Towards Expressive Perception and Generation in Human-Computer Conversational Interaction

Yaohua Bu, Runnan Li, Zhenwei You
Beijing University of Posts and Telecommunications, Beijing, China
Email: buyaohua@bupt.edu.cn, runnan.li@bupt.edu.cn

The construction of a natural interactive human-computer interaction has become an integral component of intelligent system development, constituting a core subject within the field of human-computer interaction. Grounded in the utilization of human perception theories, this study proposes a robust method for speech emotion recognition and a model for inferring user emotional state changes, thereby achieving a human-computer interaction experience characterized by both listening and articulating, and conveying sentiments effectively. Simultaneously, focusing on the audio-visual modalities within human-computer interaction, this research explores methods for constructing interactive feedback in auditory and visual modalities within real human-computer dialog scenarios. This aims to synchronize textual, verbal, and visual expressions in human-computer interaction, enhancing its naturalness, improving user experience, and augmenting satisfaction and pleasure during interaction. The research contributions are as follows:

Introducing a method that robustly identifies user emotional states in real human-computer speech dialogue scenarios through the utilization of local and global attention mechanisms. This method generates robust and distinguishable representations of speech emotion, thereby enhancing the robustness and accuracy of speech emotion recognition systems in authentic dialogue scenarios.

- **Proposing a method for inferring user emotional state changes based on a structured multitask learning framework.** This method, built upon a structured multitask learning framework, jointly models the user's current emotional state and the emotional state changes induced by system feedback, leveraging the user's input speech information and system candidate feedback to infer intelligent user interaction experiences.

- **Presenting a method for expressive speech synthesis based on human speech production mechanisms and deep neural network structures.** This method integrates research and exploration into human speech production mechanisms, proposing the use of structured output layers to model dependencies between different acoustic parameters, thereby enhancing the predictive performance of acoustic models, improving the quality and naturalness of synthesized speech, and establishing corresponding methods for synthesizing expressive control. To further enhance synthesis quality, this study employs a language encoder based on prior knowledge and proposes a monotonicity enhancement method, effectively improving speech synthesis stability and quality through the combined use of stepwise monotonic attention mechanisms and multi-head attention mechanisms.
- **Investigating methods for driving virtual human animation based on speech and modeling techniques for virtual human neural rendering.** This research proposes a novel, robust, efficient, cross-lingual, and cross-speaker speech-to-animation method based on a hybrid expert algorithm. This method, taking speech modality as input, synchronously generates high-quality facial lip-driven animations with utmost efficiency during human-computer interaction. Furthermore, this study presents a method to enhance the robustness of neural renderers, reducing lip jitter in synthesized videos, improving visual quality and lip synchronization, and enhancing the realism of synthesized virtual human videos. Overall, this achieves controllable, stable visual feedback generation effects in human-computer interaction.

Keywords: Human-Computer Interaction; Multimodal; Speech Emotion Recognition; User Emotive State Changes Inferring; Expressive Speech Synthesis; Virtual Avatar Generation