LS: DM: Dialigoue Manager

Liao, Borui, Yulong Xu, Jiao Ou, Kaiyuan Yang, Weihua Jian, Pengfei Wan, and Di Zhang. “FlexDuo: A Pluggable System for Enabling Full-Duplex Capabilities in Speech Dialogue Systems.” arXiv, February 19, 2025. FlexDuo: A Pluggable System for Enabling Full-Duplex Capabilities... .

Model Architecture

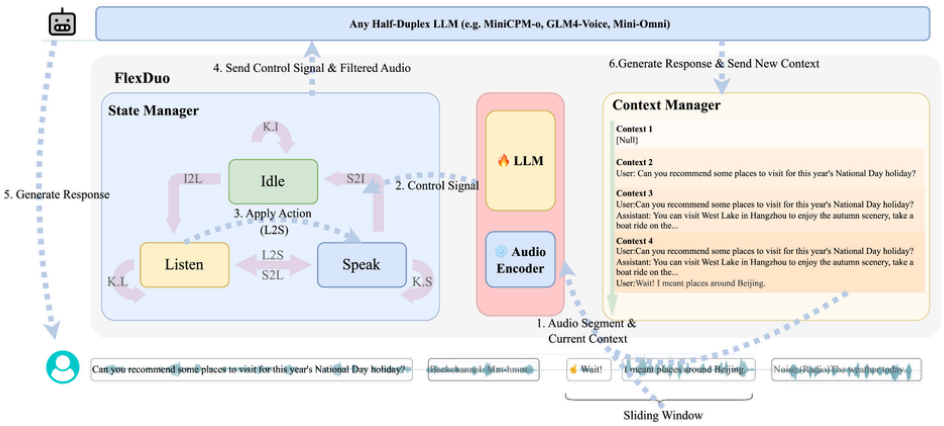

Figure 2: FlexDuo Workflow and Framework for Interaction with half-duplex LLMs. The input to FlexDuo is user audio and the conversational responses from a half-duplex LLM, while its outputs are control signals and filtered user audio.

FlexDuo uses **Qwen2-audio-7B-Instruct** as its base model with the following configuration:

- Audio Encoder: Kept frozen to maintain audio processing capabilities
- LLM component: Fine-tuned for duplex interaction alignment
- Special tokens: Added to the LLM vocabulary for dialogue actions

The overall architecture consists of three main components:

1. **State Manager**: Predicts dialogue actions using a finite state machine. Follows a finite state machine with 7 possible transition actions:
   - K.S (Keep Speaking): The assistant maintains the current Speak state.
   - K.L (Keep Listening): The user has not finished speaking.
   - K.I (Keep Idling): Environmental noise, user’s backchannels, or third-party dialogue.
   - S2L (Speak to Listen): The user interrupts the assistant.
   - S2I (Speak to Idle): The assistant finishes speaking and ends naturally.

- L2S (Listen to Speak): The assistant responds to or interrupts the user.
- I2L (Idle to Listen): The user starts speaking.

2. **Context Manager**: Maintains dialogue states and context history
3. **Sliding Window**: Monitors real-time audio input (size set to 5 in experiments)

$$W_t = \begin{cases} [W_{t-1}, a_t], & \text{if } S_{t-1} = \text{Listen} \\ [a_{t-w+1}, \ldots, a_{t-1}, a_t], & \text{others} \end{cases} \tag{2}$$

In Equation 2, a_t denotes the speech block in the sliding window at time t, w is the sliding window size, which is set to 5 in this paper.

Our duplex control module predicts the current dialogue strategy based on the context of historical dialogues, the current state, and the accumulated speech chunks in the sliding window.

Data Processing Pipeline

1. **Dataset Source**:
   - Fisher corpus (English: 831 hours, Chinese: 389 hours)
   - Contains separate recordings for two speakers in conversations
2. **Data Preprocessing**:
   - Used VAD to extract Inter-Pausal Units (IPUs)
   - Merged IPUs with gaps less than 160ms
   - Labeled completely overlapped IPUs as backchannels
   - Others labeled as valid dialogue turns
3. **Quality Control**:
   - Transcribed dialogues using ASR
   - Filtered using GPT-o1-mini dialogue evaluation
   - Human validation showed 91.6% acceptance rate
   - Final dataset: 671 hours English, 263 hours Chinese
4. **Training Data Organization**:
   - Applied sliding window to re-segment audio streams
   - Labeled based on VAD information (turn transitions, utterance status, system state)
   - Incorporated human preference in labeling process
   - Added 500ms delay for "listening phase" to distinguish backchannels from interruptions

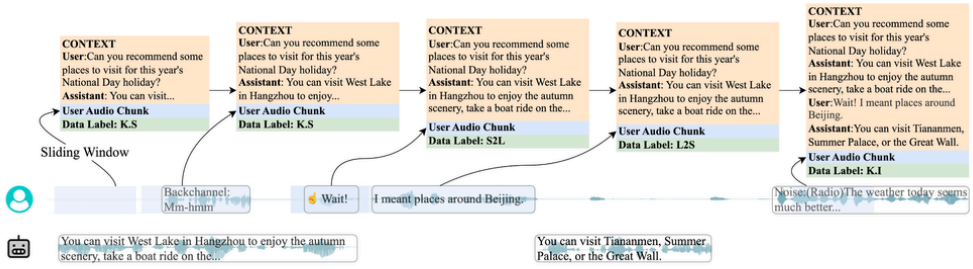

Figure 3: An example of training data construction

Training Process

1. **Configuration**:
   - Standard cross-entropy as training objective
   - Loss masking applied to historical context and user audio input

- AdamW optimizer with learning rate 1e-5
- Batch size of 3 per GPU
- 40,000 training steps with 500-step warm-up
- Hardware: 8 NVIDIA H800 GPUs
- DeepSpeed ZeRO-3 optimization

2. **Dataset Split**:
   - 10:1:1 ratio for training, validation, and test sets

- FlexDuo defines the Assistant's turn-taking as occurring strictly after the User's turn ends. However, the timing for the User's turn-taking remains unchanged to fully simulate real-world scenarios where users interrupt at specific moments.

Novel Contributions

1. **Pluggable Architecture**:
   - Decoupled design separates duplex control from dialogue system
   - Proven effective through integration with GLM4-voice and MiniCPM-o
   - Demonstrated 23.1% reduction in false interruption rate vs. VAD-controlled systems
2. **Idle State Introduction**:
   - Three-state model (Speak, Listen, Idle) vs. traditional two-state approach
   - Ablation studies validated this by showing:
     - 15.68% decrease in Turn-taking F1 score when Idle state removed
     - 13.62% increase in false interruption rate when Idle state removed
3. **Sliding Window Optimization**:
   - Controlled experiments with window sizes $w \in \{2, 4, 8\}$
   - Demonstrated trade-off between responsiveness and semantic understanding
   - Selected optimal size to balance these factors
4. **Semantic Integrity Buffer**:
   - 500ms delay mechanism for listening phase
   - Provides time for semantic judgment before state transitions
   - Empirically shown to reduce contextual noise interference

Real-World Inference Process

When deployed, FlexDuo operates as follows:

1. **Initial Setup**:
   - Integrates with an existing half-duplex LLM (e.g., GLM4-voice)
   - Initializes with Idle state
2. **Real-time Processing**:
   - Every 120ms, evaluates audio and context
   - Manages the sliding window based on current state
   - Predicts one of seven dialogue actions using the state manager
3. **Dialogue Flow Control**:
   - Filters audio in Idle state to maintain clean context
   - Passes relevant audio to the half-duplex LLM when in Listen state
   - Controls when the LLM generates responses (Speak state)
   - Interrupts LLM output when state changes to Idle or Listen
4. **Context Management**:
   - Updates dialogue context when state transitions occur
   - Maintains semantic integrity of the conversation

- Example: user backchannels (like "hmm") in the middle of a conversation are filtered out when in Idle state

5. **Response Generation**:
   - Half-duplex LLM generates responses based on filtered context
   - System can transition states mid-utterance if needed
   - Avoids false interruptions due to background noise or backchannels

Performance Validation

1. **Interaction Capability**:
   - 24.9% reduction in false interruption rate vs. integrated full-duplex systems
   - 7.6% improvement in turn-taking performance vs. integrated systems
2. **Dialogue Quality**:
   - 35.3% reduction in conditional perplexity on English Fisher dataset
   - 19% reduction in conditional perplexity on Chinese Fisher dataset
   - Demonstrates that context filtering meaningfully improves dialogue coherence
3. **Human Evaluation**:
   - 91.6% acceptance rate on the dialogue quality filtering process
   - Validates the effectiveness of the data preparation approach

Model	Turn-taking			False Interruption Rate↓		
	Assistant (Pos. F1@1/5/10)	User (Pos. F1@1/5/10)	Combined	Assistant	User	Combined
Freeze-omni	0.78/0.94/0.98	0.64/0.96/0.98	0.71	0.72	0.49	0.61
Moshi	0.28/0.57/0.65	0.26/0.60/0.76	0.27	0.37	-	-
MinMo	0.66/0.83/0.88	0.42/0.94/0.99	0.54	-	-	-
VAD+GLM4-voice	0.98/1.0/1.0	0.64/0.96/0.98	0.81	0.58	0.49	0.53
VAD+MiniCPM-o	0.98/1.0/1.0	0.64/0.96/0.98	0.81	0.58	0.49	0.53
FlexDuo+GLM4-voice	0.68/0.91/0.93	**0.89/1.0/1.0**	0.79	**0.35**	**0.25**	**0.30**
FlexDuo+MiniCPM-o	0.68/0.91/0.93	0.89/1.0/1.0	0.79	0.35	0.25	0.30

Table 1: Evaluation Metrics for Interaction Capability on English fisher data: The performance of the duplex prediction module is assessed based on turn-taking dynamics between the assistant and the user, quantified using the Positive F1 Score @offset-K metric. Additionally, the false interruption rate is measured as the proportion of effective speaking duration relative to the total speaking duration. Indicators that cannot be obtained are represented by "-".

The system's real-world performance demonstrates that FlexDuo successfully addresses the limitations of existing full-duplex systems while maintaining compatibility with half-duplex LLMs, providing a more natural conversational experience without the need for complete system retraining.

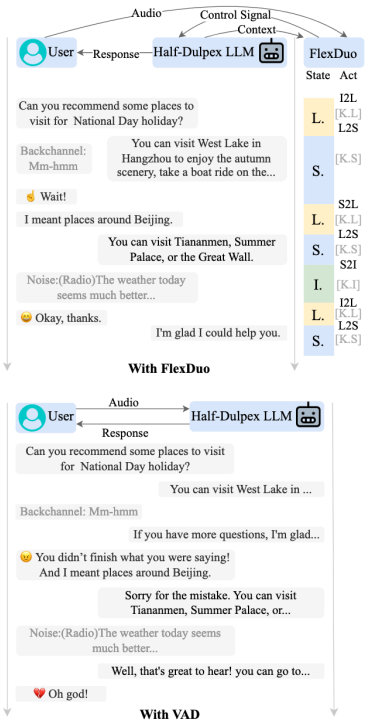

Figure 1: Performance of the half-duplex dialogue system with FlexDuo and VAD in real-world dialogue scenarios. `L.`, `S.`, and `I.` represent the Listen, Speak, and Idle dialogue states, respectively, while Act represents the dialogue strategy. Compared to the half-duplex dialogue system, FlexDuo enables natural dialogue transitions and accurate noise filtering.