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]Academia Sinica, [4]National Institute of Informatics
wangsiyi23@mails.tsinghua.edu.cn, cz277@tsinghua.edu.cn

Abstract

This paper explores a novel perspective to speech quality assessment by leveraging natural language descriptions, offering richer, more nuanced insights than traditional numerical scoring methods. Natural language feedback provides instructive recommendations and detailed evaluations, yet existing datasets lack the comprehensive annotations needed for this approach. To bridge this gap, we introduce **QualiSpeech**, a comprehensive low-level speech quality assessment dataset encompassing 11 key aspects and detailed natural language comments that include reasoning and contextual insights. Additionally, we propose the QualiSpeech Benchmark to evaluate the low-level speech understanding capabilities of auditory large language models (LLMs). Experimental results demonstrate that finetuned auditory LLMs can reliably generate detailed descriptions of noise and distortion, effectively identifying their types and temporal characteristics. The results further highlight the potential for incorporating reasoning to enhance the accuracy and reliability of quality assessments. The dataset can be found at https://huggingface.co/datasets/tsinghua-ee/QualiSpeech.

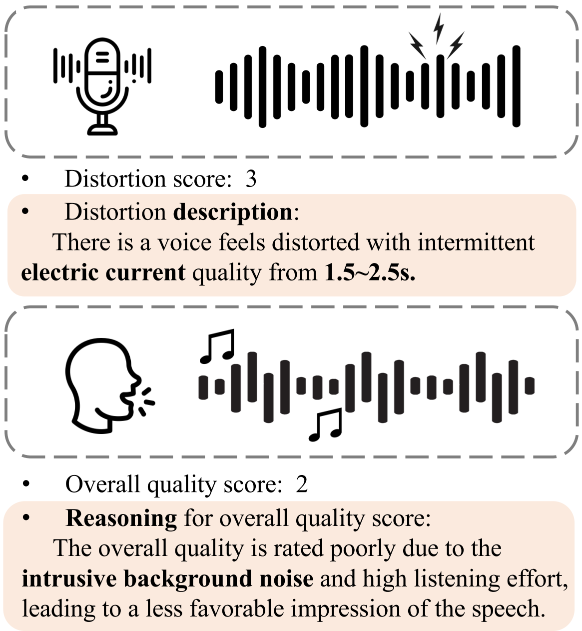

Figure 1: Examples from QualiSpeech. QualiSpeech is a comprehensive low-level speech quality assessment dataset that includes numerical scores for 7 aspects, specific descriptions for 4 aspects, and a detailed natural language comment on overall quality, incorporating reasoning and contextual insights. The examples highlight a simple distortion description alongside the reasoning section from a descriptive comment.

1 Introduction

Assessing speech quality is essential for evaluating the performance of speech synthesis systems and identifying distortions in communication networks (Cooper and Yamagishi, 2021; Mittag et al., 2021). The gold standard for speech quality evaluation remains human assessment, typically using the mean opinion score (MOS) — an average rating derived from listening tests that gauge overall audio quality (Salas et al., 2013). However, human evaluations are both time-intensive and laborious, prompting the development of automated methods for speech quality assessment. These methods, often powered by deep learning, enable rapid and scalable evaluation across large datasets (Patton et al., 2016; Lo et al., 2019). Most current approaches focus on MOS prediction, generating a numerical score that represents the perceived quality of the speech (Cooper et al., 2022; Saeki et al., 2022). While these scores allow for straightforward comparisons between samples, they do not reveal the reasoning behind a particular score, leaving the underlying quality factors unexplained.

Evaluating speech quality using natural language offers a novel and intuitive approach, enabling more nuanced and detailed feedback compared to

*Corresponding author.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 23588–23609

traditional numerical scales. Descriptive evaluations analyze multiple low-level speech features and synthesize them into an overall assessment, offering instructive insights for applications such as improving speech synthesis systems. For instance, a description like "*There is a voice that feels distorted with intermittent electric current quality from 1.5∼2.5s*" provides richer context than a standalone distortion score, as shown in Figure 1. To address the absence of datasets tailored for natural language-based speech quality assessment, we present **QualiSpeech**, the first dataset designed to capture diverse low-level speech characteristics through detailed descriptive comments.

Recent advancements in auditory LLMs have made natural language-based speech quality evaluation increasingly feasible (Tang et al., 2024a; Chu et al., 2023; Rubenstein et al., 2023). By integrating speech encoders with powerful LLM backbones, these models excel in high-level spoken language understanding tasks. However, low-level speech perception tasks remain largely underexplored in both the training and evaluation of current auditory LLMs (Huang et al., 2024; Yang et al., 2024). To address this limitation, we propose the QualiSpeech benchmark on multi-choice speech quality assessment tasks, revealing that existing auditory LLMs struggle to assess speech quality accurately. Leveraging the QualiSpeech dataset, we enhance the SALMONN-7B (Tang et al., 2024a) model with the ability to provide natural language descriptions of speech quality across multiple dimensions. This model demonstrates the ability to produce detailed and precise descriptions of noise and distortion, underscoring the advantages of using natural language for nuanced speech quality evaluation. We also demonstrate the feasibility of reasoning for speech quality assessment using text LLMs.

Our contributions can be summarized as follows:

- We introduce QualiSpeech, a database for speech quality assessment using natural language descriptions. QualiSpeech evaluates speech of both humans and various text-to-speech (TTS) synthesis systems across a comprehensive range of aspects, encompassing diverse artificial distortions and real-world scenarios. To the best of our knowledge, QualiSpeech is the first dataset designed using low-level speech perception annotated with detailed natural language descriptions for quality assessment of both synthetic and real speech.
- We also develop an auditory LLM that can assess speech quality across multiple aspects by generating detailed descriptive comments using the QualiSpeech dataset.
- We propose QualiSpeech benchmark for low-level speech understanding ability evaluation of auditory LLMs.

2 Related Work

2.1 Speech quality assessment dataset

To address the need for automatic speech quality assessment models for evaluating speech synthesis systems, the BVCC (Cooper and Yamagishi, 2021) dataset is proposed by collecting speech samples from past Blizzard Challenges (King and Karaiskos, 2016) and VCC Challenges (Yi et al., 2020) with a standardized new MOS score. More speech assessment datasets for evaluating speech synthesis models are introduced to extend application scenarios (Cooper et al., 2023) and mitigate the influence of speaker (Maniati et al., 2022). From a different perspective, the NISQA dataset (Mittag et al., 2021) focuses on real-world speech recordings and simulated distortions commonly found in communication networks.

Before the introduction of QualiSpeech, prior research works have typically treated the evaluation of synthetic and real speech as separate tasks due to their distinct characteristics. Synthetic speech is typically free of noise but often lacks naturalness, while real speech is more affected by noise than by issues of naturalness. By providing detailed aspect scores and reasoning in descriptive comments, QualiSpeech aims to facilitate the development of general speech quality assessment models capable of effectively distinguishing between these different types of speech.

Previous methods for automatic speech quality assessment have predominantly focused on MOS prediction. Deep learning-based approaches can achieve strong correlations with MOS derived from human evaluations (Patton et al., 2016; Lo et al., 2019; Cooper et al., 2022). Expanding beyond overall scores, NISQA (Mittag et al., 2021) introduced score-based prediction of specific speech quality dimensions, including noise, colouration, discontinuity, and loudness. Building on this foundation, QualiSpeech seeks to push the field further by enabling the development of more advanced speech quality assessment models. These models are designed to analyze a wider range of low-level

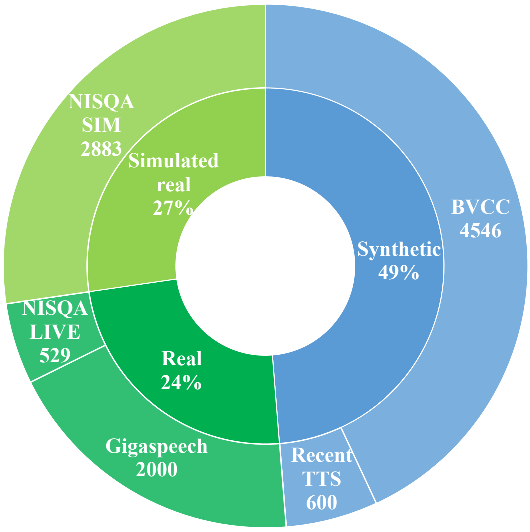

Figure 2: The dataset source of train split of QualiSpeech dataset. It has a balanced distribution of synthetic data and Real data (including simulated real data). The total size of the train split is 10,558.

speech features while providing detailed reasoning in natural language.

2.2 Auditory LLM for speech perception

Auditory LLMs (Wang et al., 2023; Tang et al., 2024a; Chu et al., 2023; Hu et al., 2024; Lu et al., 2024; Jin et al., 2024) have shown remarkable performance across a broad spectrum of high-level speech perception tasks, including speech recognition (Fathullah et al., 2024; Yu et al., 2024), translation (Wu et al., 2023; Chen et al., 2024b), understanding (Shon et al., 2024), and speaker and emotion recognition (Xu et al., 2024). By harnessing the power of LLMs to analyze speech data and generate natural language responses, auditory LLMs hold significant potential to unify diverse speech understanding tasks within a single framework (Tang et al., 2024a; Chu et al., 2024). To comprehensively evaluate these capabilities, benchmarks such as Dynamic-Superb (Huang et al., 2024), AIR-Bench (Yang et al., 2024), and AudioBench (Wang et al., 2024a) have been introduced, offering valuable insights into the strengths and limitations of auditory LLMs. However, low-level speech quality assessment tasks remain largely neglected. To fill this gap, we propose a multi-choice benchmark specifically tailored for low-level speech perception tasks, utilizing scores from seven key dimensions in QualiSpeech.

Recent efforts have expanded the range of tasks auditory LLMs can tackle, including spatial audio processing (Zheng et al., 2024; Tang et al., 2024b) and audio entailment (Deshmukh et al., 2024). The potential of using auditory LLMs for speech quality assessment has also been explored in (Wang et al., 2024b; Zezario et al., 2024; Chen et al., 2025). However, these efforts remain largely focused on MOS prediction, which fails to fully leverage the unique capability of auditory LLMs to generate rich, natural language responses. By utilizing the diverse annotations in QualiSpeech, we can develop an auditory LLM for speech quality assessment that offers a more holistic and detailed evaluation, providing nuanced descriptions of speech quality rather than relying solely on numerical scores.

3 QualiSpeech Dataset and Benchmark

3.1 Dataset

We introduce QualiSpeech, an English-language speech quality assessment dataset designed to encompass diverse scenarios and aspects. It provides the most comprehensive set of low-level speech feature annotations available to date, including numerical scores across seven dimensions and concise descriptions for four aspects. Significantly, QualiSpeech pioneers the use of natural language descriptions, offering detailed and logically structured assessments that go beyond traditional numerical scoring methods.

3.1.1 Data collection

We constructed the QualiSpeech dataset using a diverse range of sources. Figure 2 illustrates the data sources for the training split, with detailed statistics provided in the Appendix E. For synthetic speech, we utilized the BVCC dataset, which includes samples of varying quality generated by diverse systems from past Blizzard and VCC Challenges. To maintain consistency, samples shorter than two seconds were excluded. However, as BVCC was collected in 2021, it lacks data from more recent TTS models. To address this gap, we incorporated synthetic speech generated by 10 recent open-source TTS models, including ChatTTS[1], XTTS v2[2], CosyVoice[3] (Du et al., 2024) F5-TTS[4] (Chen et al., 2024a), E2 TTS (implemented by F5-TTS[4]) (Eskimez et al., 2024), OpenVoice V1[5], OpenVoice

[1]https://github.com/2noise/ChatTTS
[2]https://github.com/coqui-ai/TTS
[3]https://github.com/FunAudioLLM/CosyVoice
[4]https://github.com/SWivid/F5-TTS
[5]https://github.com/myshell-ai/OpenVoice

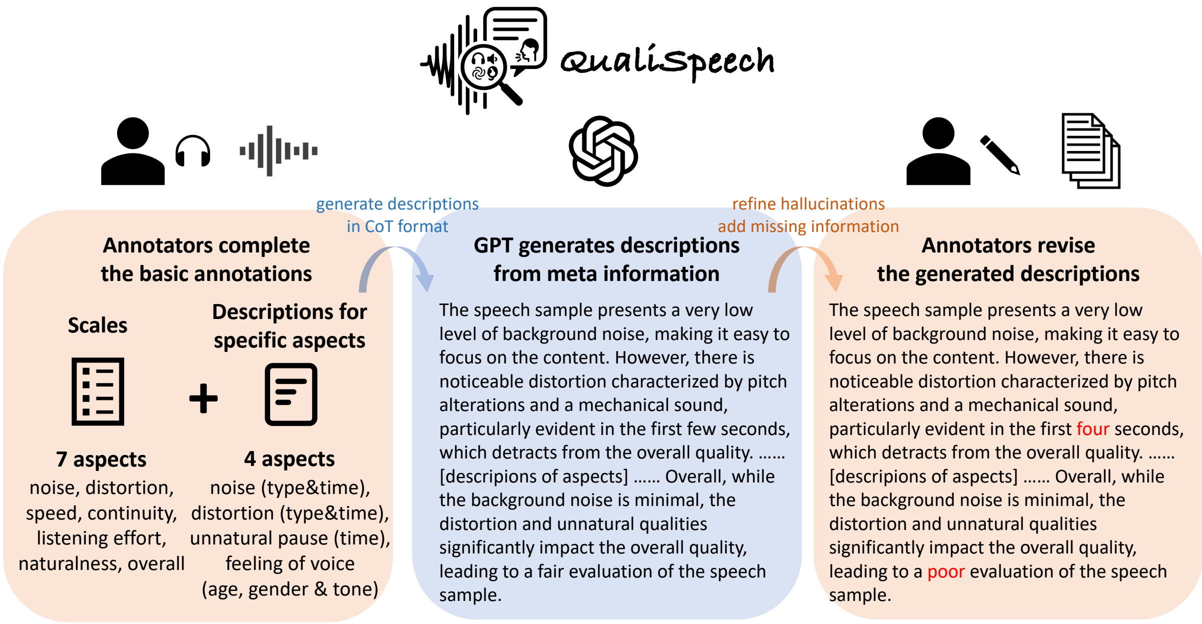

Figure 3: The annotation process of QualiSpeech dataset. In step 1, Listeners annotate basic low-level speech perception characteristics including 7 scores and 4 specific descriptions. In step 2, GPT generates natural language descriptions from annotated meta information. In step 3, annotators check and revise the generated descriptions.

V2[5] (Qin et al., 2023), Parler-TTS Mini[6], Parler-TTS Large[6] and VoiceCraft-830M[7] (Peng et al., 2024). Sentences for synthesis were sourced from the SOMOS sentence corpus (Maniati et al., 2022), which spans 10 domains, including conversational dialogue, news, and Wikipedia entries. Each TTS model generated 72 samples, distributed as 60 for training, 6 for validation, and 6 for testing. For zero-shot TTS models (except Parler-TTS), prompt audio determining the speaker of the synthesized speech was sampled from LibriHeavy (Kang et al., 2024). For Instruct-TTS Parler-TTS, text descriptions specifying speakers were either sourced from available built-in speakers or generated by GPT. The full list of text descriptions is provided in the Appendix G.

Speech synthesis models are typically trained exclusively on clean speech, making them unlikely to produce samples that are both unnatural and noisy. To address this gap, 20% of the synthetic data is mixed with noise. The noise files are sourced from the DNS Challenge dataset (Dubey et al., 2023), which draws its content from AudioSet (Gemmeke et al., 2017), Freesound (Fonseca et al., 2017), and DEMAND (Thiemann et al., 2013). The signal-to-noise ratio is uniformly sampled within the range of 0 to 15.

[6] https://github.com/huggingface/parler-tts
[7] https://github.com/jasonppy/VoiceCraft

For real speech, NISQA (Mittag et al., 2021) is exploited due to it has speech samples in rich simulated and live conditions. The simulated distortions are designed to replicate real-world transmission channels. Specifically, the simulated datasets (NISQA_TRAIN_SIM and NISQA_VAL_SIM) include 1,258 unique distortions generated by combining 9 basic distortion types. From these, one out of every five distortions is selected, resulting in 2,883 samples for the training split and 826 for the validation split. For live communication recordings, we include Skype recordings from NISQA_TRAIN_LIVE and NISQA_VAL_LIVE. NISQA_TEST_FOR and NISQA_TEST_P501, the two test sets of NISQA featuring both real and simulated conditions, are incorporated into the test split. We also source real speech from GigaSpeech (Chen et al., 2021), which includes audiobooks, podcasts, and YouTube recordings. Using UTMOS (Saeki et al., 2022), MOS scores are predicted to classify GigaSpeech S samples into four quality groups. To ensure balanced representation across varying speech qualities, an equal number of samples is selected from each group.

3.1.2 Annotation process

The annotation process, illustrated in Figure 3, comprises three main steps. The first step involves collecting detailed low-level speech features through

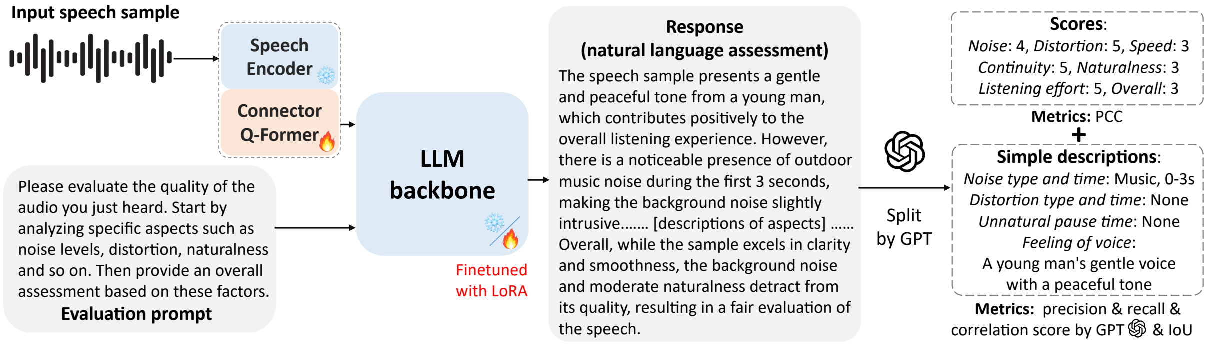

Figure 4: Finetuning auditory LLMs on QualiSpeech and the evaluation procedure of natural language speech assessment. The generated assessment will be split into scores and specific descriptions using GPT first, then PCC is calculated for scores and the correlation score generated by GPT will be used to evaluate the correctness of specific descriptions.

listening tests, which include numerical scores across seven aspects and concise descriptions for four specific aspects. Listeners are tasked with rating various low-level features of each speech sample on a five-point scale, where higher scores indicate better quality. The annotated aspects include noise, distortion, speed, continuity, listening effort, naturalness, and overall quality. Our scoring scales are designed following the guidelines in (Salas et al., 2013), with detailed scales and instructions provided in Figure 6. Additionally, the listening tests capture specific descriptions of noise (type and occurrence time), distortion (type and occurrence time), unnatural pauses (occurrence time), and vocal characteristics (perceived age, gender, and tone). These annotations aim to facilitate the generation of more detailed and context-rich natural language descriptions.

The second step is to generate natural language descriptions using GPT[8] from the annotated meta information. All annotated aspects are provided to GPT to produce descriptions in a chain-of-thought (CoT) format. This approach ensures the descriptions first analyze low-level speech features before concluding with an overall quality assessment. The third step focuses on refining the generated descriptions. Since GPT can produce hallucinations or inaccuracies, this step is essential for ensuring high-quality natural language descriptions (You et al., 2025). Annotators review and correct errors, such as inconsistencies between the descriptions and annotations or unsupported claims. They are also tasked with adding any missing aspects to ensure the descriptions fully capture all annotated information. Furthermore, annotators refine the reasoning behind the overall quality assessment, improving the logical coherence and clarity of the descriptions as necessary.

[8]The version is gpt-4o-mini-2024-07-18.

3.2 Benchmark

We also establish the QualiSpeech benchmark to assess the low-level speech perception capabilities of auditory LLMs. This multi-choice benchmark spans seven aspects of low-level speech understanding: noise, distortion, speed, continuity, listening effort, naturalness, and overall quality. It is built on the numerical scoring scales provided in the QualiSpeech dataset. Auditory LLMs are tasked with selecting the most appropriate score for a given speech sample on a specific low-level aspect, with the relevant scale provided as guidance. Open-ended question-answering is excluded from the benchmark, as current LLMs often struggle to reliably follow instructions they have not encountered before.

3.3 Evaluation metrics

For QualiSpeech benchmark evaluation, PCC (Pearson Correlation Coefficient) is used to evaluate the accuracy of predicted scores.

For QualiSpeech dataset evaluation, PCC is also employed to assess aspects represented by numerical values. For aspects in descriptions, we utilize 4 metrics to cover different evaluation dimensions. First, precision and recall are reported to assess the model's ability to accurately determine the presence of noise (or distortion, unnatural pause). From a complementary point of understanding, correlation scores generated by GPT and intersection over

Model	Aspect of low-level speech perception						
	Noise	Distortion	Speed	Continuity	Effort	Naturalness	Overall
SALMONN-7B	0.003	0.013	0.001	nan	nan	0.030	0.084
SALMONN-13B	0.001	0.002	0.025	-0.001	-0.069	0.013	0.100
Qwen-Audio-Chat	**0.014**	-0.003	0.017	0.145	**0.150**	**0.148**	**0.250**
Qwen2-Audio-7B-Instruct	-0.048	**0.056**	**0.111**	**0.201**	0.035	-0.082	0.112
WavLLM	-0.021	-0.069	0.003	-0.001	-0.007	0.005	0.071

Table 1: Open-source auditory LLMs on QualiSpeech benchmark. PCC is reported. Some correlation scores are nan because all predicted scores are the same for that aspect.

union (IoU) scores are presented. The correlation score gauges the overall relevance of the model's descriptions, while the IoU score measures the accuracy of the predicted time intervals. Note that we calculate these two scores only when the model successfully identify the noise (or distortion, natural part). This ensures that these scores will not be influenced by recall since undetected samples will receive a correlation or IoU score of 0. Precision and recall reflect holistic comprehension, while correlation and IoU focus on the ability to capture nuanced information.

Regarding to evaluation of natural language descriptions, low-level speech perception dimensions, either in numerical scores or specific descriptions, are first extracted from the natural language assessments using GPT. Subsequently, the corresponding evaluation metrics are applied to assess each dimension, as shown in Figure 4.

4 Experimental Results

4.1 Open-source auditory LLMs on QualiSpeech Benchmark

The performance of open-source auditory LLMs on the QualiSpeech benchmark is presented in Table 1. Five auditory LLMs, including SALMONN-7B, SALMONN-13B (Tang et al., 2024a), Qwen-Audio-Chat (Chu et al., 2023), Qwen2-Audio-7B-Instruct (Chu et al., 2024) and WavLLM (Hu et al., 2024), are evaluated. The results show that current open-source auditory LLMs struggle to effectively evaluate speech quality. SALMONN-7B, in particular, exhibits a strong numerical bias, predicting all listening effort and continuity scores as 4, and all noise scores as 3. This tendency towards number preference is also observed, albeit to a lesser extent, in other models. For instance, Qwen-Audio-Chat predicts 80% of overall quality scores as 3, while Qwen2-Audio-7B-Instruct assigns 66% of its scores as 3, and WavLLM predicts 86% of its scores as 4.

4.2 Building a speech quality assessment model using QualiSpeech

We utilize QualiSpeech to build a speech quality assessment auditory LLM, finetuned on SALMONN-7B (Tang et al., 2024a), with Whisper (Radford et al., 2023) and BEATs (Chen et al., 2023) serving as encoder and Vicuna (Chiang et al., 2023) as LLM backbone. We follow the default training configuration in which only the connector speech Q-former and LoRA on LLM are finetuned, as illustrated in Figure 4. First, we fine-tune SALMONN-7B on multiple-choice questions designed to select scores and provide specific descriptions focusing on individual aspects of speech, to assess whether auditory LLMs can effectively understand low-level speech features. In the subsequent phase, we enable the auditory LLMs to generate detailed and logical natural language assessments of speech quality.

4.2.1 Learning low-level speech features

Results of learning low-level speech features are shown in Table 2. For training strategy, "basic" refers to finetuning on a specific task alone, and "balance" indicates finetuning on the specific task with balanced distributed data. In this case, the proportions of each score or category (*e.g.*, noise presentation for specific descriptions of noise, and gender for voice) are kept consistent across the dataset. The term "joint" denotes joint training across all 11 low-level speech understanding tasks. All models are trained for 10 epochs.

The results show that finetuned auditory LLMs can perform better than the vanilla SALMONN-7B, suggesting that auditory LLMs can, to some extent, grasp low-level speech features. Speed is the aspect with the highest classification accuracy, while noise level classification achieves an accuracy exceeding 60%, indicating the model's ability to dis-

Training Strategy	Aspect of low-level speech perception						
	Noise	Distortion	Speed	Continuity	Effort	Naturalness	Overall
basic	**0.721**	0.553	**0.335**	0.478	0.525	0.541	0.597
balance	0.696	0.547	0.268	0.458	0.497	0.540	0.600
joint	0.693	0.595	0.240	0.525	**0.578**	0.565	0.636
joint + balance	0.696	**0.614**	0.322	**0.535**	**0.578**	**0.615**	**0.660**

(a) aspects annotated in scores

Training Strategy	Aspect of low-level speech perception												
	Noise				Distortion				Unnatural pause			Voice	
	Prec	Rec	Corr	IoU	Prec	Rec	Corr	IoU	Prec	Rec	IoU	Corr	GenderAcc
basic	**0.70**	0.50	**0.60**	**0.85**	0.64	**0.97**	**0.71**	0.78	0.55	0.78	0.39	**0.50**	0.98
balance	0.66	0.53	0.57	0.82	**0.77**	0.86	**0.71**	0.78	**0.57**	0.78	0.37	0.49	0.98
joint	0.45	0.79	0.57	0.76	0.68	0.96	0.70	0.78	0.52	**0.83**	0.41	**0.50**	0.98
joint + balance	0.40	**0.83**	0.55	0.75	0.64	**0.97**	0.70	**0.79**	0.51	0.82	**0.42**	0.49	0.98

(b) aspects annotated in descriptions

Table 2: Results of learning low-level speech features. PCC is reported for aspects annotated in scores. Correlation scores generated by GPT and IoU of the predicted time period and ground truth are reported for aspects annotated in descriptions. "Effort" denotes listening effort and "GenderAcc" denotes the accuracy of the classification of the biological gender of the speakers.

tinguish non-semantic components from semantic ones. Naturalness and overall quality, however, remain the most challenging to predict, likely due to their reliance on more subjective judgments. When comparing the "basic" and "balanced" training settings, the different data distributions show minimal impact on performance. Joint training improves the classification of certain aspects, such as continuity, without causing significant degradation, suggesting that learning multiple speech features simultaneously does not introduce conflicts. While finetuned auditory LLMs perform reasonably well at predicting speed and noise level, there is still considerable room for improvement in understanding low-level speech features.

Results for specific descriptions, which are unique to natural language assessment, are more promising. An IoU score of approximately 0.8 is achieved when describing noise or distortion, indicating that the model always identifies the correct time periods if the model recognizes noise or distortion. This underscores the potential of using auditory LLMs to generate nuanced and detailed descriptions of speech quality. The correlation scores are also satisfactory, demonstrating that the model can accurately describe the type of noise or distortion in most cases. However, the precision and recall metrics reveal that the model's ability to reliably detect the presence of noise or distortion still requires improvement. A simple balancing of data distribution, in this case, does not yield significant benefits. When it comes to identifying unnatural pauses, the results are less favourable compared to noise and distortion, suggesting rhythm is hard to learn. For voice description, while the model performs excellently in gender classification, it struggles with age and tone, resulting in a moderate correlation score of around 0.5.

We further assess the generalization ability of auditory LLMs across different data types, with the results presented in Table 8 in Appendix J.2. In this analysis, we finetune the models on one specific data type and evaluate their performance on all data types. The results show that if a model is trained only on one data type, it will exhibit a poor generalization to other domains. The findings also show that incorporating more diverse data sources leads to improved performance across all domains. The results suggest that building a robust and general speech quality assessment model for both synthetic and real data is feasible, highlighting the importance of datasets like QualiSpeech, which encompass a wide array of data sources.

4.2.2 Learning natural language descriptions

We also investigate whether auditory LLMs can benefit from reasoning in natural language. So we further finetune the checkpoint jointly trained on

Comments	Aspect of low-level speech perception						
	Noise	Distortion	Speed	Continuity	Effort	Naturalness	Overall
revised concise	0.656	**0.579**	0.212	0.452	0.496	**0.568**	**0.630**
concise with num	**0.703**	0.571	0.178	0.450	**0.513**	0.535	0.622
concise	0.642	0.559	**0.263**	**0.483**	0.511	0.520	0.582
detailed	0.686	0.518	0.250	0.459	0.475	0.486	0.572

(a) aspects annotated in scores

Comments	Aspect of low-level speech perception												
	Noise				Distortion				Unnatural pause			Voice	
	Prec	Rec	Corr	IoU	Prec	Rec	Corr	IoU	Prec	Rec	IoU	Corr	GenderAcc
revised concise	0.40	0.50	0.49	0.49	**0.80**	0.74	0.67	**0.78**	0.54	0.52	**0.34**	0.48	0.97
concise with num	**0.62**	**0.54**	**0.53**	**0.73**	0.76	**0.84**	0.66	0.77	0.55	**0.60**	**0.34**	0.48	**0.98**
concise	0.34	0.52	0.49	0.41	0.74	0.81	0.65	0.73	0.58	0.57	0.30	0.49	0.96
detailed	0.50	0.27	0.46	0.61	0.76	0.74	**0.68**	0.73	**0.60**	0.56	0.33	**0.51**	**0.98**

(b) aspects annotated in descriptions

Table 3: Results of learning natural language descriptions. PCC is reported for aspects annotated in scores. Correlation scores generated by GPT and IoU of predicted time period and ground truth are reported for aspects annotated in descriptions. “Effort” denotes listening effort and “GenderAcc” denotes the accuracy of the classification of the biological gender of the speakers.

Model	Vicuna-v1.5-7B	GPT-4o-mini
Acc	0.28	0.46

Table 4: Results on the reasoning for the overall quality score of text LLMs based on basic groundtruth low-level speech features.

all basic low-level understanding tasks to generate descriptive comments, This process involves analyzing each aspect individually before synthesizing all the dimensions to derive an overall score. All models were trained for 10 epochs. All models are trained for 10 epochs. Different formats are also explored, with detailed settings and examples provided in the Appendix F.

Results shown in Table 3 show that auditory LLMs can generate a paragraph of natural language speech quality assessment, achieving accuracy comparable to evaluating each aspect separately. The length of the generated text has minimal impact on performance. Including numerical scores within the natural language description enhances the quality of the output, likely due to the added specificity of the information provided. Revising the generated comments to eliminate any hallucinations is crucial for producing high-quality, reliable speech quality assessments.

However, the model does not achieve higher accuracy in overall score prediction through reasoning in natural language. Incorrectly predicted low-level speech aspects also interference with reasoning. To address this, we experiment with text-based LLM reasoning for predicting the overall quality score, using the groundtruth low-level speech features. The results, shown in Table 4, reveal that the LLM backbone of SALMONN-7B, Vicuna-v1.5-7B, lags behind all fine-tuned auditory LLMs, suggesting that the failure of reasoning is partly due to the LLM backbone's weak reasoning capabilities. In contrast, GPT-4o-mini outperforms all fine-tuned auditory LLMs, highlighting the possibility of reasoning when assessing speech quality.

5 Conclusion

In this paper, we present QualiSpeech, a comprehensive speech quality dataset curated from diverse sources, encompassing a wide range of aspects and incorporating natural language descriptions. We also introduce the QualiSpeech Benchmark, designed to evaluate the low-level speech understanding capabilities of auditory LLMs. Benchmark results reveal that current open-source auditory LLMs face challenges in accurately assessing speech quality. Our experiments show that natural language descriptions provide more detailed insights into noise and distortion compared to traditional methods. Generalization experiments highlight the importance of incorporating data from diverse sources to develop a robust, general-purpose speech quality assessment model suitable for all scenarios. We hope QualiSpeech will inspire further research into natural language-based speech quality assessment, enabling more fine-grained and reliable evaluations.

Acknowledgments

This work was supported by a research grant from ByteDance in 2024-2025, and Natural Science Foundation of China (62476151).

Limitations

There are some limitations of our work. First, although QualiSpeech encompasses multiple aspects and diverse sources, it inevitably leaves some aspects and sources uncovered. Second, each speech sample in QualiSpeech is annotated by only one listener due to the complexity of the annotation process, whereas MOS scoring typically involves evaluations from multiple listeners per sample. Despite this limitation, we believe our current dataset provides a valuable foundation for the community to explore and develop initial approaches. Lastly, our fine-tuned auditory LLMs do not yet fully leverage reasoning in natural language, primarily due to the limitations of the underlying LLM backbone. We hope future stronger auditory LLMs can utilize and benefit from reasoning in natural language descriptions of QualiSpeech.

Ethics Statement

All the models and datasets in this paperare publicly accessible and used under licenses. As for the generated speech samples from recent TTS systems in QualiSpeech, all open-source TTS models are used under corresponding licenses. Our dataset is provided under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License. All annotators received clear annotation rules before the annotation and fair payments upon completion. We believe the auditory LLM evaluator can only be used as a reference and should not replace human evaluators.

References

Chen Chen, Yuchen Hu, Siyin Wang, Helin Wang, Zhehuai Chen, Chao Zhang, Chao-Han Huck Yang, and Eng Siong Chng. 2025. Audio large language models can be descriptive speech quality evaluators. In *Proc. ICLR*, Singapore.

Guoguo Chen, Shuzhou Chai, Guanbo Wang, Jiayu Du, Wei-Qiang Zhang, Chao Weng, Dan Su, Daniel Povey, Jan Trmal, Junbo Zhang, et al. 2021. GigaSpeech: An evolving, multi-domain ASR corpus with 10,000 hours of transcribed audio. In *Proc. Interspeech*, Florence.

Sanyuan Chen, Yu Wu, Chengyi Wang, Shujie Liu, Daniel Tompkins, Zhuo Chen, and Furu Wei. 2023. BEATs: Audio pre-training with acoustic tokenizers. In *Proc. ICML*, Honolulu.

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

Zhehuai Chen, He Huang, Andrei Andrusenko, Oleksii Hrinchuk, Krishna C Puvvada, Jason Li, Subhankar Ghosh, Jagadeesh Balam, and Boris Ginsburg. 2024b. SALM: Speech-augmented language model with in-context learning for speech recognition and translation. In *Proc. ICASSP*, Seoul.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. 2023. Vicuna: An open-source chatbot impressing GPT-4 with 90%* ChatGPT quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2(3):6.

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*.

Erica Cooper, Wen-Chin Huang, Tomoki Toda, and Junichi Yamagishi. 2022. Generalization ability of MOS prediction networks. In *Proc. ICASSP*, Singapore.

Erica Cooper, Wen-Chin Huang, Yu Tsao, Hsin-Min Wang, Tomoki Toda, and Junichi Yamagishi. 2023. The VoiceMOS challenge 2023: Zero-shot subjective speech quality prediction for multiple domains. In *Proc. ASRU*, Taipei.

Erica Cooper and Junichi Yamagishi. 2021. How do voices from past speech synthesis challenges compare today? In *Proc. SSW*, Budapest.

Soham Deshmukh, Shuo Han, Hazim Bukhari, Benjamin Elizalde, Hannes Gamper, Rita Singh, and Bhiksha Raj. 2024. Audio Entailment: Assessing deductive reasoning for audio understanding. *arXiv preprint arXiv:2407.18062*.

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

Harishchandra Dubey, Ashkan Aazami, Vishak Gopal, Babak Naderi, Sebastian Braun, Ross Cutler, Hannes Gamper, Mehrsa Golestaneh, and Robert Aichner. 2023. ICASSP 2023 Deep Noise Suppression Challenge. In *Proc. ICASSP*, Rhodes Island.

Sefik Emre Eskimez, Xiaofei Wang, Manthan Thakker, Canrun Li, Chung-Hsien Tsai, Zhen Xiao, Hemin Yang, Zirun Zhu, Min Tang, Xu Tan, et al. 2024. E2 tts: Embarrassingly easy fully non-autoregressive zero-shot tts. In *Proc. SLT*, Macao.

Yassir Fathullah, Chunyang Wu, Egor Lakomkin, Junteng Jia, Yuan Shangguan, Ke Li, Jinxi Guo, Wenhan Xiong, Jay Mahadeokar, Ozlem Kalinli, et al. 2024. Prompting large language models with speech recognition abilities. In *Proc. ICASSP*, Seoul.

Eduardo Fonseca, Jordi Pons Puig, Xavier Favory, Frederic Font Corbera, Dmitry Bogdanov, Andres Ferraro, Sergio Oramas, Alastair Porter, and Xavier Serra. 2017. Freesound datasets: A platform for the creation of open audio datasets. In *Proc. ISMIR*, Suzhou.

Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio set: An ontology and human-labeled dataset for audio events. In *Proc. ICASSP*, New Orleans.

Shujie Hu, Long Zhou, Shujie Liu, Sanyuan Chen, Lingwei Meng, Hongkun Hao, Jing Pan, Xunying Liu, Jinyu Li, Sunit Sivasankaran, Linquan Liu, and Furu Wei. 2024. WavLLM: Towards robust and adaptive speech large language model. In *Findings of EMNLP*, Miami.

Yi Hu and Philipos C Loizou. 2007. Evaluation of objective quality measures for speech enhancement. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(1):229–238.

Chien-yu Huang, Wei-Chih Chen, Shu-wen Yang, Andy T Liu, Chen-An Li, Yu-Xiang Lin, Wei-Cheng Tseng, Anuj Diwan, Yi-Jen Shih, Jiatong Shi, et al. 2024. Dynamic-superb phase-2: A collaboratively expanding benchmark for measuring the capabilities of spoken language models with 180 tasks. *arXiv preprint arXiv:2411.05361*.

Zeyu Jin, Jia Jia, Qixin Wang, Kehan Li, Shuoyi Zhou, Songtao Zhou, Xiaoyu Qin, and Zhiyong Wu. 2024. SpeechCraft: A fine-grained expressive speech dataset with natural language description. In *Proc. ACM MM*, Melbourne.

Wei Kang, Xiaoyu Yang, Zengwei Yao, Fangjun Kuang, Yifan Yang, Liyong Guo, Long Lin, and Daniel Povey. 2024. Libriheavy: A 50,000 hours ASR corpus with punctuation casing and context. In *Proc. ICASSP*, Seoul.

Simon King and Vasilis Karaiskos. 2016. The Blizzard Challenge 2016. In *The Blizzard Challenge*.

Chen-Chou Lo, Szu-Wei Fu, Wen-Chin Huang, Xin Wang, Junichi Yamagishi, Yu Tsao, and Hsin-Min Wang. 2019. MOSNet: Deep learning based objective assessment for voice conversion. In *Proc. Interspeech*, Graz.

Ke-Han Lu, Zhehuai Chen, Szu-Wei Fu, He Huang, Boris Ginsburg, Yu-Chiang Frank Wang, and Hung-yi Lee. 2024. DeSTA: Enhancing speech language models through descriptive speech-text alignment. In *Proc. Interspeech*, Kos Island.

Georgia Maniati, Alexandra Vioni, Nikolaos Ellinas, Karolos Nikitaras, Konstantinos Klapsas, June Sig Sung, Gunu Jho, Aimilios Chalamandaris, and Pirros Tsiakoulis. 2022. SOMOS: The Samsung open MOS dataset for the evaluation of neural text-to-speech synthesis. In *Proc. Interspeech*, Incheon.

Gabriel Mittag, Babak Naderi, Assmaa Chehadi, and Sebastian Möller. 2021. NISQA: A deep CNN-self-attention model for multidimensional speech quality prediction with crowdsourced datasets. In *Proc. Interspeech*, Brno.

Brian Patton, Yannis Agiomyrgiannakis, Michael Terry, Kevin Wilson, Rif A Saurous, and D Sculley. 2016. AutoMOS: Learning a non-intrusive assessor of naturalness-of-speech. In *Proc. NIPS End-to-end Learning for Speech and Audio Processing Workshop*, Barcelona.

Puyuan Peng, Po-Yao Huang, Shang-Wen Li, Abdelrahman Mohamed, and David Harwath. 2024. VoiceCraft: Zero-shot speech editing and text-to-speech in the wild. In *Proc. ACL*, Bangkok.

Zengyi Qin, Wenliang Zhao, Xumin Yu, and Xin Sun. 2023. Openvoice: Versatile instant voice cloning. *arXiv preprint arXiv:2312.01479*.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proc. ICML*, Honolulu.

Yi Ren, Yangjun Ruan, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. 2019. Fastspeech: Fast, robust and controllable text to speech. In *Proc. NIPS*, Vancouver.

Paul K Rubenstein, Chulayuth Asawaroengchai, Duc Dung Nguyen, Ankur Bapna, Zalán Borsos, Félix de Chaumont Quitry, Peter Chen, Dalia El Badawy, Wei Han, Eugene Kharitonov, et al. 2023. AudioPaLM: A large language model that can speak and listen. *arXiv preprint arXiv:2306.12925*.

Takaaki Saeki, Detai Xin, Wataru Nakata, Tomoki Koriyama, Shinnosuke Takamichi, and Hiroshi Saruwatari. 2022. UTMOS: UTokyo-SaruLab system for VoiceMOS Challenge 2022. In *Proc. Interspeech*, Incheon.

Oscar Figuerola Salas, Velibor Adzic, and Hari Kalva. 2013. Subjective quality evaluations using crowdsourcing. In *Proc. PCS*, San Jose.

Thibault Sellam, Ankur Bapna, Joshua Camp, Diana Mackinnon, Ankur P Parikh, and Jason Riesa. 2023. SQuId: Measuring speech naturalness in many languages. In *Proc. ICASSP*, Rhodes Island.

Suwon Shon, Kwangyoun Kim, Yi-Te Hsu, Prashant Sridhar, Shinji Watanabe, and Karen Livescu. 2024. DiscreteSLU: A large language model with self-supervised discrete speech units for spoken language understanding. *arXiv preprint arXiv:2406.09345*.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. 2024a. SALMONN: Towards generic hearing abilities for large language models. In *Proc. ICLR*, Vienna.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Jun Zhang, Lu Lu, Zejun Ma, Yuxuan Wang, et al. 2024b. Can large language models understand spatial audio? *arXiv preprint arXiv:2406.07914*.

Joachim Thiemann, Nobutaka Ito, and Emmanuel Vincent. 2013. DEMAND: A collection of multi-channel recordings of acoustic noise in diverse environments. In *Poc. ICAASA*, Montreal.

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*.

Mingqiu Wang, Wei Han, Izhak Shafran, Zelin Wu, Chung-Cheng Chiu, Yuan Cao, Nanxin Chen, Yu Zhang, Hagen Soltau, Paul K Rubenstein, et al. 2023. SLM: Bridge the thin gap between speech and text foundation models. In *ASRU*, Taipei.

Siyin Wang, Wenyi Yu, Yudong Yang, Changli Tang, Yixuan Li, Jimin Zhuang, Xianzhao Chen, Xiaohai Tian, Jun Zhang, Guangzhi Sun, et al. 2024b. Enabling auditory large language models for automatic speech quality evaluation. *arXiv preprint arXiv:2409.16644*.

Matthew B Winn and Katherine H Teece. 2021. Listening effort is not the same as speech intelligibility score. *Trends in Hearing*, 25.

Jian Wu, Yashesh Gaur, Zhuo Chen, Long Zhou, Yimeng Zhu, Tianrui Wang, Jinyu Li, Shujie Liu, Bo Ren, Linquan Liu, et al. 2023. On decoder-only architecture for speech-to-text and large language model integration. In *Proc. ASRU*, Teipei.

Yaoxun Xu, Hangting Chen, Jianwei Yu, Qiaochu Huang, Zhiyong Wu, Shi-Xiong Zhang, Guangzhi Li, Yi Luo, and Rongzhi Gu. 2024. Secap: Speech emotion captioning with large language model. In *AAAI*, Vancouver.

Qian Yang, Jin Xu, Wenrui Liu, Yunfei Chu, Ziyue Jiang, Xiaohuan Zhou, Yichong Leng, Yuanjun Lv, Zhou Zhao, Chang Zhou, et al. 2024. AIR-Bench: Benchmarking large audio-language models via generative comprehension. In *Proc. ACL*, Bangkok.

Zhao Yi, Wen-Chin Huang, Xiaohai Tian, Junichi Yamagishi, Rohan Kumar Das, Tomi Kinnunen, Zhen-Hua Ling, and Tomoki Toda. 2020. Voice Conversion Challenge 2020: Intra-lingual semi-parallel and cross-lingual voice conversion. In *Proc. VCCBC*, Shanghai.

Zhiyuan You, Zheyuan Li, Jinjin Gu, Zhenfei Yin, Tianfan Xue, and Chao Dong. 2025. Depicting beyond scores: Advancing image quality assessment through multi-modal language models. In *ECCV*, London.

Wenyi Yu, Changli Tang, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. 2024. Connecting speech encoder and large language model for ASR. In *Proc. ICASSP*, Seoul.

Ryandhimas E Zezario, Sabato M Siniscalchi, Hsin-Min Wang, and Yu Tsao. 2024. A study on zero-shot non-intrusive speech assessment using large language models. *arXiv preprint arXiv:2409.09914*.

Zhisheng Zheng, Puyuan Peng, Ziyang Ma, Xie Chen, Eunsol Choi, and David Harwath. 2024. BAT: Learning to reason about spatial sounds with large language models. In *Proc. ICML*, Vienna.

A Speech quality assessment: MOS and natural language

We believe that describing speech quality in natural language and evaluating it using numerical scores, such as MOS, are complementary approaches rather than direct alternatives. Therefore, natural language assessment is not compared with traditional numerical score method in our paper. MOS score is a rough overall assessment that facilitates easy comparisons between samples and systems, while natural language provides more detailed and specific information, such as identifying the type and occurrence time of distortions. If the goal is to simply compare the quality of two speech samples, MOS remains the preferred choice. However, for more detailed and instructive feedback — such as identifying consistent issues like a clicking sound at the beginning of TTS outputs — natural language assessment is invaluable. Additionally, audio LLMs can reason in natural language, providing explanations for overall quality judgments, which paves the way for more reliable speech quality assessments.

B Choice of low-level speech aspects

We aim to build a comprehensive low-level speech quality assessment dataset, so we refer to different aspects discussed in previous works, for example, noise(Mittag et al., 2021; Hu and Loizou, 2007), distortion(Mittag et al., 2021; Hu and Loizou, 2007), speed(Ren et al., 2019), naturalness(Sellam et al., 2023; Salas et al., 2013), listening effort(Salas et al., 2013; Winn and Teece, 2021), continuity(Mittag et al., 2021), with the hope that low-level speech aspects in QualiSpeech can cover not only basic degradations, such as background noise and distortion, but also subjective perceptual assessment, including speech speed, continuity, naturalness and listening effort. We also collect specific descriptions to further enrich the description perspectives, building a speech quality assessment dataset containing comprehensive low-level speech features and natural language descriptions.

C Human annotators information

Our human annotators consist of 21 females and 4 males, all in their 20s. Due to local constraints, all annotators are native Mandarin speakers with English proficiency equivalent to an IELTS score of 7.0 or higher, although they are not native English speakers. And strong correlations between high level non-native listeners and native English listeners are reported in (Yi et al., 2020). Each annotation takes approximately six minutes to complete, and annotators are paid 6 Chinese Yuan (approximately 0.8 USD) per sample, satisfying the minimum income standards of our region.

D Label distributions of QualiSpeech

The label distributions of aspects annotated in scores are shown in Figure 5. The distributions of speed and overall quality roughly follow a normal pattern. For noise, continuity, and distortion, part of the samples are noisy or discontinuous, forming an approximately normal distribution, while the other half are clean or smooth rated as 5. Listening effort exhibits a ladder-like distribution, with the highest number of samples receiving a score of 5 and the lowest number receiving a score of 1. naturalness follows a shifted normal distribution, peaking at 2 rather than 3.

E Detailed statistics of QualiSpeech

The detailed statistics of QualiSpeech dataset shown in Table 5.

Type	Source	# Utterances Train	Valid	Test
Synthetic	BVCC	4546	975	912
	Recent TTS	600	60	60
Real	GigaSpeech	2000	200	400
	NISQA LIVE	529	106	0
	NISQA FOR	0	0	180
	NISQA P501	0	0	180
Simulated Real	NISQA SIM	2883	826	0
	NISQA FOR	0	0	60
	NISQA P501	0	0	60
Summary of synthetic		5146	1035	972
Summary of all real		5412	1132	880
Summary		10558	2167	1852

Table 5: Statistics of our QualiSpeech dataset, with respect to data sources and dataset splits.

The Pearson correlation coefficient (PCC) between overall quality scores in QualiSpeech and MOS scores in the source datasets is calculated to check the consistency of QualiSpeech with the source datasets, as shown in Table 6. It is quite relevant between these two scores, considering that overall quality scores in QualiSpeech are integers

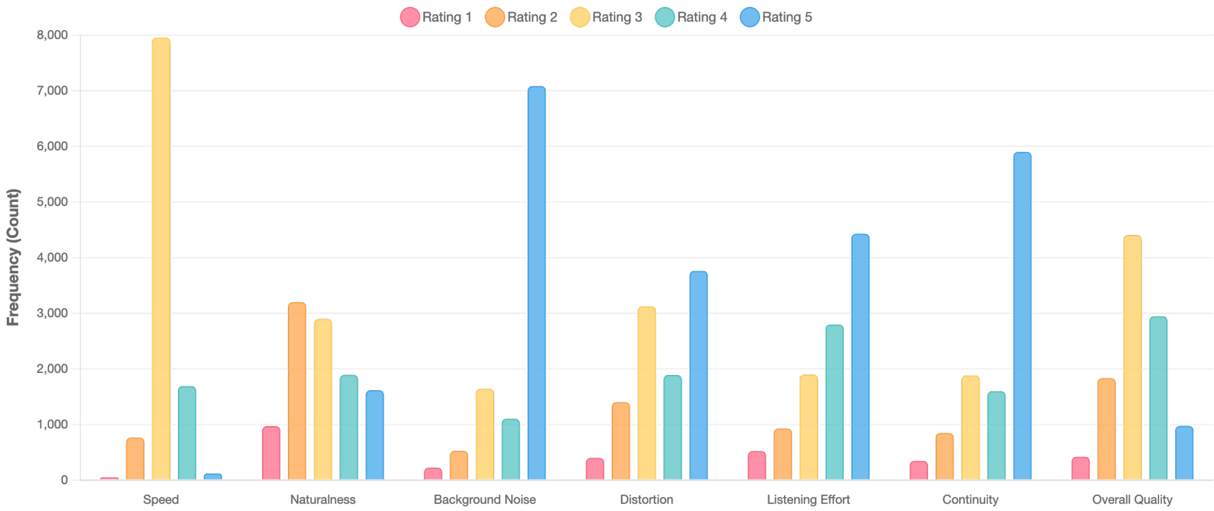

Figure 5: Label distributions of aspects annotated in scores in QualiSpeech

while MOS scores in the source datasets are decimals.

Subset	BVCC test split	NISQA TEST
PCC	0.55	0.66

Table 6: The correlation scores between overall quality scores in QualiSpeech and MOS scores in the source dataset.

F Experiment setup for descriptive comments

"Concise" and "detailed" denotes the text length of the descriptive comments. For "concise", a maximum limitation of 150 words is added for generated concise descriptions, while for detailed descriptions there is no maximum limitation of text length. For concise descriptions, the average text length is 136. While for detailed descriptions, it is 209. "Concise with num" means the annotated scores will be included in the generated descriptions and the generation has a maximum limitation of 150 words. "Revised concise" is the revised version of concise descriptions by human annotators (annotation step 3), for other descriptions, they are not revised. Below is a example of different kinds of descriptions:

Concise: *The speech sample presents a pleasant listening experience with background noise being virtually non-existent, allowing for clear audio. However, there is some noticeable pitch distortion accompanied by a mechanical sound that detracts from the overall quality, making the speech somewhat unnatural. Despite this distortion, the audio remains smooth and free of interruptions, contributing positively to the listening experience. The listener requires only a moderate amount of effort to comprehend the content, indicating that the speech is mostly accessible. The speaking speed is moderate, which complements the gentle and peaceful tone of the young man's voice, enhancing the overall impression. While the naturalness of the speech is lacking, the combination of smooth delivery and minimal background noise results in a fair overall quality, suggesting room for improvement in the clarity and authenticity of the voice.*

Revised concise: *The speech sample presents a pleasant listening experience with background noise being virtually non-existent, allowing for clear audio. However, there is some noticeable pitch distortion accompanied by a mechanical sound that detracts from the overall quality from 0 to 2 seconds, making the speech somewhat unnatural. Despite this distortion, the audio remains smooth and free of interruptions, contributing positively to the listening experience. The listener requires only a moderate amount of effort to comprehend the content, indicating that the speech is mostly accessible. The speaking speed is moderate, which complements the gentle and peaceful tone of the young man's voice, enhancing the overall impression. While the naturalness of the speech is lacking, the combination of smooth delivery and minimal background noise results in a fair overall quality, suggesting room for improvement in the clarity and authenticity of the voice.*

Concise with num: *The speech sample exhibits* *a background noise level rated at 5, indicating that it is not noticeable, which positively contributes to the overall quality. However, distortion is rated at 3, suggesting some pitch distortion and mechanical sound present throughout the voice, which detracts from the naturalness of the speech, rated at 2, indicating it sounds fairly unnatural. Discontinuity is rated at 5, reflecting a very smooth delivery, while listening effort is rated at 4, meaning only attention is necessary to understand the content. The speaking speed is moderate at 3, which is acceptable. The voice is described as a young man's gentle voice with a peaceful tone, adding a pleasant feeling. Overall, the speech quality is rated at 3, indicating a fair quality, primarily impacted by the distortion and naturalness issues, despite the smoothness and low background noise.*

Detailed: *The speech sample presents a range of qualities that contribute to its overall evaluation. Starting with background noise, it is commendable that there is no noticeable interference, allowing the listener to focus on the content without distraction. However, the distortion aspect reveals some concerns, as there is a pitch distortion accompanied by a mechanical sound that persists throughout the initial two seconds, which detracts from the natural quality of the speech. Despite this, the discontinuity is excellent, with the audio being very smooth, indicating that there are no breaks or stutters that could disrupt comprehension. In terms of listening effort, the sample requires some attention but does not demand significant effort, suggesting that the listener can grasp the meaning with relative ease. The naturalness of the speech, however, is somewhat lacking, as it does not closely resemble natural human speech, which may affect the listener's engagement. The speaking speed is moderate, which is appropriate for comprehension, neither too fast nor too slow. The voice itself is described as a young man's gentle voice with a peaceful tone, which adds a positive emotional layer to the experience. Overall, while the speech has strengths in background noise management, smoothness, and listener engagement, the distortion and naturalness issues bring down its quality. Therefore, the overall quality can be considered fair, reflecting a mix of commendable attributes and notable shortcomings.*

G Text instructions for Parler-TTS

The text instructions for Parler-TTS which specfies speaker are sampled from a corpus containing 44 sentences. 10 sentences are generated by GPT following the format of the example instruction[6], with age, gender and speed changed. 34 sentences are simple "someone's voice" in which *someone* refers to an available built-in speaker. The full list of text instructions are presented as below:

1. *A young male speaker presents a calm, steady speech at a slightly slower-than-average speed, with a deeper pitch. The recording quality is excellent, with his voice sounding close up and clear, as if he's speaking directly into the microphone.*

2. *An elderly female speaker delivers a warm, gentle speech, characterized by a slow pace and soft pitch. Her voice is clear and intimate, with the high-quality recording capturing her subtle inflections.*

3. *A middle-aged male speaker gives a confident and slightly assertive speech, maintaining a steady pace with a moderate pitch. The recording is pristine, and his voice sounds rich and detailed, as if he's speaking in a small, quiet room.*

4. *A young female speaker offers a friendly and enthusiastic speech with a faster-than-average speed and a higher pitch. Her voice is clear and lively, captured in great quality, creating a vibrant and engaging listening experience.*

5. *A middle-aged female speaker delivers a formal, measured speech at a moderate speed, with a soft but clear voice. The recording quality is high, capturing her voice with warmth and precision, as if in a professional studio.*

6. *A male teenager speaks in a casual, conversational tone, with a moderate speed and a slightly higher pitch than average. The recording is of very high quality, with his voice sounding crisp and close, capturing even his smallest breaths.*

7. *An elderly male speaker gives a slow, reflective speech with a slightly lower pitch, exuding wisdom and calm. The recording quality is excellent, making his voice sound intimate and detailed, as if he's speaking directly to the listener.*

8. *A young adult female speaks with high energy, delivering her message at a rapid pace and with a high pitch. The quality of the recording is very clear, emphasizing her upbeat tone and the subtleties in her speech.*

9. *A young adult male delivers a slightly monotone, yet articulate, speech with a moderate speed*

and pitch. The recording quality is excellent, making his voice sound clear and close, as if he's speaking in a quiet room.

10. *A middle-aged female speaker speaks with a gentle, soothing tone, maintaining a slow pace and lower pitch. Her voice is captured in exceptional quality, with every word resonating warmly and clearly, giving an intimate listening experience.*
11. *Laura's Voice.*
12. *Gary's Voice.*
13. *Jon's Voice.*
14. *Lea's Voice.*
15. *Karen's Voice.*
16. *Rick's Voice.*
17. *Brenda's Voice.*
18. *David's Voice.*
19. *Eileen's Voice.*
20. *Jordan's Voice.*
21. *Mike's Voice.*
22. *Yann's Voice.*
23. *Joy's Voice.*
24. *James's Voice.*
25. *Eric's Voice.*
26. *Lauren's Voice.*
27. *Rose's Voice.*
28. *Will's Voice.*
29. *Jason's Voice.*
30. *Aaron's Voice.*
31. *Naomie's Voice.*
32. *Alisa's Voice.*
33. *Patrick's Voice.*
34. *Jerry's Voice.*
35. *Tina's Voice.*
36. *Jenna's Voice.*
37. *Bill's Voice.*
38. *Tom's Voice.*
39. *Carol's Voice.*
40. *Barbara's Voice.*
41. *Rebecca's Voice.*
42. *Anna's Voice.*
43. *Bruce's Voice.*
44. *Will's Voice.*

H Examples in QualiSpeech

Example 1 (a low quality synthetic speech sample mixed with baby crying):

Speed: 3
Naturalness: 2
Background noise: 3
Distortion: 4
Listening effort: 3
Continuity: 2
Overall quality: 2

Feeling of voice: the calm and robotically mechnical voice of a young woman

Noise description: There is outdoor background noise like baby crying in the audio from 0 to 3.3 seconds.

Distortion description: There is a pitch distortion with mechnical sound in 2.5-3s.

Unnatural pause: 2.2-2.5s.

Descriptive comments: The speech sample presents a moderate level of background noise, with outdoor sounds, including a baby crying, noticeable from 0 to 3.3 seconds, which somewhat distracts from the clarity of the speech. There is slight distortion, particularly a mechanical pitch alteration occurring between 2.5 and 3 seconds, which affects the naturalness of the voice. The audio experiences some discontinuity, making it fairly disjointed, especially around 2.2 to 2.5 seconds, where the unnatural quality is evident. Listeners may find moderate effort required to comprehend the speech due to these factors. The speaking speed is moderate, contributing to a calm delivery, but the overall quality remains poor due to the combination of these issues. The voice, described as that of a middle-aged man, has a calm yet somewhat mechanical tone, which further detracts from the overall listening experience.

Example 2 (a high quality audiobook data):

Speed: 3
Naturalness: 5
Background noise: 5
Distortion: 5
Listening effort: 5
Continuity: 5
Overall quality: 5

Feeling of voice: A middle-aged man with bright voice that makes people feel relaxed

Noise description: not noticeable.

Distortion description: not distorted.

Unnatural pause: very smooth.

Descriptive comments: The speech sample exhibits exceptional quality across all evaluated aspects. The background noise is virtually nonexistent, allowing for a clear listening experience. There are no distortions present, ensuring that the natural sound of the speech is preserved. The audio flows very smoothly, with no interruptions or stutters, contributing to a seamless listening experience. Understanding the content requires no effort, as the speech is delivered in a way that is easily comprehensible. The naturalness of the voice is striking, closely resembling human speech in tone and clarity. The speaking speed is moderate, which balances well with the overall delivery. The voice of the speaker, described as that of a middle-aged man with a bright tone, evokes a sense of relaxation in listeners. Overall, this speech sample is of excellent quality, making it a pleasure to listen to.

I Questionnaires and prompts

Step1:
Please evaluate the quality of the speech sample from various aspects. Please read the explanations provided for each aspect carefully before scoring and select the score you find most appropriate. You may need to provide more detailed information if there is any noise, distortion or unnatural pause.

Background noise (the level of noise in the audio, reflecting the impact of background noise or other non-speech interference on audio quality.):
- 1: Very intrusive
- 2: Somewhat intrusive
- 3: Noticeable but not intrusive
- 4: Slightly noticeable
- 5: Not noticeable

If there is any noise, identify the type and when it occurs.

Distortion (the alterations in the natural sound of speech caused by distortions or unwanted modifications.):
- 1: Very distorted
- 2: Fairly distorted
- 3: Somewhat distorted
- 4: Slightly distorted
- 5: Not distorted

If there is any distortion, identify the type and when it occurs.

Listening effort (the effort required to understand the meanings of sentences.):
- 1: No meaning understood with any feasible effort
- 2: Considerable effort required
- 3: Moderate effort required
- 4: Attention necessary; no appreciable effort required
- 5: Complete relaxation possible; no effort required

Discontinuity (the discontinuity in the audio, reflecting whether there are breaks, stutters, or incoherence during playback.):
- 1: Very disjointed
- 2: Fairly disjointed
- 3: Somewhat smooth
- 4: Mostly smooth
- 5: Very smooth

If there is any unnatural pause, identify when it occurs.

Speaking speed:
- 1: Very slow
- 2: Fairly slow
- 3: Moderate speed
- 4: Fairly fast
- 5: Very fast

Naturalness (the level of how closely generated speech resembles natural human speech in terms of tone, rhythm, intonation, and clarity.):
- 1: Very unnatural
- 2: Fairly unnatural
- 3: Somewhat natural
- 4: Fairly natural
- 5: Very natural

Overall quality:
- 1: Bad
- 2: Poor
- 3: Fair
- 4: Good
- 5: Excellent

Feeling of voice (Please describe your impressions of the speaker, including but not limited to their age, gender, and tone of voice.):

Step3:
Please review the generated text to check for hallucinations, correct inaccurate content and add any missing information. Additionally, if the reasoning is weak or illogical, please make the necessary adjustments.

Figure 6: Questionnaires and instructions for annotation

```
{
    "role": "system",
    "content": "I will give you a tuple of meta information and some detailed natural language descriptions for speech quality evaluation, it contains 5 factors are rating from 1 to 5. For all these factors, higher is better. \n
    (1) Background noise: the level of noise in the audio, reflecting the impact of background noise or other non-speech interference on audio quality. 5: Not noticeable, 4: Slightly noticeable, 3: Noticeable but not intrusive, 2: Somewhat intrusive, 1: Very intrusive \n
    (2) Distortion: the alterations in the natural sound of speech caused by distortions or unwanted modifications. 5: Not distorted, 4: Slightly distorted, 3: Somewhat distorted, 2: Fairly distorted, 1: Very distorted \n
    (3) Discontinuity: the discontinuity in the audio, reflecting whether there are breaks, stutters, or incoherence during playback. 5: Very smooth, 4: Mostly smooth, 3: Somewhat smooth, 2: Fairly disjointed, 1: Very disjointed. \n
    (4) Listening-effort: the effort required to understand the meanings of sentences. 5: Complete relaxation possible; no effort required, 4: Attention necessary; no appreciable effort required, 3: Moderate effort required, 2: Considerable effort required, 1: No meaning understood with any feasible effort \n
    (5) Naturalness: the level of how closely generated speech resembles natural human speech in terms of tone, rhythm, intonation, and clarity. 5: Very natural, 4: Mostly natural, 3: Somewhat natural, 2: Fairly unnatural, 1: Very unnatural \n
    (6) Speaking speed: 5: Very fast, 4: Mostly fast, 3: Moderate speed, 2: Fairly slow, 1: Very slow \n
    (7) Feeling of voice: the natural language description of how the listener feels about the voice \n
    (8) Overall Quality: 5: excellent, 4: good, 3: fair, 2: poor, 1: bad \n
    I need you to generate a descriptive evaluation for this speech considering all aspects in one parapraph. Do not omit any aspect. Do not omit the time information."
},
{
    "role": "user",
    "content": "These are the labels of one speech sample : Background noise: {}, Distortion: {}, Discontinuity: {}, Listening-effort: {}, Naturalness: {}, Speaking speed: {}, Overall quality: {}, Feeling of voice: {}, noise descriptions: {}, distortion description: {}, unnatural pause: {}." [special prompt for different kinds of descriptions]
}
```

Special prompt for different kinds of descriptions:

Concise: *Please generate a concise and descriptive evaluation of the speech sample in the Chain of Thought (CoT) reasoning format. Begin by analyzing each aspect individually, then synthesize the findings to arrive at an overall score. Do not contain scores anymore, use natural langauge please. Limit your response to a single paragraph with no more than 150 words.*

Concise with num: *Please generate a concise and descriptive evaluation of the speech sample in the Chain of Thought (CoT) reasoning format. Begin by analyzing each aspect individually, then synthesize the findings to arrive at an overall score. Limit your response to a single paragraph with no more than 150 words.*

Detailed: *Please generate a concise and descriptive evaluation of the speech sample in the Chain of Thought (CoT) reasoning format. Begin by analyzing each aspect individually, then synthesize the findings to arrive at an overall score.*

Figure 7: Prompts used in annotation procedure step 2 for generating different kinds of descriptive comments

Speed:
How would you assess the speaking speed in the audio? Use the following scale:
5. Very fast, 4. Mostly fast, 3. Moderate speed, 2. Fairly slow, 1. Very slow.

Naturalness:
How would you rate the naturalness of the audio? Choose a number from the following scale:
5. Very natural, 4. Mostly natural, 3. Somewhat natural, 2. Fairly unnatural, 1. Very unnatural.

Noise:
How would you rate the noise level in the audio? Use the following scale to describe its impact:
5. Not noticeable, 4. Slightly noticeable, 3. Noticeable but not intrusive, 2. Somewhat intrusive, 1. Very intrusive.

Distortion:
How would you rate the level of distortion in the audio? Use the following scale to describe its quality:
5. Not distorted, 4. Slightly distorted, 3. Somewhat distorted, 2. Fairly distorted, 1. Very distorted.

Listening effort:
How would you rate the intelligibility of the audio you just listened to? Use the following scale to describe your experience:
5. Complete relaxation possible; no effort required, 4. Attention necessary; no appreciable effort required, 3. Moderate effort required, 2. Considerable effort required, 1. No meaning understood with any feasible effort.

Continuity:
How would you rate the continuity of the audio? Use the following scale to describe its smoothness:
5. Very smooth, 4. Mostly smooth, 3. Somewhat smooth, 2. Fairly disjointed, 1. Very disjointed.

Overall:
How would you rate the overall quality of the audio? Use the following scale to describe your experience:
5. excellent, 4. good, 3. fair, 2. poor, 1. bad.

Noise description:
Please analyze the noise present in this audio clip.

Distortion description:
Please analyze the distortion in this audio clip.

Unnatural pause:
Analyze the audio for unnatural pauses. For each instance, specify the exact timestamps.

Feeling of voice:
Analyze the speaker's voice and provide a concise description.

Quality Description:
Please evaluate the quality of the audio you just heard. Start by analyzing specific aspects such as noise levels, distortion, naturalness and so on. Then provide an overall assessment based on these factors.

Figure 8: Prompts for each tasks when finetuning and testing auditory LLMs.

```
{
    "role": "system",
    "content": " Here are some basic low-level speech perception aspects. For all numerical scores, higher is better. \n
    (1) Background noise: the level of noise in the audio, reflecting the impact of background noise or other non-speech interference on audio quality. 5: Not noticeable, 4: Slightly noticeable, 3: Noticeable but not intrusive, 2: Somewhat intrusive, 1: Very intrusive \n
    (2) Distortion: the alterations in the natural sound of speech caused by distortions or unwanted modifications. 5: Not distorted, 4: Slightly distorted, 3: Somewhat distorted, 2: Fairly distorted, 1: Very distorted \n
    (3) Discontinuity: the discontinuity in the audio, reflecting whether there are breaks, stutters, or incoherence during playback. 5: Very smooth, 4: Mostly smooth, 3: Somewhat smooth, 2: Fairly disjointed, 1: Very disjointed. \n
    (4) Listening-effort: the effort required to understand the meanings of sentences. 5: Complete relaxation possible; no effort required, 4: Attention necessary; no appreciable effort required, 3: Moderate effort required, 2: Considerable effort required, 1: No meaning understood with any feasible effort \n
    (5) Naturalness: the level of how closely generated speech resembles natural human speech in terms of tone, rhythm, intonation, and clarity. 5: Very natural, 4: Mostly natural, 3: Somewhat natural, 2: Fairly unnatural, 1: Very unnatural \n
    (6) Speaking speed: 5: Very fast, 4: Mostly fast, 3: Moderate speed, 2: Fairly slow, 1: Very slow \n
    (7) Feeling of voice: the natural language description of how the listener feels about the voice \n
    (8) Overall Quality: 5: excellent, 4: good, 3: fair, 2: poor, 1: bad \n
     I will give a descriptive assessment on speech sample. And you need to break down each aspect of the evaluation into scores or simple descriptions. "
},
{
    "role": "user",
    "content": "Here is a descriptive quality assessment of one speech sample. \"{}\" Please convert the text description into numerical scores or simple descriptions of different aspects based on the scales. The aspects include noise, distortion, discontinuity, listening effort, naturalness, speaking spead, overall quality, noise type and time (if any), distortion type and time (if any), unnatural pause time (if any) and feeling of voice. The scores and simple description should be put in *"
}
```

Figure 9: Prompts for breaking down description comments into basic aspects, used in evaluating natural language descriptions.

Noise type:
```
{
    "role": "user",
    "content": "Please compare the two noise descriptions below and provide a relevance score based on the type of noise. The score should range from 0 to 1, with higher scores indicating greater relevance. The noise descriptions are \"{}\" and \"{}\". Please include the relevance score in parentheses ()."}
```

Distortion type:
```
{
    "role": "user",
    "content": "Please compare the two distortion descriptions below and provide a relevance score based on the type of distortion. The score should range from 0 to 1, with higher scores indicating greater relevance. The distortion descriptions are \"{}\" and \"{}\". Please include the relevance score in parentheses ()."
}
```

Feeling of voice:
```
{
    "role": "user",
    "content": "Please compare the two voice descriptions below and provide a relevance score. The score should range from 0 to 1, with higher scores indicating greater relevance. The voice descriptions are \"{}\" and \"{}\". Please include the relevance score in parentheses ().*"
}
```

Figure 10: Prompts for calculating correlation score between two descriptions, used in evaluating specific descriptions.

J More results

J.1 Ablation study on finetuning encoders

We also investigate the influence of finetuning encoders, with results demonstrated in Table 7. The training setting follows the "joint" setting. Results show that finetuning encoders can bring further performance improvement. We suggest that with encoders joint finetuned, the low-level speech information can be transformed to a format that LLM backbone can better understand, resulting in a performance gain.

J.2 Results of generalization experiments

J.3 Holistic correlation of generated descriptions

We also try to give a holistic correlation score generated by GPT based on the entire description, the results are shown below. The results show that a holistic correlation score makes it hard to distinguish the differences in methods, and therefore correlation scores of specific descriptions are reported in Section 4.

Comments	Holistic Corr.
revised concise	0.667
concise with num	0.675
concise	0.639
detailed	0.661

Table 9: Holistic correlation of the entire generated descriptions

Finetuning Encoder	Aspect of low-level speech perception						
	Noise	Distortion	Speed	Continuity	Effort	Naturalness	Overall
freeze	0.693	0.595	0.240	0.525	0.578	0.565	0.636
Whisper	0.687	0.626	**0.355**	0.576	0.598	0.587	0.643
BEATs	**0.723**	0.645	0.308	0.536	0.597	0.595	0.654
Whisper + BEATs	0.707	**0.648**	0.325	**0.601**	**0.617**	**0.606**	**0.670**

Table 7: Results of ablation study on finetuning encoder

Training Dataset	Aspect of low-level speech perception							
	Noise	Distortion	Speed	Continuity	LE	Naturalness	Average first 6	Overall
synthetic	0.785	0.533	**0.288**	0.324	0.517	0.426	**0.479**	0.520
real	0.802	0.468	0.211	0.163	0.476	0.450	0.428	0.471
all	**0.806**	**0.543**	0.200	**0.333**	**0.521**	**0.463**	0.478	**0.543**

(a) Results on synthetic data

Training Dataset	Aspect of low-level speech perception							
	Noise	Distortion	Speed	Continuity	LE	Naturalness	Average first 6	Overall
synthetic	**0.717**	0.458	0.221	0.523	0.500	0.399	0.470	0.597
real	0.652	0.588	**0.284**	0.607	0.582	0.614	0.555	0.709
all	0.589	**0.611**	0.258	**0.691**	**0.631**	**0.618**	**0.566**	**0.724**

(b) Results on real data

Table 8: Full results of generalization experiments. "Average first 6" denotes the averaged accuracy of 6 aspects annotated in scores except for overall.

J.4 Investigation of multiple annotations

To investigate the impact of multiple annotations, we collect 2 more annotations for a subset of test split in QualiSpeech (only the data from BVCC resource). We further explore the consistency among multiple annotations and assess whether incorporating multiple annotations can benefit the testing procedure.

J.4.1 Consistency of multiple annotations

To evaluate the consistency of multiple annotations, mutual consistency metrics are calculated by averaging pairwise consistency metrics. For aspects annotated in scores, PCC is selected as consistency metric. For aspects annotated in descriptions, correlation score generated by GPT is utilized.

Results show that different annotations do not exhibit a high consistency, underscoring the inherent complexity of speech quality assessment. Notably, consistency is higher for basic degradations such as noise and distortion compared to more subjective perceptual aspects. Furthermore, compared results in Table 2 and Table 10, the finetuned auditory LLM demonstrates a promising potential to function as an annotator for speech quality evaluation.

Low-level speech aspect	Mutual PCC
noise	0.728
distortion	0.682
speed	0.316
continuity	0.604
effort	0.653
naturalness	0.458
overall	0.603

Table 10: Consistency of multiple annotations on aspects annotated in scores

Low-level speech aspect	Mutual Corr
noise	0.672
distortion	0.611
voice	0.483

Table 11: Consistency of multiple annotations on aspects annotated in descriptions

J.4.2 Multiple annotations for evaluation

We further explore the use of multiple annotations for evaluation. The model tested is the one trained under "joint" setting in Table 2. We report the evaluation metrics for each annotation individually, along with their averaged results. Furthermore, we explore two new evaluation metrics. For aspects annotated in scores, we use the mean of all annotations as the ground truth label, denoted as "Mean value" in Table 12. For aspects annotated in descriptions, we select the highest correlation score as the final result, labeled as "Best" in Table 13.

The results indicate that incorporating multiple annotations enables the use of more reliable evaluation metrics. For aspects assessed with numerical scores, the PCC with the mean value is the highest, suggesting that even when the model is trained on a single annotation, it aligns most closely with the averaged scores. This finding highlights that averaging multiple annotations will lead to more stable evaluation. For aspects annotated in descriptions, we believe "Best" is also a better evaluation metric, since generating descriptions is an open-ended task with multiple reasonable answers. We hope our pioneering dataset can inspire large-scale natural language speech quality assessment dataset with multiple annotations in the future.

Low-level speech aspect	1	2	3	Average	Mean value
noise	0.756	0.755	0.748	0.753	0.831
distortion	0.704	0.709	0.693	0.702	0.790
speed	0.183	0.310	0.346	0.280	0.393
continuity	0.642	0.635	0.608	0.628	0.731
effort	0.679	0.673	0.651	0.668	0.760
naturalness	0.445	0.403	0.412	0.420	0.522
overall	0.646	0.636	0.607	0.630	0.732

Table 12: Evaluation of aspects annotated in scores using multiple annotations

Low-level speech aspect	1	2	3	Average	Best
noise	0.498	0.499	0.515	0.504	0.665
distortion	0.581	0.580	0.581	0.581	0.702
voice	0.442	0.443	0.449	0.445	0.548

Table 13: Evaluation of aspects annotated in descriptions using multiple annotations