Interspeech 2025
17-21 August 2025, Rotterdam, The Netherlands

SpeechDialogueFactory: A Framework for Natural Speech Dialogue Generation

Minghan Wang[1], *Ye Bai*[1], *Yuxia Wang*[2], *Thuy-Trang Vu*[1], *Ehsan Shareghi*[1], *Gholamreza Haffari*[1]

[1]Department of Data Science & AI, Monash University, Australia
[2]MBZUAI, United Arab Emirates

{firstname.lastname}@monash.edu, yuxia.wang@mbzuai.ac.ae

Abstract

High-quality speech conversational datasets are essential for developing and evaluating Speech-LLMs. However, collecting real-world recordings presents significant challenges including high costs, privacy concerns, and inconsistent quality, while existing synthetic approaches often lack authenticity due to limited acoustic variety and insufficient paralinguistic information. We present SPEECHDIALOGUEFACTORY, a framework that addresses these limitations through a three-stage pipeline: generating comprehensive metadata, creating detailed scripts, and producing utterances enriched with paralinguistic features. Our framework retrieves speaker voices from a voice bank and leverages paralinguistic tags for expressive TTS. We also introduce an automated evaluation protocol that shows strong correlation with human assessments. Experimental results demonstrate that our synthesized dialogues achieve quality comparable to human recordings while offering greater flexibility and control.
Index Terms: spoken dialogue, dialogue generation, large language model, data synthesis

1. Introduction

Recent advances in large language models (LLMs) have catalyzed the development of voice-based interfaces for AI systems, positioning Speech-LLMs at the forefront of human-AI interaction research [1, 2, 3]. These models promise more natural and accessible ways for users to interact with generative AI services, potentially transforming applications from virtual assistants to educational tools. The development of Speech-LLMs requires two critical training phases: 1) initial joint pretraining to align text and speech unit modalities, establishing foundational cross-modal understanding [4, 5, 6], and 2) speech dialogue capability training analogous to text LLM's SFT stage [6, 7]. This latter phase demands high-quality speech-to-speech dialogue datasets satisfying two fundamental requirements: 1) logically coherent and fluent colloquial content, distinct from written discourse [8, 9]; 2) natural prosody with appropriate emotional expression and clear articulation [10, 11].

Current approaches to Speech-LLM training data acquisition rely primarily on human-recorded dialogues and synthetic speech generated from textual data [1, 12, 7], each presenting fundamental challenges that limit their effectiveness. Human recordings present a critical quality-cost dilemma: natural conversations, while readily available, contain excessive disfluencies, filler words, and unstructured responses [13, 14, 15, 16, 17], while professional voice actor recordings, though superior in quality, incur prohibitive costs, resulting in severe data scarcity [12, 18]. The synthetic approach, leveraging existing text LLM dialogue datasets, introduces two critical misalignments. 1) **The Content-Style Mismatch**: datasets exhibit characteristics of written rather than spoken discourse, resulting in models that communicate more like prepared statements than natural dialogue [4, 7]. 2) **The Speech-Content Mismatch**: the absence of crucial paralinguistic information in text datasets, including speech rate, prosodic variation, and emotional markers, leads to mechanically synthesized speech lacking natural rhythm and emotional expression [4, 7].

To address these challenges, we present SPEECHDIALOGUEFACTORY, a systematic framework advancing speech dialogue synthesis through 3 key innovations: 1) **Customizable Control**: the framework enables fine-grained manipulation of dialogue parameters including topics, emotional dynamics, and character backgrounds, facilitating targeted domain-specific data generation; 2) **Content Authenticity**: it ensures dialogue coherence through a three-stage content generation process, maintaining consistent alignment with initial scenario specifications via enhanced contextual guidance; and 3) **Natural Speech Flow**: the framework incorporates comprehensive paralinguistic information during generation, providing crucial guidance for TTS systems in producing naturally inflected and emotionally appropriate speech.

Our experimental evaluation demonstrates that SPEECHDIALOGUEFACTORY matches the quality of professional recordings at a fraction of the cost. The main contributions of this paper are: 1) a **novel framework** for generating high-quality speech dialogues, which integrates structured content generation with advanced speech synthesis and employs paralinguistic control to maintain dialogue coherence and natural speech characteristics; and 2) a **comprehensive evaluation protocol** that validates the quality of the generated dialogues across multiple dimensions.

2. SpeechDialogueFactory

2.1. Textual Synthesis

While LLMs could generate dialogues with minimal prompting (e.g., "Generate a natural dialogue between two people"), our experiments revealed fundamental limitations with this naive approach. The LLM often produced limited number of turns in a dialogue with degrading fluency and coherence. More critically, such direct generation yields only raw textual content, lacking the paralinguistic information essential for natural speech synthesis. To address these limitations, we decompose textual dialogue synthesis into 3 distinct stages: metadata generation, dialogue scripting, and dialogue simulation. This structured approach enables better generation control while producing guiding variables for speech synthesis, ensuring natural and controllable output.

10.21437/Interspeech.2025-2013

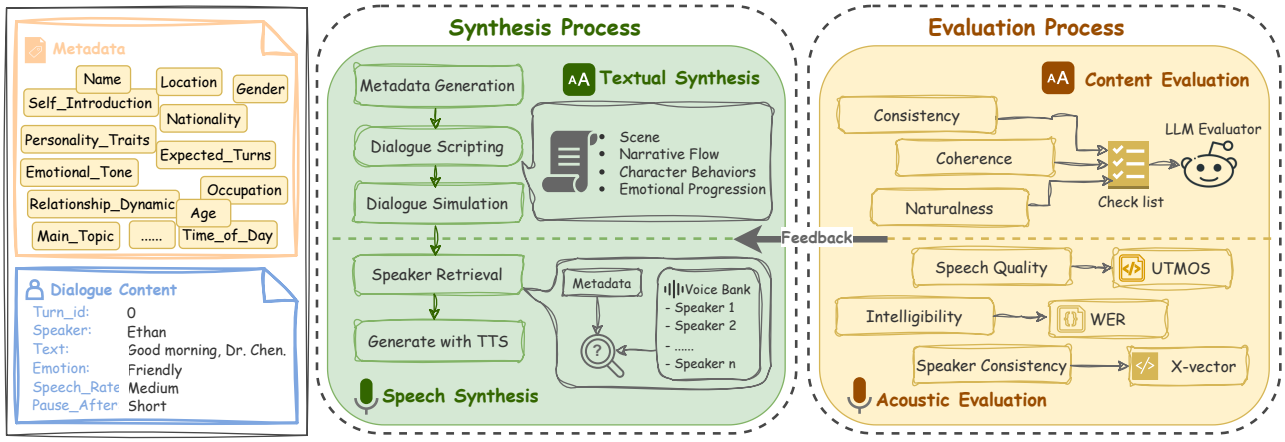


Figure 1: *Overview of* SPEECHDIALOGUEFACTORY*'s pipeline and output examples. The synthesis process (left) consists of 2 main stages: textual synthesis and speech synthesis. The evaluation process (right) combines LLM-based content assessment and acoustic evaluation (using off-the-shelf metrics). Examples demonstrate key metadata fields and utterance-level dialogue information.*

2.1.1. Metadata Generation

The first stage generates structured metadata defining 3 fundamental aspects: **dialogue settings**, **character profiles**, and **conversation context**. Dialogue settings encompass temporal and spatial contexts, expected dialogue length. Character profiles detail essential speaker attributes including names, ages, gender identifications, personalities, and cultural backgrounds. Conversation context outlines the topical progression and emotional design of the dialogue. These 3 aspects are decomposed into 26 specific fields that the LLM must populate. We employ constrained decoding [19] to ensure comprehensive coverage through well-structured JSON outputs.

However, unrestricted metadata generation by LLMs, while structured, still faces diversity limitations. To address this, we introduce a **seed-based** control mechanism that defines macro-level dialogue parameters through 4 key dimensions: interaction type (e.g. academic discussion, social gathering), goal (e.g. problem solving, information exchange), location (e.g. home, public spaces), and cultural background (e.g. Asian, Middle Eastern). These seed parameters can be either user-specified or LLM-generated. In our implementation, we bootstrapped the process by manually crafting 100 seed combinations, which were then algorithmically expanded by LLMs into a diverse set of over 1000 macro-level settings, dialogues generated in our experiments are sampled from this seed set. This approach ensures both high diversity and precise control over the metadata generation process.

2.1.2. Dialogue Scripting

While metadata structures the overall dialogue framework, directly generating dialogues from metadata often leads to narrative inconsistencies due to two key limitations. First, JSON-formatted metadata can be overly simplistic, lacking necessary contexts, giving LLMs too much freedom for interpretation. Second, the autoregressive nature of generation makes narrative flow corrections impossible once errors occur. To ensure natural narrative progression, we are inspired by [20, 21] and introduce an intermediate scripting stage before dialogue generation. The script, written in natural language for better LLM comprehension, contains 4 components:

- **Scene**: Details the dialogue setting, including spatiotemporal context, character backgrounds, and dialogue triggers.
- **Narrative Flow**: Segments the dialogue into opening, initial discussion, main discussion, and wrapping up, specifying turn counts and speaker intentions for each segment.
- **Character Behaviors**: Defines required behavioral patterns, such as showing enthusiasm during greetings or confusion during problem discussions.
- **Emotional Progression**: Maps emotional undertones across 3 phases (start, middle, end), incorporating necessary emotional transitions beyond metadata specifications to ensure dynamic emotional development.

2.1.3. Dialogue Simulation

The final stage converts metadata and dialogue scripts into concrete utterances by prompting the LLM to simulate a dialogue using previously generated context. The simulation executes in a single pass rather than through multi-round role-playing, improving computational efficiency. The output is structured as JSON objects containing both utterance text and synthesis-critical paralinguistic information, including speaking rate, emotional expression, and turn gap durations.

2.2. Speech Synthesis

To generate authentic and diverse dialogue audio, we implement a two-stage speech synthesis pipeline: speaker voice retrieval driven by character profiles, followed by TTS with voice cloning.

2.2.1. Speaker Retrieval

We leverage the Common Voice dataset [22] as our voice bank, capitalizing on its rich speaker attribute annotations. The retrieval process matches voices to character profiles by filtering speakers based on metadata-specified demographic attributes, ensuring acoustic characteristics align with the intended character personas.

2.2.2. Generate with TTS

While the field offers numerous TTS models with voice cloning capabilities [23, 24, 25], we primarily adopt COSYVOICE2 [23] and FISH-TTS [25] for their superior control over prosodic

features and consistent voice cloning quality. Our synthesis workflow combines two crucial elements: **voice prompts** derived from retrieved speaker samples, and **natural language control prompts** (only used by COSYVOICE2) constructed from speaker personality traits, utterance-specific speech rates, and emotional expressions. The final dialogue synthesis incorporates sophisticated turn-taking dynamics [26] by inserting variable-length silence intervals between utterances conditioned on the LLM planned gap length, mirroring human dialogue rhythms and patterns.

3. Quality Evaluation

While our dialogue synthesis pipeline enables scalable data generation, rigorous quality assessment remains essential. We propose a comprehensive evaluation framework examining both content and acoustic aspects of synthesized dialogues.

3.1. Automatic Evaluation

3.1.1. Content Evaluation

The evaluation of open-ended dialogue content presents unique challenges due to its inherent subjectivity and lack of absolute references [27]. Building on recent concepts of LLM-as-a-judge [28], we develop a three-dimensional evaluation approach addressing consistency, coherence, and naturalness.

Consistency Our consistency evaluation prioritizes metadata assessment, recognizing its foundational role in dialogue generation. To ensure objective assessment, we developed 15 questions (can be answered with a 0-100 score) covering major metadata fields, supplemented by two overall scores (0-100). Specifically we examine two critical aspects:

- **Input scenario consistency**: Verifies alignment (faithfulness) between generated metadata and user-provided scenario seeds. (7 questions + 1 overall score)
- **Self consistency**: Examines internal consistency within generated metadata. (9 questions + 1 overall score)

Coherence Coherence evaluation operates at the utterance level, only considering the generated dialogue as context without accessing metadata and script. Using five 0-100 scoring questions and an overall score (0-100), we assess multiple dimensions such as logical flow continuity, topic relevance between utterances.

Naturalness Our naturalness evaluation balances subjective and objective measures through 5 quantifiable criteria (0-100 scored) and an overall score (0-100). The evaluation focuses on: conversational style authenticity, contextually appropriate utterance length, emotional expression consistency and paralinguistic-semantic alignment.

While building upon existing evaluation paradigms [20], our framework advances the field through more granular and objective assessment criteria. This comprehensive approach enables effective quality control and sample filtering, ensuring dataset integrity.

3.1.2. Acoustic Evaluation

Our acoustic evaluation focus on objective assessment metrics, including speech quality, intelligibility and speaker consistency, integrating established evaluation methodologies into a comprehensive framework.

Speech Quality For speech quality assessment, we adopt the widely-used Mean Opinion Score (MOS) metrics. To balance efficiency and reliability, we primarily employ the automated UTMOSV2 [29] model for scoring. The dialogue-level MOS score is computed by averaging utterance-level scores across each conversation.

Intelligibility We evaluate intelligibility through automatic speech recognition (ASR) performance, using WHISPER-LARGE-V3-TURBO [30] to transcribe generated speech and computing Word Error Rate (WER) against reference texts. To focus specifically on TTS quality, our WER calculation excludes punctuation and normalizes case, minimizing ASR-specific biases in the evaluation.

Speaker Consistency Speaker consistency evaluation assesses the stability of voice cloning across utterances. Using the pre-trained x-vector model from SpeechBrain [31], we extract speaker embeddings for each utterance and compute cosine similarity between consecutive utterance pairs from the same speaker. We employ a threshold of 0.9 for binary consistency classification.

3.2. Human Evaluation

Our human evaluation focuses on content assessment, mirroring the automated evaluation protocol (§3.1.1) through identical checklist-based questions. We developed a Gradio-based interface to facilitate efficient scoring by evaluators.

4. Experiments

4.1. Experimental Setup

We evaluated SPEECHDIALOGUEFACTORY on 3 scales of Llama3 [32] models[1] to assess the robustness of the framework. For baseline comparison, we selected DAILYTALK [12] due to its comparable dialogue structure and comprehensive annotations. We randomly sampled 2,541 seeds to generate the same number of dialogues to DAILYTALK. For speech synthesis, we utilized dialogues generated by the LLAMA3.3-70B model, as we found that the quality of speech generated by TTS is less sensitive to content quality. COSYVOICE2 [23] and FISH-TTS [25] were both used to evaluate the robustness of the speech synthesis pipeline. The content evaluation was conducted using GPT-4O[2]. All experiments are conducted on a single computing node with 2 A100 80G GPUs. For human evaluation, we sampled 50 dialogues from each of the 4 configurations (200 total dialogues), with each of the 3 graduate-level evaluators proficient in English assessing all samples. This resulted in 600 individual content assessments were averaged across evaluators to obtain the final human evaluation scores.

4.2. Experimental Results

Table 1 presents our content generation results, where we aggregated fine-grained scoring criteria into two key aspects per dimension. Analysis reveals 3 significant findings: 1) Our evaluation framework demonstrates robust alignment with human assessment across quality levels, validating its reliability; 2) While model scale significantly impacts generation quality between 1B and 8B parameters, improvements plateau beyond 8B, suggesting a practical sweet spot for deployment; and 3) The 70B model surpasses DAILYTALK across multiple dimensions, indicating that our structured dialogue generation approach can achieve superior coherence and naturalness compared to human-written conversations.

[1] instruction tuned version for 3.3-70B, 3.1-8B and 3.2-1B

[2] gpt-4o-2024-11-20

Dataset	SpeechDialogueFactory						DailyTalk	
Generator	Llama3.3-70B		Llama3.1-8B		Llama3.2-1B			
Evaluator	GPT-4o	*Human*	GPT-4o	*Human*	GPT-4o	*Human*	GPT-4o	*Human*
Consistency	**96.9** ± 2.3	*97.6* ± 5.8	95.9 ± 3.5	95.5 ± 8.4	56.6 ± 13.1	44.9 ± 10.2	-	-
- Scenario Consistency(6)	**94.7** ± 5.6	*93.8* ± 5.8	92.9 ± 8.5	90.1 ± 12.1	21.6 ± 24.2	6.2 ± 11.7	-	-
- Self Consistency(9)	**99.0** ± 2.3	*99.3* ± 1.5	98.8 ± 2.8	99.1 ± 2.3	91.8 ± 11.2	87.0 ± 11.7	-	-
Coherence	**99.7** ± 0.5	*99.5* ± 2.1	99.6 ± 0.7	99.3 ± 2.0	99.0 ± 1.3	97.1 ± 7.6	99.2 ± 1.3	97.4 ± 6.2
- Topic Relevance(2)	**99.6** ± 0.7	*99.3* ± 2.2	99.4 ± 1.0	99.0 ± 3.0	98.4 ± 2.0	96.4 ± 10.3	98.9 ± 1.8	95.7 ± 10.2
- Logical Flow(2)	**99.8** ± 0.3	*99.7* ± 1.7	**99.8** ± 0.4	99.5 ± 1.5	99.3 ± 0.9	97.8 ± 7.2	99.4 ± 1.0	98.0 ± 4.9
Naturalness	91.6 ± 1.8	83.1 ± 7.7	**92.3** ± 2.6	*84.9* ± 9.4	90.3 ± 2.8	82.7 ± 15.2	85.9 ± 5.0	79.1 ± 12.0
- Stylistic Flow(3)	90.3 ± 2.2	84.0 ± 5.5	**90.6** ± 2.9	83.4 ± 7.6	88.6 ± 3.1	80.6 ± 10.7	89.5 ± 4.0	*84.5* ± 9.6
- Emotional Appropriateness(2)	93.6 ± 2.1	89.2 ± 7.6	**96.0** ± 2.5	*94.2* ± 6.4	94.4 ± 3.1	90.5 ± 9.3	83.1 ± 7.0	69.8 ± 22.9

Table 1: *Content evaluation results across three model scales of* SpeechDialogueFactory*, against* DailyTalk*baseline. Sub-dimension scores reflect specific questions (counted in parentheses) while top-level scores indicate overall assessment.* ***Bold*** *and* *underlined* *values denote best performance.* DailyTalk*'s consistency evaluation is omitted due to lack of metadata.*

Dataset	SpeechDialogueFactory		DailyTalk
TTS Models	CosyVoice2	Fish-TTS	
Speech Quality ↑	**3.38** ± 0.14	2.81 ± 0.18	3.28 ± 0.21
Intelligibility ↓	2.36 ± 2.85	**1.72** ± 1.66	7.05 ± 6.28
Speaker Consistency ↑	**99.90** ± 0.17	97.56 ± 3.40	-
# Speaker	36	33	2

Table 2: *Acoustic evaluation comparing two TTS models in* SpeechDialogueFactory *against the* DailyTalk *baseline. Speaker consistency evaluation is omitted for* DailyTalk *as it is recorded by two fixed human speakers.*

Table 2 summarizes the acoustic evaluation results, where SpeechDialogueFactory with CosyVoice2 demonstrates competitive performance against DailyTalk's human recordings. Both TTS models achieve higher intelligibility scores, suggesting clearer articulation compared to human recordings. Notably, while maintaining comparable quality, our framework offers substantially greater speaker diversity.

4.3. Ablation Study

4.3.1. *Effectiveness of Multi-Stage Generation*

We conducted comprehensive ablation studies focusing on content generation to validate the necessity of each pipeline component in SpeechDialogueFactory. Using the 8B model, we generated 300 dialogues under 4 distinct context-usage configurations: 1) unconstrained generation without any context, 2) metadata-only, 3) script-only and 4) the complete pipeline. All configurations were evaluated using the 70B model.

Results in Table 3 reveal the hierarchical importance of different components: metadata demonstrates the most substantial impact, particularly in controlling dialogue length and maintaining coherence and naturalness across turns. Script-only setting doesn't provide notable improvements, while the complete pipeline consistently achieves superior performance across all metrics. These findings validate our multi-stage design, suggesting that each component addresses distinct aspects of dialogue quality.

4.3.2. *Effectiveness of Input Seed*

We also conducted an ablation study examining the impact of input seeds on metadata diversity. Using our 70B model, we generated 3,000 samples under each configuration and quantified metadata field diversity using Shannon Entropy Ratio. Figure 2 demonstrates that input seeds substantially increase the diversity of generated dialogue settings.

Settings	Coherence	Naturalness	# Turns
Baseline	90.25 ± 9.20	84.87 ± 4.67	4.06 ± 1.34
+ Metadata	95.68 ± 4.68	87.03 ± 4.46	9.38 ± 2.01
+ Script	93.39 ± 8.43	84.06 ± 4.33	11.26 ± 7.72
+ Metadata & Script	**97.39** ± 3.19	**89.01** ± 3.84	11.82 ± 5.31

Table 3: *Ablation study results comparing four configurations. Coherence and naturalness scores are averaged for all corresponding questions. Dialogue turn counts are also reported.*

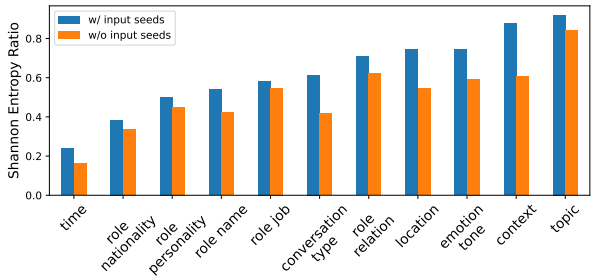


Figure 2: *The comparison of metadata fields' diversity with or without input seeds.*

5. Conclusion

We present SpeechDialogueFactory, a systematic framework for generating high-quality speech dialogues that addresses critical challenges in Speech-LLM training data acquisition. Through its multi-stage generation pipeline, our framework ensures both content authenticity and acoustic naturalness while significantly reducing production costs compared to professional voice recordings. Experimental results demonstrate the framework's effectiveness in maintaining dialogue coherence, natural speech flow, and speaker diversity, with our comprehensive evaluation methodology validates its quality across multiple dimensions. These achievements suggest that structured generation approaches can effectively address current limitations in Speech-LLM training data creation. Future work could explore extending the framework to multi-party conversations and developing more comprehensive evaluation metrics for speech dialogue quality assessment.

6. References

[1] A. Défossez, L. Mazaré, M. Orsini, A. Royer, P. Pérez, H. Jégou, E. Grave, and N. Zeghidour, "Moshi: a speech-text foundation model for real-time dialogue," 2024. [Online]. Available: https://arxiv.org/abs/2410.00037

[2] Y. Chu, J. Xu, X. Zhou, Q. Yang, S. Zhang, Z. Yan, C. Zhou, and J. Zhou, "Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models," 2023. [Online]. Available: https://arxiv.org/abs/2311.07919

[3] OpenAI, :, A. Hurst, A. Lerer, A. P. Goucher, A. Perelman, A. Ramesh, and A. C. et al., "Gpt-4o system card," 2024. [Online]. Available: https://arxiv.org/abs/2410.21276

[4] Q. Zhang, L. Cheng, C. Deng, Q. Chen, W. Wang, S. Zheng, J. Liu, H. Yu, and C. Tan, "Omniflatten: An end-to-end gpt model for seamless voice conversation," 2024. [Online]. Available: https://arxiv.org/abs/2410.17799

[5] Z. Xie and C. Wu, "Mini-omni2: Towards open-source gpt-4o with vision, speech and duplex capabilities," 2024. [Online]. Available: https://arxiv.org/abs/2410.11190

[6] A. Zeng, Z. Du, M. Liu, K. Wang, S. Jiang, L. Zhao, Y. Dong, and J. Tang, "Glm-4-voice: Towards intelligent and human-like end-to-end spoken chatbot," 2024. [Online]. Available: https://arxiv.org/abs/2412.02612

[7] X. Wang, Y. Li, C. Fu, Y. Shen, L. Xie, K. Li, X. Sun, and L. Ma, "Freeze-omni: A smart and low latency speech-to-speech dialogue model with frozen llm," 2024. [Online]. Available: https://arxiv.org/abs/2411.00774

[8] H. Zhan, Z. Li, Y. Wang, L. Luo, T. Feng, X. Kang, Y. Hua, L. Qu, L.-K. Soon, S. Sharma, I. Zukerman, Z. Semnani-Azad, and G. Haffari, "Socialdial: A benchmark for socially-aware dialogue systems," 2023. [Online]. Available: https://arxiv.org/abs/2304.12026

[9] M. Zhong, Y. Liu, D. Yin, Y. Mao, Y. Jiao, P. Liu, C. Zhu, H. Ji, and J. Han, "Towards a unified multi-dimensional evaluator for text generation," 2022. [Online]. Available: https://arxiv.org/abs/2210.07197

[10] Y. Chen, X. Yue, C. Zhang, X. Gao, R. T. Tan, and H. Li, "Voicebench: Benchmarking llm-based voice assistants," 2024. [Online]. Available: https://arxiv.org/abs/2410.17196

[11] S. Maiti, Y. Peng, T. Saeki, and S. Watanabe, "Speechlmscore: Evaluating speech generation using speech language model," 2022. [Online]. Available: https://arxiv.org/abs/2212.04559

[12] K. Lee, K. Park, and D. Kim, "Dailytalk: Spoken dialogue dataset for conversational text-to-speech," 2023. [Online]. Available: https://arxiv.org/abs/2207.01063

[13] Z. Yang, Y. Chen, L. Luo, R. Yang, L. Ye, G. Cheng, J. Xu, Y. Jin, Q. Zhang, P. Zhang, L. Xie, and Y. Yan, "Open source magicdata-ramc: A rich annotated mandarin conversational(ramc) speech dataset," 2022. [Online]. Available: https://arxiv.org/abs/2203.16844

[14] J. Barker, S. Watanabe, E. Vincent, and J. Trmal, "The fifth 'chime' speech separation and recognition challenge: Dataset, task and baselines," 2018. [Online]. Available: https://arxiv.org/abs/1803.10609

[15] Y. Saito, Y. Nishimura, S. Takamichi, K. Tachibana, and H. Saruwatari, "Studies: Corpus of japanese empathetic dialogue speech towards friendly voice agent," 2022. [Online]. Available: https://arxiv.org/abs/2203.14757

[16] J. Li, Y. Meng, C. Li, Z. Wu, H. Meng, C. Weng, and D. Su, "Enhancing speaking styles in conversational text-to-speech synthesis with graph-based multi-modal context modeling," 2022. [Online]. Available: https://arxiv.org/abs/2106.06233

[17] R. Liu, Y. Hu, Y. Ren, X. Yin, and H. Li, "Generative expressive conversational speech synthesis," 2024. [Online]. Available: https://arxiv.org/abs/2407.21491

[18] R. Zandie, M. H. Mahoor, J. Madsen, and E. S. Emamian, "Ryanspeech: A corpus for conversational text-to-speech synthesis," 2021. [Online]. Available: https://arxiv.org/abs/2106.08468

[19] B. T. Willard and R. Louf, "Efficient guided generation for large language models," 2023. [Online]. Available: https://arxiv.org/abs/2307.09702

[20] J. Chen, X. Zhu, C. Yang, C. Shi, Y. Xi, Y. Zhang, J. Wang, J. Pu, R. Zhang, Y. Yang, and T. Feng, "Hollmwood: Unleashing the creativity of large language models in screenwriting via role playing," 2024. [Online]. Available: https://arxiv.org/abs/2406.11683

[21] Z. Song, C. Wang, J. Sheng, C. Zhang, G. Yu, J. Fan, and T. Chen, "Moviellm: Enhancing long video understanding with ai-generated movies," 2024. [Online]. Available: https://arxiv.org/abs/2403.01422

[22] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," 2020. [Online]. Available: https://arxiv.org/abs/1912.06670

[23] Z. Du, Y. Wang, Q. Chen, X. Shi, X. Lv, T. Zhao, Z. Gao, Y. Yang, C. Gao, H. Wang, F. Yu, H. Liu, Z. Sheng, Y. Gu, C. Deng, W. Wang, S. Zhang, Z. Yan, and J. Zhou, "Cosyvoice 2: Scalable streaming speech synthesis with large language models," 2024. [Online]. Available: https://arxiv.org/abs/2412.10117

[24] A. Vyas, B. Shi, M. Le, A. Tjandra, Y.-C. Wu, B. Guo, J. Zhang, X. Zhang, R. Adkins, W. Ngan, J. Wang, I. Cruz, B. Akula, A. Akinyemi, B. Ellis, R. Moritz, Y. Yungster, A. Rakotoarison, L. Tan, C. Summers, C. Wood, J. Lane, M. Williamson, and W.-N. Hsu, "Audiobox: Unified audio generation with natural language prompts," 2023. [Online]. Available: https://arxiv.org/abs/2312.15821

[25] S. Liao, Y. Wang, T. Li, Y. Cheng, R. Zhang, R. Zhou, and Y. Xing, "Fish-speech: Leveraging large language models for advanced multilingual text-to-speech synthesis," 2024. [Online]. Available: https://arxiv.org/abs/2411.01156

[26] G. Skantze, "Turn-taking in conversational systems and human-robot interaction: A review," *Computer Speech & Language*, vol. 67, p. 101178, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S088523082030111X

[27] T. Ito, K. van Deemter, and J. Suzuki, "Reference-free evaluation metrics for text generation: A survey," 2025. [Online]. Available: https://arxiv.org/abs/2501.12011

[28] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. P. Xing, H. Zhang, J. E. Gonzalez, and I. Stoica, "Judging llm-as-a-judge with mt-bench and chatbot arena," 2023. [Online]. Available: https://arxiv.org/abs/2306.05685

[29] K. Baba, W. Nakata, Y. Saito, and H. Saruwatari, "The t05 system for the VoiceMOS Challenge 2024: Transfer learning from deep image classifier to naturalness MOS prediction of high-quality synthetic speech," in *IEEE Spoken Language Technology Workshop (SLT)*, 2024.

[30] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," 2022. [Online]. Available: https://arxiv.org/abs/2212.04356

[31] M. Ravanelli, T. Parcollet, P. Plantinga, A. Rouhe, S. Cornell, L. Lugosch, C. Subakan, N. Dawalatabad, A. Heba, J. Zhong, J.-C. Chou, S.-L. Yeh, S.-W. Fu, C.-F. Liao, E. Rastorgueva, F. Grondin, W. Aris, H. Na, Y. Gao, R. D. Mori, and Y. Bengio, "SpeechBrain: A general-purpose speech toolkit," 2021, arXiv:2106.04624.

[32] A. Grattafiori, A. Dubey, and A. J. et al., "The llama 3 herd of models," 2024. [Online]. Available: https://arxiv.org/abs/2407.21783