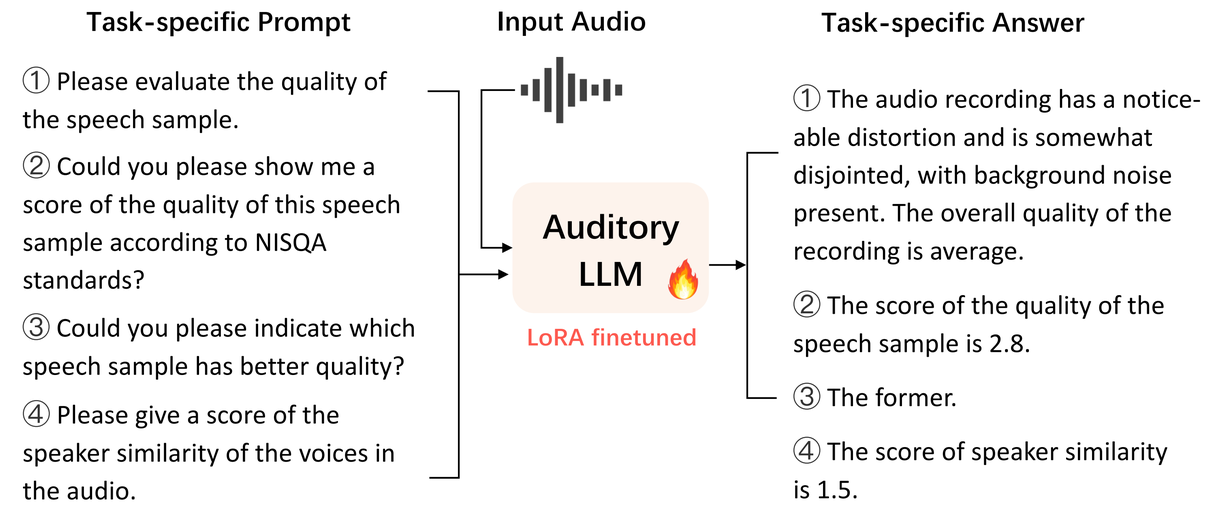
Siyin Wang[1], Wenyi Yu[1], Xianzhao Chen[2], Xiaohai Tian[2], Jun Zhang[2]
Lu Lu[2], Yu Tsao[3], Junichi Yamagishi[4], Yuxuan Wang[2], Chao Zhang[1]
[1]Tsinghua University, [2]ByteDance, [3]Acedemia Sinica [4]National Institute of Informatics
wangsiyi23@mails.tsinghua.edu.cn,cz277@tsinghua.edu.cn

ACL 2025 VIENNA JULY 27 - AUGUST 1

How can speech quality assessment benefit from auditory LLM?



Multi-task speech quality assessment !

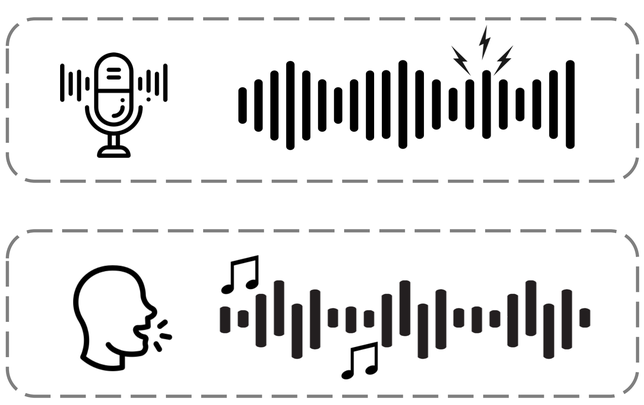


Better than SSL model !

Natural language speech quality assessment

Detailed Description
more nuanced descriptions than a simple numerical score
for example, we can identify type and time of distortion and noise using natural language

Reasoning for rating
explain why the audio sample is rated as such a score
for example, giving a reason alongside a score provides more detailed and instructive feedback



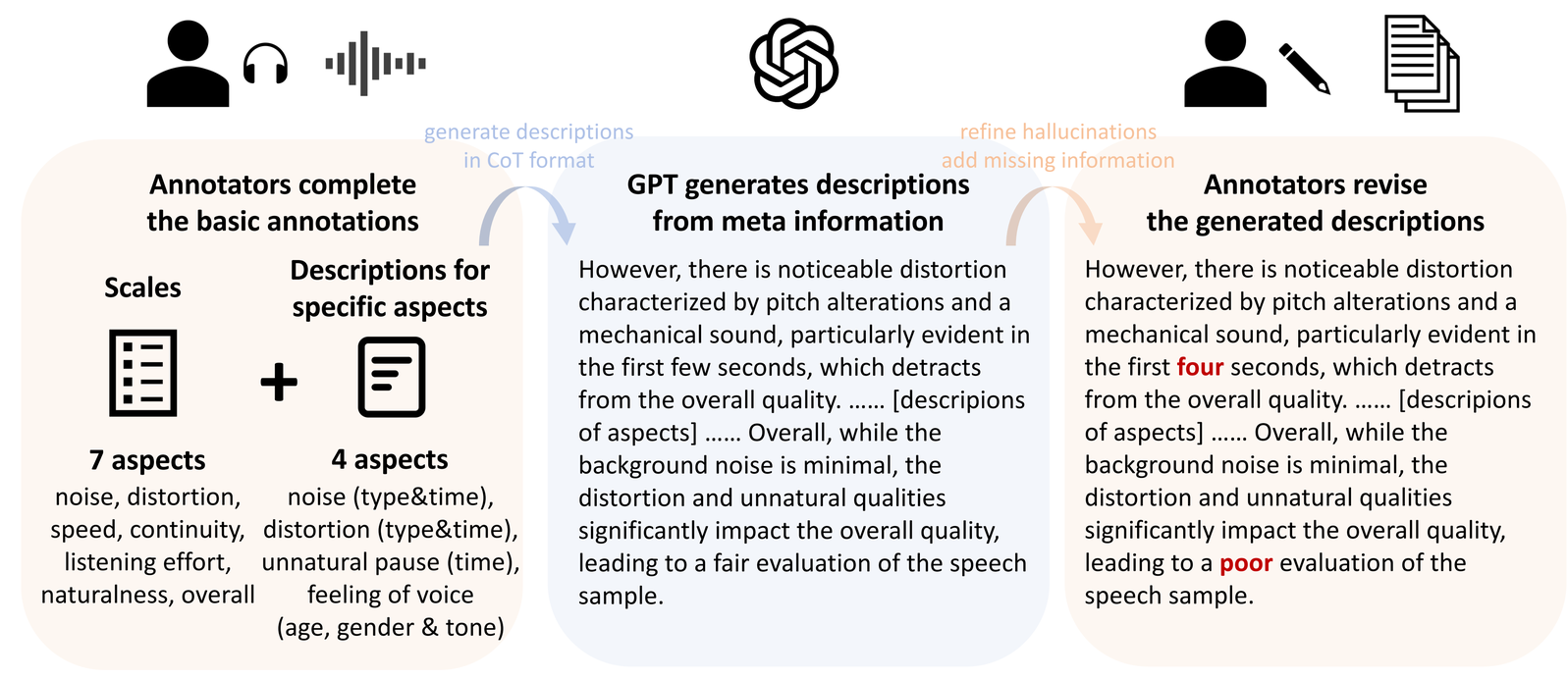
- Distortion score: 3
- Distortion **description**:
  There is a voice feels distorted with intermittent **electric current** quality from **1.5~2.5s.**
- Overall quality score: 2
- **Reasoning** for overall quality score:
  The overall quality is rated poorly due to the **intrusive background noise** and high listening effort, leading to a less favorable impression of the speech.

QualiSpeech dataset !

Annotation
3 steps, questionnaires to collect basic annotations, Leveraging GPT, then annotator double check

Evaluation
Split natural language assessment into aspects by GPT, then evaluating each aspects separately

Results
- Auditory LLMs can grasp low-level speech features.
- The model always identifies the correct time periods if the model recognizes noise or distortion
- The precision and recall metrics reveal that the model's ability to reliably detect the presence of noise or distortion still requires improvement.



Main Results

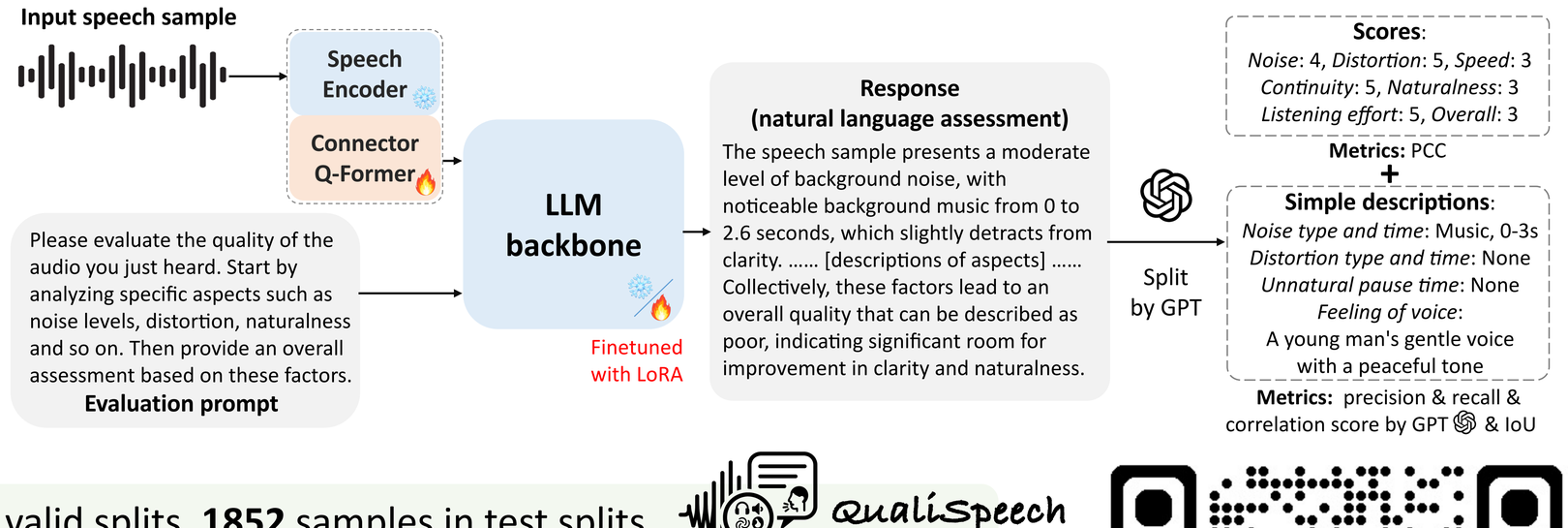
Predicting aspects in scores

Noise	Distortion	Speed	Continuity	Effort	Naturalness	Overall
0.703	0.571	0.178	0.450	0.513	0.535	0.622

Predicting aspects annotated in descriptions

Noise				Distortion			
Prec	Rec	Corr	IoU	Prec	Rec	Corr	IoU
0.62	0.54	0.53	0.73	0.76	0.84	0.66	0.77

Unnatural pause			Voice	
Prec	Rec	IoU	Corr	GenderAcc
0.55	0.60	0.34	0.48	0.98



10558 samples in train splits, **2167** samples in valid splits, **1852** samples in test splits

QualiSpeech

Covering a wide range of both **real** data and **synthetic** data, also data from recent high-performance open-source zero-shot speech generation and voice conversion models

Comprehensive annotations include 11 low-level speech aspects, from scores to simple descriptions

Also **QualiSpeech Benchmark** for evaluating low-level speech perception in auditory large language models (LLMs). **Welcome to download and try our dataset!**