university of groningen

campus fryslân

Character Identity and Emotion-Aware TTS for Otome Games

Qianqian Bian

university of groningen

campus fryslân

University of Groningen - Campus Fryslân

Character Identity and Emotion-Aware TTS for Otome Games

Master's Thesis

To fulfill the requirements for the degree of
Master of Science in Voice Technology
at University of Groningen under the supervision of
Dr. Shekhar Nayak (Voice Technology, University of Groningen)

Qianqian Bian (S6029388)

June 11, 2025

Acknowledgements

First, I would like to thank my supervisor, Dr. Shekhar Nayak, for his strong support during this project. He always gave helpful and clear advice, pointed out things I had missed, and responded quickly at every important step. Thanks to his guidance, I was able to build an interesting TTS dialogue system.

I also want to thank the friends I met during my Master's study. Many of us came from different backgrounds, and together we worked hard to learn speech technology. Even when things were difficult, we encouraged each other, kept moving forward, and tried to follow the path we truly wanted.

I also want to thank myself. I came alone to the Netherlands to study and took on the big challenge of moving from business to science. There were many sleepless nights and failed attempts, but in the end, I gained not only valuable knowledge but also rich life experience. I believe the journey never stops — we keep learning, growing, and seeing the world and ourselves in new ways. Now it's time for the next chapter. Let's move forward, and figure things out along the way.

Lastly, I thank the university for giving me access to GPU resources. In today's fast-growing world of AI, this kind of computer power made it possible for me to finish my project.

Declaration

I hereby affirm that this Master thesis was composed by myself, that the work here in is my own except where explicitly stated otherwise in the text. This work has not been submitted for any other degree or professional qualification except as specified, nor has it been published. Where other people's work has been used (from any source: printed, internet or other), this has been carefully acknowledged and referenced. During the preparation of this thesis, I used *ChatGPT, version as GPT-4o* for the following purposes: generating the visualization script for Figure 2 in Section 3.2.3; assisting with the data preprocessing script for the .txt files in Section 4.2; integrating a unified pipeline script that combines Whisper, Yi-AI, and CosyVoice2 for the proposed use case; and generating Wilcoxon test plot (Figure 3 in Section 5.1) as well as the boxplot scripts for Figures 4 and 5. All generated scripts were subsequently reviewed, verified, and used by myself to produce the final figures.

In compliance with university guidelines, I acknowledge that use of AI for code generation, experimental design, or data interpretation requires clear justification and understanding of its limitations. Throughout the process, I provided detailed and purpose-specific prompts to the AI tool and took full responsibility for validating the outputs before usage. While such tools improved development efficiency, they required significant effort in verifying correctness and contextual fit. This process not only demanded careful academic judgment but also strengthened my ability to justify methodological decisions and maintain critical oversight in tool-assisted workflows.

Name:

Qianqian Bian

Date: 2025.6.11

Abstract

This study presents a character-based speech synthesis system designed for interactive otome games. The system captures players' speech using Whisper for transcription, generates dialogue using the Yi-1.5-6B-Chat language model based on character-specific prompts, and synthesises responses with a CosyVoice2-based text-to-speech (TTS) model adapted to two distinct in-game characters.

The TTS component is fine-tuned on the Mandarin male subset of the Emotional Speech Database (ESD), which includes five speakers across five emotion styles. From this dataset, two character voices are constructed to represent distinct romantic non-player characters (NPCs). Their vocal styles are guided by both prompt-based instructions and explicit speaker identity control.

System performance is evaluated along two dimensions. First, a subjective Mean Opinion Score (MOS) for character consistency was conducted using 37 listener responses, of which 32 were retained after filtering out five surveys that showed inconsistent or extreme mismatch ratings. The final average MOS score was 3.55, suggesting moderate perceived consistency between the synthesized voice and the intended character identity. Second, speaker similarity (SS) was computed using cosine similarity between embeddings of reference and generated speech, resulting in an average score of 0.83.

These results demonstrate that combining prompt-driven dialogue generation with instruct-based vocal style control enables expressive, character-consistent speech interactions.

Contents

1 Introduction

The female-oriented game market, particularly romantic video games (RVGs) known as otome games, has witnessed unprecedented growth in recent years. Both the number of female gamers and the revenue generated by female-targeted games have surged, highlighting the significant commercial potential of this genre. In both China and the United States—the two largest gaming markets globally—women make up a substantial share of domestic player bases, with female gamers accounting for 48% of players in China and 45% in the U.S. (Statista, 2021). These figures highlight the strong presence of female gamers in countries that dominate global game consumption. In 2024, Asia contributed approximately 23% of otome games' revenue (Otome, 2025). Recent commercial successes further underscore this trend. *Love and Deep Space*, a project developed by Papergames, achieved remarkable financial results, generating approximately 80 million USD in monthly revenue in the Asian market and 9.67 million USD in the Western market as of August 2024 (Tower, 2024). This rapid rise demonstrates a growing demand among female players for emotionally immersive and interactive entertainment experiences.

This shift aligns with broader societal changes. As women gain increased visibility and agency in education and professional sectors, their emotional and expressive needs are evolving (Ganzon, 2022). Unlike past decades where gender roles were more rigidly defined, contemporary women seek interactive media experiences that offer emotional resonance, intimacy, and autonomy. Otome games provide such spaces, where players—mostly women—can explore affection, agency, and fantasy through emotionally scripted relationships with idealized virtual characters (Andlauer, 2018)(Zhang, 2024). These games typically place the player in the role of a female protagonist surrounded by romantic male leads, with narratives unfolding through branching storylines, emotional dialogue, and aesthetic worldbuilding (Sellier, 2024). However, recent research also highlights a potential psychological cost: prolonged engagement with romantic video games may reduce players' desire for real-life relationships, primarily due to the development of highly satisfying parasocial romantic relationships with in-game characters (Wu, Cai, & Mensah, 2024). This underscores the importance of designing interaction systems that foster emotional fulfillment without detaching players from real-world intimacy.

Originating in Japan with early titles such as *Angelique* (1994), otome games have since evolved into a transnational media phenomenon, particularly flourishing across East Asia. Over the past two decades, the genre has expanded beyond traditional console formats into mobile applications, anime adaptations, character merchandising, and fan-produced content—supported by a robust ecosystem of global fan communities and participatory fan labor (Ganzon, 2019) (Andlauer, 2018). Despite these developments, one notable limitation remains across most commercial otome games: their continued dependence on pre-recorded voice content. Emotional phone calls, character dialogues, and other scripted interactions are typically pre-written, voiced by professional actors, and embedded as fixed assets triggered by specific in-game events. This format, while polished in presentation, restricts the possibility for spontaneous or adaptive communication. Even in widely popular titles such as *Love and Deep Space*, where immersive design is prioritized, character interaction remains one-directional and non-dynamic, failing to replicate the natural flow of emotional conversation. This structural constraint highlights a gap between narrative immersion and interactional flexibility—an area that has received limited attention in existing academic literature.

To overcome the limitation of scripted audio playback and enable emotionally responsive interactions, I propose a dialogue system specifically designed for otome games. This system supports natural spoken communication between players and in-game characters via a three-stage pipeline: automatic speech recognition (ASR) transcribes the player's utterance, a compact large language model (LLM) interprets and generates the reply, and a text-to-speech (TTS) model synthesizes the response with the target character's voice and emotion.

This project specifically targets Mandarin Chinese, and Whisper (Radford et al., 2022) offers strong support for this language, having been trained on a large multilingual dataset that includes Chinese speech. Its robustness to noise and high accuracy on conversational input make it well-suited for player interactions. Moreover, lightweight variants such as Whisper-base can run in real time on standard GPUs, enabling both correct and low-latency transcription during conversations. As the first stage in the interaction pipeline, Whisper ensures accurate and efficient conversion of spoken Mandarin into text, allowing less time consuming.

Although recent advances in compact large language models (LLMs) have significantly reduced the deployment cost of interactive AI agents, not all models are equally suited for character consistant applications such as otome games. For example, Qwen, a bilingual Chinese-English model developed by Alibaba, demonstrates strong performance across general-purpose tasks and supports tool-use and planning capabilities (Bai et al., 2023) However, its conversational output in Chinese tends to be overly general, lacking the stylistic richness and emotional nuance needed for romance-driven dialogue. Additionally, Phi-2, a 2.7B model released by Microsoft Research, is good at common-sense reasoning and educational benchmarks with a surprisingly small parameter count (Eldan, Lee, & Nguyen, 2023). Yet, its monolingual English orientation limits its suitability for the current project, which targets the Chinese otome game market—a segment that contributes disproportionately high revenue globally. While English-language users are certainly relevant, Chinese-speaking players have been the primary drivers of commercial success in this filed, as evidenced by the dominance of titles like *Love and Deep Space* in the Asian market. Consequently, a language model lacking Chinese capabilities cannot meet the linguistic and emotional requirements essential for localized, interactive romantic dialogue in this context.

In contrast, Yi-1.5-6B-Chat (AI et al., 2025), presents a better solution. It supports both Chinese and English efficiently, and its structured prompt control is essential for shaping character consistancy in interactive dialogue. The Yi model was trained on a 3.1T bilingual dataset with an emphasis on quality over scale, and its instruction-tuned variants achieve competitive human preference ratings compared to GPT-3.5. Its lightweight architecture can be quantized to 4-bit precision and run on consumer-grade devices without noticeable performance loss, making it a good choice for character consistant and emotion-aware interactions in otome games.

In evaluating the text-to-speech system (TTS) for emotion-aware speech generation, several commonly used models were reviewed. FastSpeech2 (Ren et al., 2022) offers efficient, non-autoregressive synthesis with support for prosodic variation through pitch, duration, and even global style tokens (GST). However, its expressive range is primarily tuned for tasks such as audiobook narration and news reading. In practice, it lacks the stylistic flexibility and emotion-driven dialogue structure required to support romantic, persona-based conversations typical of otome games.

OpenVoice (Qin, Zhao, Yu, & Sun, 2024) is optimized for rapid speaker adaptation and timbre cloning. While it allows for controllable variation in speed and pitch, it does not enable text-conditioned emotional synthesis. This limitation prevents it from capturing dynamic, context-sensitive emotional shifts driven by dialogue content, which are core to emotionally responsive non-player characters (NPCs) in otome games.

By contrast, *CosyVoice2* (Du, Chen, et al., 2024) provides a fine-grained emotional synthesis framework that supports both the speaker's identity and they and the categorical emotions. It allows for direct control over emotion types and intensities within a single speaker identity, enabling transitions between emotional states across multi-turn interactions. These features make it particularly well suited to simulate personalized emotional conversations in otome games. Based on these considerations, I adopt CosyVoice2 as the speech synthesis module in this system.

Now that the motivation for this research has been presented, the structure of this thesis is as follows:

1.1 Research Questions and Hypotheses

In light of the preceding discussion, this research addresses the following question:

> **Can fine-tuning a CosyVoice2 (TTS) model on 4.86 hours of multi-style male speech (five emotion-based voice styles) from the ESD dataset maintains speaker similarity (SS $>$ 0.75) and perceived character consistency (MOS $>$ 3.5) in speech synthesis for Chinese otome games, when synthesizing dialogue for two distinct characters using instruction-following responses from Yi-1.5-6B-Chat (LLM), following speech recognition of the player's spoken input by Whisper (ASR)?**

This main question can be broken down into the following sub-questions:

- **Character Consistency:** Can prompt-based conditioning of the CosyVoice2 model produce speech that listeners perceive as consistent with the intended character identity, as measured by MOS (average score greater than 3.5)?
- **Speaker Similarity:** Does the generated speech preserve speaker identity features under different emotion styles and prompts, as indicated by a speaker similarity score greater than 0.75?

This study hypothesizes that the Yi-1.5-6B-Chat language model, guided by input transcriptions from Whisper and combined with a fine-tuned CosyVoice2 model, can generate speaker-consistent and emotionally appropriate speech across two character types. The goal is to build a character consistant and emotion-aware dialogue TTS system in otome games.

2 Literature Review

This section presents a structured and critical review of previous research relevant to dialogue generation in the context of otome games. In particular, the review focuses on the history of otome games and how recent developments in Automatic Speech Recognition (ASR), Large Language model (LLM) and Text-To-Speech (TTS) can be applied to support speaker consistent and emotion-aware in gaming environments. The aim is to establish a theoretical and technical foundation for the proposed Whisper + Yi-1.5-6B-Chat + CosyVoice2 pipeline.

The literature review is organized around three central themes. First, I examine the background and commercial potential of otome games, as well as the limitations of interaction caused by the reliance on pre-recorded audio. Second, I review the capabilities and limitations of LLMs in generating speaker-consistent, emotionally adaptive dialogue. Third, I survey recent developments in emotional speech synthesis, particularly focusing on models with speaker instructing and emotion control capabilities suitable for otome games.

To facilitate the review process, relevant works were selected using a targeted keyword strategy. The literature search was guided by three keywords: (1) otome games, including terms such as otome games, romantic video games, female-oriented games, and player–character bonding; (2) dialogue large language models, including controllable LLMs and stylistic prompts; and (3) instructing and emotional speech synthesis, including terms such as emotional TTS, multi-style speaker modeling, expressive TTS, speaker consistency, and emotional controllability.

Some studies were excluded from the review. First, works that were not related to speech synthesis or LLM-based dialogue systems. Second, fan discussions without technical analysis were not included. Third, studies on TTS or LLM were excluded if they did not include experiments or could not be applied to interactive systems.

This structure provides a focused overview of how prior work has informed the integration of expressive dialogue and voice synthesis into non-player characters (NPCs) interaction frameworks. The following subsections review each of the three major themes in detail, presenting key findings and outlining current gaps in the literature.

The literature review is organized into six subsections reflecting the core components of the proposed system. Subsection 2.1 outlines the search strategy and criteria used to identify relevant studies. Subsection 2.2 introduces the concept of emotional interactivity in otome games, highlighting narrative design, player psychology, and the limitations of pre-recorded audio. Subsection 2.3 reviews automatic speech recognition (ASR) technologies for dialogue systems, with a focus on models in Mandarin Chinese. Subsection 2.4 examines large language models (LLMs) for character consistent and emotion-aware dialogue generation, particularly in Mandarin Chinese. Subsection 2.5 discusses recent advances in emotional text-to-speech (TTS) models, emphasizing voice cloning and prompt-based control. Finally, Subsection 2.6 summarizes the key findings of the review and outlines the research gaps addressed in this work.

2.1 Search Strategy and Selection Criteria

- **Otome Games:** otome games, romantic video games, female-oriented games, player–character bonding

- **Dialogue Modeling:** dialogue large language models, controllable LLMs and stylistic prompts
- **Emotional Speech Synthesis with instructions:** emotional TTS, multi-style speaker modeling, expressive TTS, speaker consistency

Inclusion Criteria

1. Studies addressing emotional expressiveness and speaker consistency in TTS systems
2. Research on conversational LLMs, particularly in relation to character consistent or emotion-aware generation with support for Chinese dialogue
3. Papers on otome games or female-oriented romance games with an emphasis on emotional interaction design

Exclusion Criteria

1. Studies unrelated to speech synthesis or LLM-based dialogue modeling
2. Descriptive fan commentary
3. TTS or LLM studies lacking experimental evaluations or practical application to interactive systems

2.2 Emotional Interactivity in Otome Games

Otome games, also known as romantic visual games (RVGs), have experienced significant expansion across East Asia and globally over the past two decades. Originating in Japan with early games such as *Angelique* (1994), the genre has grown into a transnational media phenomenon supported by mobile applications, anime adaptations, and robust fan ecosystems (Ganzon, 2019). Their enduring appeal lies in emotionally scripted narratives, where players—typically female—engage in romantic relationships with idealized male characters in visually rich, choice-driven settings.

This growth aligns with broader societal shifts. As women gain increased agency in professional and educational sectors, their media consumption increasingly reflects a desire for emotionally resonant, intimate, and autonomous experiences (Ganzon, 2022). Otome games serve as interactive spaces where affection, fantasy, and agency converge. In China and the United States—the two largest gaming markets—female gamers account for 48% and 45% of the player base respectively (Statista, 2021), reflecting the potential economy demand. In 2024, otome games in Asia contributed approximately 23% of global mobile gaming revenue (Otome, 2025). Titles such as *Love and Deep Space* illustrate this success, generating 80 million USD in monthly revenue in the Asian market and 9.67 million USD in the West as of August 2024 (Tower, 2024).

Despite this commercial success, most otome games still rely on pre-recorded voice content. Emotional phone calls and character dialogues are typically fixed, scripted, and professionally voiced in advance, embedded as static assets tied to specific events. This polished but rigid format limits spontaneity interaction, creating a gap between immersive narrative and interactive flexibility. Even in widely acclaimed otome games, character responses remain one-directional and non-adaptive, failing to simulate the natural flow of conversation.(Andlauer, 2018)

This study argues that emotion-driven, Text-to-Speech (TTS) technologies can bridge this gap. By replacing static audio with dynamically synthesized, emotionally expressive speech tailored to context and player input, TTS systems can deepen narrative immersion and enhance emotional authenticity. In otome games, such a system has the potential to transform non-player characters (NPCs) from scripted scenarios into responsive, emotionally present partners—enabling players to experience various kinds of interactions without sacrificing flexibility or intimacy.

This motivates the present study, which explores the integration of emotional TTS into otome game non-player characters (NPCs), with the goal of enhancing interactional authenticity and emotional value in gameplay.

2.3 Automatic Speech Recognition for Dialogue Systems

In this project, Automatic Speech Recognition (ASR) plays a key role in interpreting players' spoken input into semantically meaningful text. To address the challenges of low-latency transcription and robustness under variable acoustic conditions, this work adopts Whisper (Radford et al., 2022), an open-source Transformer-based ASR model trained on approximately 680,000 hours of weakly supervised audio-text pairs across multiple languages and tasks.

Whisper supports 99 languages and demonstrates consistent performance across zero-shot evaluation benchmarks. Within the model family, the medium and large configurations (769M and 1.55B parameters) yield competitive results without the need for dataset-specific fine-tuning. For example, Whisper Large-v2 achieves a 2.5% word error rate (WER) on LibriSpeech test-clean and achieves up to 74.7% relative error reduction on out-of-distribution datasets such as CHiME-6 and VoxPopuli, compared with similarly sized supervised models.

As the project targets Mandarin-speaking users, the model's support for Chinese is particularly relevant. Whisper's training data includes over 23,000 hours of Mandarin speech, contributing to its performance in spontaneous and emotionally expressive speech settings. This linguistic coverage, combined with the model's multitask training on transcription, translation, voice activity detection, and language identification, supports its integration into dialogue systems.

Furthermore, Whisper exhibits robustness under noisy conditions. Its performance under low signal-to-noise ratios (e.g., under pub noise at SNR $<$ 10 dB) remains more stable than many LibriSpeech-tuned models, supporting its applicability to mobile or casual environments.

In summary, Whisper provides a suitable balance of generalization ability, transcription accuracy, and latency for integration into interactive, speech-driven dialogue systems, particularly in Mandarin-language game contexts.

2.3.1 Mandarin-Focused System Design

Given that 48% of otome game players are based in China, Mandarin Chinese is selected as the primary language for the speech recognition component of the system. The Whisper model is adopted due to its inclusion of over 23,000 hours of Mandarin audio in its training data, which contributes to its ability to transcribe a wide range of speaking styles, including informal and emotionally expressive speech. This choice supports more accurate recognition in the linguistic and cultural context of the intended user group, and reduces the need for extensive adaptation or domain-specific tuning.

2.4 Large Language Models for Emotion-Aware Dialogue

Recent advances in large language models (LLMs) have facilitated the development of dialogue systems on resource-constrained devices. However, their suitability for character consistent and emotion-aware applications, such as romantic interactions in otome games, remains an open question. A key consideration in model selection is the ability to balance inference efficiency with the capacity to generate character appropriate, emotionally adaptive dialogue in Mandarin Chinese.

Qwen(Bai et al., 2023), developed by Alibaba, is a bilingual Chinese-English LLM trained on 3 trillion tokens. While it exhibits strong performance across a range of general NLP tasks and supports advanced agent functionalities, its dialogue generation in Mandarin tends to lack the stylistic depth required for character consistent applications. Qwen-Chat ranks competitively among open-source models but often produces generic or emotionally neutral outputs in romantic settings, limiting its applicability to otome-style interactions.

Phi-2(Eldan et al., 2023), a 2.7B-parameter model from Microsoft Research, achieves performance with significantly larger models across commonsense and reasoning benchmarks. It is trained on 1.4 trillion tokens and excels in compact inference scenarios. However, it is primarily English-oriented and lacks the necessary coverage and tuning for Chinese conversational or narrative generation, which constrains its relevance to the targeted user base.

Yi-1.5-6B-Chat(AI et al., 2025), developed by 01.AI, offers a more suitable foundation for this task. It is pretrained on a 3.1 trillion-token corpus including high-quality Chinese data, and fine-tuned with a curated instruction dataset of under 10K examples annotated and verified by engineers. Yi models support both bilingual output and parameter-efficient deployment; notably, the 6B version supports 4-bit quantization while retaining over 90% of full-precision performance. On benchmarks such as C-Eval and CMMLU, which emphasize Chinese language understanding, Yi-6B outperforms several larger models. Its ability to retain stylistic consistency and respond coherently to emotionally framed prompts makes it appropriate for emotionally aware interaction in otome game scenarios.

In summary, considering performance on Mandarin benchmarks, stylistic alignment, and deployment constraints, Yi-1.5-6B-Chat is selected as the dialogue backbone for this project without further fine-tuning. It offers a practical balance between computational efficiency and dialogue quality, making it suitable for interactive, emotion-driven applications.

2.5 Emotional Text To Speech Models with Instructions

OpenVoice (Qin et al., 2024) adopts a modular design that decouples speaker tone color cloning from prosody and style generation. It uses a base speaker TTS model to generate expressive speech conditioned on user-defined style parameters (e.g., emotion, rhythm, intonation), and a tone color converter to transplant the target speaker's timbre into the generated audio. This approach allows for zero-shot voice cloning across languages, with the advantage of fast inference due to its fully feed-forward architecture. However, OpenVoice lacks integrated dialogue-level context modeling and does not support text-instructed emotional control, which limits its utility in narrative-driven applications such as otome games. Without explicit conditioning from textual prompts, the expressiveness often relies on the choice of base speaker model, making it difficult to sustain affective coherence over conversational interactions.

CosyVoice2 (Du, Chen, et al., 2024) introduces a unified instruction-driven framework that advances controllable text-to-speech (TTS) synthesis in both streaming and non-streaming settings. Building upon the earlier CosyVoice model (Du, Wang, et al., 2024), it adopts a semantic-acoustic decoupling approach, combining a large language model with a chunk-aware flow matching decoder to achieve low-latency, high-fidelity generation.

A core innovation of CosyVoice2 is its support for instruction-based synthesis. The model accepts structured prompts that embedding fine-grained prompts directly into the input text. These include high-level natural language instructions—such as emotion (e.g., “happy”, “serious”), role-playing style, dialect, and speaking rate—as well as token-level prosodic controls, including vocal bursts ([laughter], [breath]) and emphasis tags (<strong>, <laughter>). Instructions are prepended to the synthesis target text using a special **`<|endofprompt|>`** token, enabling consistent modulation of style and affect across utterances.

To support this functionality, CosyVoice2 simplifies its architecture by removing speaker embeddings and explicit text encoders, instead using a pre-trained large language model (Qwen2.5-0.5B) as the backbone for the text-speech prediction. Semantic tokens are generated using a Finite Scalar Quantization (FSQ) module, which achieves full codebook utilization while reducing entanglement with speaker identity. These tokens are then transformed into mel-spectrograms by a causal flow matching decoder, which is designed to handle both offline and streaming scenarios through chunk-aware masking strategies.

Through this design, CosyVoice2 enables expressive and context-aware voice synthesis with prompt-controllable attributes, making it particularly suitable for character-consistent applications, voice-driven agents, or dialogue systems in narrative games.

2.5.1 Model Selection Justification

CosyVoice2 is adopted as the backbone TTS model due to its unified instruction-based framework, which aligns closely with the needs of interactive otome game scenarios. Unlike modular approaches such as OpenVoice, which decouple timbre cloning from prosody modeling and rely on predefined style parameters, CosyVoice2 can be easily controled by the promptsa as input. This allows for fine-grained modulation of emotional tone, speaking style, and character consistent through natural language instructions.

Its architecture supports both streaming and non-streaming synthesis within a single model, offering a possibility to be deployed as mobile application in the future. Additionally, CosyVoice2 supports goal-directed fine-tuning through customized annotations as prompts, enabling flexible control over speaking style and emotional intent. These features make it particularly suitable for this character consistent and emotion-aware system.

2.6 Conclusion

This review examined recent advances in ASR, LLMs, and TTS with respect to their applicability in emotionally interactive dialogue systems, particularly in otome game scenarios. While significant progress has been made in ASR (e.g., Whisper), large language models (LLMs) (e.g., Yi-1.5-6B-Chat), and emotional speech synthesis (e.g., CosyVoice2), most existing systems have yet to support both character consistent and emotion-aware interactions.

Notably, prior work has often addressed these components in isolation, leaving a gap in integrated architectures that combine emotion-aware speech synthesis with stylistically adaptive language models and low-latency speech recognition. As a consequence, this research aims to address this gap by constructing a dialogue system for otome games, in which player speech is interpreted via Whisper, responded to by Yi-based LLM, and synthesized through CosyVoice2 with affective conditioning. The proposed system prioritizes emotional fidelity, responsiveness, and mobile deployability—factors essential to immersive, voice-driven romantic gameplay.

Table 1: Summary of Key Literature

Reference	**Key Findings**	**Theme**
Whisper (Radford et al., 2022)	Describes Whisper, a multilingual speech recognition model trained on 680,000 hours of transcribed audio. It demonstrates stable performance across accents and noise conditions, with strong semantic transcription accuracy, though it does not include emotion recognition capabilities.	Multilingual ASR
Yi (AI et al., 2025)	Introduces Yi-1.5-6B-Chat, a bilingual language model with support for stylistic variation and quantized deployment. It is conducted for contextually appropriate and emotionally styled Mandarin dialogue.	Stylistic LLM
Cosyvoice2 (Du, Chen, et al., 2024)	Proposes CosyVoice 2, a low-latency TTS system combining instruction-conditioned synthesis with flow-based modeling. It supports speech generation with emotional and role-specific control, suitable for interactive character-driven scenarios.	Emotion-Aware TTS

3 Methodology

This section outlines the methodology used to investigate the effectiveness of my proposed emotional speech synthesis system in an interactive game setting. First, in Subsection 3.1, I describe the datasets used for fine-tuning. Next, Subsection 3.2 introduces the CosyVoice2-based instruction-following synthesis model and explains how it interacts with Whisper and a lightweight language model to generate emotion-conditioned speech. Section 3.3 outlines subjective and objective evaluation metrics, with a focus on MOS and speaker similarity. Finally, Subsection 3.4 presents the experimental procedures, including deployment settings.

3.1 Emotional Speech Database (ESD)

For this study, I utilized a Mandarin Emotional Speech Database (ESD)(Zhou, Sisman, Liu, & Li, 2022), a publicly available corpus that provides parallel recordings across five emotional categories: neutral, happy, angry, sad, and surprised. Each utterance in the dataset is phonetically transcribed and labeled with its corresponding emotion, making it well-suited for supervised training in emotional text-to-speech (TTS) systems. Importantly, I did not fine-tune the model using explicit emotion classification keywords, as CosyVoice2 has already been pre-trained with emotion-labeled data, allowing inference prompts to guide expressive speech synthesis.

The Mandarin portion of ESD includes recordings from 10 speakers (5 male and 5 female), with approximately 2,000 utterances per speaker. For the purpose of this project, I selected a 4.86-hour subset of male speech to reflect stylized character profiles designed for use in an otome game. The recordings were sampled at 16 kHz and manually verified to ensure audio clarity and labeling consistency. Each audio file is paired with its corresponding text and emotion annotation, creating a reliable resource for emotion-conditioned fine-tuning. To better align character-specific speaking styles with system outputs, I used speaker ID as a prompt during fine-tuning of both the FLOW and the LLM module in CosyVoice2.

Although neutral corpora such as LJ Speech offer extensive training data, they lack emotional diversity. Moreover, the baseline model relies on large-scale proprietary datasets with hundreds of hours of audio, which are not publicly available. In contrast, ESD provides a publicly accessible and balanced emotional dataset with consistent speaker variation and parallel sentence structure between emotions. These features make it particularly suitable for character control and emotional modeling under limited data conditions.

Using the Mandarin ESD data, I aim to simulate the character consistent and emotion-aware dialogues in otome games. Associating each emotion with a fictional character persona allows the model to generate speech that reflects both affective emotion and narrative identity, thereby enhancing the immersive quality of in-game voice interaction.

3.2 Model Framework

This section introduces the model architecture of the proposed instruction TTS system, which enables synthesising for emotionally expressive game characters. The system consists of three core modules: Whisper for speech recognition, Yi-1.5-6B-Chat for dialogue generation, and

CosyVoice2 for emotional speech synthesis. Each module is chosen and set up to work together in the proposed emotional dialogue system.

3.2.1 CosyVoice2 Architecture

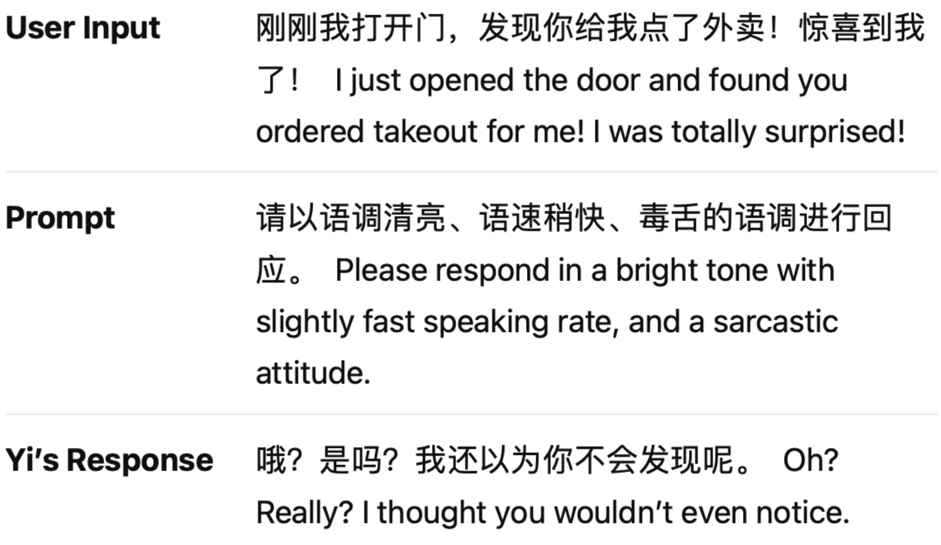
CosyVoice2 serves as the backbone of speech synthesis pipeline. It follows the original architecture as proposed, featuring a non-autoregressive flow-based model that generates mel-spectrograms from character and prompt embeddings. Unlike FastSpeech2-based approaches, CosyVoice2 does not rely on a duration predictor but uses invertible flows to model the complex distribution of mel features.

During training, I introduce speaker identity embeddings as prompt conditions, allowing the model to learn distinct vocal styles aligned with specific in-game character personalities. Emotion supervision is not explicitly added during fine-tuning, as the pre-trained model already encodes emotional variation. At inference time, emotion prompts such as 'happy' or 'surprised' and the speaker ID are used to control the generated style and emotion. The final waveform is reconstructed using HiFi-GAN, a high-quality neural vocoder.

3.2.2 Dialogue Generation with Whisper and Yi-1.5-6B

To enable interactive responses, the system begins with Whisper, an automatic speech recognition model trained on audio and text from many different languages. I use the base version of Whisper with the `--language "zh"` setting to transcribe the player's spoken input in Mandarin Chinese. This transcribed text is passed to Yi-1.5-6B-Chat, a Chinese language model optimized for instruction-following tasks.

Yi takes a structured prompt with two parts: the user's input and a style instruction. The user input gives the conversation context. The style instruction tells the model how to respond, such as the tone or speed. This helps the model give replies that match the desired style and emotion. For example, a dialogue generation request may be structured as follows:



User Input	刚刚我打开门，发现你给我点了外卖！惊喜到我了！ I just opened the door and found you ordered takeout for me! I was totally surprised!
Prompt	请以语调清亮、语速稍快、毒舌的语调进行回应。 Please respond in a bright tone with slightly fast speaking rate, and a sarcastic attitude.
Yi's Response	哦？是吗？我还以为你不会发现呢。 Oh? Really? I thought you wouldn't even notice.

Figure 1: Example of a structured dialogue prompt and model response

This generated response is then passed directly to the TTS module for synthesis.

3.2.3 Interaction Pipeline

The ideal system flow is as follows:
User microphone input → Whisper ASR → Yi-LLM dialogue generation → CosyVoice2 synthesis → Audio playback.

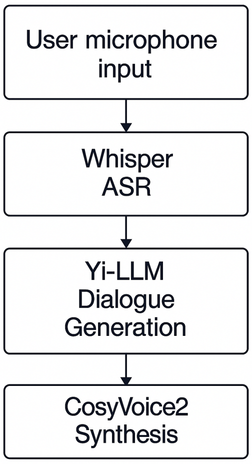


Figure 2: Pipeline

This pipeline is currently functional in a GPU-based environment. Each module's integration into a mobile system on a device remains a future deployment goal.

3.3 Evaluation – Subjective and Objective Metrics

To evaluate the quality and character consistency of the synthesized speech in the proposed TTS system, I adopt both subjective and objective evaluation methods. These assessments are designed to determine whether the system-generated speech aligns with the expected emotions and speaker identity of each fictional character in the otome game scenario.

3.3.1 Subjective Test for Character Consistency

The primary subjective evaluation in this study uses a Mean Opinion Score (MOS) test to assess character consistency, following the protocol commonly adopted in non-parallel emotional voice conversion tasks (Gao, Chakraborty, Tembine, & Olaleye, 2019). This MOS test focuses on measuring the perceived consistency between a character's intended persona and the synthesized voice, extending beyond standard evaluations of naturalness or intelligibility.

In this test, participants are presented with a textual description of a fictional character (e.g., "a sarcastic but soft-spoken romantic lead") and asked to rate how well the corresponding audio matches this persona on a 5-point Likert scale. Each synthesized sample is evaluated independently by multiple raters, and the final score is calculated as the mean of all ratings.

In addition to standard trials, two extra test items were included to assess mismatched emotional styles, allowing listeners to rate samples that deliberately contradict the given character description. This design enhances the ability to detect style misalignment in instruction-based synthesis tasks (Camp, Kenter, Finkelstein, & Clark, 2023).

The questionnaire is targeted at native speakers of Mandarin Chinese, as emotional nuances and role-based identity are often culture- and language-specific. Participants are required to self-report their native language status, familiarity with otome games, and age group before proceeding. Only respondents who confirm they are native Mandarin speakers are included in the analysis.

Each participant listens to a set of synthesized speech samples corresponding to two character profiles, each with three emotional speaking styles (e.g., calm, surprised, happy). Additionally, for each character, one mismatched emotion sample is included (e.g., a calm voice paired with a cheerful character description), acting as a control condition to test whether participants can detect emotional-incongruent outputs.

To ensure the reliability of the test, these mismatch control samples are used to assess attention and rating consistency. Responses with abnormally high scores for mismatched pairs are excluded.

3.3.2 Objective Test for Speaker Similarity

To complement the subjective evaluation, I also assess speaker similarity using an embedding-based metric. This objective metric quantifies how closely the synthesized voice matches the original speaker identity from the training data (Mandarin ESD). I use the VoiceEncoder module from the Resemblyzer toolkit to extract speaker embeddings(Jia et al., 2019). And a pre-trained speaker verification model is used to extract speaker embeddings from both the reference and synthesized audio(Wan, Wang, Papir, & Moreno, 2020). The cosine similarity between these embeddings is then computed to estimate how well the speaker characteristics are preserved .

$$\mathrm{SS} = \cos(\mathbf{e}_{\mathrm{ref}}, \mathbf{e}_{\mathrm{gen}}) = \frac{\mathbf{e}_{\mathrm{ref}} \cdot \mathbf{e}_{\mathrm{gen}}}{\|\mathbf{e}_{\mathrm{ref}}\| \cdot \|\mathbf{e}_{\mathrm{gen}}\|} \tag{1}$$

where

SS denotes the speaker similarity score between the reference and generated speech;

$\mathbf{e}_{\mathrm{ref}}$ is the speaker embedding extracted from the reference (ground-truth) utterance;

$\mathbf{e}_{\mathrm{gen}}$ is the speaker embedding extracted from the generated (synthesized) utterance;

$\cdot$ represents the dot product between two vectors;

$\| \cdot \|$ denotes the L_2 norm of a vector;

$\cos(\cdot)$ indicates cosine similarity, ranging from -1 to 1, with higher values indicating greater speaker similarity.

Evaluation Relevance This two evaluation approaches capture both character consistency and speaker similarity of the generated speech, which are essential for delivering emotionally expressive and consistent character voices. By using MOS and speaker similarity together—with the added reliability check via mismatch samples can assess whether the system successfully synthesizes speech that is not only intelligible and expressive but also faithful to each character's narrative role.

3.4 Training Setup

In this subsection, I describe the setup used to fine-tune a CosyVoice2-based emotional speech synthesis model. The model aims to support generation of expressive Mandarin speech for interactive scenarios in otome games. The codebase is implemented in PyTorch and trained on a single NVIDIA A100 GPU.

3.4.1 Model Configuration

- **Architecture**: The model adopts the original CosyVoice2 architecture (Du, Chen, et al., 2024), a non-autoregressive TTS system that combines a flow-based decoder with speaker prompts. Its major components include a transformer-based encoder, a conditional flow-matching decoder, and a neural vocoder based on HiFi-GAN with harmonic modeling. The system is configured with a chunk size of 16 tokens and `num_decoding_left_chunks` set to 2. This configuration was chosen to better accommodate the relatively short utterances and expressive emotional content in the target dataset, thereby reducing the risk of attention masking errors during generation.

- **Pre-trained Weights**: The model is initialized from the CosyVoice2 checkpoint, which was trained on a large-scale multilingual dataset comprising 130,000 hours of Chinese, 30,000 hours of English, and additional hours of Japanese and Korean speech data. This large-scale pretraining, as described in the original paper, enables CosyVoice2 to model expressive prosody and speaker variation across languages. The pretraining process included both supervised text-speech alignment using Paraformer and SenseVoice models, as well as internal force-alignment to filter out low-quality samples and improve punctuation precision.

- **Fine-tuning Strategy**: In this study, only the FLOW module and the LLM component are updated. The other components are frozen during fine-tuning. Speaker identity is encoded as a prompt vector and injected into the model to guide speech style and affective tone. Although no explicit emotion labels are used during fine-tuning, the pretrained model had already been exposed to emotion-labeled data. As a result, emotional variation can be controlled at inference time by providing natural language prompts describing the desired emotion.

3.4.2 Training Process

The fine-tuning dataset consists of 4.86 hours of Mandarin male speech from 10 speakers in the ESD corpus, covering five emotional styles. Audio is resampled to 16 kHz and truncated to a maximum of 2000 frames. Each utterance includes character-level transcriptions and speaker ID annotations.

Training uses the Adam optimizer with a learning rate of 1e-5, 2500 warm-up steps, and gradient clipping at 5. A dynamic batching strategy is applied based on maximum frame length, and training is conducted for 21 epochs. The objective is to minimize mel-spectrogram reconstruction loss using the frozen CosyVoice2 feature extractor.

3.4.3 Inference Setup

At inference, the model is conditioned on a ground truth audio as the instruction and speaker ID prompts. A latent variable is sampled from a Gaussian prior and passed through the inverse flow to produce mel-spectrograms, which are then converted to waveform using the HiFi-GAN vocoder.

3.5 Objective

This study aims to build a emotional speech synthesis system desighed for interactive character-based dialogue in otome games. The system integrates automatic speech recognition (ASR), large language model (LLM) response generation, and emotional text-to-speech (TTS) synthesis, and is fine-tuned on expressive Mandarin data to reflect multi-style emotional profiles. The primary objectives of this research include:

- Adapting the CosyVoice2 architecture to perform prompt-based emotion-aware speech synthesis using emotional speech data from the Mandarin ESD corpus, with fine-tuned speaker identity and style embeddings.
- Evaluating the perceptual consistency between character descriptions and generated speech through MOS scores and embedding-based speaker similarity, emphasizing role alignment and emotional coherence in TTS output.
- Demonstrating the feasibility of integrating ASR, LLM, and TTS modules into a unified pipeline that supports voice interaction for dynamic storytelling and character immersion.

This research helps build a character-consistent systems that offering emotional conversations. This direction is particularly relevant for character-driven applications in otome games.

4 Experimental Setup

4.1 Overview

In this section, I provide a detailed breakdown of the experimental setup used for my study on emotional speech synthesis in Mandarin using the CosyVoice2 model. The subsections cover the following aspects:

Model Fine-Tuning Configuration: This part describes the baseline checkpoint, fine-tuning strategies, and dataset splitting procedures used to adapt CosyVoice2 to multi-style emotional speech synthesis.

Experimental Setup: This part introduces the hardware and software environment for model training and inference, including the GPU specifications, dependencies, and runtime framework.

Training and Evaluation Process: This subsection outlines the model training stages, including checkpoint selection and loss monitoring, followed by both subjective and objective evaluation methods used to assess emotional style appropriateness and speaker similarity.

Each subsection is designed to provide a thorough understanding of the methodologies and tools applied in this study, ensuring transparency and reproducibility of the results[1].

4.2 Model Fine-Tuning Configuration

4.2.1 Baseline Model – CosyVoice2 Pretrained Checkpoint

The baseline model used in this study is the publicly released CosyVoice2 checkpoint (Du, Chen, et al., 2024), which was pretrained on a large-scale multilingual corpus, including 130,000 hours of Chinese, 30,000 hours of English, and additional Japanese and Korean data. During pretraining, explicit emotion labels were used to train the model on emotion-controllable synthesis. As a result, the model is capable of generating expressive speech conditioned on natural language prompts describing style, emotion, or speaker features. This pretrained checkpoint serves as a strong general-purpose TTS foundation with support for prompt-based stylistic control.

4.2.2 Fine-Tuned Model – CosyVoice2 Adapted to Multi-Style Emotional Mandarin Speech

The fine-tuned model builds upon the CosyVoice2 baseline and is trained on a 4.86-hour subset of the Mandarin Emotional Speech Database (ESD), consisting exclusively of male speakers. The dataset covers five emotional categories: neutral, happy, angry, sad, and surprised.

During fine-tuning, only the FLOW module and the LLM are updated. The encoder and neural vocoder remain frozen to preserve the pretrained model's general synthesis capabilities.

Speaker identity is incorporated as a learned prompt embedding and is prepended to the input text during both training and inference. Although no explicit emotion labels are used in the training set, emotional variation is controlled at inference time using instruction-style

[1] `https://github.com/Qianqian1220/HeartEcho`

prompts (e.g., “Please speak in an angry tone”), which the pretrained CosyVoice2 model is able to interpret due to its exposure to labeled emotional data during pretraining.

4.2.3 Data Splitting and Composition

The fine-tuning corpus is divided into two non-overlapping subsets[2]:

- **Training Set**: 70% of the data (∼3.4 hours), used to update the flow and speaker prompt components during fine-tuning.
- **Test Set**: 30% (∼1.46 hours), used to monitor reconstruction loss and assist in selecting the final model checkpoint.

All audio was resampled to 16 kHz to fit within GPU memory constraints during batch processing.

4.2.4 Hardware Environment

All experiments were conducted on an NVIDIA A100 GPU with 16 GB of memory, supporting long-sequence dynamic batching and high-throughput training.

4.2.5 Software Environment

The system was developed using Python 3.10 and PyTorch 2.6.0 with CUDA 12.6 support. Key libraries include TorchAudio, Transformers, Gradio, Librosa, and openai-whisper. Platform-specific acceleration was enabled via ONNX Runtime and TensorRT. Full version specifications and dependency details are provided in the project repository[3].

4.2.6 Hyperparameter Tuning

No automated hyperparameter tuning tools (e.g., Optuna) were used. All hyperparameters, including learning rate, batch configuration, and gradient clipping threshold, were manually configured based on CosyVoice2 defaults and preliminary trial runs.

4.3 Training and Evaluation Process

4.3.1 Training Procedure

The training procedure for the CosyVoice2-based emotional speech synthesis model is outlined as follows:

Step 1: Load the Pretrained CosyVoice2 Checkpoint

The base model used in this study is the publicly released CosyVoice2 checkpoint, which was pretrained on approximately 130,000 hours of Chinese and 30,000 hours of English speech data.

[2] The data splitting script is available at `https://github.com/Qianqian1220/HeartEcho/blob/main/split_male.py`

[3] Dependency details available at `https://github.com/Qianqian1220/HeartEcho/blob/main/requirements.txt`

This large-scale multilingual training, which included explicit emotion annotations, equips the model with strong capabilities for zero-shot and instruction-based synthesis.

Step 2: Prepare the Fine-Tuning Dataset
For this study, a 4.86-hour subset of the Mandarin Emotional Speech Database (ESD) was used. This subset consists entirely of male speakers and covers five emotional categories: neutral, happy, angry, sad, and surprised. Each audio sample is paired with a manually transcribed Chinese sentence in character format, along with its corresponding speaker identity. All audio files were resampled to 16 kHz and truncated to a maximum of 2000 frames to ensure training stability and GPU efficiency.

Step 3: Fine-Tune the Model Only two components were updated during fine-tuning: the FLOW module, which transforms semantic speech representations into acoustic latent features. And the LLM, which maps input prompts and tokens to semantic speech representations. The encoder and neural vocoder (HiFi-GAN) were kept frozen.

Speaker identity was encoded as a 192-dimensional learned vector and used as a prompt prepended to the input text. No explicit emotion labels were used in the fine-tuning dataset. Instead, emotional control was introduced at inference time using natural language instruction prompts (e.g., “Please speak in a happy emotion”).

The model was trained for 21 epochs using the CosyVoice2 training pipeline. The learning rate was set to 1e$-$5 with 2500 warm-up steps, and gradient clipping was applied with a maximum norm of 5. A dynamic batching strategy was employed with a maximum of 2000 frames per batch. Training and validation were conducted on an NVIDIA A100 GPU with 16 GB of memory.

Performance Monitoring: Throughout training, internal logs such as training loss and GPU utilization were tracked using TensorBoard. Although no validation loss was computed, these logs were useful for monitoring convergence and ensuring model stability. [4]..

Step 4: Model Saving The model was trained for 21 epochs. No formal validation set was used during training. The checkpoint selection was based on manual inspection of synthesized outputs at regular intervals. After reviewing the perceptual quality and stability of samples, the checkpoint at epoch 21 was chosen for final evaluation and downstream inference[5].

All training was performed using PyTorch 2.6.0, with configuration and runtime management handled via Hydra and OmegaConf. The config file is modular and reproducible for future experiments[6].

4.3.2 Evaluation Method

The evaluation of the CosyVoice2-based emotional speech synthesis model is conducted using both subjective and objective metrics to assess the generated speech in terms of character consistency and speaker similarity.

[4] The logs for both modules are available at `https://github.com/Qianqian1220/HeartEcho/tree/main/examples/libritts/cosyvoice2/tensorboard`

[5] The fine-tuned checkpoint is available at `https://drive.google.com/file/d/1zG7VHnZFytfUFswkkNb2dY-6-l2j9tYP/view?usp=sharing` and `https://drive.google.com/file/d/1_yj1Q1OxBoGvBTMqcaf8q-M_1OJdEqdp/view?usp=sharing`

[6] The corresponding configuration files are available at `https://drive.google.com/file/d/1WPBnKOUBsllatRo7UnK4dEYh17A1kI2g/view?usp=sharing`

Subjective Evaluation – Character Consistent:
The primary subjective metric is a 5-point Likert scale measuring how well the synthesized speech matches a given character persona. Each test item presents a predefined character description along with a synthesized utterance. Participants were asked:

"Do you think this voice matches the character's described style?"

Two distinct character types were used in the evaluation:

- **Cool and Polite Type**: Calm and quiet, with a deep and smooth voice. Always polite and respectful. He may seem cold and doesn't talk much, but shows care and kindness when it really matters.
- **Awkwardly Warm Type**: Clear mid-high voice, speaks a bit fast, shows emotions openly but in a subtle way. Seems sharp-tongued and playful on the outside, often says teasing or biting things, but deep down really cares a lot about the other person.

These descriptions were shown to participants prior to listening, so their evaluation focused on whether the voice conveyed the appropriate stylistic and emotional traits as described.

Objective Evaluation – Speaker Similarity:
Speaker similarity was computed by measuring the cosine similarity between embeddings extracted from synthesized utterances and reference recordings. These embeddings are test by `VoiceEncoder`. This objective metric quantifies the model's ability to maintain a consistent speaker identity across various prompts and emotions.

In conclusion, this evaluation framework provides both subjective and objective insights into the model's ability to produce speech that is emotionally expressive and character-consistent.

5 Results

5.1 MOS for Character Consistency

To assess whether the synthesized speech maintained character consistency across different emotion styles, a Mean Opinion Score (MOS) test was conducted. Each sample was evaluated by 37 listeners, from which 32 valid responses were retained after filtering out 5 surveys containing at least one extreme mismatch rating of 5 for the final two mismatch samples.[7]

Table 2 summarizes the average MOS scores across eight experimental conditions. These conditions include three emotion styles (Surprise, Happy, Calm) for each of the two character profiles (Speaker A and Speaker B), as well as two mismatch conditions where character identity was intentionally mismatched. The mismatch samples were designed to assess the degree of alignment between the voice and the intended character personality, serving as a reference point for evaluating character consistency. Speaker A's Surprise style received the highest MOS score of 3.78, followed by Speaker B's Calm (3.69) and Happy (3.59) styles, suggesting generally strong consistency across these settings. The two mismatch conditions received the lowest MOS ratings (2.50 for Speaker A and 2.72 for Speaker B), indicating that listeners were able to detect a lack of alignment between the voice and character personality when mismatches occurred.

Table 2: Mean Opinion Scores (MOS) by Condition

Condition	**MOS**
SpeakerA_Surprise	3.78
SpeakerA_Happy	3.50
SpeakerA_Calm	3.31
SpeakerB_Surprise	3.44
SpeakerB_Happy	3.59
SpeakerB_Calm	3.69
SpeakerA_Mismatch	2.50
SpeakerB_Mismatch	2.72

To assess whether perceived identity consistency in each emotion condition was affected by the speaker, a Wilcoxon signed-rank test was conducted between Speaker A and Speaker B for each emotion. Results showed a statistically significant preference for Speaker A over B in the Calm condition ($p = 0.0209$), while differences in Surprise ($p = 0.1522$) and Happy ($p = 0.0976$) were not statistically significant.

These results indicate that speaker identity had small impact on perceived character consistency in most emotional conditions, with a significant effect observed only in the calm

[7] The MOS test was conducted via SurveyMars: `https://surveymars.com/q/7nritOKog`. The resulting scoring data is available at: `https://docs.google.com/spreadsheets/d/1eGEOjqzgoOdVhCeb_PDA_nzz_ARhfrAH/edit?usp=sharing&ouid=107678446489797918453&rtpof=true&sd=true`

emotion. This implies that emotional perception was generally robust across speakers, except in calm emotion.

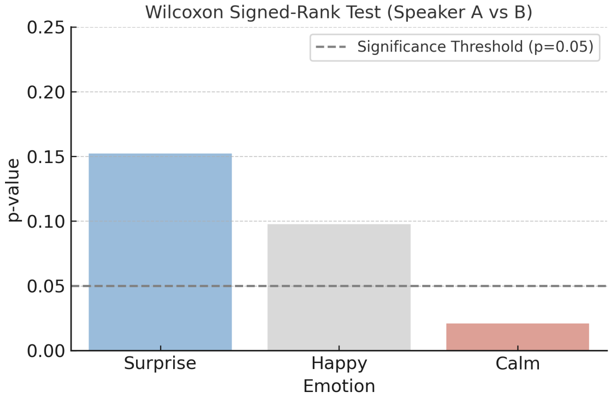


Figure 3: Wilcoxon test

Moreover, a Kruskal–Wallis test was conducted to assess whether MOS scores differed significantly across emotion styles within each speaker. The test returned p-values of 0.0675 for SpeakerA and 0.4947 for SpeakerB. Although neither result met the conventional threshold for statistical significance ($p < 0.05$), the findings suggest that listeners' evaluations of character consistency were generally stable across different emotional styles. This indicates that emotional variation did not significantly affect identity perception within each speaker.

This observation is further supported by the distributions shown in Figure 4 and Figure 5. Upon further examination, the six groups also share the same median and third quartile (Q3).

For Speaker A, the MOS scores for *Surprise*, *Happy*, and *Calm* show clear differences in how listeners judged character consistency for each style.

Surprise shows the most favorable distribution. The median score is exactly 4.0, with an upper whisker extending to 5, indicating that a substantial portion of listeners rated this style as highly consistent with the intended character persona. The absence of low-end outliers further suggests broad agreement among raters. This supports the hypothesis that prompt-based conditioning can achieve strong perceived consistency in more expressive styles.

Happy also has a median of 4.0, but it contains a low-end outlier rated at 1, indicating that at one listener perceived a strong mismatch between voice and character. Nonetheless, the majority of scores are tightly centered, and the third quartile reaches 4, suggesting that for most listeners, the happy speech style was also convincing. This aligns with the research hypothesis, though slightly less robustly than Surprise.

Calm, in contrast, presents the lowest median score of 3.0, which belows the 3.5 threshold defined in the hypothesis for character consistency. While the upper quartile still includes high ratings (up to 5), the median and presence of multiple low-end ratings (including one outlier of 1) reflect a more varied perception among listeners. This indicates that the calm emotional style may be more challenging for the model to express in a character-consistent way, potentially due to its lower acoustic variation or reduced expressive features. This partially contradicts the hypothesis and suggests that further refinement is needed for subtle styles.

Taken together, these results partially support the hypothesis: both *Surprise* and *Happy* styles exceed the expected MOS threshold, with *Surprise* showing the most consistent alignment with character identity. However, the *Calm* style falls short of the 3.5 average, highlighting an area for future improvement in emotional nuance and prosody control.

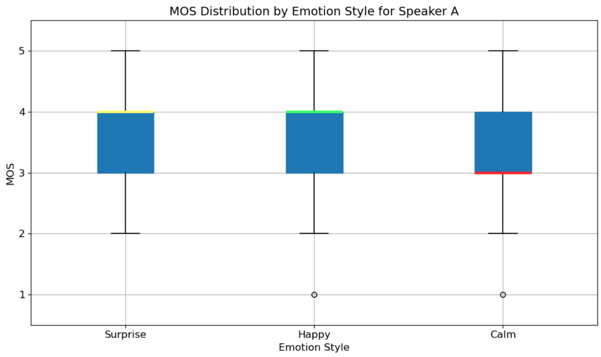


Figure 4: MOS Distribution by Emotion Style for Speaker A

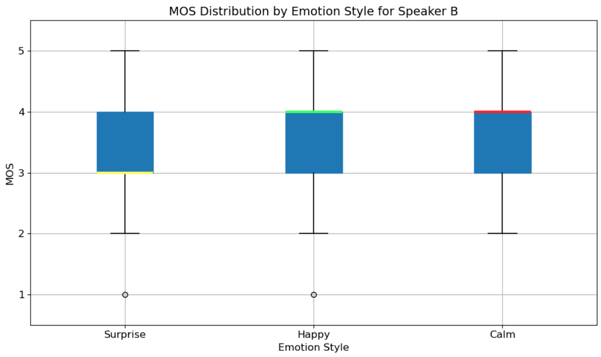


Figure 5: MOS Distribution by Emotion Style for Speaker B

For Speaker B, the MOS distributions for *Surprise*, *Happy*, and *Calm* show moderate variation, with all three styles sharing similar interquartile ranges. Both *Surprise* and *Happy* contain low-end outliers rated at 1, while *Calm* shows no outliers and a more stable shape. The median score for *Surprise* is 3.0, and which just belows the 3.5 threshold, indicating mixed evaluations from listeners. In contrast, both *Happy* and *Calm* have medians of 4.0 and average scores above 3.5, reflecting more consistent agreement with the intended character identity. These results suggest that while all three emotion styles produce fairly concentrated ratings, only *Happy* and *Calm* meet the expected standard of character consistency (MOS $>$ 3.5). Compared to Speaker A—where *Surprise* performed best—Speaker B shows a different pattern, with *Surprise* being the least convincing. This indicates that emotional conditioning was only partially successful for Speaker B, also with room for improvement in generating more expressive and distinct speech styles.

In summary, the MOS results reveal speaker-dependent differences in emotional effectiveness. For Speaker A, *Surprise* and *Happy* styles achieved strong perceived character consistency, while *Calm* fell below the hypothesized threshold. In contrast, Speaker B showed the opposite trend: *Happy* and *Calm* were rated consistently above 3.5, whereas *Surprise* received lower and more varied evaluations. These contrasting patterns suggest that the effectiveness of emotional conditioning is not uniform across speakers, and may depend on the interplay between vocal characteristics and emotional expressiveness. While both speakers demonstrated the potential of prompt-based control to guide style-specific synthesis, further refinement is needed—particularly for subtler emotions like *Calm* in Speaker A and more dynamic expressions like *Surprise* in Speaker B—to ensure reliable character alignment across different voice profiles.

5.2 Speaker Similarity

5.2.1 Formula

To measure whether the TTS system preserved speaker identity across emotions, I computed speaker similarity scores using the `VoiceEncoder` from the `resemblyzer` library. All audio files were preprocessed using `preprocess_wav` to normalize sampling and silence, then embedded into fixed-length 256-dimensional speaker vectors. Similarity was computed using the cosine similarity formula:

$$\text{similarity}(A, B) = \frac{A \cdot B}{\|A\|\|B\|} \tag{2}$$

The resulting score ranges from -1 to 1, where values closer to 1 indicate stronger similarity in speaker timbre and vocal traits.

5.2.2 Test 1: Identity Preservation Under Style-Matched Conditions

The first test evaluated whether the TTS output retained speaker identity when the ground truth and synthesized utterances came from the same speaker and shared the same emotional style. Six pairs were tested (three for Speaker A and three for Speaker B), covering *surprise*, *happy*, and *calm* emotions. The results are summarized below:

Pair	Speaker	Emotion	Similarity
1	A	surprise	0.84
2	A	happy	0.88
3	A	calm	0.71
4	B	surprise	0.81
5	B	happy	0.89
6	B	calm	0.82

Table 3: Speaker similarity scores under identity and emotion conditions.

Speaker A: Identity was well preserved under *surprise* and *happy* conditions, with similarity scores of 0.84 and 0.88, respectively. However, a significant drop was observed in the *calm* condition (0.71), indicating that low-energy, less prosodically rich speech may challenge the model's ability to reproduce stable speaker embeddings.

Speaker B: Scores remained consistently high across all three emotions, with *calm* reaching 0.82 and *happy* peaking at 0.89. This suggests that the system captured and maintained Speaker B's vocal identity more robustly, even in emotionally neutral expressions.

These results confirm that identity retention is under emotionally rich conditions, but more vulnerable when the emotional content is low or less distinctive.

5.2.3 Test 2: Identity Suppression Under Cross-Speaker Mismatch

To verify that high similarity scores in the main test reflected true speaker identity rather than acoustic coincidence, I conducted a contrastive test using two mismatched groups. Here, the synthesized utterance came from the other speaker, with matched emotional style but intentionally exchanged identities.

Pair	Ground Truth Speaker	TTS Speaker	Emotion	Similarity
7	A	B	happy	0.60
8	B	A	surprise	0.58

Table 4: Speaker similarity scores under cross-speaker mismatch conditions.

Both scores were significantly lower than the identity-matched cases. This confirms that the model's embeddings remain sensitive to speaker-level traits, and that cross-speaker generation leads to measurable degradation in similarity, even under shared emotional styles.

This test is different from the MOS mismatch settings, which focus on character consistency. In contrast, speaker similarity emphasizes voice identity. The lower scores here show that the model preserves distinct speaker voices even when emotional styles are changed.

5.2.4 Conclusion

The speaker similarity evaluation shows that the fine-tuned TTS model can retain speaker identity across expressive conditions, with high consistency in the *surprise* and *happy* styles for both speakers. However, *calm* speech presented more difficulty, particularly for Speaker A. Crucially, identity was not preserved under cross-speaker synthesis, confirming that the model does not collapse speaker representations when instructed to mimic another style. These findings support the model's capacity for identity retention and highlight the need for further improvement in neutral emotional settings.

5.3 Discussion

The MOS and speaker similarity results together offer insight into how effectively the fine-tuned CosyVoice2 model maintains character consistency and speaker identity.

MOS scores reveal clear differences in perceived character appropriateness. For SpeakerA, the *surprise* style received the highest score (3.78), with ratings concentrated around the upper end and no low-end outliers, indicating strong agreement among listeners. In contrast, the *calm* style scored the lowest (3.31), with all ratings below the neutral threshold of 3.5. This suggests that the calm voice, while not heavily low, was consistently perceived as less aligned with the intended persona. SpeakerB, on the other hand, received more stable ratings across styles, with an average score of 3.57 and a narrow range of only 0.25, indicating a more robust preservation of character profiles regardless of emotion.

The speaker similarity (SS) results mirror this pattern. SpeakerA achieved high SS in *happy* (0.88) and *surprise* (0.84), but showed a noticeable drop in *calm* (0.71), confirming that low-energy speech made it harder for the model to preserve distinctive speaker features. SpeakerB again displayed stable identity retention across all styles, suggesting stronger generalization for this voice under the same fine-tuning setup.

Mismatch samples further validate these findings. When character identity and voice were intentionally mismatched, both MOS and SS dropped substantially (MOS = 2.50 and 2.72; SS = 0.60 and 0.58). This confirms that listeners were sensitive to inconsistencies between style and identity, and the model successfully preserved voice boundaries even under style-matched but speaker-swapped conditions.

In sum, the system performs under expressive, high-energy conditions and for speaker profiles with consistent vocal features. However, neutral or low-energy styles like *calm* remain a challenge for both characteristic and acoustic identity retention. These results highlight the need for improved modeling strategies adapted to finer emotional expressions and less dynamic voices.

6 Conclusion

This chapter summarizes the key challenges faced during the development of the emotional TTS system, reflects on current limitations, and outlines practical directions for future work. The section is divided into three parts: Challenges, Limitations and Recommendations, and Future Work.

6.1 Challenges

Throughout the project, I encountered several practical challenges—mainly related to data selection, training stability, and system compatibility with evolving open-source tools.

Data Selection and Training Stability: The initial choice of training data was the male subset of the CASIA (Chinese Academy of Sciences, 2023) emotional dataset. However, its limited size and relatively constrained emotional range proved insufficient for meaningful fine-tuning. The model exhibited poor convergence, and outputs lacked emotional variation and identity retention. Replacing this with the Mandarin ESD dataset significantly improved both training stability and perceptual output quality. The ESD's broader emotion coverage and larger sample size enabled the model to better learn emotion-specific acoustic features and speaker-dependent traits.

Data Split and Generalization: In early experiments, a 9:1 training-test split caused overfitting, as the small validation set failed to expose generalization errors. The model performed well on training samples but produced degraded outputs for unseen inputs. Switching to a 7:3 split improved robustness by enabling more reliable validation and loss tracking, particularly during instruction-based fine-tuning. This adjustment was crucial for ensuring stable synthesis performance beyond the training set.

Emotion Modeling Difficulties: Low-energy emotions, especially *calm*, presented consistent challenges. Unlike expressive emotions like *surprise*, which include clear pitch and intensity contours, *calm* utterances have minimal prosodic variation. This limited the model's ability to ground emotional expression and often resulted in flat or identity-degraded outputs. The difficulty underscores a fundamental limitation in current prosody modeling under low-variance conditions.

Model Maintenance and Framework Updates: As CosyVoice2 is an actively evolving open-source system, many of its key features, configuration interfaces, and dependencies were updated during the training process. Initially, the related config file was not available, but later updates introduced breaking changes on May. Adapting the pipeline to these updates while preserving reproducibility required continual version tracking, dependency patching, and code modification. Nevertheless, the changes, such as updated flow modules and prompt conditioning—enhanced overall system flexibility and control.

6.2 Limitations and Recommendations

While the final system met its core objectives, producing character-aligned emotional speech in Mandarin—several technical and methodological limitations remain.

Limited Initial Data: The early CASIA subset was insufficient to support stylistically rich synthesis. While ESD improved results, it still reflects constraints in diversity, as all recordings are studio-quality and lack spontaneous or context-aware interaction patterns. Future work should prioritize diverse and expressive datasets, possibly including crowd-sourced or dialogue-embedded emotional speech to improve generalization in real-world scenarios.

Limited Emotion Labels: The ESD dataset provides five emotional labels: neutral, happy, angry, sad, and surprised. This limited set is not fine-grained enough to support nuanced stylistic variation. For instance, it performs poorly in recognizing the *calm* emotion, which is a low-energy style often found in character dialogue. Richer annotation schemes or continuous affective ratings could allow the model to represent emotional subtypes more faithfully.

Prosody and Phrasing Limitations: Listener feedback indicated that pause placement and prosodic phrasing were occasionally unnatural, particularly for low-arousal styles. The current model lacks explicit control over prosodic units and syntactic boundaries. This limits its capacity to deliver rhythmically convincing and emotionally grounded speech. Incorporating a prosody predictor or syntax-aware phrasing module could improve immersion and control.

Low-Variance Emotion Degradation: Both objective and subjective evaluations indicated that speaker identity degraded under low-energy emotion conditions, especially for Speaker A. While the model retained identity in expressive conditions, it failed to preserve sufficient speaker embeddings in *calm* utterances. This highlights the entanglement between emotion and identity, and the need for more robust representation learning methods for prosodically flat input.

Evaluation Bias: MOS scores, while informative, introduce subjectivity and listener bias. Factors such as player expectations, age, or voice style preference may influence ratings. Moreover, the current evaluation focused solely on character consistency, overlooking other important dimensions such as expressiveness, naturalness, and emotional clarity. A more comprehensive assessment would benefit from incorporating multiple evaluation angles to better capture the overall performance of the system.

Overly Formal Output: Several users noted that outputs sounded scripted or overly polished. This is due to the lack of spontaneous phrasing and absence of colloquial or dialectal variations. Given the interactive nature of otome games, where natural-sounding, emotionally reactive speech is crucial, future systems should incorporate stylistic diversity through prompt tuning or data augmentation with spontaneous corpora.

Style-Identity Entanglement: The system sometimes entangled speaker identity with emotional style, especially in low-arousal speech. For example, calm utterances from Speaker A not only sounded emotionally flat but also exhibited weakened speaker similarity. This indicates that current embeddings do not fully disentangle speaker identity from acoustic style features. Better modeling of these attributes, such as through multiobjective loss or contrastive learning, could improve synthesis consistency across emotional contexts.

6.3 Future Work

Building on the system developed in this study, several future directions are worth exploring.

Expanding to Multi-Character Style Modeling: Supporting multiple character profiles would increase the system's flexibility and realism. This would require more diverse training data and better disentanglement of speaker and emotion attributes during synthesis.

Improving Emotion Control: Current outputs reflect basic emotional intent, but lack finer control. Future models could support continuous emotion scaling or multidimensional emotional prompts, allowing for more expressive and precise speech generation.

Diversifying Evaluation Metrics: Current evaluations focus primarily on character consistency, offering a limited view of system performance. Future work should assess additional dimensions such as naturalness, emotional intensity, and emotional recognizability. These aspects are crucial for understanding how well the system conveys expressive intent and maintains voice quality. Incorporating both perceptual ratings and objective emotion classification can provide a more nuanced and complete evaluation framework.

Optimizing for Mobile Deployment: The current model runs on GPU and is not ready for real-time deployment on phones or tablets. Future work should look into quantization, lightweight inference architectures, and streaming synthesis to reduce latency and memory use—without compromising output quality.

References

AI, ., Young, A., Chen, B., Li, C., Huang, C., Zhang, G., … Dai, Z. (2025, January). *Yi: Open Foundation Models by 01.AI.* arXiv. Retrieved 2025-06-03, from http://arxiv.org/abs/2403.04652 (arXiv:2403.04652 [cs]) doi: 10.48550/arXiv.2403.04652

Andlauer. (2018). Pursuing One's Own Prince: Love's Fantasy in Otome Game Contents and Fan Practice. *Mechademia: Second Arc*, *11*(1), 166. Retrieved 2025-06-03, from https://www.jstor.org/stable/10.5749/mech.11.1.0166 doi: 10.5749/mech.11.1.0166

Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., … Zhu, T. (2023, September). *Qwen Technical Report.* arXiv. Retrieved 2025-06-03, from http://arxiv.org/abs/2309.16609 (arXiv:2309.16609 [cs]) doi: 10.48550/arXiv.2309.16609

Camp, J., Kenter, T., Finkelstein, L., & Clark, R. (2023, August). MOS vs. AB: Evaluating Text-to-Speech Systems Reliably Using Clustered Standard Errors. In *INTERSPEECH 2023* (pp. 1090–1094). ISCA. Retrieved 2025-06-03, from https://www.isca-archive.org/interspeech_2023/camp23_interspeech.html doi: 10.21437/Interspeech.2023-2014

Chinese Academy of Sciences. (2023). *Casia chinese emotional speech corpus.* Retrieved from https://gitcode.com/Universal-Tool/a813d/ (Accessed on June 10, 2025. Unofficial redistribution of the CASIA corpus.)

Du, Z., Chen, Q., Zhang, S., Hu, K., Lu, H., Yang, Y., … Yan, Z. (2024, July). *CosyVoice: A Scalable Multilingual Zero-shot Text-to-speech Synthesizer based on Supervised Semantic Tokens.* arXiv. Retrieved 2025-05-09, from http://arxiv.org/abs/2407.05407 (arXiv:2407.05407 [cs]) doi: 10.48550/arXiv.2407.05407

Du, Z., Wang, Y., Chen, Q., Shi, X., Lv, X., Zhao, T., … Zhou, J. (2024, December). *CosyVoice 2: Scalable Streaming Speech Synthesis with Large Language Models.* arXiv. Retrieved 2025-06-03, from http://arxiv.org/abs/2412.10117 (arXiv:2412.10117 [cs]) doi: 10.48550/arXiv.2412.10117

Eldan, R., Lee, Y. T., & Nguyen, A. (2023). *Phi-2: The surprising power of small language models - microsoft research.* Retrieved from https://www.researchgate.net/publication/385654002_Phi-2_The_surprising_power_of_small_language_models (Technical Report, Microsoft Research)

Ganzon, S. C. (2019, November). Growing the Otome Game Market: Fan Labor and Otome Game Communities Online. *Human Technology*, *15*(3), 347–366. Retrieved 2025-06-03, from https://www.researchgate.net/publication/337646376_Growing_the_Otome_Game_Market_Fan_Labor_and_Otome_Game_Communities_Online doi: 10.17011/ht/urn.201911265024

Ganzon, S. C. (2022). *Playing at romance: Otome games, globalization and postfeminist media cultures.* Retrieved 2025-06-06, from https://spectrum.library.concordia.ca/id/eprint/990916/ (Concordia University Research Repository)

Gao, J., Chakraborty, D., Tembine, H., & Olaleye, O. (2019, September). Nonparallel Emotional Speech Conversion. In *Interspeech 2019* (pp. 2858–2862). ISCA. Retrieved 2025-06-10, from https://www.isca-archive.org/interspeech_2019/gao19b_interspeech.html doi: 10.21437/Interspeech.2019-2878

Jia, Y., Zhang, Y., Weiss, R. J., Wang, Q., Shen, J., Ren, F., … Wu, Y. (2019, January). *Trans-*

fer Learning from Speaker Verification to Multispeaker Text-To-Speech Synthesis. arXiv. Retrieved 2025-06-03, from http://arxiv.org/abs/1806.04558 (arXiv:1806.04558 [cs]) doi: 10.48550/arXiv.1806.04558

Otome. (2025). *Otome games market trends and growth 2025.* Otome.com. Retrieved from https://otome.com/2025/02/25/otome-games-market-trends-and-growth/

Qin, Z., Zhao, W., Yu, X., & Sun, X. (2024, August). *OpenVoice: Versatile Instant Voice Cloning.* arXiv. Retrieved 2025-06-03, from http://arxiv.org/abs/2312.01479 (arXiv:2312.01479 [cs]) doi: 10.48550/arXiv.2312.01479

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2022, December). *Robust Speech Recognition via Large-Scale Weak Supervision.* arXiv. Retrieved 2025-06-03, from http://arxiv.org/abs/2212.04356 (arXiv:2212.04356 [eess]) doi: 10.48550/arXiv.2212.04356

Ren, Y., Hu, C., Tan, X., Qin, T., Zhao, S., Zhao, Z., & Liu, T.-Y. (2022, August). *FastSpeech 2: Fast and High-Quality End-to-End Text to Speech.* arXiv. Retrieved 2025-06-03, from http://arxiv.org/abs/2006.04558 (arXiv:2006.04558 [eess]) doi: 10.48550/arXiv.2006.04558

Sellier, H. (2024). Mobile Otome Games: Desire and Suspense as Economic Strategy. , *3*. Retrieved from https://shs.hal.science/halshs-04613757v1/document

Statista. (2021). *Distribution of video gamers in the united states from 2006 to 2021, by gender.* Retrieved from https://www.statista.com/statistics/232383/gender-split-of-us-computer-and-video-gamers/

Tower, S. (2024). *State of mobile gaming 2024.* Author. Retrieved from https://sensortower.com/state-of-gaming-2024

Wan, L., Wang, Q., Papir, A., & Moreno, I. L. (2020, November). *Generalized End-to-End Loss for Speaker Verification.* arXiv. Retrieved 2025-06-11, from http://arxiv.org/abs/1710.10467 (arXiv:1710.10467 [eess]) doi: 10.48550/arXiv.1710.10467

Wu, Y., Cai, W., & Mensah, S. A. (2024, August). "We Found Love": Romantic Video Game Involvement and Desire for Real-Life Romantic Relationships Among Female Gamers. *Social Science Computer Review*, *42*(4), 892–912. Retrieved 2025-06-03, from https://journals.sagepub.com/doi/10.1177/08944393231217940 doi: 10.1177/08944393231217940

Zhang, Z. (2024, July). The Rise of Otome Games in China: Exploring the Social and Psychological Traits of Otome Game Players. *Lecture Notes in Education Psychology and Public Media*, *59*(1), None–None. Retrieved 2025-06-03, from https://www.ewadirect.com/proceedings/lnep/article/view/14667 doi: 10.54254/2753-7048/59/20241720

Zhou, K., Sisman, B., Liu, R., & Li, H. (2022, February). Emotional voice conversion: Theory, databases and ESD. *Speech Communication*, *137*, 1–18. Retrieved 2025-06-10, from https://linkinghub.elsevier.com/retrieve/pii/S0167639321001308 doi: 10.1016/j.specom.2021.11.006