OmniFlatten: An End-to-end GPT Model for Seamless Voice Conversation

Qinglin Zhang, Luyao Cheng, Chong Deng, Qian Chen, Wen Wang,
Siqi Zheng, Jiaqing Liu, Hai Yu, Chaohong Tan, Zhihao Du, Shiliang Zhang
Tongyi Lab
{qinglin.zql, shuli.cly, dengchong.d, tanqing.cq, w.wang}@alibaba-inc.com

Abstract

Full-duplex spoken dialogue systems significantly surpass traditional turn-based dialogue systems, as they allow simultaneous bidirectional communication, closely mirroring human-human interactions. However, achieving *low latency* and *natural interactions* in full-duplex dialogue systems remains a significant challenge, especially considering human conversation dynamics such as interruptions, backchannels, and overlapping speech. In this paper, we introduce a novel End-to-End GPT-based model **OmniFlatten** for full-duplex conversation, capable of effectively modeling the complex behaviors inherent to natural conversations with low latency. To achieve full-duplex conversation capabilities, we propose a multi-stage post-training scheme that progressively adapts a text large language model (LLM) backbone into a speech-text dialogue LLM, capable of generating text and speech in real time, without modifying the architecture of the backbone LLM. The training process comprises three stages: modality alignment, half-duplex dialogue learning, and full-duplex dialogue learning. In all training stages, we standardize the data using a flattening operation, which enables unifying the training methods and the GPT backbone across different modalities and tasks. Our approach offers a simple modeling technique and a promising research direction for developing efficient and natural end-to-end full-duplex spoken dialogue systems. Audio samples of dialogues generated by OmniFlatten can be found at this web site [1].

1 Introduction

Traditional turn-based spoken dialogue systems only support half-duplex communication, that is, the communication is conducted bidirectionally between user and system (assistant) but *not simultaneously*. These systems, while effective in many real-world applications, often fall short when they come to handle interruptions, backchannels, and overlapping speech, which reflect the spontaneous nature of human-human conversation. Conversely, full-duplex spoken dialogue systems allow *simultaneous* two-way communication, closely mirroring human-human conversations. Full-duplex spoken dialogue systems facilitate **more natural and efficient communications than turn-based dialogue systems**, by *speaking, listening, and thinking at the same time*. However, achieving **low latency** and **natural interactions** in full-duplex dialogue systems remains a significant challenge.

Recent efforts in developing spoken dialogues systems have been driven by advancements in large language models (LLMs) and can be roughly categorized into *collaborative systems* and *end-to-end (E2E) systems*. Collaborative systems interface LLM-based dialogue modules with external ASR or TTS modules for speech understanding and generation. For example, Qwen-audio (Chu et al., 2024) takes speech input, outputs text, and converts them to verbal responses via TTS; Mini-Omni2 (Xie and Wu, 2024b) investigated a command-based interruption approach to enable full-duplex conversation capabilities. In contrast, some E2E systems (Zhang et al., 2023; Xie and Wu, 2024a; Fang et al., 2024; Zeng et al., 2024) directly model speech-to-speech dialogues based on speech-text multimodal models; yet most of these models are turn-based dialogue models and do not support full-duplex conversation. Notable recent progresses in developing E2E full-duplex spoken dialogue systems include SyncLM (Veluri et al., 2024) and Moshi (Défossez et al., 2024). Specifically, Moshi models multiple streams of user's speech input and Assistant's text and speech output in parallel, simplifying modeling full-duplex dialogues. However, this parallel framework is not natively supported by GPT models, hence requires sophisticated designs. Regarding SyncLM,

[1] https://omniflatten.github.io/

similar to our approach, it also learns to predict interleaved chunks of user and assistant speech units to acquire real-time full-duplex conversation capabilities. However, different from the simple deduplication strategy by SyncLM to improve the model's semantic capabilities, we explore explicit text token prediction to achieve this goal.

To address the challenges of achieving *natural interactions* and *low latency* in full-duplex spoken dialogue systems, we introduce a novel E2E GPT-based model **OmniFlatten** for full-duplex speech conversation. OmniFlatten is capable of effectively learning the complex behaviors inherent to natural conversations and facilitates human-like interactions with low latency. We propose a **multi-stage progressive post-training scheme** to adapt a text LLM backbone into a robust speech-text dialogue model. The multi-stage post-training process begins with supervised multi-task fine-tuning of the text LLM backbone, using ASR and TTS tasks, to achieve speech-text modality alignment and obtain a multimodal LLM capable of accurately interpreting and generating both speech and text. After obtaining the speech-text LLM, we fine-tune it using interleaved and flatten dialogues through progressive stages of half-duplex dialogue learning and full-duplex dialogue learning. Notably, throughout all training stages, we standardize dialogue data with this flattening operation, which enables unifying the training methods and the GPT backbone across different modalities and tasks.

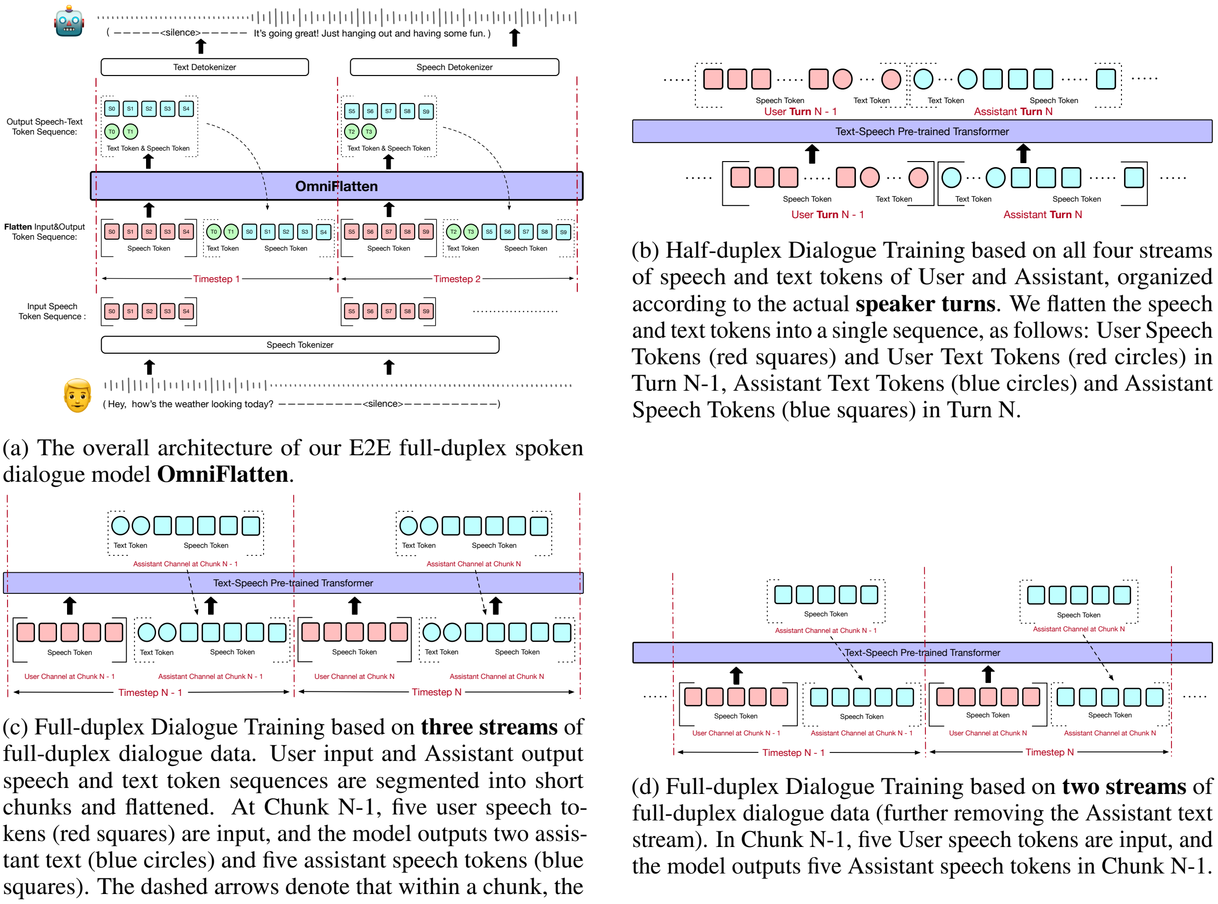
Our contributions can be summarized as follows:

- We propose a novel End-to-End GPT-based model **OmniFlatten**, capable of effectively modeling the complex behaviors inherent to natural human-like dialogues, with low latency. We propose a **multi-stage post-training scheme** that successfully adapts a text-based foundation LLM into a robust speech-text dialogue model, by performing supervised multi-task fine-tuning based on ASR and TTS for speech-text modality alignment, then conducting fine-grained chunking of speech and text streams of dialogues and flattening them into a single sequence to progressively train the model to acquire half-duplex and full-duplex conversation capabilities. Notably, **OmniFlatten does not make any structure modifications to the GPT model, nor relies on computationally intensive pre-training**. Our approach offers a simple modeling technique and a promising research direction for developing efficient and natural E2E full-duplex dialogue systems.
- We design **a data synthesis and simulation pipeline to synthesize full-duplex spoken dialogue data** for training and evaluation.
- We evaluate the dialogue quality generated by OmniFlatten using high-performing LLMs as evaluators, and evaluate the turn-taking performance, including assistant turn-taking and user turn-taking, as well as the run-time efficiency. The results demonstrate that the dialogues generated by OmniFlatten exhibit reasonable quality, with both modality alignment and half-duplex learning stages improving the model's full-duplex dialogue capabilities.

2 Related Work

In recent years, the field of spoken dialogue models has seen significant advancements driven by the prominent technological advancements in LLMs and particularly speech-text multimodal models. Qwen-audio2 (Chu et al., 2024) and SALMONN (Tang et al., 2024) support voice input but only output text; hence, they rely on external TTS systems to synthesize speech output. SpeechGPT (Zhang et al., 2023), LauraGPT (Du et al., 2024b), Mini-Omni (Xie and Wu, 2024a), LLaMA-Omni (Fang et al., 2024) and GLM-4-Voice (Zeng et al., 2024) are capable of comprehending both speech and text input and generating output in both speech and text; however, they are predominantly turn-based dialogue models and do not support full-duplex conversation. LSLM (Ma et al., 2024) explores full-duplex scenarios by integrating TTS models to perform end-to-end modeling of turn-taking task, thereby enabling the model to continuously listen while speaking and stop at any moment.

Recent progresses in developing end-to-end full-duplex spoken dialogue systems include dGSLM (Nguyen et al., 2023), VITA (Fu et al., 2024), the open-sourced Moshi (Défossez et al., 2024), and SyncLM (Veluri et al., 2024). VITA implements a duplex scheme with two separate modules: one module generates responses to user queries while another module continuously monitors environmental inputs to selectively provide updated interactions. Moshi models multiple streams of user's speech input and system's text and speech output in parallel, allowing for conceptually and practically simple handling of full-duplex dialogues. This approach, compared to its predeces-


(------<silence>------ It's going great! Just hanging out and having some fun.)
Text Detokenizer
Speech Detokenizer
Output Speech-Text Token Sequence:
S0 S1 S2 S3 S4
T0 T1
Text Token & Speech Token
S5 S6 S7 S8 S9
T2 T3
OmniFlatten
Flatten Input&Output Token Sequence:
Speech Token
Text Token
Timestep 1
Timestep 2
Input Speech Token Sequence :
Speech Tokenizer
(Hey, how's the weather looking today? ---------<silence>----------)
Speech Token
Text Token
User Turn N - 1
Assistant Turn N
Text-Speech Pre-trained Transformer
Assistant Channel at Chunk N - 1
Assistant Channel at Chunk N
User Channel at Chunk N - 1
User Channel at Chunk N
Timestep N - 1
Timestep N

(a) The overall architecture of our E2E full-duplex spoken dialogue model **OmniFlatten**.

(b) Half-duplex Dialogue Training based on all four streams of speech and text tokens of User and Assistant, organized according to the actual **speaker turns**. We flatten the speech and text tokens into a single sequence, as follows: User Speech Tokens (red squares) and User Text Tokens (red circles) in Turn N-1, Assistant Text Tokens (blue circles) and Assistant Speech Tokens (blue squares) in Turn N.

(c) Full-duplex Dialogue Training based on **three streams** of full-duplex dialogue data. User input and Assistant output speech and text token sequences are segmented into short chunks and flattened. At Chunk N-1, five user speech tokens (red squares) are input, and the model outputs two assistant text (blue circles) and five assistant speech tokens (blue squares). The dashed arrows denote that within a chunk, the model appends the predicted Assistant text and speech tokens into input to complete autoregressive decoding.

(d) Full-duplex Dialogue Training based on **two streams** of full-duplex dialogue data (further removing the Assistant text stream). In Chunk N-1, five User speech tokens are input, and the model outputs five Assistant speech tokens in Chunk N-1.

Figure 1: The overall architecture of our **OmniFlatten** and the three dialogue learning stages.

sors, offers a more robust solution for managing simultaneous voice inputs and outputs, thereby facilitating more natural and efficient interactions. However, this parallel framework is not supported natively by GPT-based models and hence requires sophisticated designs such as acoustic delay and inner monologue (Défossez et al., 2024).

Another related concurrent work is SyncLM (Veluri et al., 2024), which, like our model, employs time-chunking for full-duplex interaction but uses a deduplication strategy for audio sequences. This strategy, while simplifying modeling, leads to errors in audio reconstruction. In contrast, our approach maintains the integrity of discretized audio tokens, thereby preserving audio quality. Additionally, we enhance full-duplex dialogue efficacy through direct modality alignment using ASR and TTS, and progressive learning. Our model concurrently generates both text and audio tokens, unlike SyncLM which only produces speech tokens.

3 Methodology

Figure 1a depicts the overall architecture of our E2E full-duplex dialogue model OmniFlatten. We employ audio tokenizer to discretize each user input and assistant output speech stream of a dialogue into a discrete speech token sequence. We then interleave the speech token sequences with the corresponding text token sequences and flatten them into a single sequence. Our approach employs a multi-stage progressive training process to adapt a text-based LLM into a robust end-to-end full-duplex spoken dialogue model, through modality alignment and dialogue learning.

3.1 Audio Tokenization And Detokenization

We adopt the speech tokenizer used in CosyVoice[2] (Du et al., 2024a; An et al., 2024) since through supervision from multilingual ASR, this speech tokenizer converts speech to semantic

[2] https://github.com/FunAudioLLM/CosyVoice/tree/main

tokens, hence benefiting speech understanding and content consistency for speech generation. The tokenizer utilizes an encoder and a Vector Quantization (VQ) layer to discretize audio into speech tokens with a single codebook of 4096 codes. For detokenization, we adopt the same Optimal-transport Conditional Flow Matching model (OT-CFM) used in CosyVoice. OT-CFM transforms the speech token sequence into Mel spectrogram, which is used to generate the final audio output with the HifiGAN vocoder (Kong et al., 2020). Prior works (Lipman et al., 2023; Tong et al., 2024) show that OT-CFM outperforms diffusion probabilistic models with simpler gradients, easier training, and faster generation.

3.2 Modality Alignment

We start with post-training a pre-trained text LLM backbone to obtain a speech-text multimodal model for speech understanding and speech generation. We use Qwen2-0.5B[3] as our base model due to its small model size for low compute resource consumption and its competitive performance for models with this small size. We perform supervised fine-tuning (SFT) using paired speech-text data for ASR and TTS tasks. For each speech-text pair $< S_{seq}, T_{seq} >$, we construct ASR training samples as $[ASR][SOS]$S_seq$[EOS][SOT]$T_seq$[EOT]$, and TTS training samples as $[TTS][SOT]$T_seq$[EOT][SOS]$S_seq$[EOS]$, where [ASR] and [TTS] denote ASR and TTS task IDs; [SOS], [EOS], [SOT], [EOT] are special tokens denoting the Start and the End of the Speech sentence or the Text sentence, respectively. The speech-text multimodal model is used for the subsequent Dialogue Learning.

3.3 Dialogue Learning

Building upon the speech-text multimodal model, we conduct dialogue learning in three stages, including half-duplex dialogue training using both speech and text streams of turn-based dialogue data, and then full-duplex dialogue training based on fine-grained chunking and alignment of speech and text sequences. Specifically, during full-duplex dialogue training, we first remove the input text stream and use the remaining three streams for training, and then further remove the output text stream and use the remaining two streams for training, in order to gradually eliminate the dependence on text information, focus on speech-to-speech generation.

3.3.1 Half-duplex Dialogue Training

Half-duplex dialogue agents are special and simpler cases of full-duplex dialogue agents, where human and assistant take turns to speak and there is no overlapping speech, that is, during the speaker's turn, the listener is fully silent. Since there is no overlapping speech in the ASR and TTS data used for learning modality alignment, half-duplex dialogue training is more consistent with the aligned multimodal model than full-duplex dialogue training which requires the model to handle turn-taking, backchannel, and overlapping speech. Adopting the concept of curriculum learning, we first conduct half-duplex dialogue training then full-duplex dialogue training. During half-duplex dialogue training, we train the model to essentially perform ASR on the speech tokens of user and obtain the text content, next generate a textual response for assistant based on the text content of user, and then predict the speech tokens for assistant's textual response by basically executing a TTS task. This pattern is extended to multiple turns of a dialogue, as illustrated in Figure 1b.

3.3.2 Full-duplex Dialogue Training

Training on Three-Stream Data In order to meet the real-time requirements of a full-duplex conversational agent, we remove the user text stream from the four-stream data and train the model with the remaining three steams. In order to handle overlapping speech, we introduce *chunking* and *relaxed speech-text token alignment* based on the chunks; in this way, we do not require strict token-level alignment between speech and text. Specifically, to prepare the training data for this stage, we chunk the speech and text token sequences of dialogues with fixed chunk sizes and then interleave the three-stream data and *flatten* them into a single sequence for training, following the order of *input speech*, *output text*, and *output speech*, as depicted in Figure 1c. The chunking and flattening operation enables our model to stream input speech tokens and output text and speech tokens in real time. Notably, given the higher efficiency of text, the size of the text chunks is generally smaller than that of the speech chunks. In this work, we set the text chunk size to 2 tokens and the speech chunk size to 10 tokens. This approach ensures that the output text does not excessively precede the speech content,

[3] https://huggingface.co/Qwen/Qwen2-0.5B

thereby both minimizing the discrepancies with the aforementioned 4-stream data format and maximizing preservation of TTS capabilities. After the end of the text content, we use the special character $silent\_text\_token$ to pad the text stream and use $silent\_speech\_token$ to pad the silent regions of the output speech stream.

Training on Two-Stream Data In order to further reduce latency and eliminate dependence on intermediate text and hence focus on speech-to-speech generation, we further remove the assistant output text stream and retain only the user input and assistant output speech streams. Figure 1d depicts this training process on chunked two-stream data.

4 Experiments

4.1 Data

Modality Alignment Dataset The objective of the training stage for modality alignment (Section 3.2) is to help the model learn the correspondence between speech tokens and text tokens and enable the model to acquire two key capabilities: ASR and TTS. To achieve this goal, we combine a set of TTS and ASR datasets that consist of both open-source and proprietary data. The open-source datasets comprise both Mandarin and English data, including Aishell-3 (Yao Shi, 2015), LibriTTS (Zen et al., 2019), TED-LIUM (Hernandez et al., 2018), VoxPopuli (Wang et al., 2021), Librispeech (Panayotov et al., 2015), MLS (Pratap et al., 2020) and Wenetspeech (Zhang et al., 2022). Notably, we only use 15% of the training set of Wenetspeech that has high quality transcripts, for the ASR task. We incorporate several proprietary ASR and TTS datasets. In sum, the datasets for the speech-text modality alignment include about 100K hours of audio. Among all data, approximate 30% are open-source data and 70% are proprietary data.

Simulated Voice Chat Dataset For constructing the voice chat data for Dialogue Learning (Section 3.3), we design a data synthesis and simulation pipeline to synthesize dialogue data, as shown in Figure 2. Firstly, we collect a substantial amount of high-quality open-source textual dialogue data for subsequent speech synthesis, including Alpaca (Peng et al., 2023), Moss (Sun et al., 2024), BelleCN (Ji et al., 2023), and ultraChat (Ding et al., 2023). We then use heuristic rules and filter out samples that are inappropriate for TTS, for example, samples that comprise a high percentage of non-text elements such as code and mathematical expressions, samples with more than 200 words in English or Chinese, and also samples containing rare or unusual symbols. In the end, we retain approximately 390K multi-turn sessions of turn-based dialogues (half-duplex dialogues).

Secondly, we use CosyVoice to synthesize speech from turn-based dialogue texts. Regarding timbre selection, the voice timbre for the user channel was sampled from Librispeech (Panayotov et al., 2015) and 3D-Speaker (Zheng et al., 2023b) datasets, whereas a fixed timbre was employed for the assistant channel.

Thirdly, upon obtaining the audio for each turn, we designed a simulation approach to place the audio at specific times, thereby emulating the user-assistant interaction dynamics typically seen in real human-machine interactions. This method encompasses several key situations: For case (1), the user finishes asking a question, followed by the assistant's immediate response. For case (2), the user tries to interrupt, leading the assistant to stop speaking abruptly, we make the user's next turn begin before the system finishes playing the assistant's current audio response. In these cases, the assistant's audio is truncated. In our approach, we ensure that at least one-quarter of the assistant's audio is completed and then we uniformly select the point after one-quarter of assistant's audio to truncate. For case (3), Once the assistant has finished speaking, it maintains silence while waiting for the user's speaking. We configure the user's audio to delay playback until the conclusion of the assistant's audio, followed by inserting a silence period adhering to a normal distribution with mean of 0.85 and standard deviations of 0.6.

Finally, to mimic real-world scenarios of the user audio channel, we also sample background noise from the MUSAN noise dataset (Snyder et al., 2015) and add noise into the user audio channel with the signal-to-noise ratio (SNR) between 15 dB and 30 dB. Based on this data synthesis and simulation pipeline, we generate a total of **2,000 hours of multi-channel spoken dialogue data**. We randomly select 1% of this dataset as the validation set and another 1% as the test set, and the remaining data are used as the training set.

4.2 Training and Inference Setup

We use Qwen2-0.5B (Yang et al., 2024) as the base model. During the modality alignment training phase, the maximum sequence length is set to 1024 tokens. In the dialogue learning phase, the maxi-

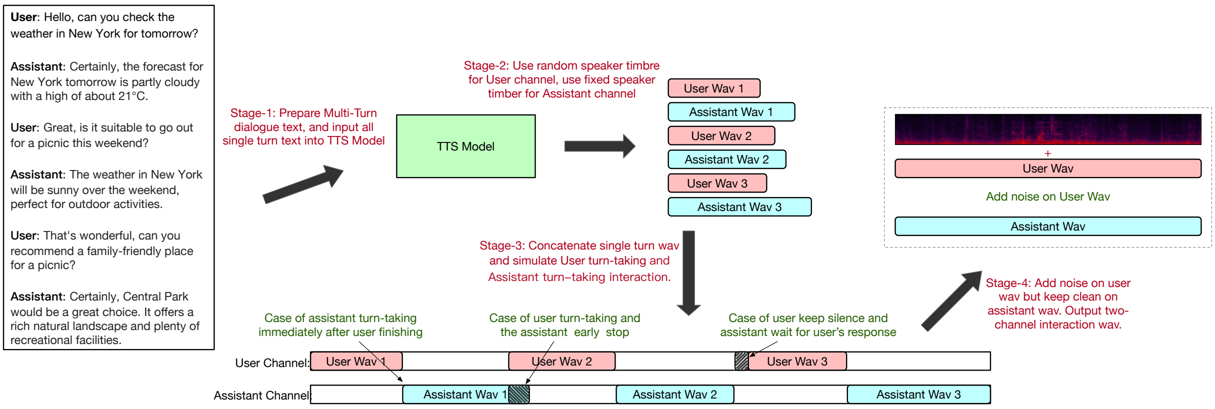

User: Hello, can you check the weather in New York for tomorrow?
Assistant: Certainly, the forecast for New York tomorrow is partly cloudy with a high of about 21°C.
User: Great, is it suitable to go out for a picnic this weekend?
Assistant: The weather in New York will be sunny over the weekend, perfect for outdoor activities.
User: That's wonderful, can you recommend a family-friendly place for a picnic?
Assistant: Certainly, Central Park would be a great choice. It offers a rich natural landscape and plenty of recreational facilities.
Stage-1: Prepare Multi-Turn dialogue text, and input all single turn text into TTS Model
TTS Model
Stage-2: Use random speaker timbre for User channel, use fixed speaker timber for Assistant channel
User Wav 1
Assistant Wav 1
User Wav 2
Assistant Wav 2
User Wav 3
Assistant Wav 3
+
User Wav
Add noise on User Wav
Assistant Wav
Stage-3: Concatenate single turn wav and simulate User turn-taking and Assistant turn–taking interaction.
Stage-4: Add noise on user wav but keep clean on assistant wav. Output two-channel interaction wav.
Case of assistant turn-taking immediately after user finishing
Case of user turn-taking and the assistant early stop
Case of user keep silence and assistant wait for user's response
User Channel
User Wav 1
User Wav 2
User Wav 3
Assistant Channel
Assistant Wav 1
Assistant Wav 2
Assistant Wav 3

Figure 2: The simulation process of dialogue learning data.

Table 1: ASR results on Librispeech and Wenetspeech test datasets. For Librispeech, we report the WER metric. For Wenetspeech, we report the CER metric.

Model	Librispeech		Wenetspeech	
	test-clean	test-other	test-meeting	test-net
Whisper-S	3.13	7.37	25.62	16.66
Whisper-L	1.82	3.5	18.87	10.48
VITA	8.14	18.4	12.15	16.53
OmniFlatten	7.91	19.21	26.1	19.0

Table 2: TTS results on LibriTTS and AIShell-3. For LibriTTS, we utilize whisper-large-v3 model to perform recognition on TTS outputs and assess the WER metric. For AIShell-3, we deploy the paraformer-zh model to recognize TTS results and evaluate CER metric.

Model	LibriTTS(WER)	AIShell-3(CER)
Original	2.66	2.52
ChatTTS	8.32	3.87
CosyVoice	2.89	3.82
OmniFlatten	4.51	4.46

mum sequence length is extended to 8192 tokens. We use the standard cross-entropy loss as our training objective in all stages. Additionally, during the dialogue learning phase, we apply loss masking on the user channel, as we observe that this operation enhances the stability of model training, probably due to the presence of noisy audio input in the user channel. We employ the AdamW optimizer with a weight decay of 0.1, β_1 of 0.9, and β_2 of 0.95. The maximum learning rate is set to 2e-05, with warm-up and cosine-decay. We train the model with 5 epochs, with the best model selected based on loss on the validation set. Each batch contains 100M tokens.

During inference, to obtain the assistant textual response prediction from the model, we use the ground truth user channel speech in the test set as the fixed speech input and alternately fill in the predicted assistant speech and text with the fixed speech chunk and text chunk sizes.

4.3 Evaluations

Evaluation of ASR and TTS Performance after Modality Alignment. The Modality Alignment training stage (Section 3.2) aims to help the model learn the correspondence between speech tokens and text tokens and acquire ASR and TTS capabilities; hence, we evaluate the effectiveness of this training stage by evaluating the ASR and TTS performance of the multimodal-aligned model.

For TTS evaluation, we use the model to generate speech tokens based on the input text, which are then synthesized into audio. We follow the experimental setup outlined in the CosyVoice paper: the synthesized audio is subsequently recognized using the Whisper-Large-V3 model (Radford et al., 2023) [4] and Paraformer-zh model (Gao et al., 2023) [5], for English and Chinese datasets, respectively, and the resulting ASR transcripts are scored against the input text to compute Word Error Rate (WER) for English datasets and Character Error Rate (CER) for Chinese datasets. WER/CER could measure the synthesis accuracy and robustness of a model's TTS capabilities, and also reflect the audio quality to a significant extent.

For ASR evaluation, we discretize the input speech into discrete speech tokens, then use the model to decode the speech tokens into text. We compare the speech-text aligned multimodal model after Modality Alignment (denoted by **OmniFlat-**

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

[5] https://huggingface.co/funasr/paraformer-zh

ten) with Whisper-small [6] and Whisper-Large-V3 models. In addition, we have conducted a comparison of ASR capabilities with those reported by VITA. For TTS evaluation, We conduct comparisons between our approach and two competitive TTS systems, namely, ChatTTS[7] and CosyVoice.

As shown in Table 1 and 2, OmniFlatten demonstrates considerable performance in both ASR and TTS tasks. These results indicate that the Modality Alignment training stage effectively transitions the unimodality text-based LLM to speech-text multimodal model with reasonable speech comprehension and generation capabilities for the following dialogue learning. We also evaluate the audio quality of the TTS output from OmniFlatten after the Modality Alignment training stage, using UTMOS (Saeki et al., 2022). On the LibriTTS test set, OmniFlatten achieved a UTMOS score of 4.27, demonstrating high audio quality. On the AIShell-3 test set, OmniFlatten achieved a UTMOS score of 3.77, indicating relatively good audio quality. The UTMOS scores on AIShell-3 are lower than those on LibriTTS, which might be attributed to the larger proportion of English data compared to Chinese data in our modality alignment training data. These UTMOS results demonstrate that the audio generated by OmniFlatten has a relatively high level of naturalness and audio quality.

Evaluation of Full-duplex Conversation Capability after Modality Alignment, Half-duplex Dialogue Learning, and Full-duplex Dialogue Learning. As described in Section 3.3.2, the training stage of full-duplex dialogue learning on three-stream data helps the model acquire full-duplex dialogue capacity and this model generates both speech and text. Prior works (Zheng et al., 2023a) demonstrate that competitive text-based LLMs can be reliable evaluators for various natural language generation tasks, as scores assigned by LLM evaluators on generated text show high correlations with human evaluations. Therefore, we evaluate the full-duplex dialogue capabilities of OmniFlatten, by prompting a competitive text LLM to evaluate the **semantics of dialogues** and assign a score on the predicted assistant response from the model after training on three-stream data. We also compared three models: LLaMA-Omni, Moshi, and GLM-Voice in both Chinese and English chat capabilities. To investigate the effects of adding audio modality on the chat capability of models, we examined the text-only chat performance of Qwen2-0.5B-Instruct and Qwen2-7B-Instruct.

The scoring mechanism involves designing a specific prompt and utilizing a competitive text LLM, QWen-max model [8], to rate the responses from the model on a scale of 1 to 10 points. The specific prompt we use for LLM scoring is detailed in Appendix A. We carefully design the prompt to evaluate *fluency* and *coherence* of the predicted assistant response. Given that the full-duplex model may be disrupted by user speech in multi-turn dialogues, which could affect the assessment of chat capabilities, we exclusively focus on assessing the single-turn chat performance of the model. Given the dual output of speech and text by the model, our evaluation encompasses scoring the text output as well as converting the speech output into text using an ASR model specific to the output language, similar to the methodology described in the TTS evaluation section. The outcomes are reported separately as LLM Score (Text) and LLM Score (Speech).

To analyze the impact of the modality alignment training stage (Section 3.2) and the half-duplex dialogue learning stage (Section 3.3) on the full-duplex conversation capability of OmniFlatten, we compare the LLM scores on the predicted assistant responses from the following models: (1) Trained directly on three-stream data (**OmniFlatten directly 3-stream**). (2) Trained with modality alignment and full-duplex three-stream dialogue data (**OmniFlatten 3-stream w/o half-duplex**). (3)Trained with modality alignment, half-duplex dialogue data, and full-duplex three-stream dialogue data (**OmniFlatten 3-stream full process**). (4) Trained with modality alignment, half-duplex dialogue data, and full-duplex three-stream dialogue data and then on two-stream data (**OmniFlatten 2-stream full process**).

The results, as displayed in Table 3, reveal that the multi-stage training approach enhances performance when comparing several 3-streaming models with OmniFlatten. Specifically, the OmniFlatten directly 3-stream model, lacking speech-text alignment, scores lowest; the absence of a half-duplex stage further diminishes its chat capabilities. The 2-stream experiments show a significant performance drop when models generate only speech

[6] https://huggingface.co/openai/whisper-small

[7] https://github.com/2noise/ChatTTS

[8] https://help.aliyun.com/zh/model-studio/developer-reference/use-qwen-by-calling-api

Model	Params	En		Zh	
		Score(Text)	Score(Speech + ASR)	Score(Text)	Score(Speech + ASR)
Qwen2-0.5B-Instruct	0.5B	6.75	-	6.98	-
Qwen2-7B-Instruct	7B	8.37	-	8.09	-
LLaMA-Omni	8B	6.01	5.50	4.17	3.89
Moshi	7B	3.92	3.46	-	-
GLM-Voice	9B	6.97	6.40	7.02	6.69
OmniFlatten directly 3-stream	0.5B	2.99	2.59	4.94	3.95
OmniFlatten 3-stream w/o half-duplex	0.5B	3.89	3.54	5.25	4.76
OmniFlatten 3-stream full process	0.5B	4.88	3.92	5.6	5.15
OmniFlatten 2-stream full process	0.5B	-	2.19	-	3.06
Ground Truth Response	-	7.65	-	6.83	-

Table 3: Performance results of speech and text chat capabilities in both Chinese and English test datasets.

Models	Assistant Turn-taking Acc@K 1/5/10/25 (%)	Average Assistant Turn-taking Response Time (ms)	User Turn-taking Acc @K 1/5/10/25 (%)	Average User Turn-taking Response Time (ms)
Moshi	2.9/18.8/38.5/55.1	553	0.0/6.2/14.8/45.7	753
OmniFlatten	20.6/53.6/ 66.3/71.7	193	10.9/30.9/41.8/51.8	287

Table 4: Assistant Turn-taking and User Turn-taking accuracy at the k-th token (Acc@K) and Response Time.

outputs, with these models occasionally incorporating unclear semantic content in speech endings, adversely affecting scores. When comparing models that produce both speech and text, the performance in the speech modality consistently falls short of that in the text modality, indicating a slight compromise in accuracy despite the ability to synchronize speech with text outputs.

OmniFlatten exhibits superior chat performance on the English test set compared to Moshi, which frequently resorts to counter-questioning the user instead of providing direct responses. In contrast, compared with the 7B parameter LLaMA-Omni, OmniFlatten lagging scores on the English dataset while leading in the Chinese dataset. The proficiency in Chinese for LLaMA-Omni predominantly stems from the speech-text alignment process, despite the absence of Chinese data in its dialogue learning process. Comparatively, against the 9B parameter GLM-Voice model, OmniFlatten's indicators are relatively underperforming, a situation we attribute to potential data leakage during GLM-Voice's training process involving our test set. Additionally, comparing with the Qwen2-0.5B-Instruct results, the inclusion of the speech modality results in a noticeable decline in chat capabilities.

Turn-taking Performance and Runtime Efficiency. To assess the naturalness of OmniFlatten's full-duplex interactions, we evaluate the assistant's ability to respond promptly after the user finishes speaking and to cease speaking when the user attempts to interrupt. In real-world scenarios, timely responses by machines prompt specific human interactive behaviors. For instance, if a machine fails to answer a question quickly, a human might repeat the question. Likewise, if a machine does not stop talking when interrupted, the human may try to interrupt again. Our current testing approach involves using a predefined user speech input while setting a 1.5-second threshold. We consider it a failure of user turn-taking or assistant turn-taking if, exceeded by either, the machine fails to respond (i.e., start or stop answering) within the threshold. Based on this principle, we define the following metrics:

Assistant Turn-taking Acc@k: This metric is defined as whether assistant correctly predicts a non-silence token at the k-th token after the end of a semantically meaningful speech token from user, indicating that assistant starts speaking.

User Turn-taking Acc@k: This metric is defined as whether assistant correctly outputs a silence token at the k-th token after a semantically meaningful speech token is input from user while assistant is speaking. This metric indicates that assistant has successfully responded to the turn-taking attempt from user by stopping assistant's speech and listening.

The evaluation results are shown in Table 4. We compared the response times of OmniFlatten and Moshi during interactions involving user turn-taking and assistant turn-taking scenarios. It was observed that OmniFlatten demonstrates a shorter average response time compared to Moshi in both user and assistant turn-taking contexts. However, in terms of the user turn-taking metric, although OmniFlatten leads in response time and accuracy, neither model exhibits a particularly high success rate in achieving user turn-taking within 25 tokens.

5 Conclusion

In this paper, we introduce an end-to-end full-duplex spoken dialogue model OmniFlatten based on synthesizing full-duplex spoken dialogue data and designing a multi-stage progressive training paradigm for modality alignment and dialogue learning. Our approach offers a straightforward full-duplex modeling scheme as it does not change the architecture of the backbone text-based LLM nor relies on computationally intensive pre-training. Empirical evaluations show that the proposed approach is promising for developing end-to-end models to handle full-duplex interactions.

In future work, we plan to refine the data synthesis pipeline and better simulate the complex interaction patterns, such as user backchannels. Additionally, we will explore more potentials of this modeling scheme by extending to full-duplex interaction involving more modalities, such as vision.

6 Limitations

We acknowledge the limitations of this paper. Our work is based on training with the qwen2-0.5B model and utilizes a relatively small scale of dialogue training data. Compared to other similar models that have been scaled up, our model exhibits considerable potential for enhancement in chat capabilities. Particularly, the response speed of our model, especially in scenarios involving user turn-taking, requires further optimization. We have identified certain conflicts between the goal of early stopping by the assistant and some training objectives in speech-text alignment. Moreover, our current model is not yet capable of handling more complex human-machine interaction phenomena such as user back-channel or even assistant back-channel, indicating that the naturalness of human-machine interaction could be further enhanced.

References

Keyu An, Qian Chen, Chong Deng, Zhihao Du, Changfeng Gao, Zhifu Gao, Yue Gu, Ting He, Hangrui Hu, Kai Hu, Shengpeng Ji, Yabin Li, Zerui Li, Heng Lu, Haoneng Luo, Xiang Lv, Bin Ma, Ziyang Ma, Chongjia Ni, Changhe Song, Jiaqi Shi, Xian Shi, Hao Wang, Wen Wang, Yuxuan Wang, Zhangyu Xiao, Zhijie Yan, Yexin Yang, Bin Zhang, Qinglin Zhang, Shiliang Zhang, Nan Zhao, and Siqi Zheng. 2024. Funaudiollm: Voice understanding and generation foundation models for natural interaction between humans and llms. *Preprint*, arXiv:2407.04051.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, Chang Zhou, and Jingren Zhou. 2024. Qwen2-audio technical report. *Preprint*, arXiv:2407.10759.

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. Technical report, Kyutai.

Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. 2023. Enhancing chat language models by scaling high-quality instructional conversations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 3029–3051. Association for Computational Linguistics.

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. 2024a. Cosyvoice: A scalable multilingual zero-shot text-to-speech synthesizer based on supervised semantic tokens. *Preprint*, arXiv:2407.05407.

Zhihao Du, Jiaming Wang, Qian Chen, Yunfei Chu, Zhifu Gao, Zerui Li, Kai Hu, Xiaohuan Zhou, Jin Xu, Ziyang Ma, Wen Wang, Siqi Zheng, Chang Zhou, Zhijie Yan, and Shiliang Zhang. 2024b. Lauragpt: Listen, attend, understand, and regenerate audio with gpt. *Preprint*, arXiv:2310.04673.

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

Chaoyou Fu, Haojia Lin, Zuwei Long, Yunhang Shen, Meng Zhao, Yifan Zhang, Shaoqi Dong, Xiong Wang, Di Yin, Long Ma, Xiawu Zheng, Ran He, Rongrong Ji, Yunsheng Wu, Caifeng Shan, and Xing Sun. 2024. Vita: Towards open-source interactive omni multimodal llm. *Preprint*, arXiv:2408.05211.

Zhifu Gao, Zerui Li, Jiaming Wang, Haoneng Luo, Xian Shi, Mengzhe Chen, Yabin Li, Lingyun Zuo, Zhihao Du, and Shiliang Zhang. 2023. Funasr: A fundamental end-to-end speech recognition toolkit. In *24th Annual Conference of the International Speech Communication Association, Interspeech 2023, Dublin, Ireland, August 20-24, 2023*, pages 1593–1597. ISCA.

François Hernandez, Vincent Nguyen, Sahar Ghannay, Natalia A. Tomashenko, and Yannick Estève. 2018. TED-LIUM 3: Twice as much data and corpus repartition for experiments on speaker adaptation. In

Speech and Computer - 20th International Conference, SPECOM 2018, Leipzig, Germany, September 18-22, 2018, Proceedings, volume 11096 of *Lecture Notes in Computer Science*, pages 198–208. Springer.

Yunjie Ji, Yan Gong, Yong Deng, Yiping Peng, Qiang Niu, Baochang Ma, and Xiangang Li. 2023. Towards better instruction following language models for chinese: Investigating the impact of training data and evaluation. *Preprint*, arXiv:2304.07854.

Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. 2020. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. In *NeurIPS*.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. 2023. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Ziyang Ma, Yakun Song, Chenpeng Du, Jian Cong, Zhuo Chen, Yuping Wang, Yuxuan Wang, and Xie Chen. 2024. Language model can listen while speaking. *CoRR*, abs/2408.02622.

Tu Anh Nguyen, Eugene Kharitonov, Jade Copet, Yossi Adi, Wei-Ning Hsu, Ali Elkahky, Paden Tomasello, Robin Algayres, Benoît Sagot, Abdelrahman Mohamed, and Emmanuel Dupoux. 2023. Generative spoken dialogue language modeling. *Trans. Assoc. Comput. Linguistics*, 11:250–266.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210.

Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. 2023. Instruction tuning with GPT-4. *CoRR*, abs/2304.03277.

Vineel Pratap, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve, and Ronan Collobert. 2020. Mls: A large-scale multilingual dataset for speech research. *ArXiv*, abs/2012.03411.

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, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR.

Takaaki Saeki, Detai Xin, Wataru Nakata, Tomoki Koriyama, Shinnosuke Takamichi, and Hiroshi Saruwatari. 2022. UTMOS: utokyo-sarulab system for voicemos challenge 2022. *CoRR*, abs/2204.02152.

David Snyder, Guoguo Chen, and Daniel Povey. 2015. MUSAN: A Music, Speech, and Noise Corpus. ArXiv:1510.08484v1.

Tianxiang Sun, Xiaotian Zhang, Zhengfu He, Peng Li, Qinyuan Cheng, Xiangyang Liu, Hang Yan, Yunfan Shao, Qiong Tang, Shiduo Zhang, Xingjian Zhao, Ke Chen, Yining Zheng, Zhejian Zhou, Ruixiao Li, Jun Zhan, Yunhua Zhou, Linyang Li, Xiaogui Yang, Lingling Wu, Zhangyue Yin, Xuanjing Huang, Yu-Gang Jiang, and Xipeng Qiu. 2024. Moss: An open conversational large language model. *Machine Intelligence Research*, 21(5):888–905.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. 2024. Salmonn: Towards generic hearing abilities for large language models. *Preprint*, arXiv:2310.13289.

Alexander Tong, Kilian Fatras, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Guy Wolf, and Yoshua Bengio. 2024. Improving and generalizing flow-based generative models with minibatch optimal transport. *Trans. Mach. Learn. Res.*, 2024.

Bandhav Veluri, Benjamin N Peloquin, Bokai Yu, Hongyu Gong, and Shyamnath Gollakota. 2024. Beyond turn-based interfaces: Synchronous llms as full-duplex dialogue agents. *Preprint*, arXiv:2409.15594.

Changhan Wang, Morgane Rivière, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Miguel Pino, and Emmanuel Dupoux. 2021. Voxpopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 993–1003. Association for Computational Linguistics.

Zhifei Xie and Changqiao Wu. 2024a. Mini-omni: Language models can hear, talk while thinking in streaming. *Preprint*, arXiv:2408.16725.

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

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. 2024. Qwen2 technical report.

Xin Xu Shaoji Zhang Ming Li Yao Shi, Hui Bu. 2015. Aishell-3: A multi-speaker mandarin tts corpus and the baselines.

Heiga Zen, Viet Dang, Rob Clark, Yu Zhang, Ron J. Weiss, Ye Jia, Zhifeng Chen, and Yonghui Wu. 2019. Libritts: A corpus derived from librispeech for text-to-speech. In *20th Annual Conference of the International Speech Communication Association, Interspeech 2019, Graz, Austria, September 15-19, 2019*, pages 1526–1530. ISCA.

Aohan Zeng, Zhengxiao Du, Mingdao Liu, Kedong Wang, Shengmin Jiang, Lei Zhao, Yuxiao Dong, and Jie Tang. 2024. Glm-4-voice: Towards intelligent and human-like end-to-end spoken chatbot. *Preprint*, arXiv:2412.02612.

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. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE.

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. *Preprint*, arXiv:2305.11000.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023a. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Siqi Zheng, Luyao Cheng, Yafeng Chen, Hui Wang, and Qian Chen. 2023b. 3d-speaker: A large-scale multi-device, multi-distance, and multi-dialect corpus for speech representation disentanglement. *Preprint*, arXiv:2306.15354.

A Prompt LLM for Dialogue Quality Evaluation

Auto-Evaluation Prompt

Please rate the given dialogue context and response from the voice dialogue system based on the following criteria (1-10 points), and provide a brief evaluation:

1. Relevance: Is the response relevant to the query? Is the content related?
2. Accuracy: Does the response correctly address the user's query and provide accurate information?
3. Completeness: Does the response comprehensively cover all aspects of the query?
4. Conversational Nature: Is the response easy to understand, concise, clear, and fluent?

Context: context
Response: response

Output in JSON format:

```
{
"Strengths": "Positive aspects of the response",
"Weaknesses": "Negative aspects of the response",
"Overall Evaluation": "Overall assessment of the response",
"Total Score (out of 10, directly provide the score)": ""
}
```