Voicing Personas: Rewriting Persona Descriptions into Style Prompts for Controllable Text-to-Speech

Anonymous ACL submission

Abstract

In this paper, we propose a novel framework to control voice style in prompt-based, controllable text-to-speech systems by leveraging textual personas as voice style prompts. We present two persona rewriting strategies to transform generic persona descriptions into speech-oriented prompts, enabling fine-grained manipulation of prosodic attributes such as pitch, emotion, and speaking rate. Experimental results demonstrate that our methods enhance the naturalness, clarity, and consistency of synthesized speech. Finally, we analyze implicit social biases introduced by LLM-based rewriting, with a focus on gender. We underscore voice style as a crucial factor for persona-driven AI dialogue systems.

1 Introduction

The rapid advancements in large language models (LLMs) have greatly increased the demand for interactive AI applications such as personalized chatbots, metaverse dialogue systems, and virtual influencers. Persona research has thus emerged as a core component; a persona refers to a fictional character that encapsulates various identity traits, such as personality, background, and interests, and is widely used to generate coherent and immersive conversations (Tseng et al., 2024; Liu et al., 2024b). In particular, incorporating personas that reflect specific backgrounds or talking styles (e.g., “25-year-old US college student who likes to travel and cook") into LLM-based dialogue systems enhances both response consistency and realism (Piercy et al., 2025; Kühne et al., 2024).

In parallel, text-to-speech (TTS) technology has achieved remarkable gains; prompt-based control now provides fine-grained manipulation of prosodic features, such as pitch, emotion, speaking rate, and vocal intensity (e.g., "male, British accent, low-pitch, fast speaking speed") (Guo et al., 2023; Lacombe et al., 2024). These controllable TTS systems can easily generate expressive and emotionally resonant speech that closely resembles natural human conversation. Unlike previous TTS models that depended on reference-style audio snippets to convey voice characteristics, prompt-based style specification via natural language affords greater flexibility, controllability, and human interpretability. However, despite these gains, studies bridging textual personas and TTS style prompts remain scarce, and current models frequently fail to preserve speaker characteristics and clarity.

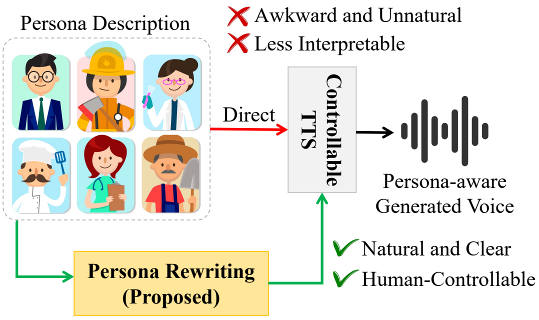
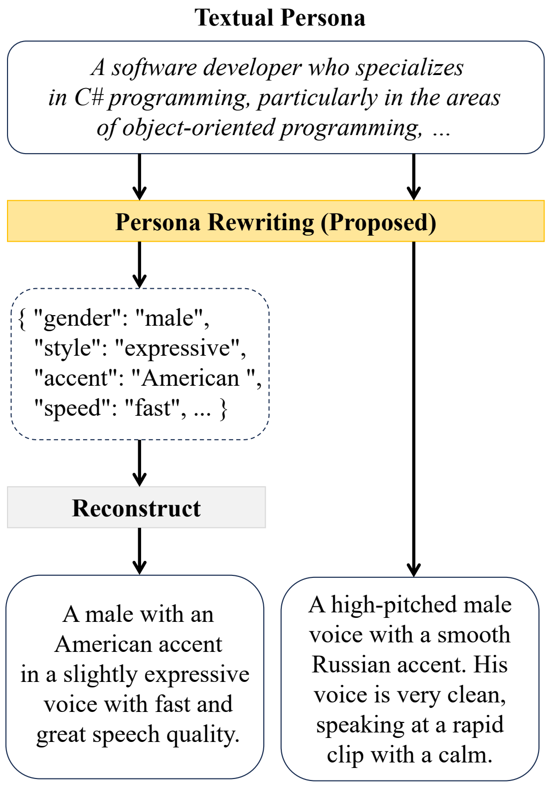


Figure 1: Illustration of the proposed method. Persona descriptions designed for text-based LLMs are suboptimal for controllable TTS models.

In this paper, we present a novel, plug-and-play method to control voice style in off-the-shelf TTS models using persona descriptions. Specifically, we introduce two **persona rewriting** techniques that convert a persona description to optimized style prompts (see Figure 1). Our experiments demonstrate considerable gains across four key speech-quality metrics. By extending persona research beyond text-only dialogue into the domain of speech, our approach offers a new direction for personalized and engaging AI conversation.

We summarize our main contributions as follows:

- We demonstrate that directly applying textual persona descriptions as TTS style prompts


Textual Persona
A software developer who specializes in C# programming, particularly in the areas of object-oriented programming, ...
Persona Rewriting (Proposed)
{ "gender": "male", "style": "expressive", "accent": "American ", "speed": "fast", ... }
Reconstruct
A male with an American accent in a slightly expressive voice with fast and great speech quality.
A high-pitched male voice with a smooth Russian accent. His voice is very clean, speaking at a rapid clip with a calm.

(a) Closed-ended Prompt (b) Open-ended Prompt

Figure 2: Proposed prompt rewriting process. (a) Closed-ended prompting converts textual persona traits into a structured attribute format (e.g., gender, speed, accent), while (b) Open-ended prompting produces free-form, natural language style descriptions.

yields suboptimal results.

- We propose two persona rewriting techniques that improve the naturalness, clarity, and pronunciation accuracy.
- We investigate potential biases that may arise during the rewriting process, revealing how LLMs can inject implicit social biases into voice style.

2 Related Work

2.1 Prompt-based TTS

Prompt-based TTS has emerged as a controllable and intuitive alternative to conventional synthesis methods. Models such as PromptTTS (Guo et al., 2023) and PromptTTS-2 (Shimizu et al., 2024) have demonstrated that natural language descriptions of both content ("what is spoken") and style ("how it is spoken") can effectively guide TTS systems without reliance on reference audio. Furthermore, multimodal approaches including VALL-E (Wang et al., 2023a), Parler-TTS (Lyth and King, 2024), and MM-TTS (Guan et al., 2024) have advanced expressiveness by integrating textual and acoustic cues. However, existing studies predominantly focus on formal styles, and the challenge of converting persona character descriptions into effective vocal style prompts remains insufficiently explored.

2.2 LLM Personalization Using Persona

LLMs have increasingly incorporated personas to ensure consistent tone, background, and attitude in dialogue generation (Tseng et al., 2024). Previous approaches, such as PersonaChat (Zhang et al., 2018), rely on fixed persona sentences to preserve character identity. Recent methods improve robustness and adaptability by incorporating retrieval augmentation (Chae et al., 2023), , and dynamic identity modeling techniques (Shea and Yu, 2023), large-scale persona datasets (Ge et al., 2024) and contrastive learning (Ji et al., 2025). While these techniques have improved textual coherence and user engagement, they remain limited to the text modality. Despite conceptual parallels between textual personas and vocal style attributes, the application of persona information to speech synthesis has been largely overlooked. Our study bridges this gap by adapting persona-based representations for vocal style control in TTS systems.

3 Method

Our method comprises two complementary persona rewriting strategies that convert persona descriptions into style prompts compatible with TTS models. Figure 2 illustrates the overall workflow. Please see the Appendix for more details, including prompts for the LLM.

3.1 Closed-ended Prompting

This approach guides the LLM to extract information based on a predefined set of attributes, such as gender, age, tone, speed, and pitch. Given a persona description, the LLM infers values for each attribute and outputs them in a structured JSON format. The structured data is then converted into a prompt-style description suitable for TTS input using an LLM. The method offers strong consistency and controllability, making it well-suited for applications requiring predictable and interpretable outputs. This reliability also makes closed-ended prompting particularly effective for analyzing LLM behavior.

Method	UTMOS(↑)	WER(↓)	PQ(↑)	CE(↑)
Baseline	2.847±0.539	0.222±0.151	6.739±1.132	5.734±0.742
Closed-ended Prompting (**Ours**)	**2.942**±0.532	**0.174**±0.145	**6.858**±1.113	**5.816**±0.613
Open-ended Prompting (**Ours**)	2.884±0.530	0.179±0.146	6.766±1.141	5.738±0.677

Table 1: Quantitative evaluation results across four metrics, including naturalness (UTMOS), pronunciation accuracy (WER), technical audio quality (PQ), and expressive richness (CE).

3.2 Open-ended Prompting

In contrast, this approach allows LLM to craft its own style descriptions without imposing constraints on the output format. As a result, the model generates more creative and nuanced natural language prompts (see Figure 2(b)). This flexibility makes open-ended prompting particularly suitable for emotionally expressive interfaces or narrative-driven content generation, where expressive, story-driven, or character-rich voice styles are preferred.

4 Experimental Results

4.1 Setup

We utilize Parler-TTS (Lyth and King, 2024), a natural language-controllable TTS model, because it offers a flexible interface for fine-grained style manipulation without reference audio. We sample persona prompts from the Persona-1M (Ge et al., 2024) dataset. This large-scale persona corpus is a synthetic dataset constructed for persona research in LLMs. To keep the linguistic content consistent, we use the transcripts provided by the LJ-Speech (Ito and Johnson, 2017), a widely used TTS dataset.

4.2 Style Preset

For closed-ended prompting, we define a fixed list of style attributes by examining the most frequent labels in Parler-TTS's training data. Our preset includes options for gender, tone (e.g., analytical, warm, engaging), speaking rate (e.g., slow, normal, fast), and pitch (e.g., low, medium, high).

4.3 Evaluation Metrics

We measure synthesized speech quality and accuracy using four complementary metrics: UTMOS (Saeki et al., 2022), Word Error Rate (WER), Production Quality (PQ), and Content Enjoyment (CE) from Audiobox Aesthetics (Tjandra et al., 2025). These metrics evaluate synthesized speech from multiple perspectives, including perceived naturalness, pronunciation accuracy, overall quality, and contextual coherence. Please see the Appendix for more details.

	Male	Female	Others
Before rewriting	10%	9%	81%
After rewriting	64%	33%	3%

Table 2: Gender distribution change after rewriting. 'Others' include not specified and uncertain samples.

4.4 Quantitative Evaluation

We prepare a total of 1,000 text–persona input pairs and generate speech using three persona rewriting approaches: Baseline (i.e., no rewriting), Open-ended prompting, and Closed-ended prompting. Table 1 shows that the closed-ended prompting achieves the best performance across all four metrics. In particular, it reduces WER by 5% compared to the baseline, increases UTMOS by 0.1 points, and delivers corresponding gains in PQ and CE. These results confirm that structuring persona information into optimized style prompts facilitates the generation of speech that is both clearer and more natural in style.

5 Analysis

5.1 Bias in Style Attributes

We analyze potential biases in LLM-based prompt-to-style conversion by classifying outputs of the closed-ended prompting across five dimensions: gender, accent, tone, speaking rate, and pitch (see Figure 3). Regardless of explicit gender cues in the input persona, the LLM assigns male voices 64% of the time, female voices 33%, and leaves 3% unspecified (see Table 2). For accent, 90% of outputs default to North American or British variants, while other regional accents are highly underrepresented. Additionally, 61% of the generated prompts favor a fast speaking rate.

These imbalances reveal clear biases toward male representation and Western accents, suggesting that the underlying model or data carries implicit gender and regional biases. Consequently, the

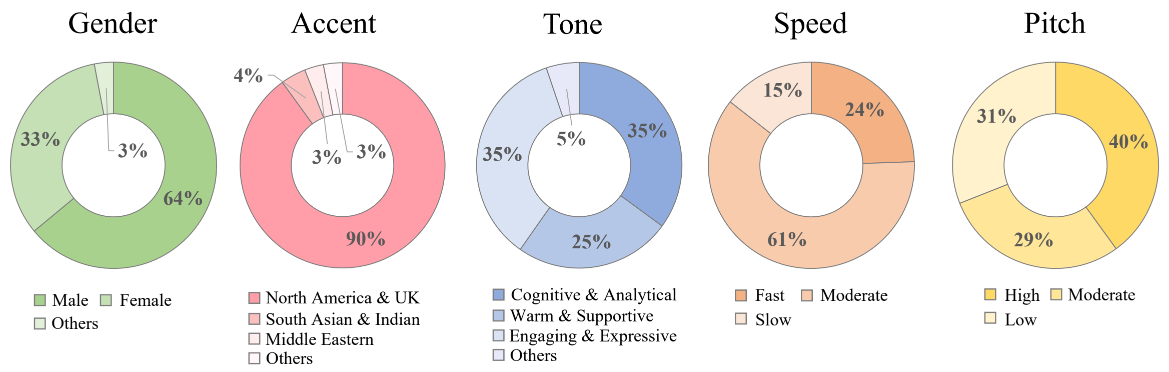


Figure 3: Distribution of speech style attributes generated via LLM-based prompt rewriting. The attributes are categorized into five dimensions: gender, accent, tone, speed, and pitch.

Gender	C&A	W&S	E&E	Others
Male	**41.3%**	14.3%	36.5%	7.9%
Female	23.5%	**44.1%**	32.4%	0.0%

Table 3: Speech tone profiles by gender. See Figure 3 'Tone' (3rd circle) for the abbreviations.

Gender	Fast	Normal	Slow
Male	35.1%	**43.3%**	21.6%
Female	22.7%	36.4%	**40.9%**

Table 4: Speech speed profiles by gender.

Gender	High	Moderate	Low
Male	27.0%	31.7%	**41.3%**
Female	**61.8%**	23.5%	14.7%

Table 5: Speech pitch profiles by gender.

model risks misrepresenting and excluding multicultural and diverse user groups.

5.2 Correlation Between Style Biases

We further analyze interactions between these biases by conditioning style distributions on gender. Tables 3, 4, and 5 demonstrate LLM's implicit bias, some of which aligns with common stereotypes. Male voices co-appear most often with '*Cognitive & Analytical*' tone and '*Low*' pitch. On the other hand, female voices disproportionately co-occur with '*Warm & Supportive*' tone and '*High*' pitch. Note that each value in these tables represents the relative proportion within the given gender.

These findings indicate that LLMs not only exhibit bias in speech style but also perpetuate deeper gender stereotypes. Notably, an indirect route—inferring persona characteristics and converting them to style attributes—surfaces the entrenched prejudices. This suggests that bias mitigation efforts should also consider latent inference mechanisms, not just explicit persona injection. While recent studies have shown that explicitly assigning personas to LLMs amplifies bias (Gupta et al., 2024; Liu et al., 2024a), our results demonstrate that even reasoning about persona descriptions without explicit assignment is sufficient to introduce bias.

5.3 Future Research Directions

Building on our findings, future work should assess existing bias-mitigation methods (Bai et al., 2022; Smith et al., 2022; Raza et al., 2024; Bai et al., 2024) within our rewriting frameworks to reduce gender and regional skew. Moreover, As voice interfaces proliferate, persona specifications should include nuanced voice-style dimensions beyond character attributes. To support this, we advocate for the creation of comprehensive benchmarks that measure both speech-style fidelity and demographic parity across multiple dimensions (gender, accent, age, etc.).

6 Conclusion

In this work, we introduced two persona rewriting techniques that transform textual persona descriptions into the TTS style prompts. Experiments showed that our methods provide considerable improvements in speech quality. Furthermore, we analyzed how this approach reflects the inherent stereotypes embedded in the LLM. This work represents an initial exploration of persona-driven TTS, highlighting the need for future research to comprehensively address factors not only the voice quality but also the ethical and cultural considerations.

Limitations

While this work presents a promising approach for integrating textual personas into prompt-based TTS systems, there are several limitations. First, all experiments are conducted on a single model, Parler-TTS. Although our method can be applied to other off-the-shelf LLM and TTS models, its generalizability to different architectures remains to be evaluated. Second, the absence of standardized benchmarks for persona-based TTS makes it difficult to compare prior results or reproduce experiments. For example, end-to-end performance—from textual persona to synthesized voice—is hard to quantify because outputs are inherently subjective. Nevertheless, our method shows that persona-based TTS performance can still be partly evaluated in terms of speech quality and fairness.

Ethics Statements

We acknowledge several ethical considerations associated with the integration of textual personas into prompt-based TTS systems. One concern is that the persona rewriting process can reinforce social stereotypes and cultural biases. We emphasize the need for bias detection and mitigation strategies to support fair and inclusive voice style representations. Another concern is the potential misuse of this technology. Techniques such as synthetic voice spoofing, deepfake generation, or impersonation fraud may become more feasible when combined with persona-based LLM and speech interfaces. Although such risks exist, our research would also help mitigate them by improving the identifiability and transparency of synthesized speech.

References

Xuechunzi Bai, Angelina Wang, Ilia Sucholutsky, and Thomas L. Griffiths. 2024. Measuring implicit bias in explicitly unbiased large language models. In *NeurIPS 2024 Workshop on Behavioral Machine Learning*.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, and 1 others. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Thomas Bott, Florian Lux, and Ngoc Thang Vu. 2024. Controlling emotion in text-to-speech with natural language prompts. *arXiv preprint arXiv:2406.06406*.

Hyungjoo Chae, Yongho Song, Kai Ong, Taeyoon Kwon, Minjin Kim, Youngjae Yu, Dongha Lee, Dongyeop Kang, and Jinyoung Yeo. 2023. Dialogue chain-of-thought distillation for commonsense-aware conversational agents. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5606–5632.

Tao Ge, Xin Chan, Xiaoyang Wang, Dian Yu, Haitao Mi, and Dong Yu. 2024. Scaling synthetic data creation with 1,000,000,000 personas. *arXiv preprint arXiv:2406.20094*.

Wenhao Guan, Yishuang Li, Tao Li, Hukai Huang, Feng Wang, Jiayan Lin, Lingyan Huang, Lin Li, and Qingyang Hong. 2024. Mm-tts: Multi-modal prompt based style transfer for expressive text-to-speech synthesis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18117–18125.

Zhifang Guo, Yichong Leng, Yihan Wu, Sheng Zhao, and Xu Tan. 2023. Prompttts: Controllable text-to-speech with text descriptions. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Shashank Gupta, Vaishnavi Shrivastava, Ameet Deshpande, Ashwin Kalyan, Peter Clark, Ashish Sabharwal, and Tushar Khot. 2024. Bias Runs Deep: Implicit reasoning biases in persona-assigned LLMs. In *The Twelfth International Conference on Learning Representations*.

Keith Ito and Linda Johnson. 2017. The lj speech dataset. https://keithito.com/LJ-Speech-Dataset/.

Ke Ji, Yixin Lian, Linxu Li, Jingsheng Gao, Weiyuan Li, and Bin Dai. 2025. Enhancing persona consistency for llms' role-playing using persona-aware contrastive learning. *arXiv preprint arXiv:2503.17662*.

Ye Jia, Yu Zhang, Ron Weiss, Quan Wang, Jonathan Shen, Fei Ren, Patrick Nguyen, Ruoming Pang, Ignacio Lopez Moreno, Yonghui Wu, and 1 others. 2018. Transfer learning from speaker verification to multispeaker text-to-speech synthesis. *Advances in neural information processing systems*, 31.

Katharina Kühne, Erika Herbold, Oliver Bendel, Yuefang Zhou, and Martin H Fischer. 2024. "ick bin een berlina": dialect proficiency impacts a robot's trustworthiness and competence evaluation. *Frontiers in Robotics and AI*, 10:1241519.

Deuksin Kwon, Sunwoo Lee, Ki Hyun Kim, Seojin Lee, Taeyoon Kim, and Eric Davis. 2023. What, when, and how to ground: designing user persona-aware conversational agents for engaging dialogue. *arXiv preprint arXiv:2306.03361*.

Yoach Lacombe, Vaibhav Srivastav, and Sanchit Gandhi. 2024. Parler-tts. https://github.com/huggingface/parler-tts.

Shuhua Li, Qirong Mao, and Jiatong Shi. 2024. Pl-tts: A generalizable prompt-based diffusion tts augmented by large language model. In *Proc. Interspeech 2024*, pages 4888–4892.

Yunpeng Li, Yue Hu, Yajing Sun, Luxi Xing, Ping Guo, Yuqiang Xie, and Wei Peng. 2023. Learning to know myself: A coarse-to-fine persona-aware training framework for personalized dialogue generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 13157–13165.

Jungwoo Lim, Myugnhoon Kang, Yuna Hur, Seung Won Jeong, Jinsung Kim, Yoonna Jang, Dongyub Lee, Hyesung Ji, DongHoon Shin, Seungryong Kim, and 1 others. 2022. You truly understand what i need: Intellectual and friendly dialog agents grounding persona and knowledge. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1053–1066.

Andy Liu, Mona Diab, and Daniel Fried. 2024a. Evaluating large language model biases in persona-steered generation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9832–9850.

Jiongnan Liu, Yutao Zhu, Shuting Wang, Xiaochi Wei, Erxue Min, Yu Lu, Shuaiqiang Wang, Dawei Yin, and Zhicheng Dou. 2024b. Llms+ persona-plug= personalized llms. *arXiv preprint arXiv:2409.11901*.

Dan Lyth and Simon King. 2024. Natural language guidance of high-fidelity text-to-speech with synthetic annotations. *arXiv preprint arXiv:2402.01912*.

Cameron W Piercy, Gretchen Montgomery-Vestecka, and Sun Kyong Lee. 2025. Gender and accent stereotypes in communication with an intelligent virtual assistant. *International Journal of Human-Computer Studies*, 195:103407.

Shaina Raza, Ananya Raval, and Veronica Chatrath. 2024. MBIAS: Mitigating bias in large language models while retaining context. In *Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 97–111.

Takaaki Saeki, Detai Xin, Wataru Nakata, Tomoki Koriyama, Shinnosuke Takamichi, and Hiroshi Saruwatari. 2022. Utmos: Utokyo-sarulab system for voicemos challenge 2022. In *Proc. Interspeech 2022*, pages 4521–4525.

Ryan Shea and Zhou Yu. 2023. Building persona consistent dialogue agents with offline reinforcement learning. *arXiv preprint arXiv:2310.10735*.

Reo Shimizu, Ryuichi Yamamoto, Masaya Kawamura, Yuma Shirahata, Hironori Doi, Tatsuya Komatsu, and Kentaro Tachibana. 2024. Prompttts++: Controlling speaker identity in prompt-based text-to-speech using natural language descriptions. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 12672–12676. IEEE.

Eric Michael Smith, Melissa Hall, Melanie Kambadur, Eleonora Presani, and Adina Williams. 2022. "I'm sorry to hear that": Finding new biases in language models with a holistic descriptor dataset. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9180–9211.

Andros Tjandra, Yi-Chiao Wu, Baishan Guo, John Hoffman, Brian Ellis, Apoorv Vyas, Bowen Shi, Sanyuan Chen, Matt Le, Nick Zacharov, and 1 others. 2025. Meta audiobox aesthetics: Unified automatic quality assessment for speech, music, and sound. *arXiv preprint arXiv:2502.05139*.

Yu-Min Tseng, Yu-Chao Huang, Teng-Yun Hsiao, Wei-Lin Chen, Yu Huang, Chao-Wei andM̃eng, and Yun-Nung Chen. 2024. Two tales of persona in LLMs: A survey of role-playing and personalization. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16612–16631.

Chengyi Wang, Sanyuan Chen, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, and 1 others. 2023a. Neural codec language models are zero-shot text to speech synthesizers. *arXiv preprint arXiv:2301.02111*.

Hongru Wang, Rui Wang, Fei Mi, Yang Deng, Zezhong Wang, Bin Liang, Ruifeng Xu, and Kam-Fai Wong. 2023b. Cue-CoT: Chain-of-thought prompting for responding to in-depth dialogue questions with LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12047–12064.

Dongchao Yang, Songxiang Liu, Rongjie Huang, Chao Weng, and Helen Meng. 2024. Instructtts: Modelling expressive tts in discrete latent space with natural language style prompt. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213.

Xuehao Zhou, Mingyang Zhang, Yi Zhou, Zhizheng Wu, and Haizhou Li. 2024a. Accented text-to-speech synthesis with limited data. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:1699–1711.

Yixuan Zhou, Xiaoyu Qin, Zeyu Jin, Shuoyi Zhou, Shun Lei, Songtao Zhou, Zhiyong Wu, and Jia Jia. 2024b. Voxinstruct: Expressive human instruction-to-speech generation with unified multilingual codec language modelling. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 554–563.

A Biases After Rewriting

These examples illustrate potential biases in the resulting speech styles, as discussed in Section 5.1. Specifically, they highlight how gender, accent, and tonal preferences are disproportionately assigned even in the absence of such cues in the input text. In other words, the examples below demonstrate instances where the persona does not specify attributes such as gender or region, yet the LLM inferred these details during persona rewriting (top: before rewriting, bottom: after rewriting).

> A researcher or geographer who is interested in the geology and ecology of the Monterey Bay area, particularly the study of sedimentary rock formations and the marine ecosystem. They are likely to have a strong background in GIS and geospatial analysis, and may use the Monterey Bay DEM layer to perform site-specific analysis or to support larger-scale research projects. They may also have a particular interest in using the DEM layer in conjunction with other datasets, such as bathymetry or imagery, to create detailed maps or models of the bay's geology and ecology.
>
> A clean, **high-pitched female** voice with a neutral **American** accent, speaking **slowly** in a **warm and supportive** tone and very clear audio.

> A nurse who is interested in using virtual simulation and virtual reality in nursing education, and is looking for ways to transform traditional clinical or didactic teaching and learning. They are a seasoned nurse educator with expertise in maternal-child health, public health nursing, and medical-surgical nursing, and have worked as a National Teacher Trainer and the West Coast Territory Nurse Consultant for Institutional Partnerships. They are passionate about building diagnostic competency and clinical judgment skills through virtual simulation and have volunteered as a nurse at RotoCare Free Medical Clinic and is on the board of The Send It Foundation, providing outdoor experiences to young adults battling.
>
> A **female** voice with a **medium pitch**, **American** accent, **warm and supportive** tone, speaking **slowly** and very clear audio.

> A financial expert who is interested in Canadian currency, its history, and its role in global finance. They are knowledgeable about the Canadian dollar's value, the country's economic systems, and the stability of the Canadian legal and political systems. They are likely to be involved in financial analysis, currency trading, or investment management.
>
> A **male** voice with a **low pitch**, **Canadian** accent, **cognitive and analytical** tone, speaking at a **moderate** speed and very clear audio.

> A programmer who is interested in the history and evolution of programming languages, and is looking for insights and perspectives from the creators and developers of these languages. They are likely to be a researcher, teacher, or student who is involved in the use or development of programming languages today, and is interested in learning about the motivations, processes, and decisions that led to the creation of different programming languages. They may also be interested in learning about the lesser-known languages that made significant contributions to programming language evolution.
>
> A **male** voice with a **medium pitch** and **American** accent, using a cognitive and analytical tone, delivered at a **moderate** speed with very clear audio.

> A veterinarian interested in the impact of sepsis and the role of natural killer (NK) cells in the immune system. This person is particularly interested in the study of lidocaine administration in dogs with sepsis and the potential effects on NK cell populations and survival rates. They are interested in participating in canine health research by providing samples or enrolling in clinical trials.
>
> A **female** voice with a **high pitch**, neutral **American** accent, **warm and supportive** tone, speaking **fast** and very clear audio.

B Prompt Templates for LLM

To support the reproducibility of our experiments and to clarify the structure of input prompts used in style generation, we provide the templates below.

These templates guided both close-ended and open-ended prompt formats and are critical in standardizing the voice style outputs analyzed in Section 4. Presenting them here enables transparency in our methodology and helps contextualize the examples of stylistic bias provided in the appendix. We use the ChatGPT API [1], where the system prompt remained consistent across both prompt types, while the user instructions varied depending on whether a closed-ended or open-ended prompt was used. The detailed templates are presented below.

System prompt

You are an assistant designed to convert persona descriptions into expressive voice style prompts for text-to-speech synthesis. Your goal is to extract and synthesize vocal characteristics—such as tone, speed, pitch, and accent—based on the given persona.

Instructions for closed-ended prompting

Given a {Persona Description}, select the most appropriate vocal attributes from the predefined list below. Then, compose a single-sentence voice style prompt using the selected attributes.
<Constraints>
- The output must be one sentence.
- No explanations should be included.
- The sentence must end with: “and very clear audio.”
<Vocal Attributes>
- Gender: male, female
- Accent: American, British, Indian, Middle Eastern, etc.
- Speaking rate: fast, slow, modern
- Pitch: low-pitch, moderate-pitch, high-pitch
- Tone: analytical, warm, expressive, calm, friendly, etc.

Instructions for open-ended prompting

Given a {Persona Description}, generate a brief 1–2 sentence voice style prompt based solely on inferred vocal characteristics.
<Constraints>
- Describe voice characteristics only.
- End the sentence with: "and very clear audio."
<Examples>
- She is a Canadian, and speaks with a gentle, empathetic tone, demonstrating great speech quality with a calm and measured pace, and very clear audio.
- He speaks with a calm and knowledgeable tone, conveying his passion for wildlife conservation with great speech quality. His American accent adds to his relatable demeanor, and very clear audio.

C Evaluation Metric Details

UTMOS (Saeki et al., 2022) is an automatic speech quality evaluation system that replaces the conventional Mean Opinion Score (MOS) metric. It predicts the perceptual quality of speech samples, traditionally derived from human listeners, rating on a 5-point Likert scale. WER (Word Error Rate) is calculated by running automatic speech recognition (ASR) on the synthesized speech and comparing the transcribed output with the original text. This metric provides an objective measure of intelligibility and pronunciation clarity by quantifying the proportion of errors in the recognized text. PQ and CE are metrics adapted from Audiobox Aesthetics (Tjandra et al., 2025). These scores automatically assess speech quality in a manner aligned with human aesthetic judgment. In particular, Production Quality (PQ) evaluates the technical aspects of the audio, while Content Enjoyment (CE) captures the perceived expressiveness and artistic quality of the spoken content as experienced by the listener.

D Data and Experiment Details

D.1 Parler-TTS

We employ `Parler-TTS Mini v0.1`[2]model, an open-source, lightweight TTS model. The model was trained on approximately 10,500 hours of speech data. The model consists of approximately 647M parameters and is based on a Transformer architecture optimized for high-quality speech synthesis. This model is publicly available under the Apache 2.0 license.

D.2 Persona-1M

We utilize Persona-1M dataset [3], which was developed to support research on persona-driven data generation. This dataset comprises a diverse set of synthetic samples, including 50,000 math problems, 50,000 logical reasoning problems, 50,000 instructional prompts, 10,000 knowledge-rich texts, 10,000 game NPCs, and 5,000 functional tool descriptions. Furthermore, the dataset includes a subset of the Persona Hub [4], containing 200,000 sample personas and 370 million refined elite personas. This data is publicly available under the CC-by-NC-SA-4.0 license.

[1] https://openai.com/index/openai-api/

[2] https://github.com/huggingface/parler-tts

[3] https://github.com/tencent-ailab/persona-hub

[4] https://huggingface.co/datasets/proj-persona/PersonaHub

D.3 Hardware Environment

All experiments were conducted using three NVIDIA Tesla V100 GPUs. This setup provided sufficient computational resources for both inference and fine-tuning tasks.

D.4 Inference Setting

We used the official code of `Parler-TTS` [5] model. The hyperparameters (temperature, top-k, top-p, and repetition-penalty) were left at their default values. The model generated audio at a sampling rate of 24,000 Hz, and each input was processed individually with a batch size of 1. The output waveforms were saved in float32 WAV format without any additional post-processing.

E More Related Work

E.1 Prompt-based TTS

Prompt-based TTS systems have increasingly adopted natural language prompts to control vocal attributes such as pitch, tone, and emotion. This shift allows for more intuitive and flexible voice generation compared to traditional methods. The studies below illustrate key advancements in prompt design and multimodal control. InstructTTS (Yang et al., 2024) extends this by leveraging descriptive prompts (e.g., "sad and low voice") to produce emotionally expressive speech via a diffusion-based synthesis model, using a VQ-VAE-based latent space for nuanced prosody and emotion control. VoxInstruct (Zhou et al., 2024b) proposes a unified prompting interface that integrates both content and style in a single sentence, enabling multilingual and stylistic control through a multilingual codec language model. Bott et al. (Bott et al., 2024) introduce emotional embeddings derived from natural language emotion descriptions, allowing nuanced adjustments in emotional intensity and affective style during synthesis. PL-TTS (Li et al., 2024) combines a LLaMA-2-based language model with a diffusion-based TTS system to interpret complex prompts and control diverse speech attributes such as speed, pitch, and volume. These advancements represent a shift toward more intuitive and general-purpose TTS interfaces. Natural language prompts are evolving into versatile control mechanisms for speaker identity, prosody, and emotion, enabling increasingly interactive and creative speech generation. However, prompt-based control from loosely structured or persona-style descriptions remains an underexplored but promising research direction.

E.2 LLM Personalization Using Persona

Persona has become a key strategy for improving consistency and engagement in LLM-based dialogue systems. While early work relied on static templates, recent studies explore more dynamic methods to integrate persona traits into model behavior. These approaches aim to enhance coherence, factuality, and user alignment. The following works highlight major techniques developed to improve persona consistency and control.

Shea and Yu proposed a refinement method using an offline reinforcement learning framework, where a pretrained dialogue model is optimized through a reward mechanism based on persona consistency. This reduces contradictions in persona expression and enhances overall dialogue coherence. Similarly, Lim et al. introduced a retrieval-augmented generation (RAG) strategy that integrates user persona information with external knowledge bases. This nuanced approach mitigates factual errors and hallucinations, enabling more accurate, context-rich, and engaging interactions.

Research such as DOCTOR (Chae et al., 2023) and Cue-CoT (Wang et al., 2023b) demonstrates that directly incorporating persona attributes, such as profession and personality, into the model can ensure consistent speaking style without additional fine-tuning. Ji et al. proposed a novel training framework, Persona-Aware Contrastive Learning (PCL), designed to enhance persona consistency in role-playing scenarios. The method outperforms previous models in both automatic and human evaluations. Similarly, Li et al. introduced a two-stage Coarse-to-Fine training framework aimed at improving persona consistency for personalized response generation, which proves effective in enhancing dialogue quality. Kwon et al. developed a fine-grained control strategy involving weighted dataset mixing and negative persona data augmentation. This method allows for selective modulation of persona expression timing and presence, contributing to more natural and fluid dialogue generation. More recently, Persona-Hub (Ge et al., 2024) compiled a large-scale archive of nearly one million persona entries, laying the foundation for optimized character data across diverse application scenarios.

[5] https://github.com/huggingface/parler-tts

E.3 Speaker Identity in TTS Systems

Prior research suggests that voices with clearly defined identities tend to elicit greater user engagement and trust. For instance, Zhou et al. showed that fine-tuning TTS models on various English accents (e.g., British, Australian) led to improvements in prosodic accuracy and naturalness, with listener evaluations indicating significantly higher perceived quality. Jia et al. provided empirical evidence that incorporating demographic attributes such as gender and nationality into speaker embeddings can directly enhance the expressiveness and fidelity of modern TTS systems. Piercy et al. conducted user interaction experiments with virtual assistants varying in gender (male/female) and accent (American/Indian English), and found that personas with clearly communicated identity traits were associated with higher levels of user trust and engagement. Kühne et al. reported that users familiar with a regional dialect evaluated a robot speaking in that dialect as more trustworthy, demonstrating that shared cultural or linguistic identity between user and persona can significantly enhance perceived credibility and immersion.

These findings indicate that enriching persona texts with explicit gender and country attributes is not merely a stylistic modification but a meaningful enhancement that improves both the naturalness of synthesized speech and the overall user experience, particularly in terms of trustworthiness and immersion.