API Machine Learning Workshop

Eefje Karremans - s4181794

December 18, 2025

Abstract

This note evaluates Text-to-Speech (TTS) engines, comparing a Tacotron baseline with CapTTS and EmoCapTTS. Using caption based emotional synthesis,Assesment is on naturalness, intelligibility and expressiveness. CapTTS produces high quality emotional speech, without voice cloning. Captions moderately influence prosody without affecting duration,

1 Introduction

Recent TTS engines produce natural, intelligible speech while allowing voice selection and emotional expression. Applications include audiobooks and virtual assistants. This note evaluates TTS for expressiveness, intelligibility and runtime efficiency using a fixed evaluation text.

2 Methodology

Two expressive TTS engines were considered: **EmoVoice**[1] and **CapTTS**. EmoVoice emphasizes emotional control but lacks a standalone TTS script; so, CapTTS was used for experiments.

Baseline (A)

Tacotron2-DDC [2] was used to produce neutral speech from the evaluation text (*tacotron.wav*). This serves as a reference for naturalness and intelligibility.

CapTTS / EmoCapTTS (B & C)

CapTTS supports caption expressive synthesis. EmoCapTTS[3]extends this with emotional cues. Captions guide pitch, rate and energy while the transcript remains unchanged (*CaptTSS_Natural.wav*). Three captions were evaluated: i)"A comfortable narrator telling a brave and inspiring story" ii)"A female describing the worst pain of her life with tears in her eyes" iii)"Parents trying to talk to their toddler" Emotional outputs were saved as *CaptTSS_Emotion_i.wav*. Unfortunately, voice cloning from a reference recording is not supported.

3 Results

Natural speech (*CaptTSS_Natural.wav*) is clear, female, calm, with minor mumbles at 21s; present in Emotions.
Emotion 1 male, slower, mildly narrative, with final mumbles. **Emotion 2** female, weakly sad, minor mumbles (12s/21s). **Emotion 3** female, weak engagement, flat intonation, medium intelligibility.
Efficiency Models had similar runtime, indicating that Emo-model has minimal affect computational efficiency.

4 Conclusion

CapTTS and EmoCapTTS generate intelligible, caption based speech. Emotional captions influence prosody moderately but do not guarantee nuanced expression. Voice cloning is unsupported. These models are practical for expressive TTS applications. Subtle artifacts and limitations in emotional rendering or clarity shows possible improvements. github.com/EeveeCreations/ML_API

Voice	Gender	Expressiveness	Intelligibility
Natural	Female	Neutral	High, mumbles at 21s
Emotion_1	Male	Mild	High, mumbles at 21s
Emotion_2	Female	Weak sadness	High, minor mumbles at 12s/21s
Emotion_3	Female	Weak engagement	Medium, mumbles at 21s

Table 1: CapTTS evaluation across gender, expressiveness and intelligibility. Duration was unaffected.

References

[1] G. Yang, C. Yang, Q. Chen, Z. Ma, W. Chen, W. Wang, T. Wang, Y. Yang, Z. Niu, W. Liu, *et al.*, "Emovoice: Llm-based emotional text-to-speech model with freestyle text prompting," *arXiv preprint arXiv:2504.12867*, 2025.

[2] Resemble AI, "Chatterbox-TTS." `https://github.com/resemble-ai/chatterbox`, 2025. GitHub repository.

[3] H. Wang, J. Hai, D. Chong, K. Thakkar, T. Feng, D. Yang, J. Lee, L. M. Velazquez, J. Villalba, Z. Qin, S. Narayanan, M. Elhiali, and N. Dehak, "Capspeech: Enabling downstream applications in style-captioned text-to-speech," 2025.