InteractSpeech: A Speech Dialogue Interaction Corpus for Spoken Dialogue Model

Yifu Chen[*]
Zhejiang University

Shengpeng Ji[*]
Zhejiang University

Ziqing Wang
Beijing University Of Technology

Hanting Wang
Zhejiang University

Zhou Zhao[†]
Zhejiang University

Abstract

Spoken Dialogue Models (SDMs) have achieved significant progress in recent years, yet they continue to face challenges in handling nuanced interactional phenomena. A significant bottleneck hindering further advancement is the scarcity of publicly available, high-quality datasets meticulously designed to train and evaluate these fine-grained interactive capabilities. We introduce InteractSpeech, a 150-hour English speech interaction dialogue dataset designed to empower spoken dialogue models with nuanced real-time interaction capabilities, such as handling interruptions and backchannels. InteractSpeech was created by synthesizing interactive dialogues from text using advanced speech synthesis, and by filtering real-world spoken dialogues for interactive segments. The dataset features precise speaker timestamps and annotations for diverse dialogue interactions, underpinned by a formal framework for interaction dynamics. We demonstrate InteractSpeech's utility by fine-tuning a LLaMA 3-8B model on its textual scenarios and, crucially, by training a speech understanding model that accurately classifies key interactional events directly from audio. This highlights the dataset's value in developing models capable of more natural and responsive conversational turn-taking. Audio samples are available at `https://interactspeech.github.io/`.

1 Introduction

Recently, spoken dialogue models such as GPT-4o (OpenAI, 2024) and Moshi (Défossez et al., 2024) have garnered significant attention in the speech domain. These end-to-end spoken dialogue models (Ji et al., 2024a; Xie and Wu, 2024a; Fang et al., 2024; Chen et al., 2024; Xie and Wu, 2024b; Wang et al., 2024; Chen et al., 2025a,b) can not only generate natural, human-like speech responses but also demonstrate an advanced understanding and generation of acoustic features beyond text, such as timbre, emotion, and style (Wenan et al., 2012; Yunhui et al., 2018). Their realistic conversational interactivity and low-latency dialogue experiences further distinguish them among the traditional spoken dialogue models (Huang et al., 2024). However, for spoken dialogue models to achieve truly natural and effective human-computer interaction, they must master the subtle art of conversational turn-taking and interaction dynamics. This includes appropriately handling interruptions, providing timely backchannels, and managing conversational flow—capabilities that are often underdeveloped. A primary bottleneck is the lack of high-quality, open-source datasets specifically designed to train and evaluate these fine-grained interactive abilities. Existing resources present several challenges: 1) Some prominent models, like Moshi (Défossez et al., 2024), still rely on **outdated** datasets such as the Fisher corpus (Cieri et al., 2004) for fine-tuning, which may not reflect contemporary conversational patterns or provide sufficient detail for interaction modeling. 2) Recent open-source interactive datasets (Ma et al., 2024) often focus on **command-based** interruptive words rather than the diversity of real-world conversational interruptions and other interactional phenomena. Given these limitations, we propose InteractSpeech, an open-source, multi-turn, and highly diverse interactive dialogue dataset. InteractSpeech is specifically designed to facilitate the development and evaluation of models that can understand and participate in complex spoken interactions. It provides detailed annotations crucial for modeling the interactional dynamics we formalize within this work, moving beyond simple speech recognition or generation.

InteractSpeech is a mixed dataset consisting of synthetic and real-world data, containing approximately 150 hours of spoken dialogue data. Com-

[*] Equal contribution.
[†] Corresponding author.

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 8024–8033
November 4-9, 2025 ©2025 Association for Computational Linguistics

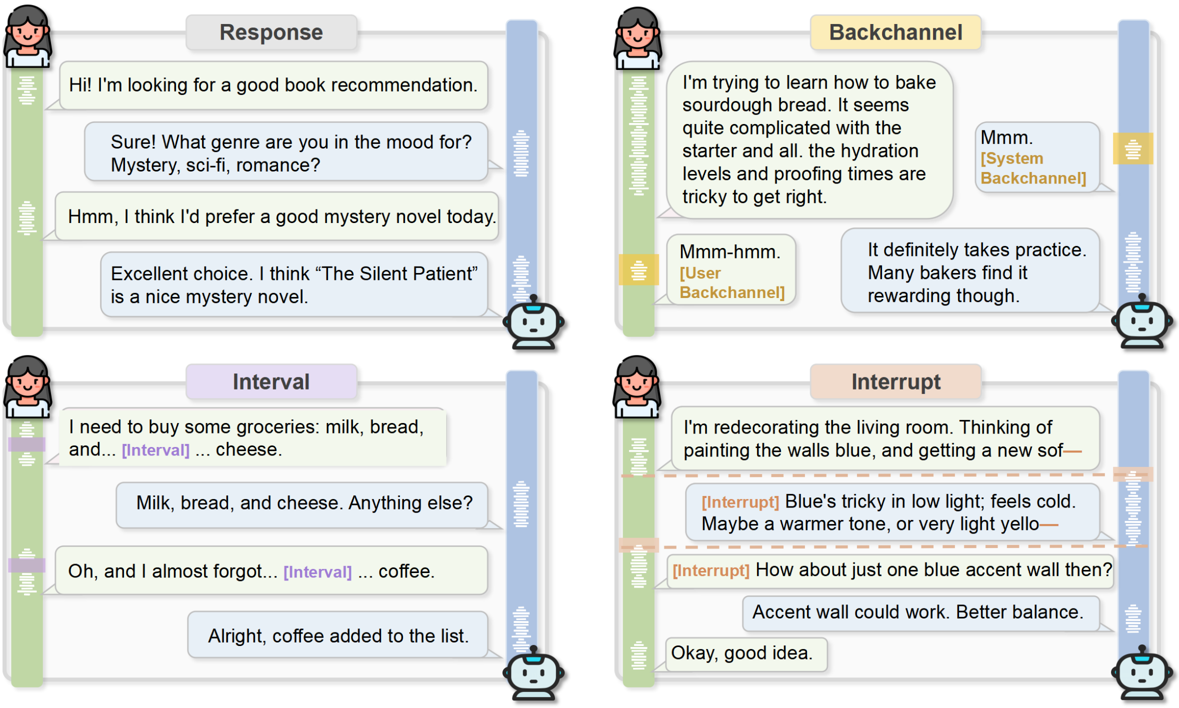

Response
Hi! I'm looking for a good book recommendation.
Sure! What genre are you in the mood for? Mystery, sci-fi, romance?
Hmm, I think I'd prefer a good mystery novel today.
Excellent choice. I think "The Silent Patient" is a nice mystery novel.
Backchannel
I'm trying to learn how to bake sourdough bread. It seems quite complicated with the starter and all. the hydration levels and proofing times are tricky to get right.
Mmm. [System Backchannel]
Mmm-hmm. [User Backchannel]
It definitely takes practice. Many bakers find it rewarding though.
Interval
I need to buy some groceries: milk, bread, and... [Interval] ... cheese.
Milk, bread, and cheese. Anything else?
Oh, and I almost forgot... [Interval] ... coffee.
Alright, coffee added to the list.
Interrupt
I'm redecorating the living room. Thinking of painting the walls blue, and getting a new sof—
[Interrupt] Blue's tricky in low light; feels cold. Maybe a warmer tone, or very light yello—
[Interrupt] How about just one blue accent wall then?
Accent wall could work. Better balance.
Okay, good idea.

Figure 1: An illustrative interaction in the InteractSpeech, with red text denoting the locations of interruptions.

pared to previous spoken dialogue datasets in Table 1, InteractSpeech contains more realistic and diverse topics and over 10 types of interuptions events, such as **disagreements, supplements, or requests for information, as well as subtle cues like backchannels and various forms of interruptions**. Furthermore, InteractSpeech supports multi-turn interactions and mutual interaction between speakers, where AI can also interrupt humans, as well as preserving the modal expressions in dialogue scenarios. An exemplary illustration of dialogue interaction with interruptions is presented in Figure 1. During multi-turn dialogues, AI can provide feedback at any time based on the specific conversation content. To obtain diverse and comprehensive dialogue interaction data, we constructed corresponding interactive text using multi-stage prompt programming and language models, and then generated multi-turn dialogue multi-speaker dual-track speech using advanced speech synthesis models (Du et al., 2024a; An et al., 2024; Du et al., 2024b). Due to the scalability of the synthesis system, we annotated the start and end time of each speaker in InteractSpeech. Moreover, this pipeline can adjust the speaker overlap ratio according to the different requirements of the dialogue system. Furthermore, we curated open-source dialogue speech data and filtered interactive segments using overlap detection and speaker timbre consistency models. Finally, **We had speech experts manually evaluate interactive scenarios of the generated text in the InteractSpeech to ensure the quality**. We evaluated InteractSpeech comprehensively. We analyzed its topic distribution and audio quality. To assess its utility for building text-level interaction understanding, and considering the lack of open-source end-to-end spoken dialogue systems, we fine-tuned a pre-trained LLaMA 3-8B (Touvron et al., 2023) model on the textual component of InteractSpeech to evaluate the reasonableness of generated interaction scenarios. More importantly, to demonstrate its value for speech-level interaction modeling, we trained a dedicated speech understanding model to classify various interactional events directly from the audio data in InteractSpeech. The primary contributions of this work are as follows:

- We propose InteractSpeech, a novel open-source dialogue dataset specifically designed for spoken dialogue interaction scenarios, supporting multi-turn dialogue, containing diverse interruptions events and high-quality audio, and annotated to facilitate the modeling of detailed interaction dynamics as defined in our formal interaction framework.
- We provide a detailed pipeline for creating In-

teractSpeech. Through multi-stage prompt programming, advanced speech synthesis, real-world data filtering, and human expert validation, we ensured the diversity, quality, and reasonableness of the interaction scenarios.

- We thoroughly evaluated InteractSpeech, demonstrating its audio quality and the reasonableness and confirming its suitability for developing more nuanced interactive systems.

2 Related Work

Modeling Conversational Interactions. The pursuit of human-like conversational AI has spurred research into modeling complex interaction dynamics. Early work in conversation analysis (Reimann et al., 2024) laid the foundation for understanding turn-taking mechanisms. Recently, end-to-end spoken dialogue models (Défossez et al., 2024; Ji et al., 2024a; Xie and Wu, 2024a; Fang et al., 2024; Chen et al., 2024, 2025a; Ji et al., 2024b, 2025) are increasingly aiming for full-duplex capabilities, enabling them to process speech and generate responses concurrently. This inherently requires a deeper understanding of interactional cues to manage barge-ins, provide timely feedback, and maintain conversational flow. For example, models like those aiming for "listen while speaking" (Ma et al., 2024) implicitly or explicitly model when to yield or take the turn. However, the ability of these models to truly understand and appropriately react to diverse interactional events is often limited by the data they are trained on. Our work contributes by providing InteractSpeech, a resource specifically designed to train and analyze these interactional modeling capabilities, supported by a formalization of key interactional events critical for robust spoken dialogue systems.

3 InteractSpeech

To develop spoken dialogue models capable of nuanced, real-time interaction, it is crucial to first define and understand the fundamental dynamics of human conversation. Subsequently, datasets must be constructed to reflect these dynamics, enabling models to learn appropriate interactive behaviors. In this section, we first introduce our framework for **Modeling Interaction Dynamics** (Section 3.1), which provides the conceptual basis for InteractSpeech. We then detail the **Overall Structure of InteractSpeech** (Section 3.2), highlighting how it embodies these interactional principles. Finally, we present the **Data Creation Pipeline** (Section 3.3) used to build this resource.

Travel
Booking tickets
Public transport
Restaurants
Career
Technology
Entertainment
Cooking
Family
Online shopping
Booking hotels
Health
Food
Relationships
Holidays
Personal finance
Movies
Learning
Daily conversations
really
Thank
Yes
time
need
feel
Sorry
help
going
make
want
well
Oh
much
work
something
great
good
question
new
Ask
clarifying
Introduce
Express
related
emotion
disagreement
urgency
Provide
Raise
information
additional
premises
perspective
assumptions

(a) topic

(b) content

(c) interaction

Figure 2: Three different word cloud sets about InteractSpeech

3.1 Modeling Interaction Dynamics

To rigorously assess a Spoken Dialogue Model's (SDM's) capacity for managing real-time conversational dynamics, we establish a formalized representation of dyadic spoken interactions between a human speaker (denoted H) and the model (denoted M). The interaction is conceptualized as an alternating sequence of speech and silence spans, each anchored by precise temporal boundaries.

Let a **Speech Utterance** by participant $X \in \{\text{H}, \text{M}\}$ be denoted by U_X^i, representing the i-th continuous interval $[t_{\text{start}}(U_X^i), t_{\text{end}}(U_X^i)]$ during which X produces vocalized speech. Correspondingly, an **Individual Silence Period** for participant X is denoted by $S_X^j = [t_{\text{start}}(S_X^j), t_{\text{end}}(S_X^j)]$, representing the j-th continuous interval where X is vocally inactive. A period of **Joint Silence**, where both participants are simultaneously inactive, is defined as $S_{\text{HM}}^k = S_{\text{H}}^j \cap S_{\text{M}}^l$, representing the k-th such interval.

A **Conversational Turn**, T_X, for participant X is defined as a maximal span during which X holds the speaking floor. A turn T_X typically comprises one or more Speech Utterances $(U_X^i, U_X^{i+1}, \dots)$ by X, potentially interspersed with X's own Individual Silence Periods (S_X^j) that occur within the turn. The boundaries of T_X are delimited either by a Joint Silence (S_{HM}) or by the initiation of a Speech Utterance from the other participant.

Within this temporal structure, we distinguish two functionally distinct types of silence based on their discourse context: **Pause** ($\mathcal{P}$): An Individual

Table 1: Comparison of different datasets for interaction scenarios. ✓ is supported, ✗ is not supported, △ is partially supported.

	InteractSpeech	Switchboard(Godfrey et al., 1992)	Fisher(Cieri et al., 2004)	LibriSpeech(Panayotov et al., 2015)	
				Command FDM (Ma et al., 2024)	Voice FDM (Ma et al., 2024)
Content	**2025s**	1990s	1990s	2015s	2015s
Multi-type Interruption	**over 10**	3	4	1	1
Multi-round	✓	△	△	✗	✗
Dual-track	✓	✓	✓	✗	✗
Speaker timestamp	✓	✗	✗	✗	✗
Diversity	✓	✗	✗	✗	✗
Flexibility	✓	✗	✗	✗	✗

Silence Period S_X^j that occurs *within* participant X's own Conversational Turn T_X (i.e., between two of X's utterances U_X^i and U_X^{i+1} within the same turn). **Gap** (**$\mathcal{G}$**): A Joint Silence S_{HM}^k that occurs at a turn boundary, typically between the end of one participant's turn and the beginning of the other's, signaling a potential transition of the speaker role.

Leveraging this formalization, we categorize key interactional events initiated by the human speaker (H) that necessitate appropriate SDM (M) responses. These events are critical for evaluating an SDM's interactive competence: **Turn Initiation** (**$\mathcal{TI}$**): Event where H commences a new Speech Utterance U_{H} following a Gap $\mathcal{G}$ that succeeded M's completed turn T_{M}. **Interruption** (**$\mathcal{INT}$**): Event where H's Speech Utterance U_{H} begins during M's ongoing Speech Utterance U_{M} (i.e., $t_{\text{start}}(U_{\text{M}}) < t_{\text{start}}(U_{\text{H}}) < t_{\text{end}}(U_{\text{M}})$), thereby violating turn exclusivity. An Interruption may serve various functions, such as floor-taking or expressing disagreement. **Backchannel** (**$\mathcal{BC}$**): Event characterized by a short Speech Utterance U_{H} produced by H during M's ongoing Speech Utterance U_{M}. Crucially, a Backchannel typically signals active listenership, agreement, or acknowledgment from H without an intent to claim the primary speaking floor.

Beyond these canonical events, InteractSpeech also annotates other spontaneous user actions, such as clarification requests, expressions of urgency, or abrupt topic shifts, treating them as discrete, labeled occurrences within the interaction.

This formalized framework, delineating utterances, silences, turns, and specific interactional events ($\mathcal{TI}$, $\mathcal{INT}$, $\mathcal{BC}$, $\mathcal{P}$, $\mathcal{G}$), provides a principled temporal and functional scaffold for the InteractSpeech dataset. It enables fine-grained, consistent annotation of both user and system behaviors, thereby supporting the supervised learning and rigorous evaluation of SDMs that must coordinate not only *what* to say, but critically, *when* and *how* to manage the conversational exchange.

3.2 The Overall of InteractSpeech

InteractSpeech is a multi-turn English spoken-dialogue dataset totaling 150 h of speech and 90 k text utterances. It comprises a 148h in-domain training set, a 2h in-domain test set with manual speaker timestamps, and a 1h out-of-domain (OOD) interactive test set in which humans freely converse with state-of-the-art end-to-end spoken dialogue models (e.g., GPT-4o voice mode, Doubao, Gemini Live). All audio is dual-track with precise onset/offset times for each speaker, supporting the interaction dynamics in Section 3.1.

We benchmark InteractSpeech against Fisher (Cieri et al., 2004), Switchboard (Godfrey et al., 1992), Command FDM and Voice FDM (Ma et al., 2024) (Table 1). Unlike Fisher and Switchboard—which feature multi-turn dialogue but lack fine-grained timestamps—and Command FDM/Voice FDM—which are limited to single-turn "command → cessation" exchanges without broader context—InteractSpeech delivers rich, multi-turn interactions with contemporary topics and exact speaker timing for turn-taking and overlap analysis.

InteractSpeech unifies synthetic (112h) and real-world (38h) data. The synthetic portion uses GPT-4o to transform text dialog corpora (SODA (Kim et al., 2022), Dialogsum (Chen et al., 2021), PLACES (Chen et al., 2023), MultiWOZ 2.2 (Zang et al., 2020)) into multi-turn interactional scenarios, rendered by a state-of-the-art TTS model (Du et al., 2024a) and timestamped. Real-world recordings are filtered via the pipeline in Section 3.3 and annotated with exact timestamps. The 1 h OOD interactive test set (excluded from the 150 h total) evaluates generalization on free-form human–model dialogues.

A central feature is the 2h manually labeled in-domain test set, which includes precise timestamps for interruptions and other interactional events. We

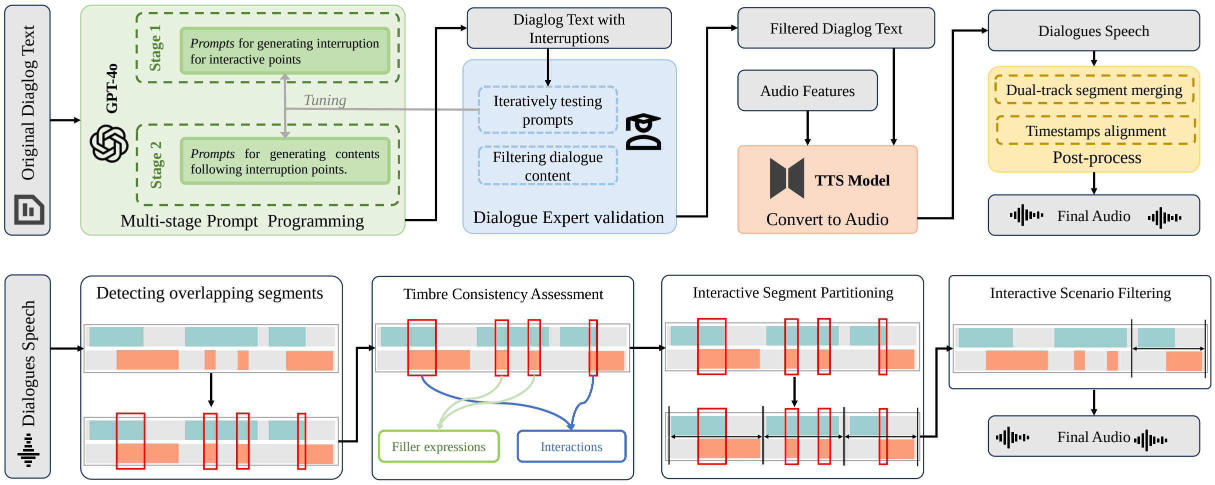

Original Diaglog Text
GPT-4o
Stage 1
Prompts for generating interruption for interactive points
Tuning
Stage 2
Prompts for generating contents following interruption points.
Multi-stage Prompt Programming
Diaglog Text with Interruptions
Iteratively testing prompts
Filtering dialogue content
Dialogue Expert validation
Filtered Diaglog Text
Audio Features
TTS Model
Convert to Audio
Dialogues Speech
Dual-track segment merging
Timestamps alignment
Post-process
Final Audio
Dialogues Speech
Detecting overlapping segments
Timbre Consistency Assessment
Filler expressions
Interactions
Interactive Segment Partitioning
Interactive Scenario Filtering
Final Audio

Figure 3: The overall pipeline about the creation process of InteractSpeech.

retain brief affirmations (e.g., “ok,” “yeah”) as negative examples to help models distinguish true interactional shifts. During end-to-end training (Figure 4), overlapping input tokens remain intact, while the model masks out non-target speakers via timestamps, ensuring only the intended speaker’s tokens are predicted.

3.3 Data Creation Pipeline

The generation process of InteractSpeech, illustrated in Figure 3, involves three main phases: synthetic data generation, real-world data filtering, and OOD interactive test set collection.

Synthetic Data Generation. We first employ a multi-stage prompt programming approach with GPT-4o to generate text-based interactive scenarios from original dialogue texts. Inspired by (Zhou et al., 2022), this multi-stage process yields better control over the interaction points and resulting dialogue flow. In the initial text generation phase, the large language model (e.g., GPT-4o) is tasked with enriching dialogues with interactional events. **Semantic Placement of Interaction Markers:** First, the model identifies suitable locations within the original dialogue text to insert markers for interruptions, backchannels, or other specified interaction points. This process is guided by explicit rules and prompts designed to ensure the semantic relevance and naturalness of these interactional cues (e.g., based on perceived urgency or speaker intent inferred from the textual context). **Generation of Interactive Content:** Following the placement of these markers, the language model generates new conversational content for the interrupting or backchanneling turn. This generation adheres to constraints on naturalness, coherence with the ongoing dialogue, and diversity of expression. Dialogue experts iteratively tested and refined these prompts and generation strategies, proceeding to large-scale deployment only after achieving over 95% satisfaction on small test sets regarding the quality and appropriateness of the generated interactive text. **Each generated textual segment was meticulously reviewed and filtered by human annotators before speech synthesis. Speech Synthesis of Interactive Scenarios.** After obtaining the curated interactive text (complete with semantically placed interaction markers and corresponding content), we use the advanced speech synthesis model CosyVoice-300M-SFT (Du et al., 2024a) to generate the speech. This model provides designated vocal characteristics, ensures voice consistency, and allows for the retrieval of precise timestamps for each audio segment. **Simulating Diverse Interaction Timings:** To simulate the variability of real-world interaction timings for the *pre-scripted* interruptions, the speech synthesis process can be configured. For instance, the audio for an interrupting speaker (whose lines are already determined by the LLM) can be set to begin after n words or within the first m seconds of the ongoing speaker’s turn. This introduces diversity in the auditory realization of the interaction, rather than implying random textual generation of the interruption itself. Finally, multiple audio segments are processed and merged into dual-track speech, with aligned text and speaker timestamps.

Real-World Data Filtering. For the real-world data component of InteractSpeech (as depicted in Figure 3), our objective was to extract segments

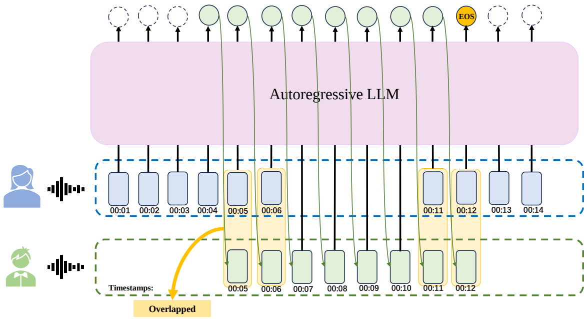


Figure 4: Autoregressive token selection using speaker timestamps to mask non-target speakers during training.

rich in natural interactional events. To achieve this, we utilized the publicly available CANDOR corpus, a large-scale, multimodal dataset of naturalistic English conversations. We then applied our specialized filtering pipeline to CANDOR to isolate and process these interaction-dense segments. **Overlap Detection:** We use the pyannote.audio toolkit[*] for speaker diarization to detect overlapping speech segments, which are common loci for interruptions. **Vocal Consistency Check:** The toolkit's vocal consistency model assesses timbre similarity to determine if a speaking turn has legitimately changed hands following an overlap. After a speech overlap segment is detected, we compare the timbre of the speaker who was speaking *before* the overlap with the timbre of the speaker who continues *after* the overlap concludes. A significant change in timbre, indicating a different speaker has taken the floor after the overlap, strongly suggests a successful *interruption* and turn-taking event. If the timbre remains consistent with the original speaker (i.e., the same speaker continues after the overlap), or if the overlapping speech was brief and did not result in a change of the primary speaker post-overlap, it is more indicative of a *backchannel* or a failed interruption attempt from the original speaker. **Interactive Segment Extraction:** We apply temporal constraints (e.g., minimum duration for overlaps, minimum/maximum length for interaction segments) to filter for high-quality, concise interactive exchanges, typically lasting several tens of seconds.

OOD Interactive Test Set Collection. We recruited human participants to engage in free-form conversations with voice-enabled dialogue systems (e.g., GPT-4o voice mode, Doubao, Gemini Live). Each session lasted approximately one hour and is *not* included in the 150 hours total. Before each dialogue, one of four interaction modes—interruption, pause, normal exchange, backchannel—was randomly assigned, and participants were instructed to maintain that mode throughout the session. Professional annotators then labeled the timestamp of every interaction event. This OOD set enables rigorous evaluation of model generalization to unseen, structured interaction patterns.

This pipeline ensures that InteractSpeech provides a richly varied, precisely annotated set of interactional events, suitable for robust training and evaluation of end-to-end spoken-dialogue models.

4 Experiments

4.1 Modeling Interaction Dynamics from Speech

To evaluate if models fine-tuned on InteractSpeech can effectively **understand** and identify specific interactional events, we adopted a binary classification approach for each event type. This assesses the model's ability to confirm or deny the presence of events defined in our framework (Section 3.1), aligning with methodologies for benchmarking audio model understanding (Arora et al., 2025).

Task Definition and Data. For each of the four primary interactional event types—Backchannel ($\mathcal{BC}$), Interruption ($\mathcal{INT}$), Gap ($\mathcal{G}$), and Pause ($\mathcal{P}$)—the task was framed as a binary classification. Given an audio segment x from InteractSpeech and a target event type $E \in \{\mathcal{BC}, \mathcal{INT}, \mathcal{G}, \mathcal{P}\}$, the model predicts if event E is present in x. The ground truth label $y \in \{0, 1\}$ is 1 (present) if event E occurs, and 0 (absent) otherwise, based on InteractSpeech annotations. A data sample consisted of (audio segment x, event query prompt q_E, ground truth binary answer y) tuples from the InteractSpeech dataset $\mathcal{D}$.

During evaluation, for an audio segment x and a queried event E, the model received an event-specific prompt q_E (e.g., "Does an Interruption occur in this audio?") and was expected to output "Yes" or "No". This output was then mapped to a binary label y for accuracy calculation.

Model Fine-tuning. We fine-tuned Qwen-2.5-Omni with true Group Relative Policy Optimization (GRPO) (Shao et al., 2024). For each audio segment x and event prompt q_E, we sample G outputs $\{a_i\}$ under the current policy π_θ and assign one simple rewards:

[*] https://github.com/pyannote/pyannote-audio

Table 2: Binary Classification Accuracy (%) for understanding specific interactional event types via prompted queries. IS = InteractSpeech Test Set; RWS = Real-World Scenarios Test Set. Overall Avg. is the macro-average accuracy across the four event types.

Model	InteractSpeech (IS) Test Set Acc (%)					Real-World Scenarios (RWS) Acc (%)				
	Backchannel	Interaction	Gap	Pause	Overall Avg.	Backchannel	Interaction	Gap	Pause	Overall Avg.
Random Guess	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0
Qwen2Audio-Instruct	52.1	48.5	53.3	51.0	51.2	51.0	47.2	51.5	49.8	49.9
Qwen-Omni (Base)	53.5	49.1	54.8	52.6	52.5	51.8	50.5	52.7	48.0	50.8
GPT-4o	55.2	51.3	56.0	54.1	54.2	55.0	49.5	55.8	51.7	53.0
Qwen-Omni-FT (Ours)	**72.3**	74.8	73.1	**73.5**	**73.4**	**70.7**	**60.2**	**60.9**	**61.3**	**63.3**
w/o Real Data	65.4	**75.6**	**74.8**	71.1	71.7	55.3	56.9	55.2	57.7	56.3
w/o Synthetic Data	68.7	61.9	59.5	60.4	62.6	67.1	53.6	51.8	56.2	57.2

$$R_i = r_{\text{acc}}(a_i) = \begin{cases} 1, & a_i = y, \\ 0, & \text{otherwise.} \end{cases} \quad (1)$$

We normalize advantages via :

$$A_i = \frac{R_i - \mu}{\sigma}, \quad (2)$$

with μ, σ the mean and standard deviation of $\{R_i\}$. The GRPO surrogate objective is

$$\begin{aligned} \mathcal{J}_{\text{GRPO}}(\theta) = \frac{1}{G} \sum_{i=1}^{G} \min\bigl(r_i(\theta) A_i, \\ \text{clip}(r_i(\theta), 1-\epsilon, 1+\epsilon)\, A_i\bigr)\,, \end{aligned} \quad (3)$$

where $r_i(\theta) = \pi_\theta(a_i \mid x, q_E)/\pi_{\theta_{\text{old}}}(a_i \mid x, q_E)$. To keep the fine-tuned policy close to the pre-trained reference π_{ref}, we add a KL-penalty. The final loss is:

$$\begin{aligned} \mathcal{L}(\theta) = & -\mathcal{J}_{\text{GRPO}}(\theta) \\ & + \lambda\, D_{\text{KL}}\bigl(\pi_\theta(\cdot \mid x, q_E) \,\|\, \pi_{\text{ref}}(\cdot \mid x, q_E)\bigr). \end{aligned} \quad (4)$$

We set $G = 4$, $\epsilon = 0.2$, $\lambda = 0.001$, and trained with AdamW for 1500 steps (learning rate 10^{-6}, batch size 2).

Baselines. We benchmarked Qwen-Omni-FT (Ours) against: **Random Guess:** 50% accuracy for binary classification. **Qwen2Audio-Instruct (Chu et al., 2024) (Zero-shot):** Evaluated without fine-tuning on InteractSpeech, using its general audio understanding capabilities to respond to prompted queries about interactional events. **Qwen-Omni (Base) (Zero-shot):** The pre-trained Qwen-Omni model, also evaluated in a zero-shot manner for this task. **GPT-4o (Audio API):** OpenAI's model queried via its audio API . We conduct further ablation studies on Qwen-Omni-FT variants trained *w/o Real Data* and *w/o Synthetic Data* to assess the contribution of InteractSpeech's data components.

Results. Table 2 summarizes the accuracy for classifying interactional events. Our Qwen-Omni-FT model significantly outperforms all baselines, achieving 73.4% overall accuracy on the InteractSpeech (IS) test set. This starkly contrasts with GPT-4o (54.2%) and near-chance level performance from Qwen2Audio-Instruct (51.2%) and Qwen-Omni (Base) (52.5%), underscoring the necessity of targeted fine-tuning on datasets like InteractSpeech to grasp these nuanced interactional cues. Crucially, Qwen-Omni-FT also demonstrates strong generalization to Real-World Scenarios (RWS) with 63.3% accuracy, again outperforming GPT-4o (53.0%). This suggests InteractSpeech effectively bridges the gap between synthetic training and real-world complexities. Notably, the model achieves a high 70.7% on RWS Backchannel detection, likely benefiting from the rich, natural backchannel occurrences in real-world data. Ablation studies further disentangle the roles of synthetic versus real-world data. Removing synthetic data incurs a large drop on IS (–10.8), indicating its foundational value for in-domain feature learning, while removing real data barely affects IS (–1.7). Conversely, on RWS the model suffers more when real examples are omitted (–7.0) than when synthetic data is removed (–6.1), underscoring that real exemplars are indispensable for robust out-of-domain generalization.

4.2 Evaluation of Speech Quality

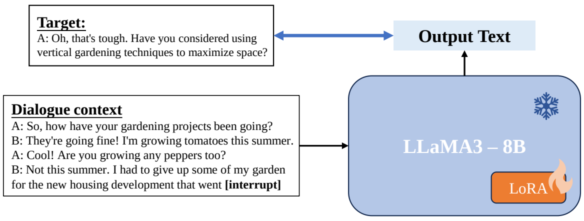
We first evaluated the audio quality of various spoken dialogue datasets, using the Mean Opinion Score (MOS) scale (ranging from 1 to 5) to assess both the naturalness and clarity of the audio samples. Ten audio segments were randomly selected from the datasets InteractSpeech, Fisher (Cieri et al., 2004), and Switchboard (Godfrey et al., 1992). The experimental results, as shown in Table 3, indicate that most dialogue datasets exhibit satisfactory audio quality. Although a portion of the InteractSpeech dataset consists of synthetic data, the audio quality remains high, owing to the use of advanced text-to-speech models (Du et al., 2024a).

Table 3: The results of MOS with 95% confidence intervals.

Datasets	MOS ↑
Switchboard (Godfrey et al., 1992)	4.09 ± 0.14
Fisher (Cieri et al., 2004)	4.21 ± 0.11
InteractSpeech	4.35 ± 0.12

4.3 Evaluation of Textual Interactive Scenarios

Considering the substantial computational cost and the current lack of readily available open-source end-to-end spoken dialogue models with full-duplex capabilities , we undertook a validation to specifically assess the effectiveness and quality of InteractSpeech's textual interaction scenarios. Specifically, we fine-tuned LLaMA3-8B model (Touvron et al., 2023) using LoRA (Hu et al., 2021) exclusively on the textual training data derived from InteractSpeech. The training process, as illustrated in Figure 5, tasked this text-based dialogue model with a challenging objective: to predict the subsequent interactive text immediately following a scripted interruption within a given dialogue. We evaluated the model's generated responses on the test set of InteractSpeech. Following the methodology outlined in SODA(Kim et al., 2022), human evaluators measured the quality of these generated responses across three critical dimensions: Naturalness , Consistency , and Specificity. As shown in Table 4, the results are compelling: the fine-tuned dialogue interaction model achieved an overall agreement rate of 86%. This high agreement rate signifies that human evaluators found the LLaMA model's generated continuations to be largely natural, consistent with the entire dialogue history and specific to the newly established conversational turn.



A: Oh, that's tough. Have you considered using vertical gardening techniques to maximize space?

A: So, how have your gardening projects been going?
B: They're going fine! I'm growing tomatoes this summer.
A: Cool! Are you growing any peppers too?
B: Not this summer. I had to give up some of my garden for the new housing development that went **[interrupt]**

Figure 5: Fine-tuning LLaMA3 with interactive conversation text.

Table 4: Results of head-to-head human evaluation between model responses on the InteractSpeech test set. The main entry shows LLaMA w/ interaction fine-tuned on InteractSpeech text. The bottom two rows are from the ablation study in Section 4.4.

Model / Prompting Method	Natural ↑	Consistent ↑	Specific ↑	Overall ↑
LLaMA w/ interaction	86%	85%	88%	86%
Single-stage prompt programming	25%	18%	22%	21%
Multi-stage prompt programming	97%	97%	99%	98%

4.4 Ablation Study on Prompt Programming

We selected the original non-interactive dialogue segments from the test set's source texts to perform an ablation study comparing single-stage and multi-stage prompt programming for generating the interactive scenarios in InteractSpeech. In the single-stage approach, instructions for generating interactive dialogues were fed directly into GPT-4o, with rule constraints limited to descriptive adjectives (e.g., *"high naturalness, appropriate positioning, limited number of interactive scenarios"*) without example-based guidance. As shown in the lower part of Table 4, the experimental results indicate that our multi-stage prompt programming approach significantly improves the quality (naturalness, consistency, specificity, and overall) of the generated interactive text compared to the single-stage approach.

5 Conclusion

In this paper, we introduced InteractSpeech, a 150-hour English spoken dialogue dataset meticulously designed to address the critical need for resources that support the development of models with nuanced real-time interaction capabilities. We presented a framework for modeling key interactional dynamics, which underpins the construction and annotation of InteractSpeech. The dataset combines synthetically generated dialogues, enriched with diverse interactional events via multi-stage prompt programming, with carefully filtered real-world spoken data. Through extensive experiments, we demonstrated the utility of InteractSpeech. We showed our fine-tuned model significantly outperformed baselines on both in-domain and out-of-domain real-world scenarios, highlighting InteractSpeech's potential to advance the development of models that can robustly perceive, interpret, and appropriately react to the dynamic flow of human conversation. We hope InteractSpeech will help further research into more natural, responsive, and human-like spoken dialogue systems.

Limitation

While InteractSpeech is designed to empower spoken dialogue models with enhanced interactive abilities, directly quantifying the improvement in holistic, end-to-end interaction modeling presents certain challenges within the current research landscape. The development and fine-tuning of full-duplex spoken dialogue systems capable of leveraging such nuanced datasets often require substantial resources and access to proprietary architectures, with limited open-source counterparts readily available for comprehensive benchmarking. Furthermore, the field is still evolving robust, standardized metrics to specifically evaluate the intricate dynamics of turn-taking, interruption handling, and backchanneling in a fully integrated conversational agent. Consequently, our current validation focuses on crucial sub-tasks: demonstrating the utility of InteractSpeech's textual scenarios for training language models to generate appropriate interactive responses, and its audio data for training models to accurately perceive key interactional events. These evaluations serve as significant intermediate steps, indicating the dataset's potential. We believe future work, building upon resources like InteractSpeech and benefiting from advancements in open-source model availability and interaction evaluation methodologies, will be better positioned to demonstrate these end-to-end improvements more directly.

Acknowledgments

This work was supported by the National Key R&D Program of China under Grant No. 2022ZD0162000, the National Natural Science Foundation of China under Grant No. 62222211, and the National Natural Science Foundation of China under Grant No. U24A20326.

References

Keyu An, Qian Chen, Chong Deng, Zhihao Du, Changfeng Gao, Zhifu Gao, Yue Gu, Ting He, Hangrui Hu, Kai Hu, and 1 others. 2024. Funaudiollm: Voice understanding and generation foundation models for natural interaction between humans and llms. *arXiv preprint arXiv:2407.04051*.

Siddhant Arora, Zhiyun Lu, Chung-Cheng Chiu, Ruoming Pang, and Shinji Watanabe. 2025. Talking turns: Benchmarking audio foundation models on turn-taking dynamics. *arXiv preprint arXiv:2503.01174*.

Maximillian Chen, Alexandros Papangelis, Chenyang Tao, Seokhwan Kim, Andy Rosenbaum, Yang Liu, Zhou Yu, and Dilek Hakkani-Tur. 2023. Places: Prompting language models for social conversation synthesis. *arXiv preprint arXiv:2302.03269*.

Qian Chen, Yafeng Chen, Yanni Chen, Mengzhe Chen, Yingda Chen, Chong Deng, Zhihao Du, Ruize Gao, Changfeng Gao, Zhifu Gao, and 1 others. 2025a. Minmo: A multimodal large language model for seamless voice interaction. *arXiv preprint arXiv:2501.06282*.

Wenxi Chen, Ziyang Ma, Ruiqi Yan, Yuzhe Liang, Xiquan Li, Ruiyang Xu, Zhikang Niu, Yanqiao Zhu, Yifan Yang, Zhanxun Liu, and 1 others. 2024. Slam-omni: Timbre-controllable voice interaction system with single-stage training. *arXiv preprint arXiv:2412.15649*.

Yifu Chen, Shengpeng Ji, Haoxiao Wang, Ziqing Wang, Siyu Chen, Jinzheng He, Jin Xu, and Zhou Zhao. 2025b. Wavrag: Audio-integrated retrieval augmented generation for spoken dialogue models. *arXiv preprint arXiv:2502.14727*.

Yulong Chen, Yang Liu, Liang Chen, and Yue Zhang. 2021. DialogSum: A real-life scenario dialogue summarization dataset. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 5062–5074, Online. Association for Computational Linguistics.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, and 1 others. 2024. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*.

Christopher Cieri, David Miller, and Kevin Walker. 2004. The fisher corpus: A resource for the next generations of speech-to-text. In *LREC*, volume 4, pages 69–71.

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*.

Zhihao Du, Qian Chen, Shiliang Zhang, Kai Hu, Heng Lu, Yexin Yang, Hangrui Hu, Siqi Zheng, Yue Gu, Ziyang Ma, and 1 others. 2024a. Cosyvoice: A scalable multilingual zero-shot text-to-speech synthesizer based on supervised semantic tokens. *arXiv preprint arXiv:2407.05407*.

Zhihao Du, Yuxuan Wang, Qian Chen, Xian Shi, Xiang Lv, Tianyu Zhao, Zhifu Gao, Yexin Yang, Changfeng Gao, Hui Wang, and 1 others. 2024b. Cosyvoice 2: Scalable streaming speech synthesis with large language models. *arXiv preprint arXiv:2412.10117*.

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*.

John J Godfrey, Edward C Holliman, and Jane McDaniel. 1992. Switchboard: Telephone speech corpus for research and development. In *Acoustics, speech, and signal processing, ieee international conference on*, volume 1, pages 517–520. IEEE Computer Society.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Rongjie Huang, Mingze Li, Dongchao Yang, Jiatong Shi, Xuankai Chang, Zhenhui Ye, Yuning Wu, Zhiqing Hong, Jiawei Huang, Jinglin Liu, and 1 others. 2024. Audiogpt: Understanding and generating speech, music, sound, and talking head. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 23802–23804.

Shengpeng Ji, Yifu Chen, Minghui Fang, Jialong Zuo, Jingyu Lu, Hanting Wang, Ziyue Jiang, Long Zhou, Shujie Liu, Xize Cheng, and 1 others. 2024a. Wavchat: A survey of spoken dialogue models. *arXiv preprint arXiv:2411.13577*.

Shengpeng Ji, Ziyue Jiang, Xize Cheng, Yifu Chen, Minghui Fang, Jialong Zuo, Qian Yang, Ruiqi Li, Ziang Zhang, Xiaoda Yang, and 1 others. 2024b. Wavtokenizer: an efficient acoustic discrete codec tokenizer for audio language modeling. *arXiv preprint arXiv:2408.16532*.

Shengpeng Ji, Tianle Liang, Yangzhuo Li, Jialong Zuo, Minghui Fang, Jinzheng He, Yifu Chen, Zhengqing Liu, Ziyue Jiang, Xize Cheng, and 1 others. 2025. Wavreward: Spoken dialogue models with generalist reward evaluators. *arXiv preprint arXiv:2505.09558*.

Hyunwoo Kim, Jack Hessel, Liwei Jiang, Peter West, Ximing Lu, Youngjae Yu, Pei Zhou, Ronan Le Bras, Malihe Alikhani, Gunhee Kim, and 1 others. 2022. Soda: Million-scale dialogue distillation with social commonsense contextualization. *arXiv preprint arXiv:2212.10465*.

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

OpenAI. 2024. Hello gpt-4o. `https://openai.com/index/hello-gpt-4o/`. Accessed: 2024-05-26.

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. IEEE.

Merle M Reimann, Florian A Kunneman, Catharine Oertel, and Koen V Hindriks. 2024. A survey on dialogue management in human-robot interaction. *ACM Transactions on Human-Robot Interaction*, 13(2):1–22.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Xiong Wang, Yangze Li, Chaoyou Fu, Yunhang Shen, Lei Xie, Ke Li, Xing Sun, and Long Ma. 2024. Freeze-omni: A smart and low latency speech-to-speech dialogue model with frozen llm. *arXiv preprint arXiv:2411.00774*.

TAN Wenan, WEN Xiang, JIANG Chuanqun, DU Yi, and HU Xiaoming. 2012. An evaluation model integrating user trust and capability for selection of cooperative learning partners. *Chinese Journal of Electronics*, 21(1):42–46.

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

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

LIU Yunhui, ZHENG Fan, GUO Ruibin, WANG Jiangliu, NIE Qiang, WANG Xin, and WANG Zerui. 2018. Robot intelligence for real world applications. *Chinese Journal of Electronics*, 27(3):446–458.

Xiaoxue Zang, Abhinav Rastogi, Srinivas Sunkara, Raghav Gupta, Jianguo Zhang, and Jindong Chen. 2020. Multiwoz 2.2: A dialogue dataset with additional annotation corrections and state tracking baselines. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI, ACL 2020*, pages 109–117.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, and 1 others. 2022. Least-to-most prompting enables complex reasoning in large language models. *arXiv preprint arXiv:2205.10625*.