VoiceForge: A Text-Driven Character Voice Generation System for Narrative Content Creation

Yunyi Ni
Nanyang Technological University
Singapore, Singapore
YUNYI001@e.ntu.edu.sg

Qirui Sun
Tsinghua University
Beijing, China
sqr22@mails.tsinghua.edu.cn

Yuan Chai
The Hong Kong University of Science and Technology (Guangzhou)
Guangzhou, Guangdong Province, China
ryan.yuanchai@outlook.com

Haipeng Mi*
Tsinghua University
Beijing, China
mhp@tsinghua.edu.cn

Abstract

In narrative content creation, generating distinctive and appropriate character voices remains challenging. While professional voice acting achieves ideal results, its high costs and complex production process limit widespread adoption. We present VoiceForge, an interactive system that enables users to intuitively generate character voices through natural language descriptions. The system converts textual character descriptions into unique voice configurations and intelligently identifies dialogue in scripts to automatically assign corresponding voices to different characters. Unlike existing platforms with preset voice libraries, VoiceForge offers flexible voice customization through text-driven generation and audio mixing capabilities. Our user evaluation comparing VoiceForge with professional voice acting and existing AI solutions shows that our system-generated voices approach professional quality in terms of character matching and speech fluency, significantly outperforming existing AI platforms. This research demonstrates an effective method for bridging the gap between character description and voice generation in narrative content creation.

CCS Concepts

• **Human-centered computing** → **Interactive systems and tools**.

Keywords

Text-to-Speech, Character Voice Generation, Voice Customization, Narrative Content Creation

ACM Reference Format:
Yunyi Ni, Qirui Sun, Yuan Chai, and Haipeng Mi. 2025. VoiceForge: A Text-Driven Character Voice Generation System for Narrative Content Creation. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '25), April 26–May 01, 2025, Yokohama, Japan.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3706599.3720140

*Corresponding Author.

CHI EA '25, Yokohama, Japan

ACM ISBN 979-8-4007-1395-8/25/04
https://doi.org/10.1145/3706599.3720140

1 INTRODUCTION

The advent of the digital media era has fundamentally transformed narrative content production and consumption patterns across multiple domains, including audiobooks, gaming, podcasting, and educational content. The global audiobook market alone reached USD 6.83 billion in 2023[6] and demonstrates consistent growth of 20% year-over-year[12], while the broader narrative content industry shows similar expansion trends. Previous studies have shown that voice influences listeners' impressions of speakers[5] and personalized voices significantly impact audience engagement[10], highlighting the critical role of voice acting in modern digital storytelling.

Currently, voice production for narrative content primarily employs three approaches, each with significant limitations. Professional voice acting, while capable of delivering nuanced performances that bring characters to life, faces scalability challenges due to high costs and complex production processes. Single AI voice narration offers convenience but compromises narrative expression through monotonous delivery. Multi-voice AI dubbing systems like iFLYTEK Smart Creation[1] partially address the monotony issue but burden creators with manual voice selection and script annotation tasks.

The technical landscape of voice synthesis has evolved rapidly, with recent advances in Text-to-Speech (TTS) technology demonstrating promising capabilities. However, significant gaps remain between technical possibilities and creative needs. Voice cloning technologies, while impressive in replication accuracy, are fundamentally limited to reproducing existing voices rather than creating new ones that align with creators' artistic visions. Additionally, maintaining voice consistency across extended narratives and effectively translating creative intent into technical parameters pose persistent challenges.

Literary works typically contain rich character descriptions that should theoretically guide voice creation, but effectively transforming these natural language descriptions into usable voice parameters has remained elusive. While early research highlighted difficulties in establishing recognizable voice terminology[11], recent

[1] https://peiyin.xunfei.cn/

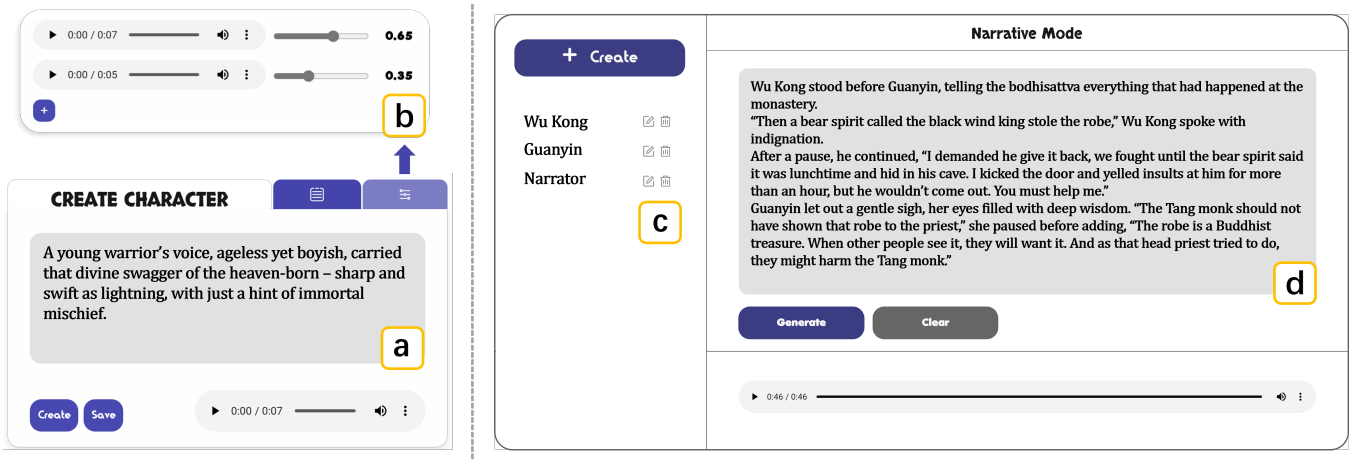

Figure 1: User interface overview: (a) Character creation area with text input for voice description; (b) Audio mixing panel for fine-tuning voice characteristics; (c) Character list displaying created voices; (d) Script input area with automated dialogue detection.

breakthroughs in Large Language Models (LLMs) have opened new possibilities for bridging this semantic gap.

To address these challenges, we present VoiceForge, an innovative interactive platform that automatically converts textual character descriptions into corresponding voice configurations and enables automated narrative content production. Our main contributions include:

(1) A novel text-to-voice mapping framework that achieves automatic conversion from character descriptions to stable voice characteristics;
(2) An intelligent script analysis and voice assignment system based on large language models, significantly improving the efficiency of multi-character narrative production;
(3) A flexible hybrid voice customization interface supporting both text description and audio feature mixing approaches to meet diverse user customization needs;
(4) Comprehensive user studies validating the effectiveness and practicality of text-driven voice generation in creative content production.

2 RELATED WORK

2.1 Voice Customization Approaches

Voice customization has traditionally been overlooked due to high costs and implementation complexities, with existing solutions primarily focused on voice changers[2, 8]. Commercial applications like Voicemod[2] offer basic real-time voice transformation through preset filters, but these systems provide limited options and often produce exaggerated voices unsuitable for narrative content.

More sophisticated approaches incorporate parametric control over voice characteristics. AVOCUS[1] combines voice search with acoustic parameter adjustment, while specialized systems like Sakai et al.[9] developed game character voice recommendations using acoustic feature estimation. However, these solutions struggle to achieve rapid, flexible voice generation that can effectively utilize textual information for voice customization.

2.2 Advances in Speech Synthesis

The evolution of TTS technology has seen several paradigm shifts, from early sequence-to-sequence models[13] to attention-based architectures like Tacotron[15], significantly improving synthesized speech naturalness. Recent breakthroughs include HuBERT[7]'s self-supervised speech representation learning and its impact on voice quality consistency.

Voice cloning technology has made particular strides, with systems like VALL-E and YourTTS[3, 14] demonstrating unprecedented capabilities in zero-shot voice synthesis. These systems can effectively capture and reproduce voice characteristics from brief audio samples, though they face challenges in maintaining stability across longer narratives and handling emotional variations. Recent research has begun exploring voice characteristic control through text descriptions, with CosyVoice[4] as an example, where its CosyVoice-300M-Instruct model generates voice timbre and emotions based on textual input. However, practice shows this model exhibits noticeable timbre shifts during extended narrative synthesis and lacks fine-grained control over emotional variations while maintaining consistent voice characteristics. While another model in its family, CosyVoice-300M, does not support text-driven voice control, it demonstrates superior stability and remarkable naturalness in speech synthesis, making it an ideal foundation for our research implementation.

3 USER INTERFACE AND INTERACTION

The VoiceForge workflow begins with character creation. In the text input area (Figure 1(a)), users describe their desired voice characteristics using natural language. The system generates a speech sample with test text, allowing users to iteratively refine their descriptions until achieving the desired voice. For cases where text description alone doesn't achieve the intended voice after multiple

[2]https://www.voicemod.net/

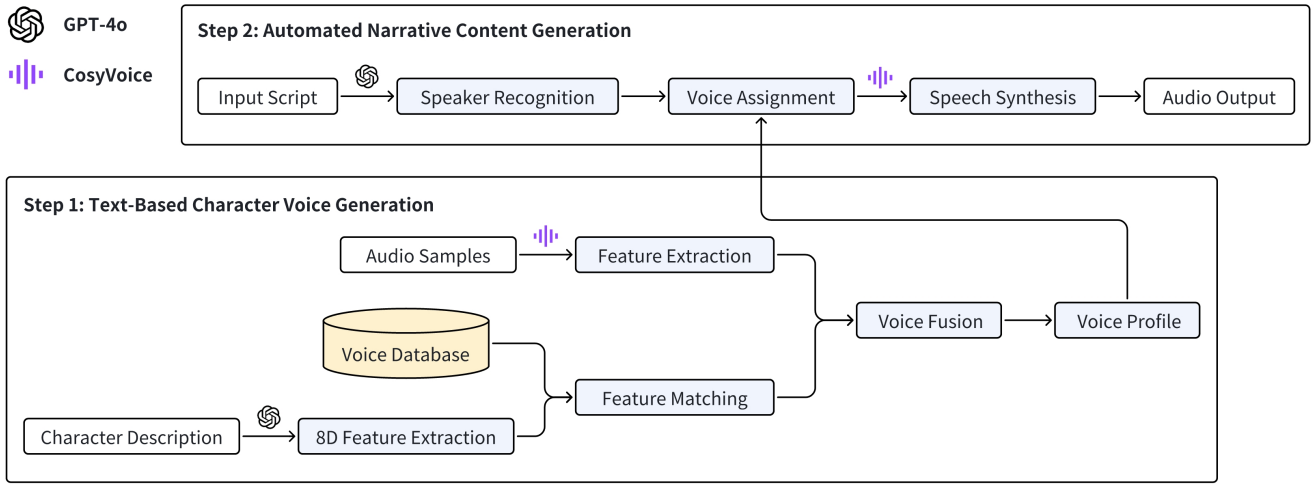

Figure 2: System architecture diagram showing the two-stage process of VoiceForge.

attempts, the system offers an auxiliary audio mixing control panel (Figure 1(b)) that allows users to blend 1-3 reference audio samples with adjustable weights for fine-tuned customization.

Created characters appear in the character list area (Figure 1(c)), where users can rename or delete them. To ensure proper voice matching, character names in the list must match exactly with character names in the script. After character creation, users can input narrative text in the script input area (Figure 1(d)). The system automatically identifies dialogue content and matches it with corresponding character voices to generate the final audio output.

4 SYSTEM OVERVIEW

As illustrated in Figure 2, the VoiceForge system achieves automated generation from character text descriptions to complete narrative audio content through 2 main steps. The first step focuses on accurately converting character descriptions into stable voice characteristics, while the second step implements automated production of multi-character narrative content. This two-stage design allows us to first establish robust voice profiles and then apply them effectively in narrative contexts.

4.1 Text-Driven Character Voice Generation

8D Feature Extraction from Character Description. To achieve precise mapping from text descriptions to voice characteristics, we leverage GPT-4o[3]'s natural language understanding capabilities to parse character descriptions into an 8-dimensional acoustic feature space $F = \{f_1, ..., f_8\}$. This feature space includes gender (treated as a continuous parameter to better represent the natural spectrum of voice characteristics), age, pitch, warmth, clarity, power, thickness, and smoothness, extending the impression value set proposed by Sakai et al. Each dimension is normalized to the [0, 1] interval to ensure feature scale consistency. This standardized feature representation lays the foundation for subsequent similarity calculations and voice matching.

[3]https://openai.com/index/hello-gpt-4o/

Voice Embedding Database Construction. At the core of our system is a carefully curated professional voice database containing 40 meticulously selected voice samples drawn from publicly accessible audio and video sources. These samples underwent rigorous preprocessing, including noise reduction, audio normalization, and acoustic artifact removal to ensure high-quality input for voice embedding. The samples cover different genders, age groups, and style characteristics, ensuring diversity in the voice feature space. For each voice sample, we maintain both a 192-dimensional speaker embedding vector generated through CosyVoice-300M's voice encoder and a corresponding manually annotated 8-dimensional feature vector following the same acoustic feature space. These vectors collectively form our reference voice library, providing rich foundation materials for subsequent voice matching and generation.

Feature Matching. To ensure optimal mapping between textual descriptions and voice characteristics, we first establish feature importance weights through LASSO regression, learning the optimal feature weights $W = \{w_1, ..., w_8\}$. For new character descriptions, the system calculates weighted similarity with each sample in the database using the following equation:

$$sim(F, D_i) = \sum_{j=1}^{8} w_j (f_j - d_{ij})^2 \quad (1)$$

where w_j represents the importance weight of the j-th acoustic feature dimension, f_j is the target feature value, and d_{ij} is the corresponding feature value of the i-th database sample. This weighted approach ensures that more perceptually significant features have greater influence on the matching process. Based on these similarity scores, the system identifies the 3 most similar voice profiles from the database for subsequent fusion.

Audio Samples Processing. For users choosing direct audio input, the system processes up to 3 reference audio samples through CosyVoice-300M's voice encoder, converting each into a speaker embedding vector. This alternative path for voice customization is particularly valuable when text descriptions alone cannot fully

capture desired voice characteristics, or when users have specific reference voices in mind. The high-dimensional embedding space allows for precise capture of voice characteristics while maintaining compatibility with our fusion framework.

Voice Fusion and Profile Generation. The final voice profile is generated through weighted combination of either the 3 most similar database voices or user-provided audio samples. For text-driven input, the weighting coefficients are determined based on the vector distances between samples in the acoustic feature space, with closer matches having stronger influence on the final voice characteristics.

4.2 Automated Narrative Content Generation

The system's second major step implements automated narrative content generation through script analysis and voice synthesis. This process utilizes GPT-4o's language understanding capabilities to perform deep parsing of the input script, accurately identifying speakers and their corresponding dialogue content. The natural language processing component is specifically tuned to handle various script formats and dialogue patterns commonly found in narrative content. Based on the recognition results, the system passes each speaker's voice embedding vector along with their corresponding dialogue text as parameters to CosyVoice-300M for speech synthesis, generating voice segments with personalized characteristics. Finally, the system concatenates all audio segments according to the script's chronological order to create complete narrative audio content. This end-to-end pipeline significantly reduces the manual effort required in traditional multi-voice production processes while maintaining consistency in character voices throughout the narrative.

5 USER STUDY

We conducted a user study with 12 participants (6 males and 6 females, aged 20-28) with experience in narrative content production, including professionals and enthusiasts from audiobooks, gaming, podcasting, and educational content creation. All participants were briefed about the study purpose and procedures, and provided consent to participate in the evaluation process. To ensure objective evaluation, we randomized and blind-tested audio samples from different sources (including professional voice acting, our tool's generation, single AI voice, and multiple AI voices).

The experiment utilized diverse text materials to represent various narrative genres and styles, including the Chinese web novel "Lord of Mysteries," the English version of the Chinese classic novel "Journey to the West," and the English classic fairy tale "The Little Mermaid." These selections were chosen to test the system's versatility across different narrative contexts. The professional voice-acted references were professionally produced versions representative of high-quality narrative content production.

For audio generated by our tool, character voices were created based on the original texts' character descriptions. For comparison, we generated additional samples using single AI voice and multiple AI voice approaches from existing mainstream content production software.

We used a 7-point scale to score each group of audio samples across four dimensions: voice-character matching, speech fluency, audience immersion, and overall satisfaction, as shown in Figure 3.

The experimental results show that the audio generated by our tool performed excellently overall, second only to professional voice acting. On the 7-point scale, this effect was most prominent in terms of speech fluency, achieving an average score of 5.34 points (SD=0.45); it also received good ratings of 4.44 points (SD=0.59) in voice-character matching and 4.69 points (SD=0.35) in overall satisfaction. Particularly in the English text group, the tool-generated effects approached professional voice acting in terms of speech fluency.

Based on the user experience data analysis, the tool received relatively positive overall evaluations. It performed particularly well in terms of usability, with the interface ease of use scoring 5.25 points (SD=1.68) and quick learning scoring 5.33 points (SD=1.93). Users expressed high recognition of the tool's user-friendliness and learning curve. In terms of voice effects, vivacity performed best, scoring 5.08 points (SD=1.68).

Approximately 70% of users highlighted the tool's efficiency, clean interface, and capability for rapid personalized voice generation and multi-character audio production. Users pointed out potential areas of improvement, such as enhancing automatic character name recognition from scripts and providing more nuanced options for voice customization and emotion control.

6 LIMITATIONS AND FUTURE WORK

Although VoiceForge demonstrates promising results in character voice generation and narrative content production, there remains room for improvement in several aspects.

Regarding our user study, we acknowledge its primary limitation in participant selection. Our sample size (n=12) and narrow age range (20-28) constrain the generalizability of our findings. While participants had relevant experience in narrative content creation, this specific demographic may not adequately represent the diverse spectrum of potential users across different age groups, professional backgrounds, and experience levels. Future studies should include a larger and more diverse participant pool to validate our findings across broader user demographics.

From our observations of current user input preferences and habits, we've identified additional technical limitations. While the current system's eight-dimensional acoustic feature space effectively captures voice characteristics from a technical perspective, this representation method shows a noticeable gap with users' natural habits of describing voices. Users tend to use concrete labels such as "monster," or "robot" to describe target voice characteristics, and large language models still face challenges in accurately mapping these descriptions to acoustic feature spaces. Future work could consider establishing a hybrid mapping system that combines manual annotation with automated learning to help the system better understand and implement users' voice requirements.

Emotional expression handling represents another significant limitation of the current system. In our practice, we found that attempts to control voice emotions often led to unexpected changes in fundamental voice characteristics, which led us to prioritize voice consistency over emotional range expansion during development.

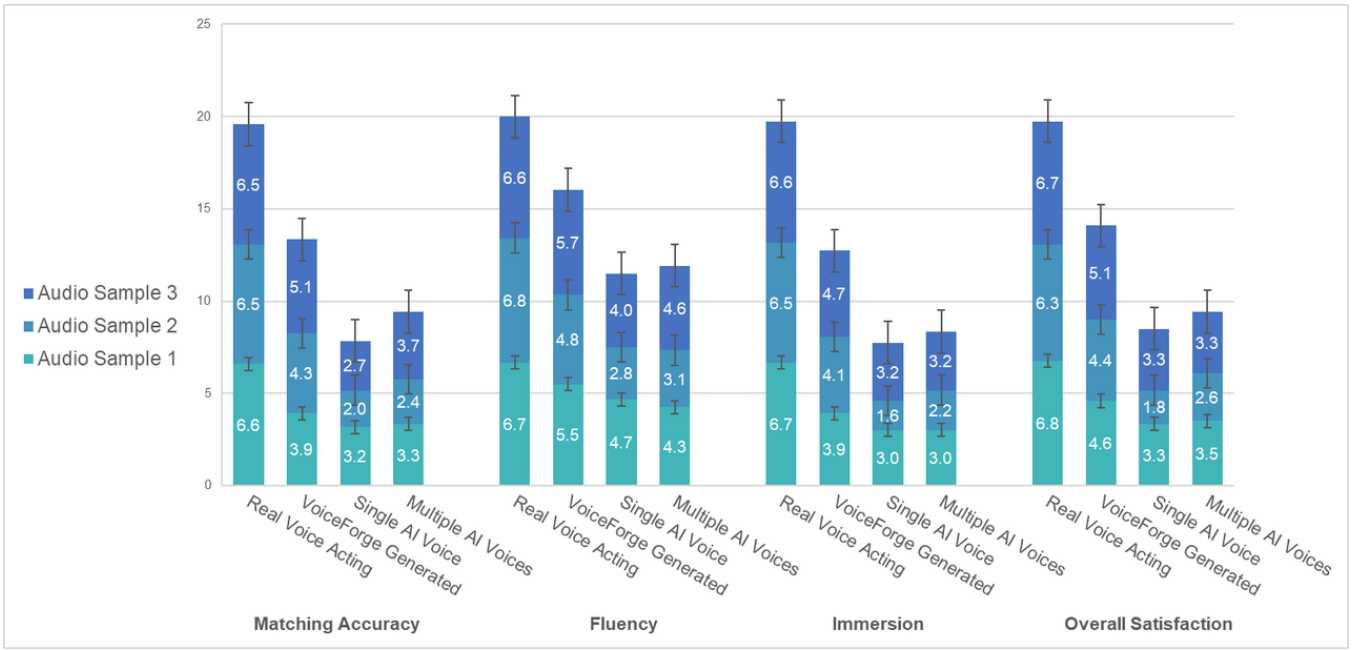

Figure 3: Comparison of different voice generation approaches across four evaluation metrics (scoring range: 1-7).

The core issue lies in our inability to find an effective method to isolate emotional components from voice vectors. One future research direction is to explore more advanced voice embedding decomposition techniques, enabling the system to achieve dynamic emotional control while maintaining stable voice characteristics, thereby enhancing the narrative expressiveness of generated content.

The extensibility of system functionality also provides broad space for future research. Dialect support represents an important development direction, as implementing specific regional accent and dialect generation can significantly enhance character authenticity in diverse narrative contexts. Meanwhile, we also anticipate developing mixed-mode voice generation capabilities, allowing the system to naturally integrate speaking and singing voices, offering richer possibilities for character performance. Additionally, exploring cross-lingual voice adaptation technology will help the system maintain consistent voice characteristics while preserving linguistic properties.

To further enhance the system's practicality, future versions could incorporate more sophisticated interactive feedback mechanisms, allowing users to iteratively optimize generated voices through intuitive control interfaces and real-time preview capabilities. This includes providing visualization tools for voice characteristic adjustment, A/B testing interfaces for voice variant comparison, and collaborative filtering recommendations based on user preferences. We believe that addressing these limitations and exploring these future directions will further bridge the gap between professional voice acting and automated voice generation, making high-quality narrative content creation more accessible to a broader range of creators.

7 CONCLUSION

This paper presents VoiceForge, a novel text-driven character voice generation system that bridges the gap between textual character descriptions and voice synthesis. Through its innovative approach combining natural language processing, acoustic feature mapping, and automated narrative content generation, VoiceForge demonstrates significant advantages in both voice quality and production efficiency. Our user studies, despite their limited sample size and demographic range, confirm that the system generates voices that closely match character descriptions while maintaining high speech fluency, approaching professional voice acting quality in several aspects. The system's intuitive interface and automated workflow significantly reduce the barriers to creating high-quality voiced narrative content. While certain limitations remain, particularly in emotional expression and dialect support, VoiceForge represents a significant step forward in making professional-quality voice content creation more accessible to a broader range of creators.

References

[1] Hyeon Jeong Byeon, Seungjin Ha, and Uran Oh. 2023. AVOCUS: A Voice Customization System for Online Personas. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI EA '23)*. Association for Computing Machinery, New York, NY, USA, Article 49, 6 pages. doi:10.1145/3544549.3585892

[2] Hyeon Jeong Byeon, Chaerin Lee, Jeemin Lee, and Uran Oh. 2022. "A Voice that Suits the Situation": Understanding the Needs and Challenges for Supporting End-User Voice Customization. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 59, 10 pages. doi:10.1145/3491102.3501856

[3] Edresson Casanova, Julian Weber, Christopher Shulby, Arnaldo Candido Junior, Eren Gölge, and Moacir Antonelli Ponti. 2023. YourTTS: Towards Zero-Shot Multi-Speaker TTS and Zero-Shot Voice Conversion for everyone. arXiv:2112.02418 [cs.SD] https://arxiv.org/abs/2112.02418

[4] Zhihao Du, Qian Chen, Shiliang Zhang, Kai Hu, Heng Lu, Yexin Yang, Hangrui Hu, Siqi Zheng, Yue Gu, Ziyang Ma, Zhifu Gao, and Zhijie Yan. 2024. CosyVoice: A Scalable Multilingual Zero-shot Text-to-speech Synthesizer based on Supervised Semantic Tokens. arXiv:2407.05407 [cs.SD] https://arxiv.org/abs/2407.05407

[5] Jerry Bryan Fuller, Tim Barnett, Kim Hester, Clint Relyea, and Len Frey. 2007. An Exploratory Examination of Voice Behavior from an Impression Management Perspective. *Journal of Managerial Issues* 19, 1 (2007), 134–151. http://www.jstor.org/stable/40601197

[6] Grand View Research. 2024. *Audiobooks Market Size, Share & Trends Analysis Report By Genre (Fiction & Non-Fiction), By Preferred Device, By Distribution Channel, By Target Audience (Kids Mode, Adult), By Region, And Segment Forecasts, 2024 - 2030.* Technical Report. Grand View Research. https://www.grandviewresearch.com/industry-analysis/audiobooks-market

[7] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.* 29 (Oct. 2021), 3451–3460. doi:10.1109/TASLP.2021.3122291

[8] Dominic Kao, Rabindra Ratan, Christos Mousas, Amogh Joshi, and Edward F. Melcer. 2022. Audio Matters Too: How Audial Avatar Customization Enhances Visual Avatar Customization. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 165, 27 pages. doi:10.1145/3491102.3501848

[9] Erika Sakai, Takayuki Itoh, and Akinori Ito. 2017. A Study on Voice Actor Recommendation for Game Characters Based on Acoustic Feature Estimation and Document Co-occurrence. *2017 Nicograph International (NicoInt)* (2017), 15–18. https://api.semanticscholar.org/CorpusID:21086827

[10] Marc Schröder. 2009. *Expressive Speech Synthesis: Past, Present, and Possible Futures.* Springer London, London, 111–126. doi:10.1007/978-1-84800-306-4_7

[11] Aatto Sonninen and Pertti Hurme. 1992. On the terminology of voice research. *Journal of Voice* 6, 2 (1992), 188–193. doi:10.1016/S0892-1997(05)80132-8

[12] Spotify. 2022. *With Audiobooks Launching in the U.S. Today, Spotify is the Home for All the Audio You Love.* https://newsroom.spotify.com/2022-09-20/with-audiobooks-launching-in-the-u-s-today-spotify-is-the-home-for-all-the-audio-you-love/

[13] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to Sequence Learning with Neural Networks. arXiv:1409.3215 [cs.CL] https://arxiv.org/abs/1409.3215

[14] Chengyi Wang, Sanyuan Chen, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, Lei He, Sheng Zhao, and Furu Wei. 2023. Neural Codec Language Models are Zero-Shot Text to Speech Synthesizers. arXiv:2301.02111 [cs.CL] https://arxiv.org/abs/2301.02111

[15] Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J. Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, Quoc Le, Yannis Agiomyrgiannakis, Rob Clark, and Rif A. Saurous. 2017. Tacotron: Towards End-to-End Speech Synthesis. arXiv:1703.10135 [cs.CL] https://arxiv.org/abs/1703.10135