Advancing Speech Language Models by Scaling Supervised Fine-Tuning with Over 60,000 Hours of Synthetic Speech Dialogue Data

Shuaijiang Zhao***, Tingwei Guo*****, Bajian Xiang, Tongtang Wan, Qiang Niu, Wei Zou**† **, Xiangang Li**

Beike Inc., Beijing, China
{zhaoshuaijiang001, zouwei026, lixiangang002}@ke.com

Abstract

The GPT-4o represents a significant milestone in enabling real-time interaction with large language models (LLMs) through speech, its remarkable low latency and high fluency not only capture attention but also stimulate research interest in the field. This real-time speech interaction is particularly valuable in scenarios requiring rapid feedback and immediate responses, dramatically enhancing user experience. However, there is a notable lack of research focused on real-time large speech language models, particularly for Chinese. In this work, we present KE-Omni, a seamless large speech language model built upon Ke-SpeechChat, a large-scale high-quality synthetic speech interaction dataset consisting of 7 million Chinese and English conversations, featuring 42,002 speakers, and totaling over 60,000 hours, This contributes significantly to the advancement of research and development in this field. The model, dataset, code and demo can be accessed at `https://huggingface.co/spaces/KE-Team/KE-Omni`.

1 Introduction

Large-language models (LLMs) hold significant promise for enhancing human-computer interaction, offering advanced conversational skills and versatility in managing diverse, open-ended user requests in various tasks and domains. Integrating speech with LLMs, namely large speech language models, enables a more natural form of interaction, allowing models to listen, process, and respond like humans. Notably, GPT-4o(OpenAI, 2024) with its real-time speech interaction capabilities has made significant strides in this direction, taking a crucial step toward realizing human-like natural speech interaction.

However, the exploration of seamless speech interaction with large speech language models remains largely absent. Achieving effective speech interaction with LLMs presents several challenges: (1) Difficulty of speech-text modality alignment. Aligning continuous, diversity speech signals with discrete text symbols poses a challenge. (2) Challenge of seamless speech interaction. Speech responses must be of high quality and low latency to ensure a fluid user experience. (3) Serious lack of speech interaction data. Acquiring large-scale speech datasets, particularly for interactive scenarios, is costly and resource-intensive, creating significant barriers to advancement.

To effectively align speech and text modalities and achieve seamless speech interaction with large language models, we introduce KE-Omni, inspired by large speech language models, like LLama-Omni(Fang et al., 2024) and SpeechGPT(Zhang et al., 2023). Unlike LLama-Omni and SpeechGPT, which are limited to English, KE-Omni is proficient in both Chinese and English. Experimental results show that KE-Omni can simultaneously generate high-quality text and low-latency speech responses.

To address the scarcity and cost of speech interaction data while protecting individuals' voices from misuse, we explore efficient synthetic data methods and have constructed the large-scale, high-quality speech interaction dataset in Chinese and English, promoting the development of this field.

In summary, this paper makes two main contributions:

- We present a novel approach to constructing Ke-SpeechChat, a large scale high-quality speech interaction dataset comprising 7 million Chinese and English conversations, featuring 42,002 speakers and totaling over 60,000 hours of audio.
- We introduce KE-Omni, a seamless large speech language model designed for real-time speech interaction in both Chinese and English, built upon the Ke-SpeechChat dataset.

2 Related work

Large speech-language models for interaction. An easy-to-implement solution is to integrate speech recognition and synthesis with a large language model (LLM), as demonstrated in (Huang et al., 2024). However, this integration presents several challenges that significantly degrade user experience, including high latency in the cascading process, non-spoken style responses, and a lack of paralinguistic communication capabilities.

*Equal contribution
†Corresponding author

Prior work such as Qwen2-Audio(Chu et al., 2024) and SALMONN(Tang et al., 2023) enhances LLMs with speech perception capabilities while relying on external text-to-speech (TTS) toolkits for speech generation. This approach has the potential to leverage paralinguistic information but hardly support duplex speech interaction.

The end-to-end method integrates both speech perception and generation within large speech-language models. SpeechGPT(Zhang et al., 2023) is a speech-text cross-modal conversational model, but it is not real-time due to Chain-of-Modality. AnyGPT(Zhan et al., 2024) is a token-based any-to-any multimodal language model, which can understand and generate speech autoregressively, but the high frame rate of speech tokenizer limits the real-time interaction. VITA(Fu et al., 2024) is a multimodal large language model that processes audio modalities and supports duplex speech interaction by requiring two models used as a monitor or a generator with role switching when the user interrupts. LLama-Omni(Fang et al., 2024) achieves low latency benefits from a streaming vocoder, however it does not support full-duplex interaction. Kyutai introduced Moshi(Défossez et al., 2024), a speech-to-speech conversational model that supports full-duplex spoken dialogue, enabling fluid and seamless interactions.

Speech Interaction Datasets. SpeechInstruct[1] (Zhang et al., 2023) contains 37,969 spoken dialogues based on the chain-of-modality mechanism. However, all speech clips are encoded into discrete units by HuBERT, limiting the exploration of speech representation. AnyInstruct[2] (Zhan et al., 2024) consists of 108,000 spoken dialogues generated by the Azure Text-to-Speech API, featuring 39 different timbres. the datasets mentioned above are entirely in English, making them hardly suitable for Chinese speech interaction research and applications. Additionally, both the scale of the datasets and the diversity of speakers are inadequate for large speech-language models.

Speech Interaction Benchmarks. Speech interaction benchmarks are scarce up to now. AIR-Bench(Yang et al., 2024) includes both foundation and chat benchmarks, featuring a variety of audio types, such as human speech, natural sounds, and music. However, the amount of speech interaction data, particularly for Chinese, is very limited.

3 Ke-SpeechChat Dataset

The success of large language models (LLMs) significantly relies on the availability of large-scale models and datasets. However, to our knowledge, open-source large-scale speech interaction datasets remain unseen, greatly hindering the advancement of speech conversation research. This scarcity can be attributed to two main factors: the high cost associated with constructing speech data and the inherent privacy risks involved.

To effectively construct large-scale and high-quality speech interaction datasets, we explore efficient synthetic data methods by leveraging advanced LLMs and TTS toolkits. To avoid privacy risks, we build a virtual voice library for speech generation, in which voices do not exist in the real world. Additionally, we inject watermarks to indicate that the data is generated by AI and prevent data from misuse.

In constructing the dialogue data, we first focus on creating text dialogue data that accurately reflects the characteristics of spoken language. We then synthesize speech from these textual dialogues. Subsequently, we perform quality assurance and filtering on the synthetic speech.

In this section, we provide a comprehensive overview of the construction process. The collection, rewriting, and post-processing of textual dialogues will be discussed in subsection 3.1, while the steps for converting textual dialogues to speech dialogues will be covered in subsection 3.2.

3.1 Textual Dialogue Data

To synthesize textual dialogue data, we leveraged various entries from open-source datasets, such as IndustryInstruction(Shi et al., 2024), LaMini-instruction (Wu et al., 2023), BELLE (BELLEGroup, 2023; Ji et al., 2023; Wen et al., 2023) (belle1M[3], belle2M[4], and belle3.5M[5]), among others.

While such open-source instruction datasets have been invaluable for various applications, they present challenges that make them inappropriate for direct application in the task of speech interaction. Firstly, many instructions in these datasets involve tasks that are not conducive to speech interaction, such as generating images, writing long articles, or creating structured text. Secondly, the format of these instructions is often too formal and detailed compared to everyday spoken language. For example, a dataset might include an instruction such as "how to install and set up a piece of software or device (e.g., a printer)," whereas in everyday conversation, one would simply ask, "How do I use the printer?" Lastly, both the instructions and responses in these datasets tend to be overly lengthy and contain special characters that cannot be pronounced, such as markdown symbols, underscores, and line breaks. These factors collectively make text-based datasets inadequate for the nuanced and dynamic nature of speech interaction.

[1] https://huggingface.co/datasets/fnlp/SpeechInstruct

[2] https://huggingface.co/datasets/fnlp/AnyInstruct

[3] https://huggingface.co/datasets/BelleGroup/train_1M_CN

[4] https://huggingface.co/datasets/BelleGroup/train_2M_CN

[5] https://huggingface.co/datasets/BelleGroup/train_3.5M_CN

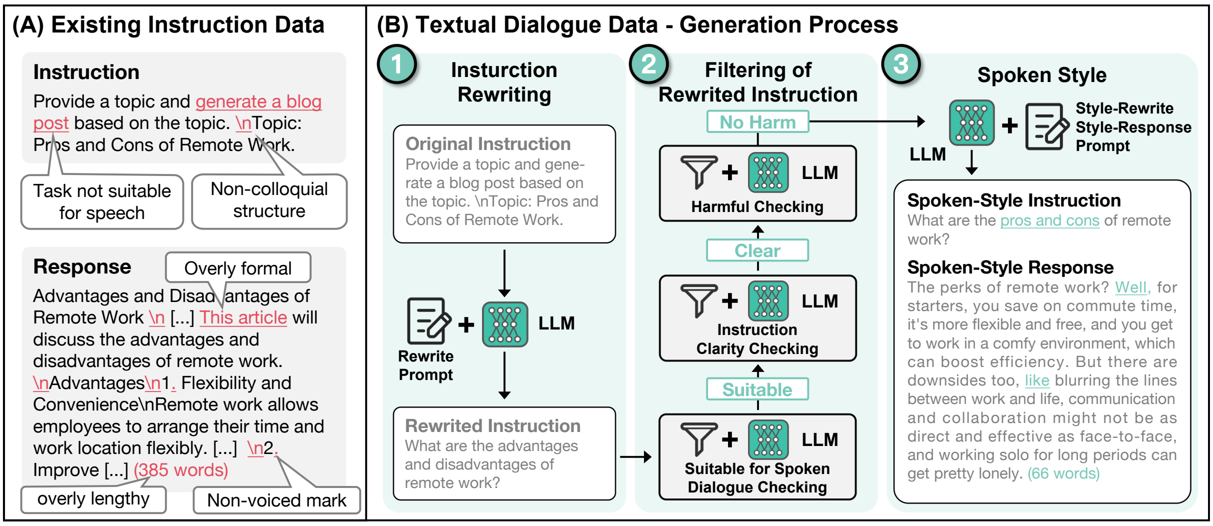

Figure 1: An overview of textual dialogue data process

To address these issues, we implement three critical stages: rewriting instructions, filtering of rewritten instructions, and spoken style post-processing. All three stages are executed using LLMs, with all prompts detailed in Appendix A.1.

3.1.1 Rewriting Instructions

The primary goal of rewriting is to transform instruction tasks to be more appropriate for verbal interactions. Our approach involves designing a specific prompt for LLM to convert the original instruction data into questions that a human might verbally ask.

We discovered that when the LLM is given the complete original instructions, they often preserve the task's original format with only minor rephrasing. This results in outputs that remain ill-suited for conversational purposes.

To tackle this, we implemented a strategy for specific types of tasks, such as classification, summarization, and other directive instructions. We removed the directive sentences from these tasks, leaving only key pieces of information. The LLM was then instructed that these fragments were incomplete and should be used as inspiration to generate new questions creatively.

This approach minimizes the LLM's tendency to adhere too closely to the original directives and encourages the creation of more natural, conversational questions. Consequently, we can effectively convert formal and structured instruction texts into queries better suited for the speech scenario.

3.1.2 Filtering of Rewritten Instructions

Following the pre-rewriting process, we filter the rewritten instructions to ensure they are suitable for spoken interactions, involving three key considerations. First, we assess whether the rewritten instructions are appropriate for verbal communication, excluding tasks that require generating long-form or structured content such as essays, lyrics, or emails. Second, we evaluate the clarity and completeness of each instruction, ensuring they include sufficient context. Instructions that are too vague or lack necessary background information, such as "What is the main content of this article?" are filtered out. Third, we assess the safety of the instructions using our internal system and Qwen2-72B-instruct.

The filtering stage ensures that the dataset mainly consists of instructions that are clear, contextually complete and safe, enhancing their suitability for conversational interactions.

3.1.3 Spoken Style Post-Processing

In the final stage, we use LLM to further modify the selected instructions for enhanced conversational quality and generate corresponding responses in a similarly natural spoken style. The LLM is instructed to adhere to a conversational tone, avoid unpronounceable content, and convert numbers and formula symbols into their verbal equivalents. Additionally, responses are kept under 100 words to ensure that no excessive information is generated in a single response. By following these guidelines, the dataset is refined to better support the training of models for natural and effective speech interactions.

In terms of LLMs, this section uses Qwen2.5-72B-Instruct[6], while sections 3.1.1 and 3.1.2 use Qwen2.5-14B-Instruct[7]. It is worth mentioning that, compared to smaller variants of Qwen2.5, such as Qwen2.5-32B-Instruct or Qwen2.5-14B-Instruct, Qwen2.5-72B-Instruct produced similar instructions but quality improved response, examples can be found in Appendix B.

[6] https://huggingface.co/Qwen/Qwen2.5-72B-Instruct

[7] https://huggingface.co/Qwen/Qwen2.5-14B-Instruct

3.2 Speech Dialogue Data

This section describes the strategy for constructing and ensuring the quality of speech dialogues derived from textual dialogues. We utilized the CosyVoice (Du et al., 2024) model, which supports custom voice profiles, to convert the textual dialogues into speech dialogues. To ensure speaker diversity, we built a large voice library that includes numerous virtual speakers sourced from open-source speech data. To maintain the quality of the synthetic speech dialogues, we transcribed the synthetic audio and calculated the Character Error Rate (CER), filtering the data based on CER to ensure the high quality of dataset.

3.2.1 Voice Library

In this section, we describe the process of constructing a virtual voice library using the premium part of WenetSpeech4TTS dataset. The workflow is illustrated in Figure 2 (A) and (B).

Data Sources. The WenetSpeech4TTS dataset is derived from WenetSpeech (Zhang et al., 2022), which consists of long audio recordings ranging from several minutes to hours, collected from the internet. WenetSpeech4TTS processes these long recordings by applying Voice Activity Detection (VAD) to segment them into shorter clips, while simultaneously measuring the DNSMOS (Reddy et al., 2022) for each segment. These short clips are then merged based on speaker cosine similarity, ensuring that each short clip is spoken by the same individual. However, WenetSpeech4TTS does not perform similarity detection between different short clips within the same long recording. This limitation is critical for our work, as we need to identify multiple segments spoken by the same person to create stable embeddings for individual voice profiles.

Real Speakers. Our first task was to categorize the "Premium" short audio clips a (i.e., those with DNSMOS $\geq$ 4.0) from WenetSpeech4TTS based on their original long audio recordings A. We filtered out long recordings that contained at least ten "Premium" short audio clips, denoted as $A^i = \{a_1^i, a_2^i, \ldots, a_n^i\}$, where i is the index of the long recording and $n \geq 10$ is the number of short clips.

Next, we extracted X-vectors for each short clip in these long recordings using WavLM (Chen et al., 2022). We then calculated the speaker similarity between every pair of short clips within A^i_{premium}. If a long recording contained at least $5\left\lfloor\frac{n}{10}\right\rfloor$ pairs of short clips with a similarity score over 0.97, we considered these short clips to be spoken by the same person. Using this method, we identified over 5000 speakers, which are gender balanced.

Virtual Speakers. For each identified voice profile, we calculated the speaking rate, defined as the average time per character, rounded to the nearest 10 ms interval. We then categorized these profiles based on their speaking rates. We randomly selected one voice and it paired with another same gender voice having the same speaking rate from the remaining profiles. The pair is weighted average to create composite voice profiles, with the specific aim of protecting privacy by generating non-existent, synthetic, virtual voice. The process can be applied to create an unlimited number of composite virtual voices.

These steps ensured that our voice library consisted of high-quality, diverse synthetic voices that are gender-balanced and suitable for various applications in speech synthesis, without corresponding to real individuals.

3.2.2 Speech Synthesis

Based on the aforementioned voice library, we utilized CosyVoice for speech synthesis. CosyVoice is a state-of-the-art text-to-speech (TTS) model known for its high-quality, natural-sounding voice synthesis and flexibility in customizing vocal characteristics. For each dialogue, we randomly selected one user voice and one agent voice for synthesis. To prevent data abuse, all synthetic speech is watermarked using AudioSeal (San Roman et al., 2024). Procedures are shown in Figure 2 (C).

3.2.3 Quality Assurance

To ensure the quality of synthetic dialogues, we transcribe the Chinese parts using Belle-whisper-large-v3-turbo-zh and the English parts using Whisper-large-v3-turbo. The Character Error Rate (CER) is computed for Chinese, while the Word Error Rate (WER) is computed for English. Dialogues with a CER exceeding 5% for Chinese and a WER exceeding 10% for English are dropped to maintain high quality. Processes are shown in Figure 2 (D).

3.3 Details of Ke-SpeechChat

3.3.1 Metadata

All the metadata information is saved to a single JSON file. The id, speakers, genders, texts, audio paths are provided for each dialogue. An example is presented in Appendix C.

3.3.2 Statistics

Detailed statistics are presented in Table 1. Statistics for Chinese and English dialogues are provided separately. The number of Chinese dialogues exceeds 5.1 million, totaling 40,884 hours. While the number of English dialogues exceeds 1.7 million, totaling 19,484 hours. The dataset features gender-balanced speakers, including 40,000 users and 2 agents, for both Chinese and English, The large scale of dialogues and speakers ensures the diversity of the dataset.

3.3.3 Partitions

We randomly split the training data into five subsets of varying sizes: XS, S, M, L, and XL. Each larger subset includes all the data from the smaller subsets, and

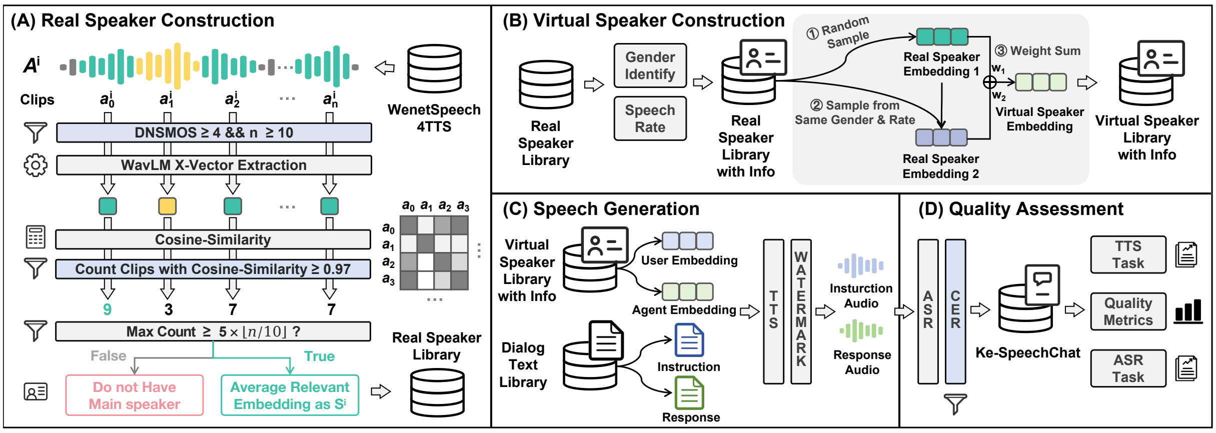

Figure 2: The construction of speech dialogue data. (A) Construction of real speakers. (B) Construction of virtual speakers. (C) Speech generation given textual dialogues and virtual speakers. (D) Quality filtering and assessment.

Table 1: Detailed statistics of Ke-SpeechChat dataset.

Items	Chinese	English
Dialogues	5195448	1772209
Max words per dialogue	300	665
Min words per dialogue	13	10
Mean words per dialogue	122	89
Duration(h)	40884	19484
Max dialogue Duration(s)	92	283
Min dialogue Duration(s)	3	3
Mean dialogue Duration(s)	28	39
User male speakers	21000	
User female speakers	21000	
Agent male speakers	1	
Agent female speakers	1	
Total Speakers	42002	

Table 2: Subsets of the speech interaction training data in Chinese and English.

Subsets	Items	Chinese	English
XS	Duration(h)	1000	500
	Dialogues	127067	45541
S	Duration(h)	4000	2000
	Dialogues	507980	181936
M	Duration(h)	10000	5000
	Dialogues	1270061	455364
L	Duration(h)	20000	10000
	Dialogues	2540907	909528
XL	Duration(h)	40884	19484
	Dialogues	5195448	1772209

all subsets contain the complete set of 42,002 speakers. The details are presented in Table 2.

3.3.4 Quality

To evaluate the quality of Ke-SpeechChat, we compared objective metrics including DNSMOS (Reddy et al., 2022) and UTMOS (Saeki et al., 2022a), with those from other datasets. Additionally, we conducted ASR and TTS tasks for further evaluation.

Quality Metrics. We calculated the DNSMOS P.835 OVRL(Reddy et al., 2022) and UTMOS (Saeki et al., 2022b) scores for the XS training subset to assess the audio quality and speech naturalness of Ke-SpeechChat. These scores were then systematically compared with those obtained from various established large scale speech datasets. It is worth noting that the UTMOS scores for the other datasets were derived from a sample of 100 hours from the corresponding datasets. This comparative analysis provides a valuable perspective on the performance of Ke-SpeechChat in relation to existing datasets, allowing us to highlight its strengths and identify areas for potential improvement.

The table 4 presents a comparative evaluation of various datasets based on two key quality metrics: DNSMOS (P.835 OVRL) and UTMOS, which is indicated by the mean scores along with standard deviations. Ke-SpeechChat achieved the highest DNSMOS score of 3.41±0.14, demonstrating its superior speech quality compared to the studio recordings like MLS. Additionally, Ke-SpeechChat scored 3.47±0.35 in UTMOS, indicating that its naturalness is comparable to that of MLS. These results position Ke-SpeechChat as a leading dataset in both audio quality and naturalness.

Overall, the results illustrate that Ke-SpeechChat outperforms several established datasets in terms of perceived audio quality, as reflected in the DNSMOS scores, while also maintaining competitive naturalness performance in UTMOS evaluations.

ASR Task. We establish speech recognition task based on Whisper-large-v3-turbo to further evaluation the quality of Ke-SpeechChat.

The training set comprises all utterances spoken by

Table 3: Datasets for ASR and TTS tasks from Ke-SpeechChat.

Datasets	Items	Chinese	English
ASR Train	Duration(h)	12165	4525
	Utterances	5195448	1774324
	Male spakers	42000	
	Female spakers	42000	
ASR Test	Duration(h)	11.3	16.2
	Utterances	5222	6977
	Male spakers	100	100
	Female spakers	100	100
TTS Train	Speakers	42002	
	Utterances	20020	20020
	Duration(h)	40.28	34.56
TTS Test	Speakers	100	
	Prompt Utterances	100	100
	Prompt Duration(h)	0.72	0.65
	Test Utterances	600	500
	Test Duration(h)	0.94	0.89

Table 4: DNSMOS scores. DNSMOS refers to DNS-MOS P.835 OVRL. The score for Ke-SpeechChat is computed on the XS training subset. The scores of GigaSpeech, WenetSpeech4TTS, MLS, and Emilia datasets are from (He et al., 2024).

Datasets	DNSMOS↑
GigaSpeech(Chen et al., 2021)	2.52 ± 0.19
WenetSpeech4TTS(Ma et al., 2024)	3.18 ± 0.22
MLS(Pratap et al., 2020)	3.33 ± 0.19
Emilia(He et al., 2024)	3.26 ± 0.14
Ke-SpeechChat(ours)	$\mathbf{3.41 \pm 0.14}$

Table 5: UTMOS scores. The score for Ke-SpeechChat is computed on the XS training subset. The scores of GigaSpeech, WenetSpeech4TTS, MLS, and Emilia datasets are computed by randomly selecting 100 hours respectively.

Datasets	UTMOS↑
GigaSpeech(Chen et al., 2021)	2.71 ± 1.15
WenetSpeech4TTS(Ma et al., 2024)	2.88 ± 0.55
MLS(Pratap et al., 2020)	$\mathbf{3.69 \pm 0.60}$
Emilia(He et al., 2024)	2.16 ± 0.75
Ke-SpeechChat(ours)	3.47 ± 0.35

users in the S subset, totaling 3,291 hours. Since the utterances are recorded in a clean acoustic environment, noise from MUSAN (Snyder et al., 2015) is added with a probability of 0.2 and an SNR ranging from 10 to 50. Additionally, speed perturbation and SpecAugment are applied with a probability of 0.5 to enhance robustness.

The training configuration includes a learning rate of 1e-7, utilizing AdamW optimization. All parameters are fine-tuned, and the training epoch is set to 1. The code and configure are at `https://github.com/shuaijiang/Whisper-Finetune`.

We construct ASR test set, which synthesized by CosyVoice, with details shown in Table 3. All speakers in test set are unseen in the training set. Each audio segment has been reviewed by professional annotators to ensure high transcription quality. Additionally, we also adopt AISHELL-1 and LibriSpeech as our test sets.

The results of speech recognition are presented in Table 6. We compare the performance of Whisper-large-v3-turbo[8], Belle-whisper-large-v3-turbo-zh[9], and our KeASR using character error rate (CER) and word error rate (WER) for Chinese and English, respectively. Despite having less training data, KeASR demonstrates highly competitive performance, particularly on the KeASR test-zh and Librispeech test sets. The results validate the high quality of the Ke-SpeechChat dataset.

TTS Task. We evaluated the performance of Ke-SpeechChat on the TTS task based on CosyVoice.

The training set included all speakers from the Ke-SpeechChat dataset, with 10 Chinese and 10 English utterances randomly selected from each speaker to ensure balanced timbre representation. This resulted in a total of 40,040 utterances, with durations of 40.28 hours for Chinese and 34.56 hours for English.

For the test set, we utilized 100 virtual speakers not included in the training set, with each generating one prompt in both Chinese and English using CosyVoice. These speakers are required to leverage zero-shot capabilities to synthesize 6 utterances in Chinese and 5 in English. Additionally, SeedTTS test-zh and test-en come from the DiDiSpeech(Guo et al., 2021) and the Common Voice(Ardila et al., 2019) respectively, are adopted as test sets. Details of training and test sets are shown in Table 3.

The training configuration includes a learning rate of 1e-5, utilizing AdamW optimization. All parameters are fine-tuned, and the training epoch is set to 1.

Character error rate (CER), word error rate (WER), UTMOS and speaker similarity (SIM) are adopted for evaluate the performance of TTS, with results presented in Table 7. Compared to CosyVoice, our KeTTS outperforms in both CER and UTMOS, while achieving comparable results in SIM. This indicates that Ke-SpeechChat has enhanced both the generation accuracy and sound quality of the base model.

Overall, the high quality of the Ke-SpeechChat dataset is verified based on the quality metrics, including DNSMOS and UTMOS, as well as the ASR and TTS tasks.

[8] `https://huggingface.co/openai/whisper-large-v3-turbo`

[9] `https://huggingface.co/BELLE-2/Belle-whisper-large-v3-turbo-zh`

Table 6: Results of ASR. Whisper refers to Whisper-large-v3-turbo, and Belle refers to Belle-whisper-large-v3-turbo-zh, and test-clean refers to LibriSpeech test-clean, test-other refers to LibriSpeech test-other.

Models	Training Data(h)	KeASR test-zh CER↓	KeASR test-en WER↓	aishell-1 CER↓	test-clean WER↓	test-other WER↓
Whisper(Radford et al., 2023)	-	7.43	13.34	8.64	4.21	5.98
Belle(BELLEGroup, 2023)	11,200	4.92	**13.17**	**3.07**	3.67	7.55
KeASR(ours)	3,291	**4.11**	13.32	6.24	**3.52**	**5.85**

Table 7: Results of zero-shot TTS. CosyVoice refers to CosyVoice-300M-Instruct.SeedTTS-zh refers to SeedTTS test-zh, SeedTTS-en refers to SeedTTS test-en.

Models	KeTTS test-zh			KeTTS test-en			SeedTTS-zh		SeedTTS-en	
	CER↓	UTMOS↑	SIM↑	WER↓	UTMOS↑	SIM↑	WER↓	SIM↑	WER↓	SIM↑
SeedTTS(Anastassiou et al., 2024)	-	-	-	-	-	-	1.12	0.796	2.25	0.762
CosyVoice(Du et al., 2024)	6.86	3.51	0.752	25.77	4.30	0.768	3.44	0.729	4.19	0.622
KeTTS(ours)	**6.28**	**3.54**	**0.754**	**25.15**	**4.33**	**0.773**	*3.20*	0.728	*3.63*	*0.623*

4 KE-Omni

This section presents the details of our large speech language model, KE-Omni. The model architecture of KE-Omni illustrated in Figure 3, it comprises three main components: a speech encoder, a large language model (LLM), and a speech decoder. Given the user‘s speech instruction, KE-Omni is designed to generate high quality text and speech response seamlessly.

4.1 Speech Encoder

We adopt the encoder of Whisper-large-v3[10](Radford et al., 2023), a widely used multilingual speech recognition model, as our speech encoder. Whisper is known for its robust performance across diverse languages, making it suitable for our application. A lightweight speech adapter facilitates speech-text modal alignment, connecting the speech encoder to the LLM.

The speech encoder processes each second of audio into 50 frames of features. The speech adapter is then employed to further compress the length of the speech feature sequence, aligning the speech modality with the LLM. We utilize a compression ratio of 5 in our speech adapter, meaning that each second of speech is ultimately converted into 10 frames of features. This enhances processing speed and reduces latency of LLM without compromising quality.

Throughout the entire training process, the parameters of the speech encoder are frozen, except for the speech adapter. This approach preserves the encoder's robust speech representation capabilities while allowing the adapter to learn the necessary transformations for effective speech-text modal alignment with the LLM.

[10]https://huggingface.co/openai/whisper-large-v3

4.2 Large Language Model

We utilize the state-of-the-art open-source LLaMA(Dubey et al., 2024) model as our large language model (LLM), which exhibits strong reasoning capabilities across multiple languages, including both Chinese and English. In KE-Omni, the LLM takes the concatenation of prompt text embeddings and the speech representations generated by the speech encoder as input. This integration allows the LLM to leverage contextual information from both text and speech modality. It then autoregressively generates a text response based on the user's speech instructions. To balance the performance and efficiency, we prefer LLaMA-3.1-8B-Instruct variant as our LLM.

4.3 Speech Decoder

The speech decoder maps text response from LLM into corresponding speech signals in a non-autoregressive manner, playing a crucial role to speech interaction. It consists of three key components: a duration predictor, a speech unit generator and a unit-based vocoder.

Similar to (Zhang et al., 2023) and (Fang et al., 2024), we adopt the pretrained HuBERT(Hsu et al., 2021) model to extract continuous representations of the speech, and convert the representations into discrete cluster indices using a K-means model.

Before generating the speech response, the duration of each text token is first predicted by the duration predictor. The duration predictor is a transformer-based model trained on the word-level timestamps extracted by Whisper. According to the duration information, the text token sequence is then upsampled to match the length of the target audio frame sequence. The duration predictor is trained in advance and kept frozen during the training process of KE-Omni.

A transformer-based speech unit generator is then performed to obtain the discrete speech units sequence in an autoregressive manner. To improve prediction

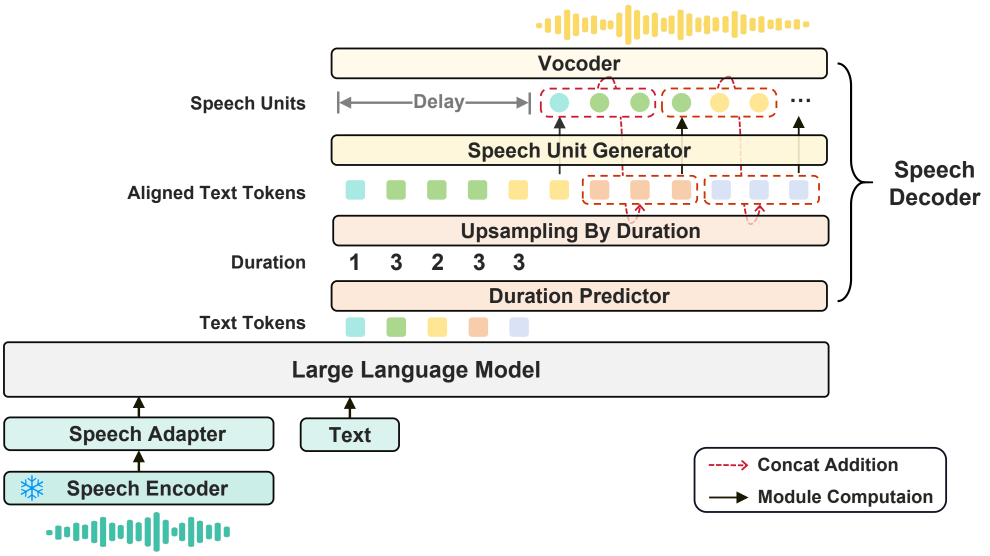

Figure 3: Model architecture of KE-Omni.

speed, we utilized a chunk-based autoregressive approach, predicting speech units chunk by chunk. Given chunk size C and the length of target speech unit sequence T, the embeddings of ith text token and the jth speech unit are concatenated as input, where $j = i - C$ and $i \in [C, T]$. A zero-embedding of speech units is used at the first timestep. To ensure the quality of speech unit generation, we introduce a delay of N steps between the extended text token sequence and the speech unit sequence. Finally, the unit-based vocoder, specifically HiFi-GAN, is performed to synthesis the waveform from these units. The HiFi-GAN vocoder is trained for the agent speakers in advance and kept frozen during the training process of KE-Omni.

5 Experiments

5.1 Setups

KE-Omni use LLaMA-3.1-8B-Instruct(Fang et al., 2024) as the LLM backbone. The duraion predictor is optimized with mean square error (MSE) loss, and the token duration obtained by Whisper is used as training target. We distribute the world-level duration evenly to individual characters, and subsequently merge them to obtain the timestep of each token used in LLM. In addition, the duration predictor take the hidden states of the last transformer layer in the LLM as the input. In speech decoder, the chunk size of the autoregressive process C is set to 5. The hyper-parameter of delay step N is set to the sum of the timesteps of the first **three** text tokens in each sample. For discrete speech units, we employ a K-means model to convert speech representations extracted by HuBERT into 4000 clusters. On this basis, separate unit-based HiFi-GAN vocoders are trained for two agent speakers respectively.

To explore the impact of data size on model performance, we trained KE-Omni models on various sub-

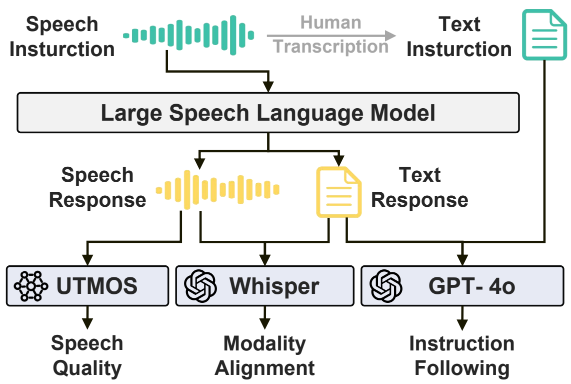

Figure 4: Evaluation methodology for large speech language model, focusing on three key capabilities: Speech-to-Text Instruction-Following(S2TIF), Modality Alignment and Speech Quality.

sets of the Ke-SpeechChat dataset separately. Each model underwent a two-stage training process: LLM fine-tuning and speech decoder training. In the first stage, all dialogues in each subset were used to train the speech adaptor and enhance the reasoning capabilities of the LLM for audio input. In the second stage, the LLM is frozen, and the dialogues were separated by agent speaker and used to train speech decoders for each agent speaker individually. Both of two stages are trained for 2 epochs, utilizing AdamW optimizer. The peak learning rate is set to 1e-4 in the first stage, while 2e-4 is used in the second stage.

5.2 Development and Test Sets

To evaluate the speech interaction ability, we construct the development and test sets, which are illustrated in Table 10. The chat-dev set consists of 2,754 spoken dialogues, totaling 14.6 hours in Chinese and 11.2 hours

Table 8: Performances of speech language models on Ke-SpeechChat chat-test.

Models	Data	Chinese					English				
		S2TIF			Modal Align	Quality	S2TIF			Modal Align	Quality
	Scale	Content↑	Style↑	Length	CER↓	UTMOS↑	Content↑	Style↑	Length	WER↓	UTMOS↑
Qwen2-Audio(Chu et al., 2024)	-	3.50	3.12	239.38	-	-	2.74	2.90	117.06	-	-
LLaMA-Omni(Fang et al., 2024)	-	-	-	-	-	-	2.85	3.70	46.8	4.65	3.95
KE-Omni(ours)	XS	3.04	3.79	90.91	38.68	2.35	2.79	3.61	72.77	56.03	2.24
	S	3.17	3.86	88.74	15.58	2.65	2.87	3.64	67.93	18.62	2.93
	M	3.89	4.24	89.29	6.67	3.29	3.45	3.96	70.02	4.41	4.08
	L	3.95	4.28	89.54	5.26	3.73	3.57	4.00	69.00	4.20	4.26
	XL	**4.12**	**4.34**	88.10	**5.18**	**3.89**	**3.61**	**4.00**	68.70	**4.15**	**4.45**

Table 9: Results of VoiceBench. The results of other models are from (Chen et al., 2024).

Models	AlpacaEval↑	CommonEval↑	OpenBookQA↑	IFEval↑	AdvBench↑
Qwen2-Audio(Chu et al., 2024)	3.74	3.43	49.45	**26.33**	**96.73**
LLaMA-Omni(Fang et al., 2024)	3.70	**3.46**	27.47	14.87	11.35
Mini-Omni2(Xie and Wu, 2024)	2.32	2.18	26.59	11.56	57.50
Moshi(Défossez et al., 2024)	2.01	1.60	25.93	10.12	44.23
KE-Omni-L(ours)	3.74	3.38	49.89	15.61	*57.69*
KE-Omni-XL(ours)	**3.78**	3.31	**59.78**	*15.81*	55.58

Table 10: Overview of the Ke-SpeechChat chat-dev and chat-test in Chinese and English.

Datasets	Items	Chinese	English
chat-dev	Duration(h)	14.6	11.2
	Dialogues	1454	1300
	User male speakers	50	50
	User female speakers	50	50
	Agent male speakers	1	
	Agent female speakers	1	
chat-test	Duration(h)	12.4	16.0
	Dialogues	1485	1460
	User male speakers	100	100
	User female speakers	100	100
	Agent male speakers	1	
	Agent female speakers	1	

in English. The chat-test set comprises 2,945 spoken dialogues, totaling 9.5 hours in Chinese and 10.5 hours in English. All speakers, except the two agent speakers in these sets are unseen in the training data. All data have been meticulously reviewed by professional annotators to guarantee the quality.

Additionally, we also evaluate our models using VoiceBench(Chen et al., 2024) , a benchmark that assesses voice dialogue systems on their general knowledge, instruction-following ability, and safety compliance. VoiceBench incorporates both synthetic and real spoken instructions to simulate diverse speaker styles, environmental conditions, and content variations.

5.3 Evaluation

In this section, we detail the methodologies and processes employed to evaluate the speech language models, as illustrated in Figure 4. We evaluate three key capabilities: Speech-to-Text Instruction-Following (S2TIF) similar to (Fang et al., 2024), modality alignment based on Character Error Rate (CER) and Word Error Rate(WER), and speech quality using UTMOS (Saeki et al., 2022b).

Speech-to-Text Instruction-Following (S2TIF). The S2TIF metric uses GPT-4 to score the response text based on transcribed instructions. It evaluates two dimensions: content and style, each rated from 1 to 5, respectively assessing whether the response covers the instructions and whether the style suits voice interaction. We used the same prompt as (Fang et al., 2024), shown in Appendix A.2. Additionally, the average length of the responses is computed to show the models' length preferences and to illustrate the difficulty of modality alignment in speech output.

Modality Alignment. We adopt the Word Error Rate (WER) and Character Error Rate (CER) metrics to evaluate the intelligibility of the speech language model's audio output and its alignment with the text output. Specifically, we used the Whisper-large-v3 model for transcription and applied text normalization to standardize the representation of numbers and symbols, removing punctuation marks before calculating the metrics.

Speech Quality. We adopted the UTokyo-SaruLab Mean Opinion Score (UTMOS) prediction system developed by (Saeki et al., 2022b) to assess the quality of the generated speech. This system generates a UTMOS score for the audio based on the naturalness and overall quality of the speech, with higher scores indicating better sound quality.

5.4 Baseline Systems

We include the following speech-language models as baseline systems: LLama-Omni, SpeechGPT and Qwen2-Audio. LLama-Omni and SpeechGPT are large speech-language model that supports both speech input and output. For they are limited to English, only the English portion of Ke-SpeechChat chat-test is used to evaluate and compare their performance. In contrast, Qwen2-Audio, as a general audio understanding model, supports both Chinese and English. Since it only support the Speech-to-Text Instruction-Following (S2TIF) task, the S2TIF portion of Ke-SpeechChat chat-test is used for performance evaluation.

5.5 Results and Analysis

Based on the evaluation methodology mentioned in Section 5.3, performances of our KE-Omni model and baseline systmes are assessed and shown in Table 8.Since both LLaMA-Omni and SpeechGPT utilize female agents for their speech responses, we selected the female agent of KE-Omni for comparison.

For the Speech-to-Text Instruction-Following (S2TIF) task, KE-Omni significantly outperforms the baseline systems, even when trained on smaller datasets. This indicates that the quality of speech dialogues is as important as the quantity.

In the modal alignment task for English, KE-Omni achieves competitive performance comparable to other baseline systems with smaller English data scales, suggesting that reliable modal alignment capabilities require a substantial amount of data. Since the second stage of training is conducted separately for each agent speaker, the amount of dialogues from the female agent accounts for roughly half of the total dialogues. It is worth mentioning that the length of responses from LLaMA-Omni tend to be shorter than those from KE-Omni, which results in less challenging speech-text modal alignment.

In VoiceBench, KE-Omni achieves competitive performance compared to other models like Qwen2-Audio, LLaMa-Omni, Mini-Omni2 and Moshi across most evaluation dimensions. For the AdvBench safety bench, which evaluates the model's ability to refuse inappropriate requests. By analyzing the results, we observe that KE-Omni tends to refuse requests without triggering the specific keywords used to compute the refusal rate.

As the scale of training data increases, the performance of KE-Omni continues to improve. Notably, since the speech decoder is trained from scratch, the initial modal alignment capability is weak but improves rapidly as the dataset reaches the L subsets.

6 Limitations

The textual dialogues in Ke-SpeechChat originate from BELLE and other open-source datasets, and then have been rewritten by Qwen2.5-72B-Instruct. Consequently, they may contain factual inaccuracies or outdated information due to the hallucination tendencies of large language models (LLMs).

The speech dialogues are generated by CosyVoice and may still exhibit pronunciation flaws despite quality filtering. The synthetic speech data in our dataset is in clean acoustic environment, free from noise and reverberation, some complex scenarios may require the addition of noise and reverberation for better robust.

The Ke-SpeechChat dataset currently contains only single-turn conversations, and work on constructing multi-turn speech dialogues is underway.

7 Conclusions

In this work, we introduced KE-Omni, a large speech language model designed for seamless speech interaction. We also presented Ke-SpeechChat, a large-scale speech interaction dataset comprising 7 million English and Chinese speech dialogues, featuring 42,002 speakers and totaling over 60,000 hours of audio. We believe both the model and dataset will significantly enhance the resources available for speech language models research, addressing the current scarcity of diverse and high-quality data. By providing a comprehensive speech interaction model and dataset construction methodology, we aim to facilitate the development and evaluation of advanced speech-language models, ultimately contributing to the growth of this field.

8 Acknowledgments

We would like to thank the KE Team members for the support and beneficial discussions. We would also like to express our special gratitude to Liang Feng and Yang Fang for providing the computing power for synthesizing the data.

References

[Anastassiou et al.2024] Philip Anastassiou, Jiawei Chen, Jitong Chen, Yuanzhe Chen, Zhuo Chen, Ziyi Chen, Jian Cong, Lelai Deng, Chuang Ding, Lu Gao, Mingqing Gong, Peisong Huang, Qingqing Huang, Zhiying Huang, Yuanyuan Huo, Dongya Jia, Chumin Li, Feiya Li, Hui Li, Jiaxin Li, Xiaoyang Li, Xingxing Li, Lin Liu, Shouda Liu, Sichao Liu, Xudong Liu, Yuchen Liu, Zhengxi Liu, Lu Lu, Junjie Pan, Xin Wang, Yuping Wang, Yuxuan Wang, Zhen Wei, Jian Wu, Chao Yao, Yifeng Yang, Yuanhao Yi, Junteng Zhang, Qidi Zhang, Shuo Zhang, Wenjie Zhang, Yang Zhang, Zilin Zhao, Dejian Zhong, and Xiaobin Zhuang. 2024. Seed-tts: A family of high-quality versatile speech generation models.

[Ardila et al.2019] Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M Tyers, and Gregor Weber. 2019. Common voice: A massively-multilingual speech corpus. *arXiv preprint arXiv:1912.06670*.

[BELLEGroup2023] BELLEGroup. 2023. Belle: Be everyone's large language model engine. `https://github.com/LianjiaTech/BELLE`.

[Chen et al.2021] 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. *arXiv preprint arXiv:2106.06909*.

[Chen et al.2022] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei. 2022. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518, October.

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

[Chu et al.2024] 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*.

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

[Du et al.2024] Zhihao Du, Qian Chen, Shiliang Zhang, Kai Hu, Heng Lu, Yexin Yang, Hangrui Hu, Siqi Zheng, Yue Gu, Ziyang Ma, Zhifu Gao, and Zhijie Yan. 2024. Cosyvoice: A scalable multilingual zero-shot text-to-speech synthesizer based on supervised semantic tokens.

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

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

[Fu et al.2024] Chaoyou Fu, Haojia Lin, Zuwei Long, Yunhang Shen, Meng Zhao, Yifan Zhang, Xiong Wang, Di Yin, Long Ma, Xiawu Zheng, et al. 2024. Vita: Towards open-source interactive omni multimodal llm. *arXiv preprint arXiv:2408.05211*.

[Guo et al.2021] Tingwei Guo, Cheng Wen, Dongwei Jiang, Ne Luo, Ruixiong Zhang, Shuaijiang Zhao, Wubo Li, Cheng Gong, Wei Zou, Kun Han, et al. 2021. Didispeech: A large scale mandarin speech corpus. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6968–6972. IEEE.

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

[Hsu et al.2021] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units.

[Huang et al.2024] Rongjie Huang, Mingze Li, Dongchao Yang, Jiatong Shi, Xuankai Chang, Zhenhui Ye, Yuning Wu, Zhiqing Hong, Jiawei Huang, Jinglin Liu, et al. 2024. Audiogpt: Understanding and generating speech, music, sound, and talking head. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

[Ji et al.2023] Yunjie Ji, Yong Deng, Yan Gong, Yiping Peng, Qiang Niu, Lei Zhang, Baochang Ma, and Xiangang Li. 2023. Exploring the impact of instruction data scaling on large language models: An empirical study on real-world use cases. *arXiv preprint arXiv:2303.14742*.

[Ma et al.2024] Linhan Ma, Dake Guo, Kun Song, Yuepeng Jiang, Shuai Wang, Liumeng Xue, Weiming Xu, Huan Zhao, Binbin Zhang, and Lei Xie. 2024. Wenetspeech4tts: A 12,800-hour mandarin tts corpus for large speech generation model benchmark.

[OpenAI2024] OpenAI. 2024. Gpt-4o:the new flagship model that can reason across audio, vision, and text in real time. Blog post.

[Pratap et al.2020] Vineel Pratap, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve, and Ronan Collobert. 2020. Mls: A large-scale multilingual dataset for speech research. *arXiv preprint arXiv:2012.03411*.

[Radford et al.2023] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

[Reddy et al.2022] Chandan KA Reddy, Vishak Gopal, and Ross Cutler. 2022. Dnsmos p.835: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors. In *ICASSP 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE.

[Saeki et al.2022a] Takaaki Saeki, Detai Xin, Wataru Nakata, Tomoki Koriyama, Shinnosuke Takamichi, and Hiroshi Saruwatari. 2022a. Utmos: Utokyo-sarulab system for voicemos challenge 2022. *arXiv preprint arXiv:2204.02152*.

[Saeki et al.2022b] Takaaki Saeki, Detai Xin, Wataru Nakata, Tomoki Koriyama, Shinnosuke Takamichi, and Hiroshi Saruwatari. 2022b. Utmos: Utokyo-sarulab system for voicemos challenge 2022.

[San Roman et al.2024] Robin San Roman, Pierre Fernandez, Hady Elsahar, Alexandre D´efossez, Teddy Furon, and Tuan Tran. 2024. Proactive detection of voice cloning with localized watermarking. *ICML*.

[Shi et al.2024] Xiaofeng Shi, Lulu Zhao, Hua Zhou, Donglin Hao, and Yonghua Lin. 2024. Industryinstruction.

[Snyder et al.2015] David Snyder, Guoguo Chen, and Daniel Povey. 2015. Musan: A music, speech, and noise corpus. *arXiv preprint arXiv:1510.08484*.

[Tang et al.2023] Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. 2023. Salmonn: Towards generic hearing abilities for large language models. *arXiv preprint arXiv:2310.13289*.

[Wen et al.2023] Cheng Wen, Xianghui Sun, Shuaijiang Zhao, Xiaoquan Fang, Liangyu Chen, and Wei Zou. 2023. Chathome: Development and evaluation of a domain-specific language model for home renovation. *arXiv preprint arXiv:2307.15290*.

[Wu et al.2023] Minghao Wu, Abdul Waheed, Chiyu Zhang, Muhammad Abdul-Mageed, and Alham Fikri Aji. 2023. Lamini-lm: A diverse herd of distilled models from large-scale instructions. *CoRR*, abs/2304.14402.

[Xie and Wu2024] Zhifei Xie and Changqiao Wu. 2024. Mini-omni2: Towards open-source gpt-4o with vision, speech and duplex capabilities. *arXiv preprint arXiv:2410.11190*.

[Yang et al.2024] 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. *arXiv preprint arXiv:2402.07729*.

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

[Zhang et al.2022] Binbin Zhang, Hang Lv, Pengcheng Guo, Qijie Shao, Chao Yang, Lei Xie, Xin Xu, Hui Bu, Xiaoyu Chen, Chenchen Zeng, Di Wu, and Zhendong Peng. 2022. Wenetspeech: A 10000+ hours multi-domain mandarin corpus for speech recognition.

[Zhang et al.2023] Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. 2023. Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities. *arXiv preprint arXiv:2305.11000*.

A PROMPTS

A.1 Prompts for Text Dialog Generation

Prompt for Rewriting of Instructions (Model: Qwen2.5-14B-Instruct)

You are an expert in crafting questions. You will receive a piece of information.

##Task:Using this piece of information as inspiration, please formulate a question that a human might ask verbally to an AI.

##Note: The information piece may be incomplete and lack semantic quality. Please do not rely much on its content; feel free to be creative.

Guidelines for the question:
1. Clear References: Ensure all necessary context and background are included in the question. Avoid using vague terms like "this" or "that".
2. Verbal Resolution: The question should be answerable verbally and should not require generating articles, images, or other content.
3. Suitable for AI: The question should be appropriate for asking an AI but should not explicitly mention AI.
4. Single Question: Include only one question, and it does not need to fully incorporate the content of the information piece.
5. Privacy: Do not include any private information such as phone numbers, websites, or social media handles.
6. Difficulty Level: For factual questions, aim for topics that are common and relatable, avoiding obscure or niche subjects. For emotional questions, provide specific details.

##`Information Piece`: {instruction}
Please output only the question you have formulated:

Prompt for Filtering of Rewritten Instructions (Model: Qwen2.5-14B-Instruct)

You will receive an instruction data from the text QA dataset. Your task is to determine if this instruction is suitable for generating an audio version to train a large speech language model.

Judgment Criteria:
- Inappropriate: The response corresponding to the instruction is not suitable for oral Q&A. That is, it requires creation or revision of written text, especially in a structured format. For example, composing music, revising written content, writing poems, short articles, lyrics, code, emails, etc.
- Appropriate: The response corresponding to the instruction is suitable for an oral answer. Examples include: factual Q&A, seeking advice, brainstorming, translating sentences, querying the source of a short poem, providing an answer to a couplet, quoting famous sayings, etc.

##`Instruction`: {instruction}

You only need to return a JSON {`"is_suitable_for_speech"`: <bool>}.

You will receive an instruction data from the text QA dataset. Your task is to determine if the instruction is clear enough.

Judgment Criteria:
- Not Clear Enough: The instruction includes ambiguities such as unclear references, vagueness, difficulty in understanding, lack of specificity, or insufficient context, requiring further inquiries to clarify the content of the issue.
- Clear: The instruction can be understood and processed without further inquiries. If the instruction includes names of people or places, it should be of common knowledge, unique, or does not require additional explanation.

##`Instruction`: {instruction}
Note: The above are the only instructions provided, without any context.

You only need to return a JSON {`"clear_enough"`: <bool>}.

Prompt for Spoken Style Post-Processing (Model: Qwen2.5-72B-Instruct)

You will receive a command. Your tasks are: 1. Rewrite the command in a conversational tone ("instruction"), and 2. Generate a conversational response ("response"). Both parts will be converted to speech using TTS, so ensure they are suitable for spoken language. Please follow these guidelines:

- Conversational Style: Natural and fluent, with sufficiently detailed responses that are not overly brief but do not exceed one hundred words.
- Avoid any content that cannot be spoken, such as underscores, brackets, line breaks, markdown symbols, enumerations, URLs, etc.
- Convert any Arabic numerals and special symbols into their spoken form in Enlish, such as fifty-five, degrees Celsius, fractions, addition and subtraction, etc.
- Tone: The instruction should mimic the attitude and manner of human, and the response should be polite and courteous.

If the original command is clear, ensure your rewritten instruction maintains the original meaning; if the command is ambiguous, clarify it in your rewrite.
After rewriting the instruction, please hide all information from the original command and ensure your response is based solely on your rewritten conversational instruction.

##`Command`: {Command}
Output only JSON {`"instruction"`:"<str>", `"response"`:"<str>"}

A.2 Prompts for Evaluation

Prompt for Speech-to-Text Instruction-Following (Model: GPT-4o)

I need your help to evaluate the performance of several models in the speech interaction scenario. The models will receive a speech input from the user, which they need to understand and respond to with a speech output. Your task is to rate the model's responses based on the provided user input transcription [Instruction] and the model's output transcription [Response]. Please evaluate the response from two perspectives: content and style, and provide a score for each on a scale of 1 to 5.

Content (1-5 points):
1 point: The response is largely irrelevant, incorrect, or fails to address the user's query. It may be off-topic or provide incorrect information.
2 points: The response is somewhat relevant but lacks accuracy or completeness. It may only partially answer the user's question or include extraneous information.
3 points: The response is relevant and mostly accurate, but it may lack conciseness or include unnecessary details that don't contribute to the main point.
4 points: The response is relevant, accurate, and concise, providing a clear answer to the user's question without unnecessary elaboration.
5 points: The response is exceptionally relevant, accurate, and to the point. It directly addresses the user's query in a highly effective and efficient manner, providing exactly the information needed.

Style (1-5 points):
1 point: The response is poorly suited for speech interaction, possibly including structured elements like lists or being overly complex, disjointed, or difficult to understand.
2 points: The response is somewhat suitable but may be too long, too short, or awkwardly phrased, making it less effective in a speech interaction context.
3 points: The response is generally suitable for speech interaction, but it may have minor issues with length, clarity, or fluency that detract slightly from the overall effectiveness.
4 points: The response is well-suited for speech interaction, with appropriate length, clear language, and a natural flow. It is easy to understand when spoken aloud.
5 points: The response is perfectly suited for speech interaction. It is the ideal length, highly clear, and flows naturally, making it easy to follow and understand when spoken.

Below are the transcription of user's instruction and models' response:
`[Instruction]`: {instruction}
`[Reponse]`: {response}

After evaluating, please output the scores in JSON format: {"`content`": content score, "`style`": style score}.
You don't need to provide any explanations.

B COMPARATION AMONG VARIENTS OF QWEN2.5

The impact of using different versions of the Qwen2.5 model on Spoken Style Post-Processing is presented in this appendix, as shown in the table below.

Command When was the Smallpox vaccine first used?
Qwen2.5-14B-Instruct Output
{"instruction":"Can you tell me when the smallpox vaccine was first used on a patient?", "response":"The smallpox vaccine was first used on a patient in **1796**. It was developed by Edward Jenner, who tested it on an eight-year-old boy named James Phipps."}
Qwen2.5-32B-Instruct Output
{"instruction":"Can you tell me when the smallpox vaccine was first used on a patient?","response":"The smallpox vaccine was first used on a patient in **1796** by Dr. Edward Jenner. He vaccinated an eight-year-old boy named James Phipps with material from cowpox blisters, which is a disease that affects cows but causes only mild symptoms in humans. This pioneering act marked the beginning of vaccination as we know it today."}
Qwen2.5-72B-Instruct Output
{"instruction": "When did doctors first use the smallpox vaccine on a patient?","response":"The smallpox vaccine was first used on a patient in **seventeen ninety-six**. Dr. Edward Jenner vaccinated a young boy named James Phipps, and this marked the beginning of vaccination as we know it today."}

Command Create a list of descriptive words for an elephant.
Qwen2.5-14B-Instruct Output
{"instruction":"Tell me some words you'd use to describe an elephant.", "response":"An elephant is typically described as massive, powerful, and intelligent. It's also known for its long trunk and large ears, which are quite distinctive. Elephants are often seen as gentle giants, showing strong family bonds and complex social behaviors."}
Qwen2.5-32B-Instruct Output
{"instruction":"Can you share some descriptive words that come to mind when you think about an elephant?","response":"Sure, when thinking about an elephant, words like large, grey, and intelligent come to mind. They're also known for their strong family bonds, long trunks, and gentle nature despite their massive size."}
Qwen2.5-72B-Instruct Output
{"instruction":"Can you think of some descriptive words to paint a picture of an elephant?","response":"Sure! I would describe an elephant as massive, majestic, and gentle. They have a distinctive trunk, large flapping ears, and sometimes impressive tusks. Elephants are also known for their intelligence, social behavior, and incredible memory."}

Command If there are 19 marbles in a bowl, with 5 of them being yellow and the rest divided between blue and red marbles in a 3:4 ratio, how many more red marbles are there compared to yellow marbles?
Qwen2.5-14B-Instruct Output
{"instruction":"There are nineteen marbles in a bowl. Five of them are yellow, and the rest are divided between blue and red marbles in a three to four ratio. How many more red marbles are there compared to yellow marbles?", "response":"Let's break it down. There are five yellow marbles, so there are fourteen marbles left for blue and red. The ratio of blue to red is three to four, which means for every seven marbles, three are blue and four are red. So, out of the fourteen marbles, ten are red and four are blue. Therefore, there are **five** more red marbles than yellow ones."}

Qwen2.5-32B-Instruct Output
{"instruction":"Imagine you have a bowl with fifty-five marbles. Five of these are yellow, and the rest are split between blue and red marbles in a three to four ratio. Can you tell me how many more red marbles there are compared to the yellow ones?","response":"Sure, let's figure this out. There are fifty-five marbles in total, with five being yellow. This leaves fifty marbles to be divided between blue and red in a three to four ratio. That means for every seven marbles, three are blue and four are red. So, there are twelve blue marbles and sixteen red marbles. Therefore, there are **eleven** more red marbles than yellow ones."}
Qwen2.5-72B-Instruct Output
{"instruction":"If there are nineteen marbles in a bowl, with five of them being yellow and the rest divided between blue and red marbles in a three to four ratio, how many more red marbles are there compared to yellow marbles?","response":"There are fourteen marbles that are either blue or red. Given the three to four ratio, there are six blue marbles and eight red marbles. So, there are **three** more red marbles than yellow marbles."}

Command What is the difference between kinetic and potential energy?
Qwen2.5-14B-Instruct Output
{"instruction":"Can you explain the difference between kinetic and potential energy?", "response":"Sure! Kinetic energy is the energy an object possesses due to its motion, like a ball rolling down a hill. Potential energy, on the other hand, is stored energy that an object has due to its position or state, such as a ball at the top of a hill, ready to roll."}
Qwen2.5-32B-Instruct Output
{"instruction":"Can you explain the difference between kinetic energy and potential energy to me?" , "response":"Sure! Kinetic energy is the energy an object has due to its motion. For example, a moving car or a flowing river both have kinetic energy. On the other hand, potential energy is the energy stored in an object due to its position or state. Think of a book on a high shelf or a stretched rubber band; they both have potential energy because they can do work if their position changes."}
Qwen2.5-72B-Instruct Output
{"instruction":"What is the difference between kinetic and potential energy?" , "response":"Sure! Kinetic energy is the energy an object has because it's moving. The faster something moves, the more kinetic energy it has. Potential energy, on the other hand, is stored energy due to an object's position or state. For example, a book on a high shelf has potential energy because it could fall and release that energy. Does that help?"}

C Dialogue Format

A Dialogue Example


```
{
    "id": "tldr_tldr_60583",
    "speaker": {
        "SPK1486m": {"role": "user", "gender": "male"},
        "agentFemale": {"role": "agent", "gender": "female"} },
    "audio": {
        "channel": 2,
        "duration": 40.18975056689342,
        "sample_rate": 22050 },
    "channel": [
        {"channel_index": 0, "language": "en"},
        {"channel_index": 1, "language": "en"} ],
    "dialog": [
        {   "channel": 0,
            "speaker": "SPK1486m",
            "text": "I'm really worried about my girlfriend ......",
            "start": 0.0,
            "end": 8.937687074829933,
            "audio_path": "tldr_tldr_60583/tldr_tldr_60583_0_mark.wav"
        },
        {   "channel": 1,
            "speaker": "agentFemale",
            "text": "It's totally understandable to feel anxious, ......",
            "start": 8.937687074829933,
            "end": 40.18975056689342,
            "audio_path": "tldr_tldr_60583/tldr_tldr_60583_1_mark.wav"
        } ]
}
```