# PERSONAPLEX: VOICE AND ROLE CONTROL FOR FULL DUPLEX CONVERSATIONAL SPEECH MODELS

*Rajarshi Roy, Jonathan Raiman, Sang-gil Lee, Teodor-Dumitru Ene,*
*Robert Kirby, Sungwon Kim, Jaehyeon Kim, Bryan Catanzaro*

NVIDIA

{rajarshir, jraiman, sanggill, tene, rkirby, sungwonk, jaehyeonk, bcatanzaro}@nvidia.com

## ABSTRACT

Recent advances in duplex speech models have enabled natural, low-latency speech-to-speech interactions. However, existing models are restricted to a fixed role and voice, limiting their ability to support structured, role-driven real-world applications and personalized interactions. In this work, we introduce PersonaPlex, a duplex conversational speech model that incorporates hybrid system prompts, combining role conditioning with text prompts and voice cloning with speech samples. PersonaPlex is trained on a large-scale synthetic dataset of paired prompts and user-agent conversations, generated with open-source large language models (LLM) and text-to-speech (TTS) models. To evaluate role conditioning in real-world settings, we extend the Full-Duplex-Bench benchmark beyond a single assistant role to multi-role customer service scenarios. Experiments show that PersonaPlex achieves strong role-conditioned behavior, voice-conditioned speech, and natural conversational responsiveness, surpassing state-of-the-art duplex speech models and hybrid large language model-based speech systems in role adherence, speaker similarity, latency, and naturalness.

***Index Terms***— Conversational Speech Model, Duplex Spoken Language Model, Role Conditioning, Voice Cloning

## 1. INTRODUCTION

Recent advances in duplex speech models have enabled real-time, low-latency speech-to-speech conversation systems that approximate natural human interaction. These systems integrate ASR, LLMs, and TTS into a unified pipeline, enabling agents to respond with high naturalness and low turn-taking delay. However, current duplex systems are generally limited to a fixed voice identity and role, restricting their use in structured or role-driven applications such as customer service, multi-character interactions, and personalized assistants. Without the ability to condition on conversational role and speaker identity, these systems fail to achieve the flexibility required in real-world human–machine interactions.

In parallel, research on voice-conditioned TTS has demonstrated progress in speaker adaptation, voice cloning, and prosody manipulation, while instruction-following LLMs have shown strong role conditioning in text-based interactions. However, these advances have not been fully realized in duplex speech systems, where latency constraints and coupled speech–text dynamics make conditioning on both role and voice more challenging. Bridging this gap requires a speech-to-speech model that integrates the conditioning capabilities of LLMs and the adaptability of modern TTS into a low-latency duplex framework.

In this work, we present PersonaPlex, a full duplex speech-to-speech conversational model that incorporates hybrid system prompts, combining text-based role conditioning with audio-based voice cloning. PersonaPlex enables zero-shot voice cloning and fine-grained role conditioning, extending duplex speech beyond generic assistants to structured domains such as customer service. We introduce a large-scale synthetic training corpus of paired prompts and user-agent conversations. We evaluate PersonaPlex on the *Full-Duplex-Bench* benchmark [1]. Since *Full-Duplex-Bench* is limited to a single assistant role, we propose *Service-Duplex-Bench*, an extension that covers real-world multi-role customer service scenarios. Our extension adds 350 customer service evaluation questions - each corresponding to a specific service role - to the 400 questions in *Full-Duplex-Bench*. In our experiments on *Full-Duplex-Bench* and *Service-Duplex-Bench* we find that PersonaPlex achieves state-of-the-art performance in role adherence, voice similarity, and dialog naturalness, while maintaining the responsiveness and turn taking abilities of duplex speech models.

## 2. RELATED WORK

Cascaded ASR-LLM-TTS systems leverage the reasoning ability of LLMs and the naturalness of dedicated TTS models, but inevitably lose paralinguistic information, reducing dialog naturalness. To address this, several approaches have been proposed to improve conversational quality. Streaming TTS models [3, 4] reduce LLM→TTS latency and support voice cloning. HumeAI's Emphatic Voice Interface[1]

[1] https://www.hume.ai/empathic-voice-interface

*Accepted for publication at ICASSP 2026.*

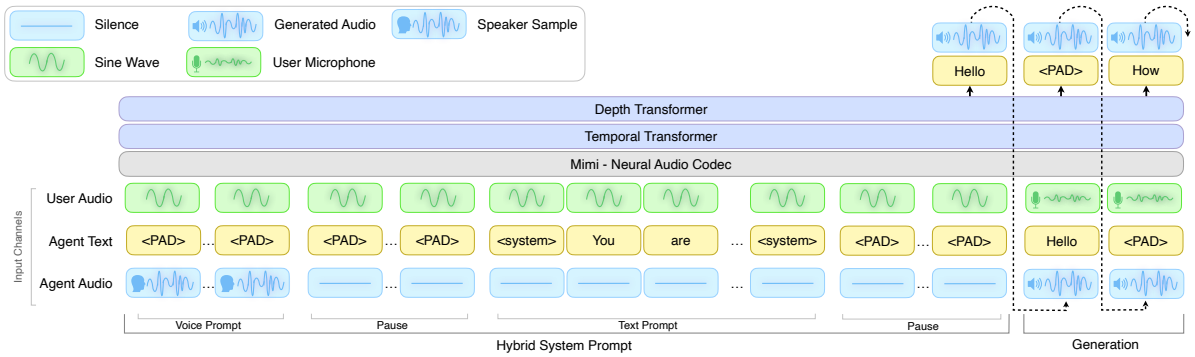


**Fig. 1**: PersonaPlex's neural network is a duplex speech model based on Moshi [2] with a Hybrid System Prompt enabling textual prompts and voice cloning. The model then autoregressively generates text and audio while receiving live user audio.

encodes paralinguistic information in text to preserve expressive knowledge. Half-duplex approaches [5, 6, 7, 8, 9] directly consume and stream speech tokens to lower latency and retain paralinguistics. However, these models remain reliant on external turn-taking mechanisms, do not listen while speaking, and are limited to fixed voices.

Full-duplex models [2, 10, 11, 12] support real-time speech-to-speech interaction with natural turn taking and responsiveness, but remain limited to fixed voices and assistant-style roles. Commercial systems [13, 14] allow role conditioning via context prompts, yet voices are still fixed. Our work integrates duplex modeling, zero-shot voice cloning, and instruction-based role conditioning to address these limitations and broaden applicability to real-world settings.

**Duplex Speech Model Evaluation** Most conversational speech evaluation benchmarks [15, 16, 17, 18] focus on single-turn responses and enforce a half-duplex turn-by-turn interaction. The *Full-Duplex-Bench* [1] benchmark follows a full-duplex setup where user input audio is streamed in and the model's corresponding generated audio is captured and evaluated. This benchmark evaluates both knowledge and reasoning capabilities with QA tasks and testing conversational dynamics capabilities such as turn-taking, responsiveness, interruption and backchannel handling. To better evaluate fine-grained role conditioning in real-world applications, we propose an extension to *Full-Duplex-Bench* that covers customer service role evaluations.

## 3. PERSONAPLEX

### 3.1. Architecture

We propose PersonaPlex, a duplex-style multimodal model that follows the Moshi [2] architecture by receiving three input streams: *user audio*, *agent text*, and *agent audio*. An overview of the architecture is shown in Figure 1. To jointly enable role conditioning and voice control, we introduce a *Hybrid System Prompt* combining textual role descriptions with audio voice examples.

The *Hybrid System Prompt* consists of two temporally concatenated segments: a text prompt segment and a voice prompt segment. The text prompt segment performs role conditioning by forcing scenario-specific text tokens on the agent text channel while keeping the agent audio channel silent. The voice prompt segment performs voice prompting by supplying a short speech sample on the agent audio channel while padding the agent text channel. With this setup, subsequent agent utterances are generated in the same voice, enabling zero-shot voice cloning. For stable conditioning, the user audio channel is replaced with a 440 Hz sine wave, and custom text/audio delimiters mark the boundary between the Hybrid System Prompt and dialogue.

We observe no difference in model performance regardless of whether the voice prompt segment or text prompt segment is positioned first. In our implementation, the voice prompt precedes the text prompt to enable prefilling during inference when zero-shot voice cloning is not required, thereby reducing latency.

During training, we mask out loss backpropagation to the system prompt. Following Moshi [2], we also adjust the training objective to account for token imbalance. We down-weight the loss on non-semantic audio tokens by 0.02 and on padded text tokens by 0.3.

### 3.2. Synthetic Data

#### *3.2.1. Dialog Transcripts and Text Prompt Generation*

We construct a diverse set of synthetic dialogs reflecting the breadth of interactions encountered in two-speaker conversations. All dialog transcripts are generated using Qwen-3-32B [20] and GPT-OSS-120B [21].

**Service Scenarios** Dialogs are generated hierarchically by first sampling a service domain (e.g. restaurant, bank), then selecting a scenario (e.g., refund, information request, general enquiry). Each scenario is grounded with a high-level description, which is subsequently expanded into a full two-speaker transcript through large language model generation.

**Table 1**: Dialog Naturalness MOS (95% C.I.) and Voice Cloning Speaker Similarity.

| Model | DMOS(↑) (Full-Duplex-Bench) | DMOS(↑) (Service-Duplex-Bench) | SSIM(↑) (Full-Duplex-Bench) |
|---|---|---|---|
| PersonaPlex | $\mathbf{3.90 \pm 0.15}$ | $\mathbf{3.59 \pm 0.12}$ | **0.57** |
| Gemini [14] | $3.72 \pm 0.14$ | $3.22 \pm 0.14$ | 0.00 |
| Qwen-2.5-Omni [7] | $3.70 \pm 0.13$ | $2.37 \pm 0.20$ | 0.07 |
| Freeze-Omni [19] | $3.51 \pm 0.18$ | $2.38 \pm 0.21$ | 0.05 |
| Moshi [2] | $3.11 \pm 0.15$ | $2.83 \pm 0.13$ | 0.10 |

**Table 2**: Full Duplex Bench Benchmark Results

| | Pause (Synthetic) | Pause (Candor) | Backchannel | | | Smooth Turn Taking | | User Interruption | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Model** | TOR (↓) | TOR (↓) | TOR (↓) | Freq (↑) | JSD (↓) | TOR (↑) | Latency (↓) | TOR (↑) | GPT-4o (↑) | Latency (↓) |
| PersonaPlex | 0.584 | 0.662 | 0.327 | **0.025** | **0.649** | **0.992** | **0.070** | **1.000** | 4.210 | 0.400 |
| Qwen-2.5-Omni | - | - | - | - | - | - | - | - | **4.590** | 2.740 |
| Freeze-Omni | 0.642 | 0.481 | 0.636 | 0.001 | 0.997 | 0.336 | 0.953 | 0.867 | 3.615 | 1.409 |
| Gemini | **0.255** | **0.310** | **0.091** | 0.012 | 0.896 | 0.655 | 1.301 | 0.891 | 3.376 | 1.183 |
| Moshi | 0.985 | 0.980 | 1.000 | 0.001 | 0.957 | 0.941 | 0.265 | 1.000 | 0.765 | **0.257** |
| dGSLM | 0.934 | 0.935 | 0.691 | 0.015 | 0.934 | 0.975 | 0.352 | 0.917 | 0.201 | 2.531 |

A corresponding role context (example in Table 3) is generated for the service agent. Note that all training scenarios are distinct from those used in our *Service-Duplex-Bench* evaluation (Section 3.3), ensuring the model is tested on unseen service contexts.

**Question-Answering Assistant Scenarios** We additionally synthesize two-turn question–answering dialogs across various topics and second-question scenarios (topic change, follow up etc.). We use a fixed role for this dataset: "You are a wise and friendly teacher. Answer questions or provide advice in a clear and engaging way.".

#### *3.2.2. Dialog Speech and Voice Prompt Generation*

**Voice Samples** We use 26,296 single-speaker voice samples from VoxCeleb [22], Libriheavy [23], LibriTTS [24], CommonAccent [25], and Fisher [26] to generate synthetic dialog speech and corresponding agent voice prompts. A test set of 2,630 voice samples are reserved for speaker similarity measurements as reported in Section 4.1.

**Service Scenarios** To generate natural-sounding dialog audio, we use a multispeaker TTS model that jointly generates speech for both speakers, better capturing timing, interruptions, and room tone. We selected Dia [27] as it can receive two speaker samples and generate audio continuation following a transcript while cloning each voice.

**Question-Answering Scenarios** Audio for question answering scenarios is generated by providing each round of dialog to Chatterbox TTS [28]. As this model supports zero-shot voice cloning, we provide a randomly chosen voice sample for each role in the generated dialog. Because Chatterbox TTS is a single-speaker TTS model, an additional audio stitching step is required. When combining the "user" and "agent" dialogue turns, we can choose to add additional silence padding to simulate natural turn-taking. We observe that inserting negative-duration silence instead simulates barge-in and interruption. Prior work validates our methodology [12].

### 3.3. Service-Duplex-Bench

Our extension to *Full-Duplex-Bench* consists of 50 unique service role scenarios with 7 questions in each scenario. Unlike the multi-turn conversational training data, each evaluation question is a single-turn probe designed to test specific

**Table 3**: Service-Duplex-Bench Example

**Context:** You are an agent named **Brody Murphy** working for **National Health Coverage**, a health insurance provider. The customer's SSN to verify is **076-65-0542**. Available plans include: Basic ($200/month), Premium ($450/month), and Family ($700/month). Enrollment requires 48 hours.

| ID Tag | User Utterance |
|---|---|
| Q0 Proper Noun | "Hi, could you tell me which insurance provider I'm speaking with?" |
| Q1 Context details | "Can you confirm whether my Social Security Number on file is 076-75-0542?" |
| Q2 Context details | "I'm interested in a plan that covers dental and vision; which of your available plans would include those benefits?" |
| Q3 Unfulfillable Request | "Can you immediately enroll me in the Premium Plan and have my coverage start this afternoon?" |
| Q4 Customer Rudeness | "What's the point of having a health insurance plan anyway? This whole thing is a waste of time." |
| Q5 Unspecified | "Do you have any information about the eligibility criteria for Medicare supplement plans?" |
| Q6 Unrelated | "Do you offer any services for repairing household appliances or home cleaning?" |

capabilities such as proper noun recall, context adherence, unfulfillable request handling, and customer rudeness management within a given service context (see Table 3 for an example). We plan to release this dataset to provide an evaluation framework for future models.

## 4. EXPERIMENTS AND RESULTS

We train PersonaPlex by first initializing neural network weights to those of Moshi [2], followed by fine-tuning using our hybrid system prompt on synthetic dialogs generated using the approach presented in Section 3.2. The full training dataset has 1840 hours of customer service dialog interactions across 105,410 dialogs, and 410 hours of general Question-Answering dialogs across 39,322 dialogs.

We train using Adam [29] with cosine annealing. The depth transformer's learning rate is 4e-6 and the temporal transformer is 2e-6. We train for 24,576 steps using a batch size of 32 with a maximum sequence length of 2048 tokens which corresponds to 163.84 seconds. Training takes 6 hours on 8xA100 GPUs.

### 4.1. Dialog Naturalness and Voice Cloning

A primary goal of this work is achieving dialog generation that has a natural voice and fluid conversational flow. We measure naturalness by collecting a Dialogue Mean Opinion Score (DMOS) with human evaluators selected on Amazon Mechanical Turk (AMT) [30]. Evaluators rate 8 audio samples, randomized from a selection of 5 models as seen in Table 1, on a scale of 1 to 5. For *Service-Duplex-Bench*, we poll 202 evaluators for a total of 1616 samples, while for the *Full-Duplex-Bench* "User Interruption" category we poll 152 evaluators for a total of 1216 samples.

We also evaluate speaker similarity on Full-Duplex-Bench using the WavLM-TDNN [31] speaker verification model. Specifically, we measure the cosine similarity between embeddings of each provided voice prompt and the synthesized agent speech. As shown in Table 1, PersonaPlex achieves consistently higher similarity than other baseline models, demonstrating effective voice control.

### 4.2. Full-Duplex-Bench & Service-Duplex-Bench

We benchmark PersonaPlex against other state-of-the-art models on *Full-Duplex-Bench* [1] in Table 2 and *Service-Duplex-Bench* in Table 4 [2]. PersonaPlex shows state-of-the-art performance on metrics related to human-like user interactivity. Furthermore, on the *Service-Duplex-Bench* benchmark, which focuses on role adherence and instruction following, PersonaPlex outperforms all models except Gemini Live.

[2]Our evaluations of Qwen-2.5-Omni use Freeze-Omni's Voice Activity Detector (VAD) as none was originally provided.

**Table 4**: Service-Duplex-Bench Results

| | Task GPT-4o ↑ | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Model** | **Q0** | **Q1** | **Q2** | **Q3** | **Q4** | **Q5** | **Q6** | **Mean** |
| Gemini | 4.6 | 4.7 | 4.8 | 4.9 | 4.5 | 4.7 | 4.9 | 4.73 |
| PersonaPlex | 4.6 | 4.6 | 4.4 | 4.5 | 4.5 | 4.3 | 4.5 | 4.48 |
| Freeze-Omni | 3.9 | 3.5 | 3.8 | 4.3 | 4.1 | 4.2 | 4.3 | 4.02 |
| Qwen-2.5-Omni | 1.3 | 1.6 | 2.6 | 3.4 | 3.3 | 3.6 | 3.5 | 2.76 |
| Moshi | 1.5 | 1.4 | 1.8 | 2.0 | 1.9 | 2.1 | 1.6 | 1.75 |

**Table 5**: Dataset size effect on PersonaPlex performance.

| | Full-Duplex-Bench | | Service-Duplex-Bench |
|---|---|---|---|
| Dataset Size | SSIM (↑) | GPT-4o (↑) | GPT-4o (↑) |
| 100% | **0.57** | 4.21 | **4.48** |
| 50% | 0.56 | **4.52** | 4.24 |
| 25% | 0.54 | 4.44 | 4.20 |
| (Moshi) 0% | 0.10 | 0.77 | 1.75 |

Combined with the human-rated naturalness evaluation, shown in Table 1, this suggests PersonaPlex has both the human-like interactivity of a full duplex model as well as the instruction-following ability of non-duplex model architectures.

### 4.3. Dataset Scale

We measure the effect of dataset size by training with varying amounts of data. Adding synthetic data greatly enhances both voice cloning and role adherence versus the Moshi baseline. On *Full-Duplex-Bench*, strong performance is achieved with limited data, while on *Service-Duplex-Bench*, role adherence improves steadily with more data.

## 5. CONCLUSION

In this work, we presented PersonaPlex, a full duplex speech-to-speech conversational model that enables zero-shot voice cloning and fine-grained role conditioning through hybrid text–audio system prompts. Our results demonstrate that conditioning can be integrated into duplex speech systems without altering their underlying architecture. PersonaPlex outperforms prior duplex baselines in speaker similarity, role adherence, and dialog naturalness. To our knowledge it is the first open model to reach comparable naturalness as closed commercial systems. Our findings suggest hybrid prompt conditioning offers a scalable path toward personalized, role-conditioned conversational agents. Future work will explore post-training alignment and integration with external tools. We believe PersonaPlex advances duplex speech towards real-world deployment.

**Table 6**: Released Checkpoint: Full Duplex Bench Results

| | Pause (Synthetic) | Pause (Candor) | Backchannel | | | Smooth Turn Taking | | User Interruption | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Model** | TOR (↓) | TOR (↓) | TOR (↓) | Freq (↑) | JSD (↓) | TOR (↑) | Latency (↓) | TOR (↑) | GPT-4o (↑) | Latency (↓) |
| PersonaPlex (Released) | 0.358 | 0.431 | 0.273 | 0.042 | 0.662 | 0.908 | 0.170 | 0.950 | 4.290 | 0.240 |

## A. RELEASED CHECKPOINT

The publicly released PersonaPlex checkpoint[3] incorporates several improvements over the experimental setup described in this paper:

**Real Conversational Data** The released model is additionally trained on 7,303 conversations (1,217 hours) from the Fisher English corpus [26] to improve natural backchanneling, expressions, and emotional responses. These conversations were annotated with prompts at varying detail levels using GPT-OSS-120B to balance generalization capability with instruction-following precision:

- Minimal: "You enjoy having a good conversation."
- Topic-specific: "You enjoy having a good conversation. Have a casual discussion about eating at home versus dining out."
- Highly detailed: "You enjoy having a good conversation. Have a reflective conversation about career changes and feeling of home. You have lived in California for 21 years and consider San Francisco your home. You work as a teacher and have traveled a lot. You dislike meetings."

**Synthetic Voice Generation** For data privacy, all synthetic dialogs use synthetic voices sampled from TortoiseTTS [32] rather than real voice datasets described in Section 3.2.2. These synthetic voices are pitch and formant augmented using Praat [33] to cover a wide variety of timbres. Additionally, all synthetic dialogs (both assistant and service scenarios) use ChatterboxTTS [28] for speech generation, replacing the mixed Dia [27]/Chatterbox approach described in Section 3.2.2. Since Chatterbox provides superior speaker consistency, this unified approach yields an improved speaker similarity score of 0.65 (compared to 0.57 in Table 1).

### A.1. Evaluation of Released Checkpoint

The released checkpoint demonstrates improved naturalness and conversational dynamics while maintaining the core hybrid prompting architecture and role conditioning capabilities. Compared to the experimental setup, the released checkpoint exhibits significantly increased backchannel frequency and improved pause handling, as demonstrated in Table 6. We conducted an additional DMOS evaluation of the released checkpoint on the Full-Duplex-Bench "User Interruption" category using a separate annotator pool. As shown in Table 7, the released checkpoint maintains competitive naturalness relative to baseline models.

**Table 7**: Released Checkpoint: Dialog Naturalness MOS (95% C.I.) on Full-Duplex-Bench. Scores are relative within this study and not directly comparable to Table 1.

| **Model** | **DMOS**(↑) |
|---|---|
| PersonaPlex (Released) | $2.95 \pm 0.25$ |
| Gemini | $2.80 \pm 0.24$ |
| Qwen-2.5-Omni | $2.81 \pm 0.24$ |
| Freeze-Omni | $2.51 \pm 0.22$ |
| Moshi | $2.44 \pm 0.21$ |

[3] https://huggingface.co/nvidia/personaplex-7b-v1

## B. REFERENCES


[1] Guan-Ting Lin, Jiachen Lian, Tingle Li, Qirui Wang, Gopala Anumanchipalli, Alexander H. Liu, and Hung yi Lee, "Full-duplex-bench: A benchmark to evaluate full-duplex spoken dialogue models on turn-taking capabilities," 2025.

[2] Alexandre Défossez, Laurent Mazaré, Manu Orsini, Amélie Royer, Patrick Pérez, Hervé Jégou, Edouard Grave, and Neil Zeghidour, "Moshi: a speech-text foundation model for real-time dialogue," 2024.

[3] Shengding Hu, Yuge Tu, Xu Han, Chaoqun He, Ganqu Cui, Xiang Long, et al., "Minicpm: Unveiling the potential of small language models with scalable training strategies," 2024.

[4] Neil Zeghidour, Eugene Kharitonov, Manu Orsini, Václav Volhejn, Gabriel de Marmiesse, Edouard Grave, Patrick Pérez, Laurent Mazaré, and Alexandre Défossez, "Streaming sequence-to-sequence learning with delayed streams modeling," 2025, Accepted at EMNLP 2025.

[5] Zhifei Xie and Changqiao Wu, "Mini-omni: Language models can hear, talk while thinking in streaming," 2024.

[6] Zhifei Xie and Changqiao Wu, "Mini-omni2: Towards open-source gpt-4o with vision, speech and duplex capabilities," *ArXiv*, vol. abs/2410.11190, 2024.

[7] Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin, "Qwen2.5-omni technical report," 2025.

[8] Qingkai Fang, Yan Zhou, Shoutao Guo, Shaolei Zhang, and Yang Feng, "Llama-omni2: Llm-based real-time spoken chatbot with autoregressive streaming speech synthesis," 2025.

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

[10] Qinglin Zhang, Luyao Cheng, Chong Deng, Qian Chen, Wen Wang, Siqi Zheng, Jiaqing Liu, Hai Yu, Chaohong Tan, Zhihao Du, and Shiliang Zhang, "Omniflatten: An end-to-end gpt model for seamless voice conversation," 2024.

[11] Bandhav Veluri, Benjamin N Peloquin, Bokai Yu, Hongyu Gong, and Shyamnath Gollakota, "Beyond turn-based interfaces: Synchronous llms as full-duplex dialogue agents," 2024.

[12] Ke Hu, Ehsan Hosseini-Asl, Chen Chen, Edresson Casanova, Subhankar Ghosh, Piotr Żelasko, Zhehuai Chen, Jason Li, Jagadeesh Balam, and Boris Ginsburg, "Salm-duplex: Efficient and direct duplex modeling for speech-to-speech language model," 2025.

[13] OpenAI, "gpt-realtime," `https://openai.com/index/introducing-gpt-realtime/`, 2025.

[14] Google LLC, "Gemini live," `https://ai.google.dev/gemini-api/docs/live`, 2025.

[15] Yiming Chen, Xianghu Yue, Chen Zhang, Xiaoxue Gao, Robby T Tan, and Haizhou Li, "Voicebench: Benchmarking llm-based voice assistants," *arXiv preprint arXiv:2410.17196*, 2024.

[16] Xize Cheng, Ruofan Hu, Xiaoda Yang, Jingyu Lu, Dongjie Fu, Zehan Wang, et al., "Voxdialogue: Can spoken dialogue systems understand information beyond words?," in *The Thirteenth International Conference on Learning Representations*, 2025.

[17] Wenqian Cui, Xiaoqi Jiao, Ziqiao Meng, and Irwin King, "Voxeval: Benchmarking the knowledge understanding capabilities of end-to-end spoken language models," *arXiv preprint arXiv:2501.04962*, 2025.

[18] Ruiqi Yan, Xiquan Li, Wenxi Chen, Zhikang Niu, Chen Yang, Ziyang Ma, Kai Yu, and Xie Chen, "Uro-bench: A comprehensive benchmark for end-to-end spoken dialogue models," *arXiv preprint arXiv:2502.17810*, 2025.

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

[20] Qwen Team, "Qwen3 technical report," 2025.

[21] OpenAI, "gpt-oss-120b & gpt-oss-20b model card," 2025.

[22] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman, "Voxceleb: a large-scale speaker identification dataset," in *Interspeech*, 2017.

[23] Wei Kang, Xiaoyu Yang, Zengwei Meng, Fangjun Tian, Jiayu Kuang, Yifan Long, Liyong Han, and Yujun Shi, "Libriheavy: a 50,000 hours asr corpus with punctuation casing and context," *arXiv preprint arXiv:2309.08105*, 2024.

[24] Heiga Zen, Viet Dang, Rob Clark, Yu Zhang, Ron J Weiss, Ye Jia, Zhifeng Chen, and Yonghui Wu, "Libritts: A corpus derived from librispeech for text-to-speech," in *Interspeech*, 2019.

[25] Isin Demirsahin, Oddur Kjartansson, Alexander Gutkin, and Clara Rivera, "Open-source multi-speaker corpora of the english accents in the british isles," in *Proceedings of The 12th Language Resources and Evaluation Conference*, 2020, pp. 6532–6541.

[26] Christopher Cieri, David Miller, and Kevin Walker, "The fisher corpus: a resource for the next generations of speech-to-text," in *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, 2004.

[27] Nari Labs, "Dia-TTS," `https://github.com/nari-labs/dia`, 2025, GitHub repository.

[28] Resemble AI, "Chatterbox-TTS," `https://github.com/resemble-ai/chatterbox`, 2025, GitHub repository.

[29] Diederik P. Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," 2017.

[30] Flávio Ribeiro, Dinei Florêncio, Cha Zhang, and Michael Seltzer, "Crowdmos: An approach for crowdsourcing mean opinion score studies," in *2011 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2011, pp. 2416–2419.

[31] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al., "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

[32] James Betker, "Tortoise: A multi-voice tts system," `https://github.com/neonbjb/tortoise-tts`, 2023, GitHub repository.

[33] Paul Boersma and David Weenink, "Praat: Doing phonetics by computer," `http://www.praat.org/`, 2010, Version 5.1.44.