Graduate School of Science and Technology Master's Thesis Abstract

Laboratory name (Supervisor)	Human-AI Interaction (Sakriani Sakti (Professor))		
Student ID	2411341	Submission date	2026 / 1 / 19
Name	ZHOU WANGZIXI		
Thesis title	Achieving Human-like Emotional Text-to-Speech with Personalized Perception and Fine-Grained Non-Verbal Control		

Abstract

Speech serves as an indispensable medium for human communication, conveying both verbal content and non-verbal vocal cues, such as emotion, attitude, and speaker identity. To enhance human-machine interaction, the pursuit of endowing AI with expressive vocal capabilities has driven decades of research in Text-to-Speech (TTS) synthesis. Historically, this field has focused primarily on naturalness, emphasizing voice quality and prosody to eliminate "robotic" artifacts. However, as AI systems become increasingly integrated into everyday life, expectations for synthetic speech are evolving. Users now demand more than intelligible text reading; they expect emotionally expressive and empathetic speech, necessitating a shift toward emotional TTS.

Although emotional TTS has achieved notable progress, it continues to face two fundamental challenges.

First, conventional TTS development has largely aimed to achieve naturalness for a general population. In contrast, emotional expression is inherently subjective and varies significantly across individuals and contexts. By overlooking this variability, most existing emotional TTS systems rely on generalized emotion labels, which amplify the subjectivity of emotion perception and fail to capture individual listeners' unique preferences and perceptual nuances.

Second, most TTS systems focus primarily on verbal vocalizations, while non-verbal vocalizations (NV) have historically been neglected. Although recent studies have begun to recognize the importance of NV, existing emotional TTS systems typically rely on coarse-grained labels—such as <laugh> or <crying>—that only specify the broad category of a vocalization. The lack of fine-grained control over specific acoustic properties, such as the rhythm of a sigh or the intensity of laughter, prevents current systems from modeling the subtle variations essential for authentic and life-like emotional expression.

This thesis addresses these limitations through two research studies:

Adaptive Personalized Emotional TTS:

To account for subjective emotion perception, we propose a Human-in-the-Loop framework based on an Interactive Genetic Algorithm (IGA). By incorporating user feedback as a fitness function, the system iteratively optimizes individualized Arousal-Valence (A-V) models. Experimental results demonstrate that this approach more accurately reflects personal and cross-cultural emotion perception than conventional one-size-fits-all models.

Fine-Grained Non-Verbal Expression Control:

We construct a fine-grained NV dataset by curating and reprocessing non-verbal utterances and introducing a new granular annotation scheme. Furthermore, we design a specialized NV processing pipeline consisting of Style, Discrete Unit, and Duration Parsers. These components enable precise modeling and control of fine-grained non-verbal cues, allowing for controllable synthesis that significantly enhances emotional expressiveness and recognition accuracy.

Overall, our results demonstrate that the proposed methods collectively advance the capabilities of emotional speech synthesis. Personalized learning effectively addresses subjective emotion perception, while fine-grained non-verbal control enhances emotional realism. This work contributes meaningfully to the development of human-like emotional TTS systems.