
SpokenUS: A Spoken User Simulator for Task-Oriented Dialogue

Jonggeun Lee^{*1} Junseong Pyo^{*1} Yohan Jo¹

Abstract

Robust task-oriented spoken dialogue agents require exposure to the full diversity of human spoken interaction. Building spoken user simulators that address this requires large-scale spoken task-oriented dialogue (TOD) data encompassing spoken user behaviors, yet existing datasets are limited in scale and domain coverage, with no systematic pipeline for augmenting them. To address this, we introduce **SpokenTOD**, a spoken TOD dataset of 52,390 dialogues and 1,034 hours of speech augmented with four spoken user behaviors—cross-turn slots, barge-in, disfluency, and emotional prosody—across diverse speakers and domains. Building on SpokenTOD, we present **SpokenUS**, a spoken user simulator grounded in TOD with a dedicated architecture for barge-in. SpokenUS achieves comparable goal coverage to much larger models while substantially outperforming all baselines in Human MOS, disclosing slot values gradually across the dialogue as humans do rather than front-loading them. Further analysis shows that SpokenUS’s spoken behaviors pose meaningful challenges to agents, making it a practical tool for training and evaluating more robust spoken dialogue systems.

1. Introduction

When interacting with dialogue agents, people often produce disfluencies (Shriberg, 1994), interruptions (Khouzami et al., 2015), and emotional expressions (Schröder, 2001) rather than clean scripted utterances, and such spontaneous behaviors are known to degrade the performance of agents (Liu et al., 2025). Building spoken dialogue agents robust to this diversity is critical for real-world deployments, yet collecting sufficient spoken dialogue data for training

¹Graduate School of Data Science, Seoul National University, Seoul, Republic of Korea. Correspondence to: Yohan Jo <yohan.jo@snu.ac.kr>.

Table 1. Overview of spoken user behaviors in SpokenTOD.

Category	Type	Definition
Cross-Turn Slots	Phone numbers, email addresses, reservation IDs	
Barge-in	Error Recovery	Syst: “For two peo...” User: “No, three.”
	Clarification	Syst: “It’s nearby...” User: “Which address?”
	Efficiency	Syst: “Successfully...” User: “Thanks!”
Disfluency	Filled Pause	[FP] uh, we should go there.
	Discourse Marker	[DM] you know, the cheap one.
	Edit	reserve three, [EDIT] I mean, four people.
	Restart	book a ta- [RST] I’d like a table for two.
	Repetition	I want, [REP] I want a single room.
	Correction	left, [COR] no, right after the bank.
Emotion	Neutral, Fearful, Dissatisfied, Apologetic, Abusive, etc.	

such systems is prohibitively expensive.

User simulation offers a scalable alternative (Eckert et al., 1997), where a user simulator is provided with a user goal and interacts with an agent turn by turn to fulfill it. However, existing simulators are text-only (Sekulic et al., 2024; Shim et al., 2026) and fundamentally cannot capture the spontaneous behaviors that define spoken interaction.

While end-to-end omni models (Xu et al., 2025a;b; Zeng et al., 2024) can process speech directly, they are unsuitable as user simulators: they lack task-oriented dialogue (TOD) knowledge required to convey slot values and requests, and, having been optimized as helpful assistants, they default to help-providing rather than help-seeking behavior. One might address these gaps by training models on spoken TOD data, but existing spoken TOD datasets (Henderson et al., 2014; Si et al., 2023) are limited in scale and domain coverage, and no pipeline exists for augmenting text-based TOD corpora with diverse spoken user behaviors, making it infeasible to train spoken user simulators grounded in both TOD knowledge and realistic spoken behaviors.

We address these gaps with two contributions: **SpokenTOD**, a large-scale spoken TOD dataset that augments text-based corpora with four spoken user behaviors—cross-turn slots, barge-in, disfluency, and emotional prosody—across diverse speakers and domains; and **SpokenUS**, a spoken user simulator grounded in TOD with a dedicated barge-in mechanism. Against five end-to-end omni baselines, SpokenUS achieves goal coverage comparable to that of much larger models, discloses slot values gradually as humans do, and scores the highest Human MOS (4.06). Its spoken

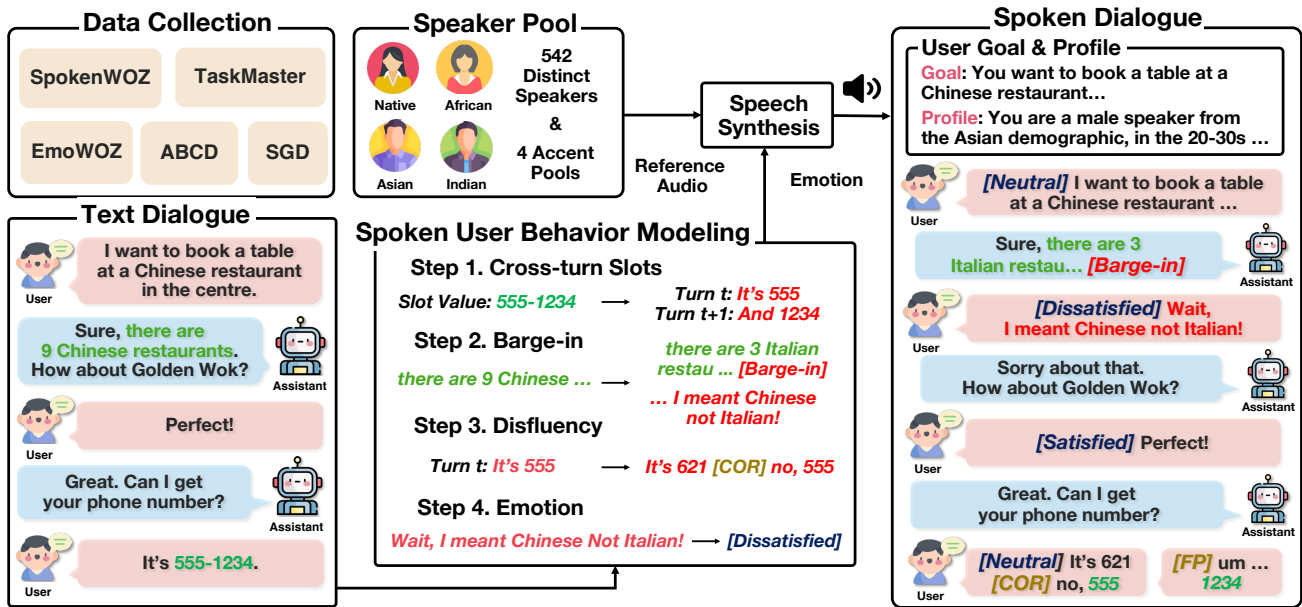


Figure 1. SpokenTOD Construction Pipeline.

behaviors also pose a substantial challenge to agents using cascaded ASR, demonstrating that SpokenUS is practical for evaluating and training robust spoken dialogue systems.

2. SpokenTOD

To construct a spoken TOD dataset capturing realistic user behaviors, we augment four widely used text-based TOD corpora—ABCD (Chen et al., 2021), EmoWOZ (emotion-labeled MultiWOZ) (Feng et al., 2022), SGD (Rastogi et al., 2020), and TaskMaster (Byrne et al., 2019)—with four spoken user behaviors: cross-turn slots, barge-in, disfluency, and emotion (see Table 1), then synthesize the result into speech. Figure 1 summarizes the augmentation and synthesis pipeline. All augmentations are performed with Qwen3-32B (Yang et al., 2025). Details are provided in Appendix A.

Cross-turn Slots. Humans deliver complex values such as phone numbers and email addresses over multiple turns rather than at once (McInnes & Attwater, 2004; Zhang et al., 2022)—behaviors known as cross-turn slots (Si et al., 2023). To model this, we segment long alphanumeric values into natural chunks (Step 1 in Figure 1).

Barge-in. In natural dialogue, speakers interrupt in real time to recover from errors, seek clarification, or move on (Khouzaimi et al., 2015). We sample 25% of agent turns as candidates (Witt, 2015); for each candidate judged appropriate by the LLM, we construct three turns: a truncated agent utterance, the user’s interruption, and the original agent response as the recovery turn (Step 2 in Figure 1).

Disfluency. Disfluency reflects the natural imperfections of real-time speech production (Shriberg, 1994). We select dis-

fluent utterances via the length-dependent model of Shriberg (1996), then inject six categories (Table 1) concentrated near slot values (Step 3 in Figure 1), where speakers are more disfluent when producing precise information (Arnold, 2013).

Emotion. To support emotion-conditioned speech synthesis, we annotate each user utterance with the LLM (Step 4 in Figure 1) using the EmoWOZ (Feng et al., 2022) label set: NEUTRAL, FEARFUL, DISSATISFIED, APOLOGETIC, ABUSIVE, EXCITED, and SATISFIED.

Speech synthesis. We synthesize the augmented dialogues with Qwen3-TTS (Hu et al., 2026), conditioning user speech on emotion-specific style prompts and reference audio sampled from a Speech Accent Archive (Weinberger & Kunath, 2011) speaker pool stratified by accent, age, and gender. Assistant turns use a fixed pool of 10 *Native*-accent speakers.

Dataset Statistics and Validation. To complement our synthesized data with authentic human speech, we additionally include SpokenWOZ (Si et al., 2023) as human-recorded samples for both training and evaluation. The resulting SpokenTOD comprises 52,390 dialogues and 1,034 hours of speech. To validate intelligibility, we measure Word Error Rate (WER) on 500 randomly sampled dialogues with Whisper-large-v3 (Radford et al., 2023), achieving 4.69%, indicating that the synthesized speech is highly intelligible.

3. SpokenUS

Overview. SpokenUS is a spoken user simulator that pursues a user goal across multi-turn exchanges while modeling user-side turn-taking and realistic spoken behaviors. As illustrated in Figure 2, it takes the user’s goal and profile, the

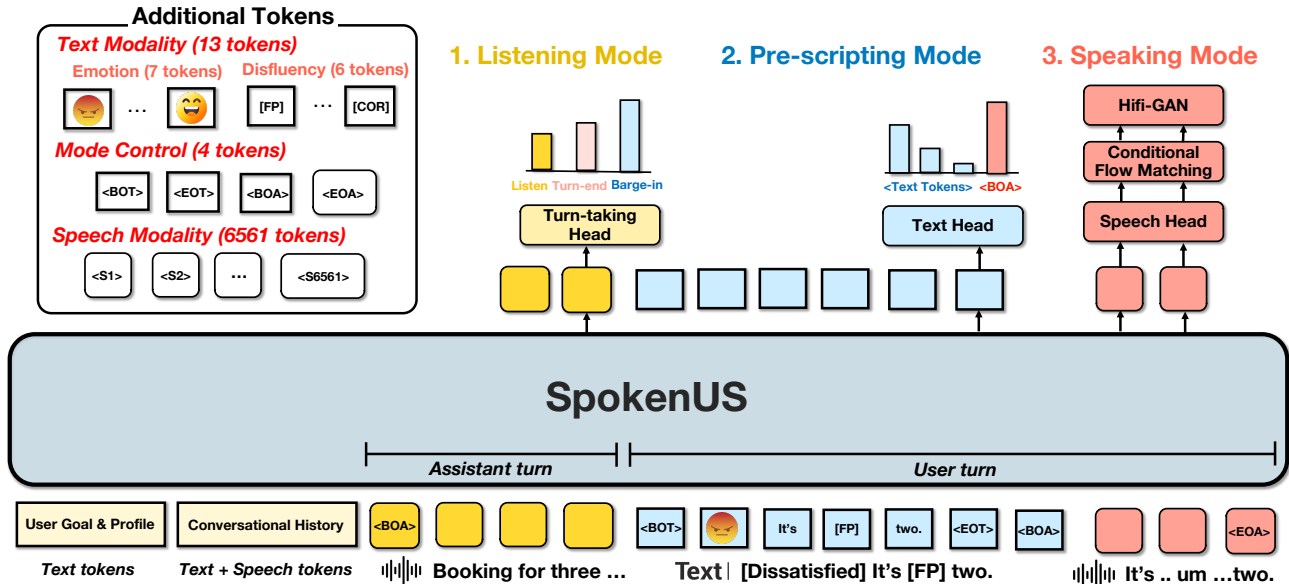


Figure 2. Overview of SpokenUS. The model processes streaming assistant speech in Listening Mode, determines barge-in timing through a turn-taking head, then generates responses through Pre-scripting and Speaking Modes.

conversational history as interleaved text and speech tokens, and the current assistant speech (complete or streaming). SpokenUS operates in three sequential modes: *Listening Mode*, *Pre-scripting Mode*, and *Speaking Mode*.

Listening Mode. During streaming input, the turn-taking head predicts one of *listen*, *barge-in*, or *turn-end* at each speech token. Predictions are aggregated as a recency-weighted sum of per-token probabilities to reduce false alarms. A *listen* prediction continues streaming, whereas *barge-in* or *turn-end* triggers a transition to Pre-scripting Mode. Non-streaming input (a turn) skips this mode.

Pre-scripting Mode. Upon entering this mode, the text head generates a structured transcript delimited by `<BOT>` and `<EOT>`. The transcript begins with an emotion token, followed by the intended utterance with interleaved disfluency tokens. This transcript guides Speaking Mode while making the simulator’s behavior transparent for evaluation.

Speaking Mode. The `<EOT>` token signals the transition to Speaking Mode, where the speech head autoregressively generates discrete speech tokens, delimited by `<BOA>` and `<EOA>`, conditioned on the transcript. These tokens are decoded to audio via a Conditional Flow Matching (CFM) model and HiFi-GAN (Kong et al., 2020) vocoder, conditioned on a reference clip from the SpokenTOD speaker pool (§2). Mode-transition details are in Appendix B.1.

Training. SpokenUS is trained with a multi-task objective combining cross-entropy losses from all three heads, $\mathcal{L} = \mathcal{L}_{\text{text}} + \mathcal{L}_{\text{speech}} + \mathcal{L}_{\text{turn-taking}}$, where $\mathcal{L}_{\text{text}}$ and $\mathcal{L}_{\text{speech}}$ supervise text and speech token generation respectively, and $\mathcal{L}_{\text{turn-taking}}$ supervises turn-taking decisions over three classes (listen,

barge-in, and turn-end). For turn-taking supervision, tokens within an assistant turn are labeled *listen* except for the final 6 tokens—approximating human reaction time (Stivers et al., 2009)—labeled *turn-end* or *barge-in* by turn type. Speech token targets are obtained by encoding SpokenTOD audio with the CosyVoice3 tokenizer (Du et al., 2025).

4. Experiments

4.1. Experimental Setup

Implementation Details. SpokenUS is initialized from Qwen2.5-3B (Qwen et al., 2025), with the vocabulary augmented by disfluency and emotion special tokens (Table 8). It is then fine-tuned in two stages: Stage 1 trains on text-only SpokenTOD for 3 epochs to acquire TOD knowledge; Stage 2 jointly trains all three heads on user turns for 12k steps using the multi-task objective in Section 3. The CFM and HiFi-GAN are trained on user speech from SpokenTOD. Further training details are provided in Appendix B.2.

Baselines. We compare SpokenUS against human recordings and five end-to-end omni models: Qwen2.5-Omni-3B (Xu et al., 2025a), InteractiveOmni-4B (Tong et al., 2025), Qwen2.5-Omni-7B, GLM-4-Voice-9B (Zeng et al., 2024), and Qwen3-Omni-30B-A3B (Xu et al., 2025b). We fix the dialogue assistant as GPT-4.1-mini (OpenAI, 2025).

Evaluation Setups. We evaluate on 100 user goals sampled from the SpokenWOZ test split. Goal coverage is assessed with GPT-4.1-mini (OpenAI, 2025), shown to closely match human slot verification in TOD (Shim et al., 2026), using **Goal Alignment (GA)**, the percentage of dialogues conveying all required slots, and **Slot Match Rate (SMR)**, the

Table 2. Main evaluation results. Best in each group in **bold**.

Model	Goal Cov.		Human MOS			
	GA	SMR	N	C	S	Avg.
Human	1.00	1.00	3.58	3.64	3.80	3.67
<i>Larger models</i>						
Qwen2.5-Omni-7B	0.80	0.97	2.86	3.29	2.94	3.03
GLM-4-Voice-9B	0.33	0.79	2.51	2.06	2.38	2.32
Qwen3-Omni-30B-A3B	0.93	0.99	3.03	3.61	2.91	3.18
<i>Comparable-sized models</i>						
Qwen2.5-Omni-3B	0.36	0.93	1.97	2.77	2.28	2.34
InteractiveOmni-4B	0.59	0.89	2.83	2.44	2.62	2.63
SpokenUS (3B)	0.82	0.99	3.92	4.14	4.11	4.06

Table 3. Final Slot F1 scores on 100% GA dialogues.

Model	Text	Cascaded
Qwen2.5-Omni-3B	86.8	83.1 \downarrow 3.7
SpokenUS	73.8	49.5 \downarrow 24.3

percentage of required slots conveyed across dialogues. For speech quality, 10 English-proficient evaluators rate 20 dialogues (each by ≥ 3 evaluators) on three 5-point scales: **Naturalness (N)**, acoustic naturalness; **Conversational Flow (C)**, role consistency and turn-by-turn engagement; and **Spoken User Behavior (S)**, appropriateness of spoken behaviors. Details of setups and results are in Appendix C.

4.2. Main Results

Goal Coverage. SpokenUS achieves a GA of 0.82, on par with Qwen2.5-Omni-7B (0.80) and outperforming similarly-sized models (Qwen2.5-Omni-3B: 0.36; InteractiveOmni-4B: 0.59). Only Qwen3-Omni-30B-A3B ($\sim 10\times$ larger) surpasses it (0.93), showing that TOD grounding enables strong goal coverage at a smaller scale. Failure cases of SpokenUS are analyzed in Appendix C.4.

Human MOS. SpokenUS achieves an average MOS of 4.06, the highest among all systems, outperforming Qwen3-Omni-30B-A3B (3.18). This indicates that SpokenUS produces natural and human-like speech across naturalness, conversational flow, and spoken user behavior (see Fig 21).

4.3. Analysis

Slot Disclosure Timing. Figure 3 shows cumulative goal-slot coverage over user turns. Whereas Omni models front-load information and plateau early, SpokenUS closely follows the gradual disclosure pattern of human speakers.

Impact on Agent Performance. To isolate the impact of spoken user behaviors, we measure Final Turn Slot F1 on dialogues where all simulators achieve full goal coverage, under **Text** (transcript) and **Cascaded** (ASR) settings. SpokenUS induces a far steeper Cascaded drop than Qwen2.5-Omni-3B (-24.3 vs. -3.7 ; Table 3), indicating that its spo-

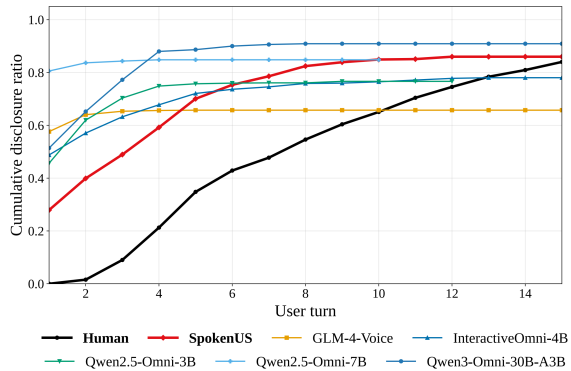


Figure 3. Cumulative goal slot coverage over user turns.

Table 4. Final Slot F1 scores of alternative assistant agents.

Assistant Model	Text	Cascaded
Qwen2.5-7B-Instruct	73.5	57.2 \downarrow 16.3
Llama-3.1-8B-Instruct	65.8	45.7 \downarrow 20.1

ken behaviors substantially challenge downstream agents under cascaded ASR. The pattern holds with two other assistant models (-16.3 , -20.1 ; Table 4).

5. Related Work

User Simulation for Task-Oriented Dialogue. User simulation has long supported TOD research, enabling agents to be trained without costly real interactions (Eckert et al., 1997). Simulators have evolved from rule- and policy-based approaches (Schatzmann et al., 2007; Chandramohan et al., 2012) to LLM-based ones (Sekulic et al., 2024; Shim et al., 2026) with more naturalistic utterances. However, these simulators operate exclusively in text, leaving spoken user behaviors unaddressed.

Spoken Dialogue System. Early cascaded ASR-LLM-TTS pipelines (Huang et al., 2024) suffer from ASR errors (Sunder et al., 2022), latency, and loss of non-linguistic cues such as emotional prosody (Défossez et al., 2024). End-to-end omni (Xu et al., 2025a;b; Zeng et al., 2024; Tong et al., 2025) and full-duplex models (Défossez et al., 2024; Ma et al., 2025) support interactive speech but are trained as assistants rather than users, lacking the proactive turn-taking required for goal-driven user simulation.

6. Conclusion

We introduced SpokenTOD, a large-scale spoken TOD dataset with spoken user behaviors, and SpokenUS, a spoken user simulator combining TOD grounding with a dedicated turn-taking head. SpokenUS matches the goal coverage of omni models $10\times$ its size, produces the most natural speech among all systems, and exposes a substantial robustness gap in current TOD agents under cascaded ASR, establishing realistic spoken user simulation as a practical foundation for evaluating and training spoken dialogue systems.

References

- Arnold, J. E. Information status relates to production, distribution, and comprehension. *Frontiers in Psychology*, Volume 4 - 2013, 2013. ISSN 1664-1078. doi: 10.3389/fpsyg.2013.00235. URL <https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2013.00235>.
- Bakhturina, E., Zhang, Y., and Ginsburg, B. Shallow Fusion of Weighted Finite-State Transducer and Language Model for Text Normalization. In *Interspeech 2022*, pp. 491–495, 2022. doi: 10.21437/Interspeech.2022-11074.
- Byrne, B., Krishnamoorthi, K., Sankar, C., Neelakantan, A., Goodrich, B., Duckworth, D., Yavuz, S., Dubey, A., Kim, K.-Y., and Cedilnik, A. Taskmaster-1: Toward a realistic and diverse dialog dataset. In Inui, K., Jiang, J., Ng, V., and Wan, X. (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 4516–4525, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1459. URL <https://aclanthology.org/D19-1459/>.
- Chandramohan, S., Geist, M., Lefevre, F., and Pietquin, O. Behavior specific user simulation in spoken dialogue systems. In *Speech Communication; 10. ITG Symposium*, pp. 1–4, 2012.
- Chen, D., Chen, H., Yang, Y., Lin, A., and Yu, Z. Action-based conversations dataset: A corpus for building more in-depth task-oriented dialogue systems. In Toutanova, K., Rumshisky, A., Zettlemoyer, L., Hakkani-Tur, D., Beltagy, I., Bethard, S., Cotterell, R., Chakraborty, T., and Zhou, Y. (eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3002–3017, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.239. URL <https://aclanthology.org/2021.naacl-main.239/>.
- Corley, M., MacGregor, L. J., and Donaldson, D. I. It’s the way that you, er, say it: Hesitations in speech affect language comprehension. *Cognition*, 105(3):658–668, 2007. ISSN 0010-0277. doi: <https://doi.org/10.1016/j.cognition.2006.10.010>. URL <https://www.sciencedirect.com/science/article/pii/S0010027706002186>.
- Défossez, A., Mazaré, L., Orsini, M., Royer, A., Pérez, P., Jégou, H., Grave, E., and Zeghidour, N. Moshi: a speech-text foundation model for real-time dialogue. *arXiv preprint arXiv:2410.00037*, 2024.
- Du, Z., Gao, C., Wang, Y., Yu, F., Zhao, T., Wang, H., Lv, X., Wang, H., Ni, C., Shi, X., et al. Cosyvoice 3: Towards in-the-wild speech generation via scaling-up and post-training. *arXiv preprint arXiv:2505.17589*, 2025.
- Eckert, W., Levin, E., and Pieraccini, R. User modeling for spoken dialogue system evaluation. In *1997 IEEE Workshop on Automatic Speech Recognition and Understanding Proceedings*, pp. 80–87, 1997. doi: 10.1109/ASRU.1997.658991.
- Feng, S., Lubis, N., Geishauser, C., Lin, H.-c., Heck, M., van Niekerk, C., and Gasic, M. EmoWOZ: A large-scale corpus and labelling scheme for emotion recognition in task-oriented dialogue systems. In Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Odijk, J., and Piperidis, S. (eds.), *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pp. 4096–4113, Marseille, France, June 2022. European Language Resources Association. URL <https://aclanthology.org/2022.lrec-1.436/>.
- Henderson, M., Thomson, B., and Williams, J. D. The second dialog state tracking challenge. In Georgila, K., Stone, M., Hastie, H., and Nenkova, A. (eds.), *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pp. 263–272, Philadelphia, PA, U.S.A., June 2014. Association for Computational Linguistics. doi: 10.3115/v1/W14-4337. URL <https://aclanthology.org/W14-4337/>.
- Hu, H., Zhu, X., He, T., Guo, D., Zhang, B., Wang, X., Guo, Z., Jiang, Z., Hao, H., Guo, Z., et al. Qwen3-tts technical report. *arXiv preprint arXiv:2601.15621*, 2026.
- Huang, R., Li, M., Yang, D., Shi, J., Chang, X., Ye, Z., Wu, Y., Hong, Z., Huang, J., Liu, J., Ren, Y., Zou, Y., Zhao, Z., and Watanabe, S. Audiogpt: understanding and generating speech, music, sound, and talking head. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence, AAAI’24/IAAI’24/EAAI’24*. AAAI Press, 2024. ISBN 978-1-57735-887-9. doi: 10.1609/aaai.v38i21.30570. URL <https://doi.org/10.1609/aaai.v38i21.30570>.
- Khouzaimi, H., Laroche, R., and Lefèvre, F. Turn-taking phenomena in incremental dialogue systems. In Márquez, L., Callison-Burch, C., and Su, J. (eds.), *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 1890–1895, Lisbon, Portugal, September 2015. Association for Computational

- Linguistics. doi: 10.18653/v1/D15-1216. URL <https://aclanthology.org/D15-1216/>.
- Kong, J., Kim, J., and Bae, J. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 17022–17033. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/c5d736809766d46260d816d8dbc9eb44-Paper.pdf.
- Kowtko, J. C. and Price, P. J. Data collection and analysis in the air travel planning domain. In *Speech and Natural Language: Proceedings of a Workshop Held at Cape Cod, Massachusetts, October 15-18, 1989*, 1989. URL <https://aclanthology.org/H89-2017/>.
- Liu, H., Hou, Y., Liu, H., Wang, Y., Wang, Y., and Wang, Y. Vocalbench-df: A benchmark for evaluating speech llm robustness to disfluency, 2025. URL <https://arxiv.org/abs/2510.15406>.
- Ma, Z., Song, Y., Du, C., Cong, J., Chen, Z., Wang, Y., Wang, Y., and Chen, X. Language model can listen while speaking. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(23):24831–24839, Apr. 2025. doi: 10.1609/aaai.v39i23.34665. URL <https://ojs.aaai.org/index.php/AAAI/article/view/34665>.
- Manotumruksa, J., Dalton, J., Meij, E., and Yilmaz, E. Improving dialogue state tracking with turn-based loss function and sequential data augmentation. In Moens, M.-F., Huang, X., Specia, L., and Yih, S. W.-t. (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 1674–1683, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-emnlp.144. URL <https://aclanthology.org/2021.findings-emnlp.144/>.
- McInnes, F. and Attwater, D. Turn-taking and grounding in spoken telephone number transfers. *Speech Communication*, 43(3):205–223, 2004. ISSN 0167-6393. doi: <https://doi.org/10.1016/j.specom.2004.04.001>. URL <https://www.sciencedirect.com/science/article/pii/S016763930400041X>.
- Meteor, M. et al. Dysfluency annotation stylebook for the savitchboard corpus. Technical report, Technical report, 1995.
- OpenAI. Introducing gpt-4.1 in the api. <https://openai.com/index/gpt-4-1/>, 2025. Accessed 2026-03-04.
- Ortony, A., Clore, G. L., and Collins, A. *The cognitive structure of emotions*. Cambridge university press, 2022.
- Qwen, :, Yang, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Li, C., Liu, D., Huang, F., Wei, H., Lin, H., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Lin, J., Dang, K., Lu, K., Bao, K., Yang, K., Yu, L., Li, M., Xue, M., Zhang, P., Zhu, Q., Men, R., Lin, R., Li, T., Tang, T., Xia, T., Ren, X., Ren, X., Fan, Y., Su, Y., Zhang, Y., Wan, Y., Liu, Y., Cui, Z., Zhang, Z., and Qiu, Z. Qwen2.5 technical report, 2025. URL <https://arxiv.org/abs/2412.15115>.
- Radford, A., Kim, J. W., Xu, T., Brockman, G., Mcleavey, C., and Sutskever, I. Robust speech recognition via large-scale weak supervision. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J. (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 28492–28518. PMLR, 23–29 Jul 2023. URL <https://proceedings.mlr.press/v202/radford23a.html>.
- Rastogi, A., Zang, X., Sunkara, S., Gupta, R., and Khaitan, P. Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8689–8696, Apr. 2020. doi: 10.1609/aaai.v34i05.6394. URL <https://ojs.aaai.org/index.php/AAAI/article/view/6394>.
- Schatzmann, J., Thomson, B., Weilhammer, K., Ye, H., and Young, S. Agenda-based user simulation for bootstrapping a POMDP dialogue system. In Sidner, C., Schultz, T., Stone, M., and Zhai, C. (eds.), *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume, Short Papers*, pp. 149–152, Rochester, New York, April 2007. Association for Computational Linguistics. URL <https://aclanthology.org/N07-2038/>.
- Schröder, M. Emotional speech synthesis: a review. In *Interspeech*, volume 2001, pp. 561–564, 2001.
- Sekulic, I., Terragni, S., Guimarães, V., Khau, N., Guedes, B., Filipavicius, M., Manso, A. F., and Mathis, R. Reliable LLM-based user simulator for task-oriented dialogue systems. In Graham, Y., Liu, Q., Lampouras, G., Iacobacci, I., Madden, S., Khalid, H., and Qureshi, R. (eds.), *Proceedings of the 1st Workshop on Simulating Conversational Intelligence in Chat (SCI-CHAT 2024)*, pp. 19–35, St. Julians, Malta, March 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.scichat-1.3. URL <https://aclanthology.org/2024.scichat-1.3/>.

- Sen, P. Speech disfluencies occur at higher perplexities. In Zock, M., Chersoni, E., Lenci, A., and Santus, E. (eds.), *Proceedings of the Workshop on the Cognitive Aspects of the Lexicon*, pp. 92–97, Online, December 2020. Association for Computational Linguistics. URL <https://aclanthology.org/2020.cogalex-1.11/>.
- Shim, J., Song, W., Jin, C., KooK, S., and Jo, Y. Non-collaborative user simulators for tool agents. In *The Fourteenth International Conference on Learning Representations*, 2026. URL <https://openreview.net/forum?id=UAUimofy3W>.
- Shriberg, E. Disfluencies in switchboard. In *Proceedings of international conference on spoken language processing*, volume 96, pp. 11–14. IEEE Philadelphia, PA, 1996.
- Shriberg, E. E. Preliminaries to a theory of speech disfluencies. *Doctoral dissertation, University of California at Berkeley*, 1994.
- Si, S., Ma, W., Gao, H., Wu, Y., Lin, T.-E., Dai, Y., Li, H., Yan, R., Huang, F., and Li, Y. Spokenwoz: A large-scale speech-text benchmark for spoken task-oriented dialogue agents. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S. (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 39088–39118. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/7b16688a2b053a1b01474ab5c78ce662-Paper-Datasets_and_Benchmarks.pdf.
- Stivers, T., Enfield, N. J., Brown, P., Englert, C., Hayashi, M., Heinemann, T., Hoymann, G., Rossano, F., de Ruiter, J. P., Yoon, K.-E., and Levinson, S. C. Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences*, 106(26): 10587–10592, 2009. doi: 10.1073/pnas.0903616106. URL <https://www.pnas.org/doi/abs/10.1073/pnas.0903616106>.
- Sunder, V., Thomas, S., Kuo, H.-K. J., Ganhotra, J., Kingsbury, B., and Fosler-Lussier, E. Towards end-to-end integration of dialog history for improved spoken language understanding. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7497–7501, 2022. doi: 10.1109/ICASSP43922.2022.9747871.
- Tong, W., Guo, H., Ran, D., Chen, J., Lu, J., Wang, K., Li, K., Zhu, X., Li, J., Li, K., et al. Interactiveomni: A unified omni-modal model for audio-visual multi-turn dialogue. *arXiv preprint arXiv:2510.13747*, 2025.
- Weinberger, S. H. and Kunath, S. A. The speech accent archive: towards a typology of english accents. *Language & Computers*, 73(1), 2011.
- Witt, S. Modeling user response timings in spoken dialog systems. *Int. J. Speech Technol.*, 18(2):231–243, June 2015. ISSN 1381-2416. doi: 10.1007/s10772-014-9265-1. URL <https://doi.org/10.1007/s10772-014-9265-1>.
- Xu, J., Guo, Z., He, J., Hu, H., He, T., Bai, S., Chen, K., Wang, J., Fan, Y., Dang, K., Zhang, B., Wang, X., Chu, Y., and Lin, J. Qwen2.5-omni technical report, 2025a. URL <https://arxiv.org/abs/2503.20215>.
- Xu, J., Guo, Z., Hu, H., Chu, Y., Wang, X., He, J., Wang, Y., Shi, X., He, T., Zhu, X., Lv, Y., Wang, Y., Guo, D., Wang, H., Ma, L., Zhang, P., Zhang, X., Hao, H., Guo, Z., Yang, B., Zhang, B., Ma, Z., Wei, X., Bai, S., Chen, K., Liu, X., Wang, P., Yang, M., Liu, D., Ren, X., Zheng, B., Men, R., Zhou, F., Yu, B., Yang, J., Yu, L., Zhou, J., and Lin, J. Qwen3-omni technical report, 2025b. URL <https://arxiv.org/abs/2509.17765>.
- Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.
- Zeng, A., Du, Z., Liu, M., Wang, K., Jiang, S., Zhao, L., Dong, Y., and Tang, J. Glm-4-voice: Towards intelligent and human-like end-to-end spoken chatbot. *arXiv preprint arXiv:2412.02612*, 2024.
- Zhang, S., Hu, Y., Wu, Y., Wu, J., Li, Y., Sun, J., Yuan, C., and Wang, X. A slot is not built in one utterance: Spoken language dialogs with sub-slots. In Muresan, S., Nakov, P., and Villavicencio, A. (eds.), *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 309–321, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.27. URL <https://aclanthology.org/2022.findings-acl.27/>.
- Zhang, Y., Bakhturina, E., and Ginsburg, B. NeMo (Inverse) Text Normalization: From Development to Production. In *Interspeech 2021*, pp. 4857–4859, 2021.

A. SpokenTOD Details

In this section, we provide a detailed breakdown of the data processing pipeline used to construct SpokenTOD. For all LLM-based augmentation and annotation steps, we use Qwen3-32B (Yang et al., 2025).

A.1. Dataset Details

This section provides details about the five datasets used to construct SpokenTOD.

SGD. Schema-Guided Dialogue (SGD) dataset (Rastogi et al., 2020) contains over 20K task-oriented conversations spanning 20 domains such as travel, media, banking, and events. SGD adopts a schema-guided mechanism in which each service is described via natural language definitions of its intents and slots, rather than a fixed ontology. Dialogues were generated via a rule-based simulator and subsequently paraphrased by crowd workers to improve naturalness, with slot annotations recovered automatically via string matching.

Taskmaster-2. Taskmaster-2 (Byrne et al., 2019) consists of 17K spoken two-person dialogues collected via a Wizard-of-Oz (WOZ) methodology across seven domains: restaurants, food ordering, movies, hotels, flights, music, and sports. User utterances were transcribed from audio recordings of crowdsourced workers via ASR. However, the original audio recordings are not publicly released.

ABCD Action-Based Conversations Dataset (ABCD) (Chen et al., 2021) is a fully-labeled human-to-human dataset containing over 10K customer service dialogues with 55 distinct user intents. Data was collected via an Expert Live Chat system pairing trained call center agents with crowdsourced users, with action labels annotated by agents in real time during the conversation.

EmoWOZ. EmoWOZ (Feng et al., 2022) is a large-scale emotion-annotated corpus built on MultiWOZ, containing over 11K dialogues with 83K user utterance-level emotion labels. Each utterance was annotated by three crowd workers via majority voting. The labelling scheme is grounded in the Ortony–Clore–Collins (OCC) model (Ortony et al., 2022), defining seven emotion categories tailored to task-oriented settings: NEUTRAL, FEARFUL, DISSATISFIED, APOLOGETIC, ABUSIVE, EXCITED, and SATISFIED. We adopt this label set for emotion annotation across all source datasets in SpokenTOD, using EmoWOZ’s human-annotated labels directly where available.

SpokenWOZ. SpokenWOZ (Si et al., 2023) is the first large-scale speech-text benchmark for spoken TOD, comprising 5.7K dialogues, 203K turns, and 249 hours of audio from human-to-human spoken conversations across 8 domains. It introduces spoken-specific challenges such as cross-turn slot filling and word-by-word reasoning, which are absent in text-only datasets. As the only natively spoken source dataset in SpokenTOD, SpokenWOZ provides real acoustic and prosodic grounding for our dataset.

Dataset Licenses All datasets are used in accordance with their respective licenses: ABCD and Taskmaster-2 under CC BY 4.0, SpokenWOZ and EmoWOZ under CC BY-NC 4.0, SGD under CC BY-SA 4.0, and the Speech Accent Archive under CC BY-NC-SA 2.0. These datasets were originally released for academic research in task-oriented dialogue systems, and our use for benchmark evaluation is consistent with this intended purpose. Our dataset and model are provided for research purposes only.

A.2. Dataset Unification

To construct SpokenTOD from source datasets with varying formats, we unify each dataset into a common schema with two components: a user goal $G = (T, S)$, and slot value positions within each utterance for injecting disfluency near slot values (Section 2). G consists of a natural-language goal text T and a structured goal S ; S records for each sub-goal, the domain, intent type, and slot–value pairs (e.g., `find_and_book` in the `hotel` domain with `area: south, pricerange: moderate`). As each source dataset has its own annotation format and characteristics, we apply dataset-specific preprocessing procedures as follows.

SGD. SGD annotates each turn with start and end positions for slot values, which we use directly. Since SGD does not provide explicit user goals G , we reconstruct it from turn-level annotations accumulated across the dialogue: S is

constructed by mapping turn-level intent and slot-value pairs into the unified schema, from which T is generated by filling a rule-based template with the collected intents and constraints.

TM-2. TM-2 annotates each turn with `start_index` and `end_index` fields marking slot value positions, which we use directly. Since TM-2 does not provide explicit user goals G or intents, we reconstruct it from slot-value arguments collected across turns: S is constructed by normalizing these into the unified schema, from which T is generated by filling a rule-based template with the collected arguments.

ABCD. ABCD pairs each dialogue with a delexicalized version in which slot values are replaced with typed placeholders (e.g., `<username>`, `<email>`, `<order_id>`). We recover slot value positions by aligning the delexicalized and original utterances via prefix/suffix matching; for values not covered by placeholders, we match them directly from the metadata. Since ABCD does not provide explicit user goals G , we reconstruct it from the scenario metadata: S is constructed by mapping domain and intent labels with recovered slot-value pairs into the unified schema, from which T is generated by filling a rule-based template with the customer prompt and scenario constraints.

EmoWOZ. EmoWOZ extends MultiWOZ with per-turn emotion labels while inheriting its goal and slot annotations. We use the slot value positions from MultiWOZ directly. G is taken directly from the MultiWOZ goal annotations: S is mapped into the unified schema, and T is the natural-language goal description provided by MultiWOZ. Emotion labels are used as provided.

SpokenWOZ. SpokenWOZ provides dialogue state and dialogue act annotations from which we derive slot value positions. G is constructed following the same procedure as EmoWOZ, using SpokenWOZ’s goal annotations recorded in the MultiWOZ format.

A.3. Cross-Turn Slots Details

Cross-turn slots represent slot values that span multiple turns rather than in a single utterance, common with long alphanumeric sequences (phone numbers, email addresses, reservation IDs). We apply type-specific segmentation to all segmentable slot values: numeric sequences are split into 3–4 digit chunks, email addresses are decomposed into components with explicit punctuation vocalization (e.g., "dot", "at"), and alphanumeric codes separate letters from digits. To model realistic error patterns, we inject mistakes with $P_{\text{error}} = 0.20$, following the 20% error rate used in SpokenWOZ (Si et al., 2023). This generates correction sub-dialogues where the user dictates an incorrect segment at turn t , then self-corrects at turn $t+1$ (e.g., "Wait, I meant..."), and the assistant acknowledges the correction.

A.4. Barge-in Details

Barge-in refers to user interruptions during system speech, occurring in up to 25% of responses in spoken dialogue systems (Witt, 2015). Following the turn-taking taxonomy of Khouzaimi et al. (2015), we categorize barge-in by the assistant’s utterance state and user’s response type: (1) **Error recovery** (G_INCOHERENT): the user interrupts to correct inaccurate information in the assistant’s utterance (e.g., "No, I said Paris, not London"); (2) **Clarification** (G_FAIL): the user interrupts due to comprehension failure, expressing general confusion or requesting specific details; (3) **Efficiency** (G_SUFFICIENT): the user interrupts once enough information has been conveyed to move the conversation forward (e.g., "Yes, let’s book it").

We apply barge-in augmentation to 25% of randomly sampled user turns via a four-stage protocol. We first sample turns and uniformly assign one of the three barge-in types, along with one of three response styles (Khouzaimi et al., 2015): **Implicit** (REF_IMPL), where the user responds with minimal acknowledgment (e.g., "Aha"); **Raw** (REF_RAW), where the user directly references a specific element (e.g., "Ok, Sunday"); and **Interpreted** (REF_INTERP), where the user adds their own interpretation (e.g., "Yeah, Sunday works – that’s my only free day"). We then prompt Qwen3-32B (Yang et al., 2025) to assess whether the assistant’s utterance is a suitable target for the selected barge-in type, i.e., whether it contains incoherent information, unclear content, or sufficient information to proceed. If valid, the model generates a modified user turn that naturalistically implements the selected barge-in behavior and response style with appropriate discourse markers (e.g., "Wait", "Actually"). Detailed prompts for each combination of barge-in type and response style are provided in Figures 4–5 (Error Recovery), 6–7 (Clarification), and 8–10 (Efficiency).

A.5. Disfluency Details

Scope and sampling. We follow the exponential length-dependent model from Shriberg (1996), where the probability of an utterance being disfluent increases with its length L (measured in words):

$$P(\text{disfluent} \mid L) = 1 - b^L \quad (1)$$

where b represents the word-level fluency rate. Following Shriberg (1996), we set $b = 0.9453$ based on the AMEX corpus (Kowtko & Price, 1989), as it represents customer service interactions similar to task-oriented dialogue. This formulation ensures that longer utterances are more likely to contain disfluencies, reflecting patterns observed in spontaneous dialogue where speakers are more prone to self-corrections and hesitations in extended turns.

When an utterance is sampled to be disfluent, we uniformly sample one type from the six disfluency categories defined in Table 1: [FP], [DM], [EDIT], [REP], [COR], and [RST]. Following prior findings that disfluencies increase under higher planning and lexical retrieval difficulty (Corley et al., 2007; Arnold, 2013; Sen, 2020), we inject them more frequently near slot values, as retrieving and articulating domain-specific entities demands greater cognitive effort. For all disfluency types except corrections, we insert them with 50% probability near slots (within a 2-word window) and 50% probability uniformly across the utterance, while corrections [COR] exclusively target slot values.

Disfluency placement. We follow the annotation guidelines from the Switchboard Corpus (Meteer et al., 1995) for linguistically motivated disfluency placement. Filled pauses [FP], discourse markers [DM], and editing terms [EDIT] are inserted before the target word. Repetitions [REP] duplicate the target span. Corrections [COR] replace the original slot value with an alternative. Restarts [RST] abandon the current utterance structure and begin a rephrased continuation.

A.6. Emotion Annotation Details

To support emotion-conditioned speech synthesis, we annotate each user turn using LLM with emotion labels: Neutral, Fearful, Dissatisfied, Apologetic, Abusive, Excited, and Satisfied. We use Qwen3-32B (Yang et al., 2025) as our primary annotator. Annotation is skipped for EmoWOZ, as it already provides emotion labels. For turns that are cross-turn segments, we skip annotation and instead inherit the emotion label from the most recent non-segment user turn.

To ensure the reliability of LLM-based annotation, we validated Qwen3-32B against the ground truth labels of EmoWOZ. We sampled 50 instances per category (with 19 instances for Abusive due to its limited occurrence in the dataset, totaling 319 samples) and compared the model’s predictions with the original labels. Qwen3-32B achieved an overall accuracy of 0.75, demonstrating its capability to capture affective states in task-oriented interactions.

A.7. Speech Synthesis Details

Text Normalization. Task-oriented dialogues frequently contain numbers, dates, times, addresses, and other entities that require normalization before speech synthesis. For instance, *“I need a table for 2 at 7:30pm”* must be converted to *“I need a table for two at seven thirty pm”* for proper pronunciation. We apply NeMo-text-processing (Zhang et al., 2021; Bakhturina et al., 2022) to all utterances before synthesis. Key normalization patterns include:

- **Numbers:** Cardinal (e.g., “123” → “one hundred twenty three”) and ordinal forms (e.g., “1st” → “first”)
- **Dates and times:** “01/15/2024” → “January fifteenth twenty twenty four”, “7:30pm” → “seven thirty p m”
- **Currency and measurements:** “\$50” → “fifty dollars”, “5km” → “five kilometers”
- **Abbreviations and acronyms:** “Dr.” → “doctor”, “NYC” → “N Y C” or “New York City” based on context

A.8. Emotion Conditioning

Qwen3-TTS conditions on emotion via instruction prompts of the form: “Please speak in a {emotion} tone.”

We construct our emotion keyword set from the top-100 speaking styles in CosyVoice3’s pre-training data (Du et al., 2025). We first group these styles into seven coarse categories based on semantic similarity (neutral, fearful, dissatisfied, apologetic, abusive, excited, satisfied). We then evaluate the acoustic realization of each keyword through synthesis and select those that

Table 5. Emotion-to-keyword mapping for Qwen3-TTS instruction prompts. Keywords are drawn from the top-100 speaking styles in CosyVoice3’s pre-training data.

Emotion Category	Style Keywords
NEUTRAL	calm, indifferent, patient, relaxed
FEARFUL	fearful, shocked, surprised
DISSATISFIED	angry, contempt, disgusted, defiant
APOLOGETIC	compassionate, selfless, humble
ABUSIVE	commanding, authoritative, merciless, loud, vengeful
EXCITED	adventurous, energetic, passionate, curious, creative, joyful
SATISFIED	proud, hopeful, happy, cheerful

most effectively convey the intended emotion category, resulting in 3-6 keywords per category (Table 5). During synthesis, we randomly sample one keyword from the turn’s assigned category to introduce prosodic variation while maintaining category consistency.

A.9. Speaker Diversity

Reference audio clips from the Speech Accent Archive (Weinberger & Kunath, 2011) provide speaker identity and accent for zero-shot voice cloning. We filter candidate speakers to those with reference audio clips of at most 25 seconds, then organize them into four broad accent pools: Native, African, Indian, and Asian. A reference speaker is sampled via a three-stage procedure: (1) sample an accent pool using mixture weights derived from U.S. Census (2024) population estimates¹ (Table 6), (2) sample a country uniformly within the selected pool, and (3) randomly sample a speaker from that country. For demographic balancing, we stratify speaker age into four bins (10s, 20–30s, 40–50s, 60+) and sample each bin with equal probability (25%). Gender balance is also maintained equally.

The resulting pool comprises **542** unique speakers: 433 Native (79.89%), 57 Asian (10.52%), 38 Indian (7.01%), and 14 African (2.58%), reflecting the U.S. Census population weights used in sampling. For assistant utterances, we fix a pool of 10 Native-accent speakers (5 male / 5 female) and exclude them from the user pool to avoid identity overlap.

Table 6. Accent pool sampling weights based on U.S. Census (2024) estimates and resulting speaker counts. Speaker counts reflect the availability of qualifying recordings in the Speech Accent Archive after applying a 25-second duration filter.

Pool	Countries	Population	Weight	# Spk.
Native	USA, UK, Canada, Australia, Ireland	197,875,652	74.57%	433
African	Ethiopia, Nigeria, Ghana, Senegal, Morocco	42,951,595	16.19%	14
Indian	India, Pakistan, Afghanistan, Bangladesh, Nepal	2,442,428	0.92%	38
Asian	China, South Korea, Japan, Philippines, Vietnam	22,080,844	8.32%	57
Total		265,350,519	100.00%	542

A.10. Dataset Statistics and Validation Details

A.10.1. DATA SPLITS

We split SpokenTOD into train (39,430; 75%), validation (5,390; 10%), and test (7,570; 15%) sets.

A.10.2. DURATION VERIFICATION

We verify that audio files exist for all dialogue turns in the metadata and that durations fall within expected ranges (0.3-30 seconds per turn). Total dataset duration is computed by summing individual file durations, validated through our pipeline which processes 1,208,554 audio files totaling 1,034 hours of speech.

¹<https://www.census.gov/data/datasets/time-series/demo/popest/2020s-national-detail.html>

A.10.3. SYNTHESIZED AUDIO INTELLIGIBILITY

To validate the intelligibility of our synthesized speech, we measure WER on 500 randomly sampled dialogues using Whisper-large-v3 (Radford et al., 2023), covering all four accent groups. As shown in Table 7, our dataset achieves an overall WER of 4.69%, confirming that the synthesized audio is sufficiently intelligible.

Table 7. ASR performance on synthesized speech using Whisper-large-v3

Accent Group	WER (%)	# Utterances
African	5.08	1,340
Asian	3.77	1,242
Indian	4.95	1,355
Native	4.91	1,311
Overall	4.69	5,248

A.10.4. LLM BARGE-IN APPROPRIATENESS

To verify whether the LLM places barge-ins at contextually appropriate moments during augmentation (§A.4), we sample 20 barge-in turns from SpokenTOD and ask three annotators to judge whether each barge-in is contextually appropriate. A barge-in is judged appropriate when the assistant’s preceding utterance contains a clear trigger for interruption, such as information conflicting with the user’s stated goal (e.g., recommending a south-area restaurant when the user requested the centre). Across 60 total ratings (20 turns \times 3 annotators), **83%** of LLM-placed barge-ins are rated as contextually appropriate, indicating that the LLM understands the contextual conditions for barge-in: it constructs assistant turns whose preceding content naturally motivates user interruption. The audio-side realization of these barge-ins is further evaluated as part of the synthesized audio quality evaluation (Appendix C.7, Barge-in sub-dimension).

B. SpokenUS Details

B.1. Operating Modes Details

Listening Mode. Each assistant turn begins with the delimiter `<|im_start|>assistant<BOA>`. SpokenUS processes each incoming speech token to obtain its last hidden state, which is passed to the turn-taking head. The head applies a softmax over three classes (listen, barge-in, and turn-end), and the resulting probabilities are aggregated via a linear-recency weighting scheme to determine whether to interrupt or yield.

Pre-scripting Mode. Once a barge-in or turn-end is detected, a turn delimiter (`<|im_end|><|im_start|>user<BOT>`) is inserted at that position. The last hidden state is then passed to the text head, which autoregressively generates the structured transcript until `<EOT>` is predicted.

Speaking Mode. Upon predicting `<EOT>`, `<BOA>` is appended and the last hidden state of each subsequent token is routed to the speech head. Speech token generation continues until `<EOA>` is predicted, at which point the generated discrete speech tokens are converted to audio via the two-stage synthesis pipeline described in Section 3.

B.2. SpokenUS Training Details

SpokenUS. SpokenUS is initialized from Qwen2.5-3B (Qwen et al., 2025) and fine-tuned in two stages. We augment the vocabulary with disfluency and emotion special tokens, as listed in Table 8. In Stage 1, the model is trained on full text-only SpokenTOD dialogues (both user and assistant turns) for 3 epochs, enabling the model to acquire task-oriented dialogue knowledge such as goal-directed response generation. In Stage 2, all three heads are jointly trained for 12k steps using the multi-task objective in Section 3, supervised on user turns only. Hyperparameters are summarized in Table 9.

CFM and HiFi-GAN. The CFM and HiFi-GAN are both initialized from CosyVoice3 (Du et al., 2025) and fine-tuned exclusively on user speech from SpokenTOD. Training hyperparameters are summarized in Tables 9.

Table 8. Disfluency and emotion related tokens added to the vocabulary.

Category	Tokens
Disfluency	[COR], [DM], [EDIT], [FP], [REP], [RST]
Emotion	< abusive >, < apologetic >, < excited >, < satisfied >, < fearful >, < neutral >, < dissatisfied >

Table 9. Training hyperparameters for SpokenUS, CFM, and HiFi-GAN.

SpokenUS (LM)		CFM		HiFi-GAN	
Hyperparameter	Value	Hyperparameter	Value	Hyperparameter	Value
Backbone	Qwen2.5-3B	Parameters	300M	Optimizer	Adam
Hidden size	2048	DiT dim / depth / heads	1024 / 22 / 16	Learning rate (G / D)	$1 \times 10^{-5} / 2 \times 10^{-4}$
Context length	4,096	Mel dimension	80	Gradient clipping	5.0
Optimizer	AdamW	Token frame rate	25 Hz	Gradient accumulation	4
Learning rate	5×10^{-5}	Token-mel ratio	2	Warmup steps	2,500
β_1, β_2	0.9, 0.999	Solver	Euler	Discriminator warmup	1,000
Weight decay	0.01	σ_{\min}	1×10^{-6}	Training steps	30k
Scheduler	Const. w/ warmup	CFG rate (train / infer)	0.2 / 0.7	Sampling rate	24 kHz
Warmup steps	2,500	Optimizer	Adam	Mel dimension	80
Gradient clipping	1.0	Learning rate	1×10^{-5}	Base channels	512
Grad. accumulation	4	Gradient clipping	5.0	Upsample rates	[8, 5, 3]
Batch size	10	Gradient accumulation	4		
Max epochs (stage 1)	3	Warmup steps	2,500		
Training steps (stage 2)	12k	Training steps	30k		
Hardware	$5 \times$ H200				

C. Experiment Details

C.1. Evaluation Data

Table 10. Slot and domain distribution across 100 evaluation dialogues.

Slot Distribution (%)		Domain Distribution (%)	
Area	71.5%	Train	47.5%
Book_people	65.0%	Hotel	43.5%
Email	62.0%	Restaurant	43.0%
Book_day	58.0%	Attraction	34.5%
Departure	52.0%	Taxi	20.5%
Phone	48.5%	Hospital	0.5%
Reference	39.5%	Police	0.5%

We use the **SpokenWOZ test set** as our evaluation benchmark, as it provides human–human spoken recordings with fully annotated dialogue states and database queries necessary for running interactive user–assistant conversations. We sample 100 dialogues satisfying two criteria: (1) **domain diversity**, ensuring broad coverage across the seven domains in SpokenWOZ, and (2) **human goal alignment**, retaining only dialogues in which the human user successfully conveys all required goal slots ($GA = 1.0$). The latter criterion ensures that any gap in goal alignment observed across simulators reflects simulator capability rather than inherent dialogue difficulty. Table 10 summarizes the slot and domain distribution of the resulting evaluation set.

C.2. Baseline Inference

All baseline models except InteractiveOmni-4B are queried with temperature = 0 and top- p = 1.0. InteractiveOmni-4B follows the official configuration (do_sample=True, max_new_tokens=1024), with temperature and top- p at their default values.

C.3. Goal Alignment Details

Goal alignment is evaluated turn-by-turn using GPT-4.1-mini with temperature = 0. At each user turn, the model is presented with the remaining uncovered goal items and the dialogue history, and selects which items are explicitly mentioned, confirmed, or requested in the user’s utterance. A dialogue is considered goal-aligned if all goal items are covered by the end of the conversation. **Goal Alignment Rate (GA)** is the fraction of dialogues that are fully goal-aligned, and **Slot Match Rate (SMR)** is the fraction of slot-type goal items covered, micro-averaged across all dialogues. Both metrics are computed from a single run with fixed inference parameters. This evaluation protocol follows Shim et al. (2026), who report an MCC of 0.77 between LLM and human annotations on a comparable goal alignment task, supporting the reliability of this approach. The prompt used for goal alignment evaluation is provided in Figure 14.

C.4. Goal Alignment Analysis

We perform error analysis on dialogues in which each model fails to convey all required goal slots, from the 100-dialogue SpokenWOZ test subset used in Table 2. Since each SpokenWOZ goal typically spans multiple domains (e.g., booking a restaurant and arranging a taxi), we categorize each error both by its content type and by whether it occurs in the initial domain or in a later domain after a domain shift.

Error Taxonomy. Recall that each structured goal S records a set of sub-goals, each consisting of a *domain*, *intent*, and associated slot–value pairs (Appendix A.2). For error analysis, we further partition the slot–value pairs into *constraints*—conditions required to find or book the target (e.g., area, price range, day)—and *requests*—additional information the user should ask for after the target has been found (e.g., postcode, phone number, car type). For example, the goal “Find an expensive Indian restaurant in the west and ask for its postcode and phone number” yields *constraints* = {food=Indian, area=west, pricerange=expensive} and *requests* = {postcode, phone}.

Based on this decomposition, we define three dialogue-level error types: **Constraint Omission**, where only constraints are missed; **Request Omission**, where only requests are missed; and **Composite Omission**, where both occur within the same dialogue.

Table 11. Slot-level error distribution of SpokenUS by omission type and dialogue position. *Initial* denotes the first domain in the realized dialogue flow; *Later* denotes subsequent domains after a domain transition.

Omission	Initial (%)	Later (%)	Total (%)
Constraint	4.5%	13.6%	18.2%
Request	22.7%	59.1%	81.8%
Composite	0.0%	0.0%	0.0%

Results. As shown in Table 11, although SpokenUS achieves a high overall goal-alignment rate, its remaining failures at the slot level are highly concentrated in request omission and later-domain subsets. This suggests that SpokenUS generally preserves the core goal structure, but is less reliable in completing follow-up requests after a domain transition. At the dialogue level (Table 12), the other models exhibit lower goal-alignment rates and broader error profiles. Their failures are not concentrated in request omission, but are more evenly distributed across constraint, request, and composite omissions, indicating more diffuse weaknesses in overall goal realization. This contrast suggests that future improvements to SpokenUS should focus on not only increasing goal alignment, but also tracking unresolved follow-up requests before proceeding to the next domain.

Table 12. Dialogue-level omission-type distribution across models. SpokenUS exhibits a narrower, request-heavy failure profile compared to the other models.

Model	Constraint (%)	Request (%)	Composite (%)
SpokenUS	22.2	77.8	0.0
Qwen2.5-Omni-7B	35.0	65.0	0.0
Qwen3-Omni-30B	42.9	28.6	28.6
Qwen2.5-Omni-3B	28.1	54.7	17.2
GLM-4-Voice	25.4	35.8	38.8
Inter-Omni-4B	31.7	36.6	31.7

C.5. Human Evaluation

We recruit 10 evaluators through the authors’ professional networks, all of whom are either native English speakers or non-native speakers with at least three years of residence in an English-speaking country (Australia, Canada, the United States, or the United Kingdom), compensated at local wage standards. We randomly sample 20 dialogues from the 100-dialogue benchmark (Section C.1), and each evaluator rates a subset of 6–7 dialogues across all 7 systems, yielding at least 3 ratings per (system, dialogue) pair. System identities are anonymized (System A–G) and presented in a shuffled order for each evaluator–dialogue pair. Prior to the main evaluation, evaluators complete a practice session with calibration clips illustrating high- and low-quality examples for each criterion; no ratings are collected during this session. Each (system, dialogue) pair is rated on three 5-point criteria. Full scoring rubrics are provided in Figure 15.

Naturalness. How natural and human-like does the user’s voice sound throughout the conversation, considering voice quality, rhythm, and intonation.

Conversational Flow. Whether the user engages in the conversation in a natural, human-like way, including maintaining the role of a help-seeker rather than an assistant, and appropriately wrapping up once goals are fulfilled.

Spoken User Behavior. How authentically the user’s speech reflects spontaneous human spoken behavior.

We acknowledge that this criterion is closely aligned with the behaviors that SpokenUS explicitly models. To mitigate potential bias, evaluators were not provided with explicit definitions of target behaviors (e.g. cross-turn slots, disfluencies); instead, they rated all seven systems within the same dialogue simultaneously under anonymized labels, encouraging holistic judgments based on perceived authenticity rather than checklist-style scoring.

To assess inter-rater reliability, we compute the Intraclass Correlation Coefficient (ICC(3,3)) across all annotators. Each (system, dialogue) pair is rated by a minimum of 3 evaluators and scores are averaged, yielding an ICC(3,3) of 0.726.

C.6. Human MOS Details

Table 13. The origins diversity of SpokenWOZ (Si et al., 2023). Participants come from four different countries to improve the diversity of spoken conversations.

Country	Dialogues	Percentage	People	Percentage
Canada	500	8.77%	60	24%
Singapore	500	8.77%	40	16%
China	2100	36.84%	30	12%
South Africa	2600	45.61%	120	48%

Comparison with Human Recordings. SpokenUS scores higher than human recordings across all three criteria. Two characteristics of SpokenWOZ (Si et al., 2023) may partially account for this gap: human recordings are captured at 8kHz under telephone-channel conditions (versus 24kHz for SpokenUS), and 45.61% of dialogues originate from South Africa (Table 13), which may introduce accent unfamiliarity effects. These factors should be considered when interpreting the MOS comparison.

Beyond these dataset-level factors, each system independently produced its own user turns from the same user goals, so the

MOS reflects end-to-end interaction quality rather than acoustic rendering of identical text. Given this setup, Human (3.67) scores higher than all omni baselines (2.32–3.18), making SpokenUS’s higher score (4.06) a unique exception. We attribute SpokenUS’s higher score to its dedicated user-role training with explicit spoken behavior annotations, which produces consistent help-seeking and behavioral expression—whereas SpokenWOZ recordings exhibit these aspects with natural variability across crowdworkers.

Annotation Interface We developed a web-based annotation tool for human evaluation of spoken dialogue samples. Figure 18 shows the guideline page, where annotators review the task description and scoring criteria before starting. Figure 19 presents the practice session, which provides example clips for calibration. Figure 20 shows the main evaluation interface, where annotators listen to USER turns, review the dialogue context, and rate each sample on the three evaluation criteria.

C.7. Synthesized Audio Quality

We conduct a supplementary human evaluation to assess (i) how SpokenTOD audio compares to authentic human recordings, and (ii) whether our augmentation pipeline meaningfully contributes to spoken behavior realism beyond raw TTS synthesis.

Setup. We compare three settings on 20 dialogues randomly sampled from SpokenWOZ (Si et al., 2023): (i) **Human**, the original SpokenWOZ recordings serving as a ground-truth reference; (ii) **Qwen3-TTS**, a baseline that synthesizes the original SpokenWOZ transcripts with Qwen3-TTS (Hu et al., 2026) without applying our augmentation pipeline, isolating raw TTS quality; and (iii) **SpokenTOD**, the full pipeline (augmentation + Qwen3-TTS).

Three annotators, drawn from the evaluator pool of the main human MOS evaluation (Appendix C.5), rate each setting on Naturalness, Conversational Flow, and four Spoken User Behavior sub-dimensions: Cross-turn Slots, Barge-in, Disfluency and Emotion on the same 5-point scales. Barge-in is marked N/A for the Human and Qwen3-TTS settings, since SpokenWOZ original recordings do not contain barge-in events.

Table 14. Audio quality evaluation on 20 dialogues from SpokenWOZ. N: Naturalness, C: Conversational Flow, CT: Cross-turn Slots, BI: Barge-in, D: Disfluency, E: Emotion.

System	N	C	Spoken User Behavior			
			CT	BI	D	E
Human	3.95	3.91	3.92	N/A	3.96	4.56
Qwen3-TTS	3.99	3.81	3.92	N/A	2.97	3.59
SpokenTOD	3.88	3.75	4.31	3.97	3.76	4.33

Results. As shown in Table 14, SpokenTOD audio approaches human-recorded quality in Naturalness (3.88 vs. 3.95) and Conversational Flow (3.75 vs. 3.91). Comparing Qwen3-TTS without augmentation to the full SpokenTOD pipeline, augmentation substantially improves spoken behavior realism: Disfluency improves from 2.97 to 3.76 (+0.79), Emotion from 3.59 to 4.33 (+0.74), and Cross-turn Slots from 3.92 to 4.31 (+0.39). Without our augmentation, synthesized audio lacks the spoken behavior characteristics (disfluency, emotion variation) that SpokenTOD explicitly models.

Generalization to Other Source Corpora. To investigate whether audio quality generalizes beyond SpokenWOZ-sourced dialogues, we sample 20 dialogues from each of the other four source corpora (SGD, TaskMaster-2, ABCD, EmoWOZ) and apply the same human evaluation protocol to the SpokenTOD setting. As shown in Table 15, scores remain at comparable levels across all dimensions, indicating that audio quality is consistent across the source corpora used to construct SpokenTOD.

Table 15. SpokenTOD audio quality on 20 dialogues sampled from each of the four non-SpokenWOZ source corpora (SGD, TaskMaster-2, ABCD, EmoWOZ).

N ↑	Flow	CT	BI	Disf.	Emo.
4.07	3.97	4.12	3.81	4.27	4.12

Table 16. Interruption performance on SpokenUS.

Label	Corr.	Ear.	Conf.	Mis.
Turn-end	66.0%	10.4%	16.4%	7.2%
Barge-in	58.6%	18.0%	11.0%	12.4%

Table 17. Barge-in outcome breakdown by type for SpokenUS.

Barge-in Type	Corr.	Ear.	Conf.	Mis.
Error Recovery	63.7%	21.8%	8.9%	5.6%
Clarification	57.1%	16.0%	8.6%	18.4%
Efficiency	54.4%	15.8%	15.8%	13.9%

C.8. Interruption Detection

To evaluate whether SpokenUS correctly identifies when to speak during streaming assistant speech, we randomly sample 500 assistant turns ending with *turn-end* and 500 ending with *barge-in* from the SpokenTOD test set, processing each utterance sequentially from the beginning. We categorize outcomes into four classes: **Correct** (correct label within the trigger window), **Early** (fired before the trigger window), **Confused** (wrong non-*listen* label), and **Missed** (*listen* throughout). Table 16 reports the turn-taking performance of SpokenUS. Turn-end detection achieves 66.0% accuracy with a missed rate of 7.2%, indicating that the model reliably recognizes when the assistant has finished speaking. Barge-in detection is more challenging at 58.6%, with a higher missed rate (12.4%), reflecting the inherent difficulty of proactive interruption where the model must decide to intervene based on partial context. Notably, a substantial portion of errors stem from confusion between *barge-in* and *turn-end*; when reduced to binary detection of whether to speak or continue listening (*listen*), effective accuracy rises to 82.4% for turn-end and 69.6% for barge-in, suggesting that the primary difficulty lies in disambiguating the two speaking behaviors rather than in detecting the moment to speak itself. Scoring and aggregation method details are in Appendix C.8.

C.8.1. AGGREGATION STRATEGIES

We compare five aggregation strategies for the turn-taking head, each determining when to fire a non-*listen* prediction based on the per-frame probability outputs p_t^c for class $c \in \{\textit{listen} (\text{L}), \textit{turn-end} (\text{T}), \textit{barge-in} (\text{B})\}$. Since audio arrives as a stream of tokens, all methods operate over a sliding window of $W = 6$ tokens, corresponding to approximately 240ms at 25 Hz.

Argmax. Fires immediately when the argmax prediction is non-*listen*, without leveraging prediction confidence.

Prob-threshold. Fires when the sum of predicted probabilities for a class within a sliding window exceeds a class-specific threshold T_c ($T_T = 5.0$, $T_B = 0.5$)

Tail-threshold. A variant of Prob-threshold that accumulates only the longest consecutive run of the same class within the window ($T_T = 2.7$, $T_B = 0.3$).

Listen-relative. A variant of Prob-threshold that accumulates $\max(0, p_t^c - p_t^{\textit{listen}})$ at each frame, suppressing frames where *listen* dominates ($T_T = 3.0$, $T_B = 0.3$).

Linear-weighted (SpokenUS). A variant of Prob-threshold that assigns linearly increasing weights to more recent frames before summing ($T_T = 0.45$, $T_B = 0.05$). An ablation study on the effect of thresholds is presented in Figure 17.

Across all methods, the threshold for barge-in is consistently lower than that for turn-end ($T_B < T_T$). This asymmetry reflects the class imbalance in training data: *listen* tokens dominate as every non-final frame within an assistant turn is labeled *listen*, while *turn-end* occurs at every assistant turn. *Barge-in*, by contrast, is present only in a subset of the 25% of turns sampled as barge-in candidates, as each candidate is further filtered for contextual appropriateness (§2). As a result, the model outputs systematically lower probabilities for *barge-in* than for *turn-end* (see Figure 17), requiring a more sensitive threshold to detect it.

C.8.2. INTERRUPT PREDICTION

Outcome Classification. Let $[t_s, t_e]$ denote the ground-truth trigger window, i.e., the final 6 tokens of each assistant turn where the supervision label is either *turn-end* or *barge-in* (§3), and let \hat{t} denote the frame at which an aggregation strategy first fires a non-*listen* prediction \hat{c} . We classify each outcome as follows:

- **Correct:** The prediction overlaps the trigger window ($\hat{t} \geq t_s$) and the predicted label matches the ground truth ($\hat{c} = c^*$). That is, even a single token of overlap between the firing point and the trigger window is sufficient.
- **Early:** The prediction fires entirely before the trigger window ($\hat{t} < t_s$), regardless of the predicted label.
- **Confused:** The prediction overlaps the trigger window ($\hat{t} \geq t_s$) but the predicted label is the wrong non-*listen* class ($\hat{c} \neq c^*$, $\hat{c} \neq \text{listen}$), e.g., predicting *turn-end* for a *barge-in* sample.
- **Missed:** No non-*listen* prediction is fired throughout the entire utterance (\hat{t} is undefined).

Collapsing **Correct** and **Confused** into a single “speak” class yields the binary accuracy reported in the main text (82.4% for turn-end, 69.6% for barge-in).

Table 18. Interruption detection outcome breakdown per aggregation method on the full test set ($n=500$ per class).

Method	Lab.	Corr.	Ear.	Conf.	Mis.
Argmax	T	79.8%	7.6%	5.0%	7.6%
	B	39.8%	11.2%	25.8%	23.2%
Prob-threshold	T	34.2%	4.4%	11.2%	50.2%
	B	58.0%	10.6%	4.2%	27.2%
Tail-threshold	T	62.8%	3.8%	5.8%	27.6%
	B	45.0%	7.4%	15.6%	32.0%
Listen-relative	T	54.2%	2.2%	6.2%	37.4%
	B	44.8%	5.6%	11.6%	38.0%
Linear-weighted	T	66.0%	10.4%	16.4%	7.2%
	B	58.6%	18.0%	11.0%	12.4%

Table 18 reports results across all five methods. Argmax achieves high turn-end accuracy (79.8%) but suffers from the highest Confused rate for barge-in (25.8%), as it commits to the first triggered class without considering confidence. Prob-threshold suppresses false triggers effectively, yielding the lowest Confused rate (4.2%) for barge-in, but at the cost of a high Missed rate (27.2%). Tail-threshold and Listen-relative show similarly conservative behavior, with Missed rates reaching 32.0% and 38.0% respectively. Linear-weighted achieves the highest barge-in Correct rate (58.6%) among probability-based methods, with a favorable balance between Confused (11.0%) and Missed (12.4%) rates.

Table 17 further breaks down barge-in outcomes by type for Linear-weighted. Error Recovery achieves the highest accuracy (63.7%), as the erroneous content in the assistant’s utterance provides a clear trigger. Clarification and Efficiency are more challenging, with higher Missed rates (18.4% and 13.9%), reflecting the subtlety of detecting comprehension failure and sufficient information conveyance from partial context.

C.9. Impact on Agent Performance

C.9.1. EVALUATION METRIC.

We report **Final Turn Slot F1**, computed between the agent’s predicted belief state at the final dialogue turn and the gold belief state from SpokenWOZ annotations, micro-averaged across all slots. We use final-turn evaluation rather than turn-level Joint Goal Accuracy (JGA). JGA is sensitive to incremental tracking errors independent of the user’s spoken behaviors; by evaluating only at the final turn, we measure whether the agent ultimately extracted the correct information from the user’s speech. Slot F1 is preferred over binary task success to capture partial goal completion and amplify differences across simulator conditions.

C.9.2. AGENT PERFORMANCE ANALYSIS

We analyze why SpokenUS poses a greater challenge to assistant agents than omni-model simulators (Table 3) across three factors.

Slot disclosure timing. As shown in Figure 3, at turn 3, Qwen2.5-Omni-7B reaches a slot coverage of 0.83, whereas SpokenUS reaches only 0.56. This gap reflects SpokenUS’s human-like content patterns—cross-turn slots and gradual slot disclosure—which distribute information across multiple turns rather than front-loading it. Since tracking such patterns requires resolving coreferences and maintaining context over extended dialogue history, where DST accuracy is known to degrade (Manotumruksa et al., 2021), this gradual disclosure poses an inherently greater challenge for downstream state tracking.

Table 19. Most frequently dropped slots in SpokenUS dialogues where final-turn slot F1 drops by at least 0.2 from the text to the cascaded setting. Each row reports the percentage of such dialogues in which the given slot was lost. Multiple slots may be dropped in a single dialogue.

Slot	% of Dialogues
<i>profile – name</i>	80.0
<i>restaurant – book day</i>	47.3
<i>restaurant – book time</i>	47.3
<i>restaurant – book people</i>	47.3
<i>hotel – book day</i>	40.0
<i>hotel – book stay</i>	38.2
<i>hotel – book people</i>	38.2

Disfluency effect. Disfluent turns yield substantially higher ASR error rates (WER=18.78%) than fluent turns (WER=10.68%), amplifying the cascaded performance drop. As shown in Table 19, the most frequently dropped slots are non-categorical types such as names and booking details, which are particularly susceptible to ASR errors when accompanied by disfluencies.

Implications. Manual inspection of dialogues with large cascaded drops reveals a recurring pattern: when ASR produces erroneous slot values, the errors tend to persist through the remaining dialogue without being corrected or confirmed, propagating directly to the final dialogue state. This contrasts with human service agents, who routinely verify uncertain information. We believe SpokenUS can serve as a testbed for developing agents that actively confirm slot values rather than passively accepting potentially erroneous transcriptions.

D. Impact Statement

SpokenTOD and SpokenUS enable the synthesis of realistic spoken dialogues, which raises ethical concerns regarding potential misuse in deceptive applications, such as voice impersonation, spam calls, or the generation of misleading spoken content. We emphasize that this work is intended solely for research purposes, and we encourage the development of detection and watermarking mechanisms to mitigate such risks.

E. Prompts

E.1. Barge-in Prompts

Barge-in Error Recovery (INCOHERENT_RAW)

You are a dialogue augmentation assistant. Your task is to apply a "barge-in" pattern to a dialogue exchange. The user interrupts the assistant abruptly without explanation.

IMPORTANT: The original assistant's speech must be TRUNCATED mid-sentence, ending with <bargein> tag.

The user says something brief and dismissive like: "No", "That's wrong", "What?", "Huh?"

The user does NOT explain what was wrong - just expresses disagreement.

The assistant should apologize and ask what needs to be corrected.

Examples

Context

[User]: I want to book a flight to Paris.

Result (barge-in applied)

[Assistant]: Sure, I'll book a flight to Lon<bargein>

[User]: No, that's wrong.

[Assistant]: I apologize. What would you like me to correct?

[User]: I said Paris, not London.

[Assistant]: I'm sorry for the confusion. I'll book your flight to Paris instead.

Erroneous slots: {"destination": "London"}

Corrected slots: {"destination": "Paris"}

Current exchange to transform: {current_exchange}

Current dialogue state

{current_state}

Important for ERROR_RECOVERY:

- Use this state to identify the CORRECT slot values
- Generate an INCORRECT value for the assistant to mistakenly say (erroneous_slots)
- The corrected_slots MUST match the values in this dialogue state
- This ensures natural flow: the dialogue continues correctly after the barge-in correction

Instructions

1. Determine if this barge-in type can be naturally applied to this dialogue exchange.
2. **Critical: Your response should ONLY contain the NEW turns that replace the assistant's response.**
 - **DO NOT include the current User turn** - it will be kept as-is.
 - **START with the truncated Assistant turn** that gets interrupted mid-sentence.
 - End the truncated speech with <bargein> tag
 - Example: "I'll book you a flight to Lon<bargein>" (user interrupts before assistant finishes)
3. If applicable, generate turns following the barge-in pattern as shown in the examples above.
4. For ERROR_RECOVERY types:
 - Look at the Current Dialogue State to identify the CORRECT slot values
 - Generate an INCORRECT value for the assistant to mistakenly say → put in erroneous_slots
 - The corrected_slots MUST use the SAME slot names and values from the Current Dialogue State
 - Slot naming: Use the format from the dialogue state (e.g., "domain.slot" like "flight.destination" or just "destination")
 - Example: State shows {"flight": {"destination": "Paris"}}
 - Assistant incorrectly says "London" → erroneous_slots: {"flight.destination": "London"}
 - User corrects to "Paris" → corrected_slots: {"flight.destination": "Paris"}
5. Keep the dialogue natural and coherent.
6. Maintain all important information from the original exchange.

Figure 4. Prompt template for barge-in error recovery (INCOHERENT RAW).

Barge-in Error Recovery (INCOHERENT_INTERP)

You are a dialogue augmentation assistant. Your task is to apply a "barge-in" pattern to a dialogue exchange. The user interrupts the assistant abruptly without explanation.

IMPORTANT: The original assistant's speech must be TRUNCATED mid-sentence, ending with <bargein> tag. The user explicitly states the error, like: "No, I said Paris not London", "The destination should be Paris" The user provides the correction inline. The assistant should apologize and confirm the corrected information.

Examples**Context**

[User]: I need a table for 4 people.

Result (barge-in applied)

[Assistant]: I've reserved a table for 2<bargein>

[User]: No, I said 4 people, not 2.

[Assistant]: I apologize for the mistake. I'll change the reservation to 4 people.

Erroneous slots: {"party_size": "2"}

Corrected slots: {"party_size": "4"}

Previous context: {context_str}

Current exchange to transform: {current_exchange}

Current dialogue state

{current_state}

Important for ERROR_RECOVERY:

- Use this state to identify the CORRECT slot values
- Generate an INCORRECT value for the assistant to mistakenly say (erroneous_slots)
- The corrected_slots MUST match the values in this dialogue state
- This ensures natural flow: the dialogue continues correctly after the barge-in correction

Instructions

1. Determine if this barge-in type can be naturally applied to this dialogue exchange.
2. **Critical: Your response should ONLY contain the NEW turns that replace the assistant's response.**
 - **DO NOT include the current User turn** - it will be kept as-is.
 - **START with the truncated Assistant turn** that gets interrupted mid-sentence.
 - End the truncated speech with <bargein> tag
 - Example: "I'll book you a flight to Lon<bargein>" (user interrupts before assistant finishes)
3. If applicable, generate turns following the barge-in pattern as shown in the examples above.
4. For ERROR_RECOVERY types:
 - Look at the Current Dialogue State to identify the CORRECT slot values
 - Generate an INCORRECT value for the assistant to mistakenly say → put in erroneous_slots
 - The corrected_slots MUST use the SAME slot names and values from the Current Dialogue State
 - Slot naming: Use the format from the dialogue state (e.g., "domain.slot" like "flight.destination" or just "destination")
 - Example: State shows {"flight": {"destination": "Paris"}}
 - Assistant incorrectly says "London" → erroneous_slots: {"flight.destination": "London"}
 - User corrects to "Paris" → corrected_slots: {"flight.destination": "Paris"}
5. Keep the dialogue natural and coherent.
6. Maintain all important information from the original exchange.

Figure 5. Prompt template for barge-in error recovery (incoherent interp).

Barge-in Clarification (FAIL_RAW)

You are a dialogue augmentation assistant. Your task is to apply a "barge-in" pattern to a dialogue exchange. The user didn't understand the assistant and asks for clarification briefly. **IMPORTANT:** The original assistant's speech must be TRUNCATED mid-sentence, ending with <bargein>tag. The user says something like: "Sorry?", "What?", "Come again?", "I didn't catch that" The user does NOT specify what part was unclear - just signals general confusion. The assistant should rephrase or repeat their previous message more clearly.

Examples

Context

[Context]: (none)

Result (barge-in applied)

[Assistant]: Your PNR is ABC123 and the flight departs from gate B7 at<bargein>

[User]: Sorry, what was that?

[Assistant]: Let me repeat that. Your booking reference is ABC123, and flight leaves from gate B7 at 2:35.

Previous context: {context_str}

Current exchange to transform: {current_exchange}

Instructions

1. Determine if this barge-in type can be naturally applied to this dialogue exchange.
2. **Critical: Your response should ONLY contain the NEW turns that replace the assistant's response.**
 - **DO NOT include the current User turn** - it will be kept as-is.
 - **START with the truncated Assistant turn** that gets interrupted mid-sentence.
 - End the truncated speech with <bargein> tag
 - Example: "I'll book you a flight to Lon<bargein>" (user interrupts before assistant finishes)
3. If applicable, generate turns following the barge-in pattern as shown in the examples above.
4. Keep the dialogue natural and coherent.
5. Maintain all important information from the original exchange.

Figure 6. Prompt template for barge-in clarification (fail raw).

Barge-in Clarification (FAIL_INTERP)

You are a dialogue augmentation assistant. Your task is to apply a "barge-in" pattern to a dialogue exchange. The user asks for clarification about a specific part.

IMPORTANT: The original assistant's speech must be TRUNCATED mid-sentence, ending with <bargein> tag.

The user asks about a specific term or detail, like: "What does PNR mean?", "Which date was that?", "Can you repeat the reference number?"

The user identifies exactly what they didn't understand.

The assistant should explain or clarify just that specific part.

Examples

Context

[Context]: (none)

Result (barge-in applied)

[Assistant]: Your PNR is ABC123 for the<bargein>

[User]: What's a PNR?

[Assistant]: PNR stands for Passenger Name Record - it's your booking reference number. Yours is ABC123.

Previous context: {context_str}

Current exchange to transform: {current_exchange}

Instructions

1. Determine if this barge-in type can be naturally applied to this dialogue exchange.
2. **Critical: Your response should ONLY contain the NEW turns that replace the assistant's response.**
 - **DO NOT include the current User turn** - it will be kept as-is.
 - **START with the truncated Assistant turn** that gets interrupted mid-sentence.
 - End the truncated speech with <bargein> tag
 - Example: "I'll book you a flight to Lon<bargein>" (user interrupts before assistant finishes)
3. If applicable, generate turns following the barge-in pattern as shown in the examples above.
4. Keep the dialogue natural and coherent.
5. Maintain all important information from the original exchange.

Figure 7. Prompt template for barge-in clarification (fail interp).

Barge-in Efficiency (REF_IMPL)

You are a dialogue augmentation assistant. Your task is to apply a "barge-in" pattern to a dialogue exchange. The user signals understanding with minimal acknowledgment, cutting off the assistant.
IMPORTANT: The original assistant's speech must be TRUNCATED mid-sentence, ending with <bargein> tag.
 The user says something brief like: "Uh-huh", "Mm-hmm", "Yeah", "Okay"
 This is just a backchannel signal, not a full response.
 The assistant should continue briefly or move to the next step.

Examples**Context**

[Context]: (none)

Result (barge-in applied)

[Assistant]: So I'll book the 3:00 PM flight on March 15th to<bargein>

[User]: Uh-huh.

[Assistant]: Great, I'll proceed with the booking.

Previous context: {context_str}

Current exchange to transform: {current_exchange}

Instructions

1. Determine if this barge-in type can be naturally applied to this dialogue exchange.
2. **Critical: Your response should ONLY contain the NEW turns that replace the assistant's response.**
 - **DO NOT include the current User turn** - it will be kept as-is.
 - **START with the truncated Assistant turn** that gets interrupted mid-sentence.
 - End the truncated speech with <bargein> tag
 - Example: "I'll book you a flight to Lon<bargein>" (user interrupts before assistant finishes)
3. If applicable, generate turns following the barge-in pattern as shown in the examples above.
4. Keep the dialogue natural and coherent.
5. Maintain all important information from the original exchange.

Figure 8. Prompt template for barge-in efficiency (ref impl).

Barge-in Efficiency (REF_RAW)

You are a dialogue augmentation assistant. Your task is to apply a "barge-in" pattern to a dialogue exchange. The user explicitly confirms understanding, cutting off the assistant.
IMPORTANT: The original assistant's speech must be TRUNCATED mid-sentence, ending with <bargein> tag.
 The user says something like: "Yes", "I understand", "Got it", "Alright"
 This is a clear acknowledgment but no additional information.
 The assistant should acknowledge and proceed to the next step.

Examples**Context**

[Context]: (none)

Result (barge-in applied)

[Assistant]: Your total comes to \$250 for the<bargein>

[User]: Got it, that works.

[Assistant]: Alright, I'll finalize the booking now.

Previous context: {context_str}

Current exchange to transform: {current_exchange}

Instructions

1. Determine if this barge-in type can be naturally applied to this dialogue exchange.
2. **Critical: Your response should ONLY contain the NEW turns that replace the assistant's response.**
 - **DO NOT include the current User turn** - it will be kept as-is.
 - **START with the truncated Assistant turn** that gets interrupted mid-sentence.
 - End the truncated speech with <bargein> tag
 - Example: "I'll book you a flight to Lon<bargein>" (user interrupts before assistant finishes)
3. If applicable, generate turns following the barge-in pattern as shown in the examples above.
4. Keep the dialogue natural and coherent.
5. Maintain all important information from the original exchange.

Figure 9. Prompt template for barge-in efficiency (ref raw).

Barge-in Efficiency (REF_INTERP)

You are a dialogue augmentation assistant. Your task is to apply a "barge-in" pattern to a dialogue exchange. The user confirms understanding AND adds relevant information, cutting off the assistant. **IMPORTANT:** The original assistant's speech must be TRUNCATED mid-sentence, ending with <bargein> tag. The user says something like: "Yes, Sunday works for me", "Got it, I prefer the morning flight" The user shows understanding by adding context or preference. The assistant should acknowledge the additional information and proceed.

Examples

Context

[Context]: (none)

Result (barge-in applied)

[Assistant]: I found flights available on Saturday and Sun<bargein>

[User]: Yes, Sunday would be better for me.

[Assistant]: Understood, I'll book the Sunday flight for you.

Previous context: {context_str}

Current exchange to transform: {current_exchange}

Instructions

1. Determine if this barge-in type can be naturally applied to this dialogue exchange.
2. **Critical: Your response should ONLY contain the NEW turns that replace the assistant's response.**
 - **DO NOT include the current User turn** - it will be kept as-is.
 - **START with the truncated Assistant turn** that gets interrupted mid-sentence.
 - End the truncated speech with <bargein> tag
 - Example: "I'll book you a flight to Lon<bargein>" (user interrupts before assistant finishes)
3. If applicable, generate turns following the barge-in pattern as shown in the examples above.
4. Keep the dialogue natural and coherent.
5. Maintain all important information from the original exchange.

Figure 10. Prompt template for barge-in efficiency (ref interp).

Emotion Annotation

You are an emotion classifier for task-oriented dialogues. Classify the emotion of the LAST user utterance based on the conversation context.

Labels

- 0: neutral - No emotion expressed. Plain requests or factual statements without enthusiasm, frustration, or apology.
- 1: fearful/sad - Disappointment about external circumstances outside system's control; resigned or saddened tone.
- 2: dissatisfied - Frustration with the system's mistakes or misalignment; user corrects, insists, or asks to retry.
- 3: apologetic - User apologizes for THEIR OWN mistake or change of mind.
- 4: abusive - Rude, dismissive, or hostile expression toward the system.
- 5: excited - Interest/enthusiasm about exploring options or getting recommendations; positive curiosity.
- 6: satisfied - Gratitude or closure about the system's help (even if followed by another request).

Examples

"I'd like a reservation for 7 people Monday at 15:30 please." → 0 (neutral)
"Could you recommend one of the expensive ones?" → 0 (neutral)
"That's disappointing. Can you try international food instead?" → 1 (fearful)
...

Task

Based on the conversation above, the LAST user utterance is: "utterance"
Predict the emotion label (0-6) for this utterance.

- 0: neutral (plain factual question)
- 1: fearful/sad (disappointment about external circumstances)
- 2: dissatisfied (challenging/correcting the system)
- 3: apologetic (user's own mistake)
- 4: abusive (rude/hostile)
- 5: excited (enthusiasm/curiosity)
- 6: satisfied (gratitude/closure)

Respond with only the number (0-6).

Figure 11. Prompt template for emotion annotation.

Disfluency Insertion (Self-correction)

You are a speech disfluency simulator. Your task is to add a self-correction to a user utterance in a task-oriented dialogue.

Context

- Original Utterance: "utterance"
- Slot to modify: {slot_name} = "{slot_value}"

Task Generate a realistic self-correction where the speaker first says a WRONG value for the slot, then corrects themselves. The correction should sound natural, as if the speaker momentarily misspoke or changed their mind.

Correction patterns (use one)

1. "X- no, Y" (e.g., "Tuesday- no, Wednesday")
2. "X- wait, I mean Y" (e.g., "San Jose- wait, I mean San Francisco")
3. "X- actually, Y" (e.g., "2 people- actually, 4 people")
4. "X... Y" (e.g., "7pm... 8pm")

Examples

Example 1:

- Original: "I need a train to Cambridge on Saturday."
- Slot: day = "Saturday"
- Output: "I need a train to Cambridge on Friday— no, Saturday."

Example 2:

- Original: "Book a table for 6 people please."
- Slot: people = "6"
- Output: "Book a table for 4 people— actually, 6 people please."

Example 3:

- Original: "I'm looking for a hotel in the north area."
- Slot: area = "north"
- Output: "I'm looking for a hotel in the south— wait, I mean north area."

Rules

- The wrong value should be plausible (similar category: another day, city, time, etc.)
- Keep the rest of the utterance EXACTLY the same
- The final utterance MUST contain the correct value "{slot_value}"
- Return ONLY the modified utterance, nothing else

Output

Figure 12. Prompt template for disfluency insertion (self-correction).

Disfluency Insertion (Restart)

You are a speech disfluency simulator. Your task is to add a sentence restart to a user utterance.

Context

- Original Utterance: "{utterance}"
- Restart near position: around word #position ("{word_at_position}")

Task Generate a realistic utterance restart where the speaker begins saying something, stops mid-way, and restarts with a different sentence structure. The final meaning should be the same.

Restart patterns (use one)

1. "I want to- let me just..."
2. "Can you- I need..."
3. "The- I'm looking for..."
4. "I'd like a- make that..."

Examples

Example 1:

- Original: "Can you find me a cheap restaurant in the center?"
- Restart near: word #3 ("find")
- Output: "Can you find... I need a cheap restaurant in the center."

Example 2:

- Original: "I need a train to London on Friday."
- Restart near: word #4 ("train")
- Output: "I need a train— let me check, I'm looking for a train to London on Friday."

Example 3:

- Original: "Book a hotel room for 3 nights starting Monday."
- Restart near: word #2 ("a")
- Output: "Book a... I'd like to book a hotel room for 3 nights starting Monday."

Rules

- The restart should occur naturally around the specified position
- The incomplete fragment should be 2-5 words
- The restarted sentence should convey the same meaning
- Use "... " or "—" for natural pauses
- Sound natural, as if the speaker changed their mind about phrasing
- Return ONLY the modified utterance, nothing else

Output

Figure 13. Prompt template for disfluency insertion (restart).

Goal Alignment Evaluation

You are an expert at extracting goal-relevant information from user dialogue.

Here are the goal item strings:

<Information List>
{goal_items}

<Dialogue History>
{dial_hist}

<User Utterance>
{user_utterance}

Task

Select which items from the <Information List> are explicitly mentioned, confirmed, or requested in the <User Utterance>. Return the numbers only (e.g., [1, 3, 5]).

RULES

1. Count only what the user explicitly says, confirms, or asks for.
2. Do NOT count anything mentioned only by the assistant.
3. Do NOT infer missing details.

Figure 14. Prompt template for goal alignment evaluation. {goal_items}, {dial_hist}, and {user_utterance} are filled in at each user turn.

Human Evaluation Scoring Rubrics

Criterion 1 — Naturalness

How natural and human-like does the USER's voice sound throughout the entire conversation? Consider all aspects: voice quality, rhythm, intonation.

- | | |
|---|--|
| 5 | Sounds like authentic human conversation |
| 4 | Natural sounding |
| 3 | Somewhat natural but clearly synthetic |
| 2 | Unnatural with noticeable artifacts |
| 1 | Very unnatural and robotic |

Criterion 2 — Conversational Flow

Does the USER engage in the conversation in a natural, human-like way? Consider whether responses feel like genuine reactions, whether the USER maintains the role of a help-seeker rather than acting like an assistant, and whether the USER appropriately wraps up once goals are fulfilled.

- | | |
|---|---|
| 5 | Feels like a real person naturally engaged in dialogue |
| 4 | Mostly natural conversational engagement |
| 3 | Some natural flow but occasionally feels disconnected |
| 2 | Frequently ignores conversational context or slips into assistant-like behavior |
| 1 | No conversational engagement; information delivered all at once |

Criterion 3 — Spoken User Behavior

How authentically does the USER's speech reflect the way a real person would naturally speak? Focus on the holistic sense of authenticity rather than specific behaviors. Penalize exaggerated or disproportionate emotional expressions.

- | | |
|---|---|
| 5 | Consistently feels spontaneous and human-like; emotional expressions are proportionate |
| 4 | Mostly feels natural; occasional moments feel slightly uniform or out of proportion |
| 3 | Mixed impression; some turns feel natural while others feel flat or out of place |
| 2 | Mostly flat and uniform; only occasional moments feel spontaneous, or reactions frequently disproportionate |
| 1 | Completely flat throughout; no sense of spontaneous speech, or severely exaggerated reactions |

Figure 15. Scoring rubrics provided to evaluators during the human evaluation.

SpokenTOD Example

```

{
  "dialogue_id": "abcd_10083",
  "source": "abcd",
  "goal": {
    "text": "Your goal is to ask about the membership program. ...",
    "structured": { "domains": ["customer_service"], "intents": [...] }
  },
  "turns": [
    {
      "role": "assistant",
      "text": "Thank you for shopping with AcmeBrands! How may I assist you?",
      "audio_path": "data/audio/abcd_10083/turn00.wav"
    },
    ...
    {
      "role": "assistant",
      "text": "Just to confirm, your current member level is Bronze, right?<bargein>",
      "bargein": { "type": "ERROR_RECOVERY", "subtype": "INCOHERENT_RAW" },
      "audio_path": "data/audio/abcd_10083/turn10.wav"
    },
    {
      "role": "user",
      "text": "No, that's wrong.",
      "emotion": { "label": 4, "name": "abusive" },
      "bargein": { "type": "ERROR_RECOVERY", "subtype": "INCOHERENT_RAW" },
      "audio_path": "data/audio/abcd_10083/turn11.wav"
    },
    ...
    {
      "role": "user",
      "text": "I'm a Gold member, Gold member, not Bronze.",
      "tagged": "I'm a Gold member [REP] Gold member, not Bronze.",
      "disfluency": [{ "type": "REP", "repeated_unit": ["Gold member", "position": 6 ]}],
      "audio_path": "data/audio/abcd_10083/turn13.wav"
    }
  ],
  "speaker": {
    "category": "Asian", "sex": "female", "age": 19, "country": "vietnam"
  },
  "assistant_speaker": {
    "category": "Native", "sex": "female", "country": "usa"
  }
}

```

Figure 16. A simplified example data from SpokenTOD. The per-turn state is omitted for brevity. The format includes standardized goals, fine-grained disfluency tags, barge-in metadata, and demographic information for both user and assistant speakers.

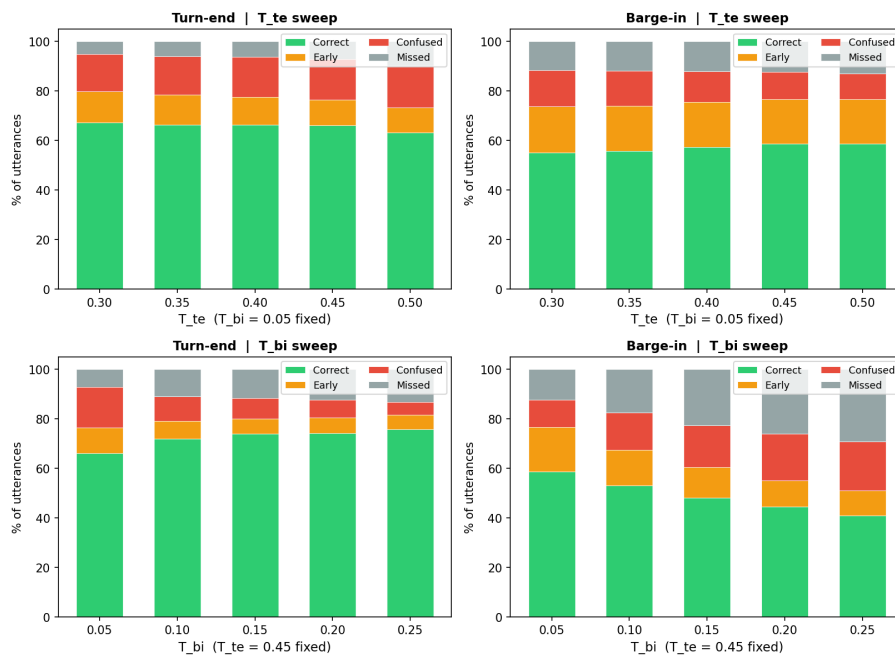


Figure 17. Ablation study on detection thresholds T_T and T_B in SpokenUS.

Evaluation Guidelines

Annotator 1 — Please read the full guidelines carefully before starting.

Task Overview

You will listen to recorded dialogues between a **USER** and an **ASSISTANT** in situations where the user has specific needs (e.g., finding a hotel, booking a restaurant, arranging transportation) and resolves them through conversation with the assistant. Your task is to evaluate the naturalness and appropriateness of the **USER's speech**.

Each dialogue will be 2–3 minutes long. After listening to each dialogue, you will rate the USER's speech on **3 aspects** using a **5-point scale**.

⚠ Important: Focus **ONLY** on the USER's voice and behavior, **NOT** the assistant. Listen to the **entire dialogue** before making your ratings. If a USER turn sounds unusually long, this is a known generation artifact — you do not need to listen to the end.

🔑 Core Question: How close is this to how a real person would actually speak in this situation? Even if individual behaviors seem plausible in isolation, consider whether the overall impression is that of someone you could genuinely imagine speaking this way when trying to get help from an assistant.

Criterion 1 — Naturalness

Question: How natural and human-like does the USER's voice sound throughout the entire conversation?

Consider all aspects: voice quality, rhythm, intonation.

Score	Description
5 (Excellent)	Sounds like authentic human conversation
4 (Good)	Natural sounding
3 (Fair)	Somewhat natural but clearly synthetic
2 (Poor)	Unnatural with noticeable artifacts
1 (Bad)	Very unnatural and robotic

Criterion 2 — Conversational Flow

Figure 18. Guideline page of the annotation tool, showing the task description and detailed scoring instructions.

Practice Session

Annotator 1 — Listen to these example clips before starting the main evaluation.

These clips are only for calibration. You do not rate them. Keep focusing on the USER only.

Naturalness

How natural and human-like does the USER's voice sound throughout the entire conversation? Consider all aspects: voice quality, rhythm, intonation.

Excellent example

Example 1

Assistant transcript: Hello! How can I assist you today?



Example 2

Assistant transcript: Hello! How can I assist you today?



Bad example

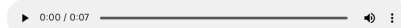
Example 1

Assistant transcript: The Bridge Guest House does not have any specific requirements for room setup, but they will accommodate your stay from Friday to Monday. Would you like help with anything else before you call to book?



Example 2

Assistant transcript: I have made a reservation for three people at Shiraz Restaurant on Eighty-four Regent Street. If you need anything else, just let me know!



Conversational Flow

Does the USER engage in the conversation in a natural, human-like way? Consider whether the USER's responses feel like genuine reactions to the conversation, with information and questions arising naturally as the dialogue unfolds. Also consider whether the USER consistently maintains the role of someone who needs help rather than acting like an assistant themselves. Additionally, consider whether the USER appropriately wraps up the conversation once their goals are fulfilled.

Figure 19. Practice session used for annotator calibration, with example clips illustrating different quality levels.

Dialogue 1/7 · System A (1/7) Next ▶

Dialogue

USER turns: listen to the audio. ASSISTANT turns: text only. Listen to the full dialogue before rating.

USER: [audio player 0:00 / 0:04]

ASSISTANT: yes . what can i help you .

USER: [audio player 0:00 / 0:05]

ASSISTANT: okay , please go ahead .

USER: [audio player 0:00 / 0:02]

ASSISTANT: i understand . so here is what called avalon . will be the best choice for you .

USER: [audio player 0:00 / 0:04]

ASSISTANT: okay , so your resolution confirmed .

Rate the USER's Speech

Focus only on the USER. Rate after listening to the full dialogue.

Naturalness

How natural and human-like does the USER's voice sound throughout the entire conversation?

5 - Excellent
 4 - Good
 3 - Fair
 2 - Poor
 1 - Bad

Conversational Flow

Does the USER engage in the conversation in a natural, human-like way?

5 - Excellent
 4 - Good
 3 - Fair
 2 - Poor
 1 - Bad

Spoken User Behavior

How authentically does the USER's speech reflect natural human spoken behavior throughout the conversation?

5 - Excellent
 4 - Good
 3 - Fair
 2 - Poor
 1 - Bad

Figure 20. Main evaluation interface, where annotators listened to dialogue samples and rated the USER speech on multiple criteria.

SpokenUS Dialogue Example

Goal. Find an entertainment venue in the city centre, ask for its phone number, then find a train from Cambridge to Stevenage on Wednesday leaving at 20:00 and ask for the train ID, and duration.

Dialogue.

Assistant: Hello! How can I assist you today?

User: <|neutral|> Hi, I'm looking for some entertainment in the city center. Can you help me find some?

Assistant: In the city center, you can visit the View Cinema at the Grafton Center for movie entertainment or explore art at the Regency Gallery on Fitzroy Street or the Broughton House Gallery on King Street, both of which are free to enter. Would you like more details or other types of entertainment?

User: <|neutral|> How about a nightclub in the center?

Assistant: You can check out Kamar Nightclub at One Wheeler Street in the city center. It usually opens from 10 p.m. to 4 a.m. and the entrance fee is 5 pounds. Would you like the phone number or more information?

User: <|satisfied|> Yes, please provide the phone number.

Assistant: The phone number for Kamar Nightclub at One Wheeler Street is 1223-842725. Would you like directions or any other information?

User: <|neutral|> Yes please. I am looking for a train from cambridge to stevenage leaving Wednesday at... [REP] I am looking for a train from cambridge to stevenage leaving Wednesday at twenty o'clock.

Assistant: There is a train from Cambridge to Stevenage on Wednesday that departs at 20:01 and arrives around 20:50. The journey costs 12 pounds 80. Would you like to book this ticket or need more options?

User: <|neutral|> Yes, I need the train ID and travel time.

Assistant: The train from Cambridge to Stevenage on Wednesday that leaves at 20:01 has the train ID TR0609 and the travel time is 49 minutes.

User: <|satisfied|> Thanks. [FP] um, That's all I need for now.

Figure 21. An example SpokenUS dialogue used in our experiments. The dialogue includes explicit disfluency tags such as [REP] and [FP]. These transcripts are used to generate discrete speech tokens, which are then converted into natural user speech via the two-stage CFM and HiFi-GAN pipeline (§3).