Voice Synthesis with Emotion and Style in Multilingual Support

Chaithra (Lilly) - M1261018, Khoa Nguyen - M1261026
陳効民 - M1261021, Hima - M1361016

1 Identifying the gap !!

Current TTS systems lack emotional depth, natural intonation, and multilingual support, limiting adaptability and customization for tone and style. With rising demand in industries like audiobooks and virtual assistants, innovative, adaptive voice solutions are essential.

2 Big Idea

Introducing a groundbreaking Multilingual Emotion and Style-Rich TTS System powered by XTTS-v2, where generative AI meets linguistic artistry. This system crafts emotionally resonant, adaptive speech with customizable tones, speeds, and styles, transforming experiences across education, healthcare, and entertainment into personalized and immersive journeys.

INPUT TEXT

coqui TTS

VOICE STYLE

spaCy

TorchAudio

librosa

3 Use Case Scenarios

Parent or caregiver seeking a soothing bedtime story for their child.

📖 Process

- Story Selection: Pick a bedtime story (e.g., Guess How Much I Love You).
- Customization: Adjust tone (calm, gentle) & style (slow, soft-spoken).
- Language Choice: Generate speech in English, Spanish, French, etc.
- Output: Play instantly or download as an MP3.

🌟 Outcome

Calming, empathetic narration engages the child, aiding relaxation and better sleep.

4 Methodology

STORY CONTENT

INCLUDES STORY CATEGORY (E.G., BEDTIME STORY, ADVENTURE, FAIRY TALE, SCI-FI).

PREPROCESSING

ANALYZE AND PREPARE TEXT FOR PROCESSING.
MAP SELECTED STYLE AND TONE PARAMETERS.

STYLE-INFUSED SPEECH GENERATION

coqui TTS

+

EMBED THE DIFFERENT MAPPED STYLE FOR DIFFERENT CATGEORY OF STORY INTO THE SPEECH SYNTHESIS MODEL.

VOICE CLONING AND CUSTOMIZATION

APPLY UNIQUE CHARACTER VOICES AND REFINE NARRATION STYLES.

AUDIOBOOK GENERATION

FINALIZE AS AN MP3 FILE OR STREAMING OUTPUT. DELIVER THE FINAL AUDIOBOOK READY FOR PLAYBACK OR DOWNLOAD.

5 Results !!

A fully synthesized audiobook based on text input.

Parameters:

- Story type (e.g., bedtime story, sci-fi).
- Emotional tone and style (e.g., calm, energetic).

Language preference and custom content features.

6 Future Work and Conclusion

- Developing Diverse Styles for Specific Story Categories.
- Adding Unique Voices for Story Characters.
- Seamless Character Voice Transitions.
- User-Defined Character Voice Customization.

This project leverages advanced generative AI (like XTTS-v2) and NLP to produce emotion-rich, stylelish, multilingual, and personalized speech, enhancing content delivery across education, healthcare, and entertainment. It supports dynamic voice cloning, scalable global solutions, and adaptive styles for maximum audience engagement.

長庚大學 CHANG GUNG UNIVERSITY
智慧運算學院 College of Intelligent Computing
人工智慧學系 Department of Artificial Intelligence