Decoupling Conversational Dynamics in Full-Duplex Spoken Models through Reinforcement Learning

Yuxin Li[1], Donghang Wu[1], Guan-Ting Lin[2], Hung-yi Lee[2], Chengwei Qin[3], Zhehuai Chen[4], Chen Chen[4]

[1] Nanyang Technological University · [2] National Taiwan University · [3] HKUST · [4] NVIDIA

Code

Demo

github.com/LiYuxin44/DuplexPO_official

Background and Motivation

Background: Natural conversation requires more than good answers. Full-duplex voice agents must listen while speaking, give timely backchannels, and yield when users interrupt.

The challenge

(1) Better interaction can come with weaker intelligence.

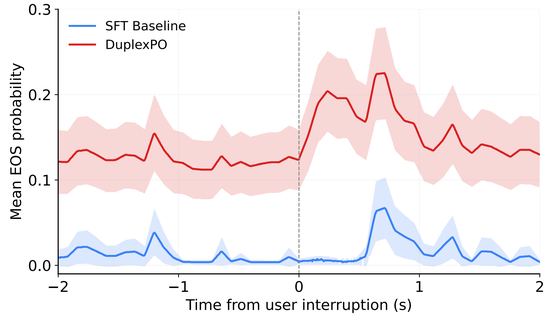
(2) Human conversations provide rich timing cues, but their content may not match helpful assistant behavior.

The Solution

We treat “when to speak” and “what to say” as distinct optimization objectives within the same model.

Boundary Token Analysis

DuplexPO increases EOS probability around user interruptions, supporting more responsive turn yielding.

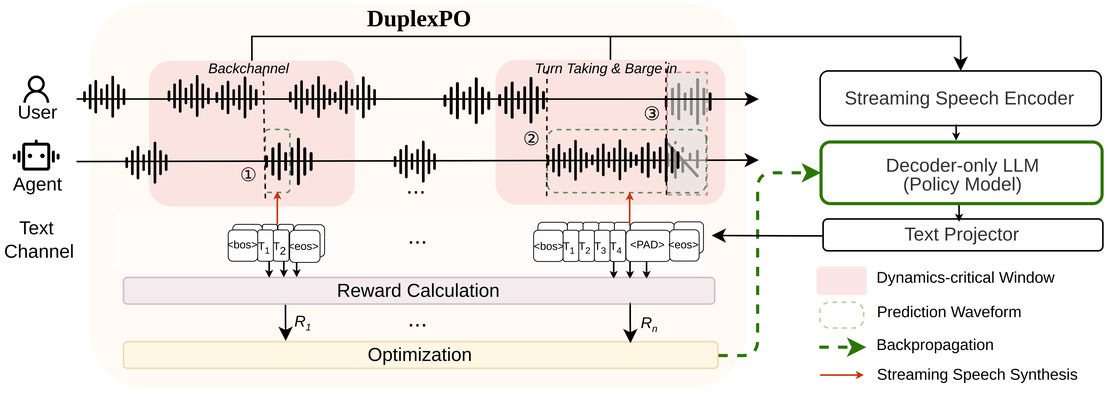


Open to collaboration!!

Wechat: Jamais_Vu4; Email: Yuxin.li@ntu.edu.sg

Method

We propose **DuplexPO**, a reinforcement learning (RL) framework. It preserves the semantic response capability of an instruction-tuned assistant, while optimizing its temporal interaction behavior over selected high-impact windows from long human conversations

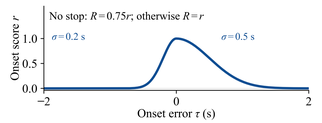


1. Sample dynamics-critical windows

2. Reward calculation Factorized Conversational Dynamics Reward (FCDR)

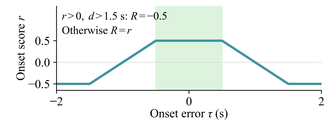
Turn-taking onset

$$r = \exp[-\tau^2/(2\sigma(\tau)^2)]$$



Backchannel onset

$$r = \max[-0.5,\ 0.5 - \max(0, |\tau| - 0.5)]$$



τ: predicted − reference onset (s); d: speech duration (∞ if no stop).

3. Window-level policy optimization

$$L_{\mathrm{GRPO}}(\theta) = -\frac{1}{N}\sum_{i,k} A_{i,k} \sum_{t \in W_{i,k}} \log \pi_\theta(y_{i,k,t} \mid h_{i,k,t}) + \beta L_{\mathrm{KL}}(\theta; \pi_{\mathrm{SFT}})$$

Group-normalized advantages; KL to the frozen SFT reference.

Results

Turn taking and backchanneling performance:

Method	Dataset	Onset MAE (↓)	Turn-taking		Backchannel	
			Init Rate (↑)	Yield Rate (↑)	Init Rate (↑)	Yield Rate (↑)
Moshi [Défossez et al., 2024]	Fisher	1.99	87.7	76.5	83.9	66.7
PersonaPlex [Roy et al., 2026]		1.50	98.6	80.3	91.4	69.2
SFT Baseline		0.98	97.8	92.1	95.7	57.1
SFT Dynamics		1.14	97.1	78.9	91.4	57.1
DuplexPO		**0.69**	**100.0**	**98.7**	**97.8**	**100.0**
Moshi [Défossez et al., 2024]	Seamless	2.16	80.1	69.0	76.2	71.3
PersonaPlex [Roy et al., 2026]		1.77	82.8	79.6	82.3	84.0
SFT Baseline		1.22	91.8	78.5	92.2	79.4
SFT Dynamics		1.34	79.7	63.0	83.2	63.4
DuplexPO		**1.03**	**98.0**	**93.6**	**99.5**	**93.3**

Better timing, knowing when to yield and more active engagement

Reasoning and Instruction Following:

Method	FD	LlamaQ	WebQ	TriQA	SDQA	AlpacaE	ComE	OBQA	MMSU
Moshi [Défossez et al., 2024]	✓	54.5	22.1	16.7	15.6	2.01	1.60	25.9	24.0
Freeze-Omni [Wang et al., 2024b]	✓	56.2	27.9	28.5	53.5	4.03	3.46	31.0	28.1
SALMONN-omni [Yu et al., 2025]	✓	73.6	43.7	56.0	-	3.22	-	-	30.0
SALM-Duplex [Hu et al., 2025]	✓	51.3	25.0	16.9	26.0	2.99	2.50	39.6	26.3
GLM-4-Voice [Zeng et al., 2024]	✗	65.7	37.0	47.5	37.0	3.97	3.42	53.4	39.8
Qwen2-Audio [Chu et al., 2024]	✗	69.7	45.2	40.3	35.7	3.74	3.43	49.5	35.7
Kimi-Audio [Ding et al., 2025]	✗	68.3	37.3	51.2	63.1	4.46	3.97	83.5	62.2
Baichuan-Audio [Li et al., 2025]	✗	74.0	40.7	53.0	45.8	4.41	4.08	71.7	53.2
STITCH-R [Chiang et al., 2025]	✗	70.0	50.3	49.6	-	2.70	-	-	-
SFT Baseline	✓	72.0	44.3	48.1	47.2	3.43	3.48	72.2	54.9
DuplexPO	✓	**75.3**	**44.5**	**49.9**	**49.8**	**3.68**	**3.74**	**73.7**	**56.2**

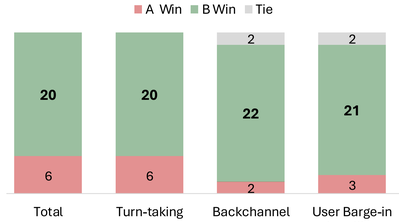
Preserves reasoning and instruction following

Latency and interruption handling (FDB-v3):

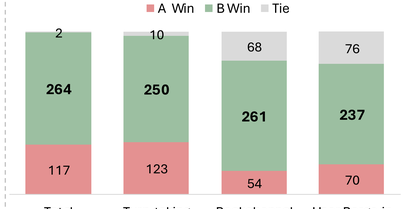
Method	Turn-Taking Dynamics		VIR (↓)	Yield (↑)
	Turn-Take (↑)	Latency (↓)		
GPT-Realtime [OpenAI, 2026]	96.0	2.65	–	–
Gemini Live 2.5 [Google, 2025]	92.0	3.84	–	–
Gemini Live 3.1 [Google, 2026]	78.0	1.59	–	–
Grok [x.ai, 2026]	94.0	3.13	–	–
Ultravox [fixie-ai, 2026]	96.0	1.90	–	–
Moshi [Défossez et al., 2024]	**100.0**	0.44	12.0	71.9
PersonaPlex [Roy et al., 2026]	98.0	1.41	15.0	66.7
SFT Baseline	99.0	7.33	8.0	64.8
SFT Dynamics	**100.0**	14.24	**4.0**	93.3
DuplexPO	**100.0**	**0.24**	5.0	**100.0**

Fewer unwanted interruptions

Pairwise behaviour comparison: A: SFT Baseline; B: DuplexPO.



(a) Pairwise Judge Evaluation on Fisher test set (n=26)



(b) Pairwise Judge Evaluation on Seamless test set (n=383)