Abstract

Speech emotion recognition (SER) systems aim to automatically recognize human emotional content in a speech signal, providing meaningful insights into a speaker's emotions, intentions, and mental state. Nowadays, these systems are utilized in various applications such as human-to-computer applications, human-to-human conversation monitoring in call centers or interview scenarios, and psychological assessments. Concerning the practical applicability, along with incrementally improving the classification accuracy across benchmarked databases, the SER research has also focused on improving the robustness of SER systems under different challenging conditions, such as noisy environments, multilingual settings, and cross-corpus domains. These issues are critical, but they are not unique to SER systems. They also exist in other speech-related tasks, such as automatic speech recognition, speaker recognition, and language recognition. Unlike speaker or language recognition, emotions in SER are complex and ambiguous, especially in real-world scenarios. Emotions require varying utterance durations for effective expression, adding complexity to the task. Moreover, emotion perception is highly subjective, varying significantly across individuals, making accurate classification even more challenging. Incorporating these emotion-specific characteristics into SER systems has the potential to make them more effective for real-world applications. However, these critical challenges remain largely unexplored in the existing SER literature.

In this thesis, by modelling the characteristics of human emotions, we aim to explicitly address crucial emotion-specific challenges for practical SER system development. We first investigate the training chunk-based duration-dependent biases in SER. To mitigate these biases, we propose a curriculum learning based approach for training SER systems, ensuring robust performance across varying utterance lengths. Secondly, we propose a mixed-emotion model and a multi-task learning framework to explore conventional SER systems beyond basic emotions

and improve robustness against elicitation mismatches between acted and spontaneous speech, respectively. Finally, we introduce a novel soft-label approach that accounts for emotion ambiguity and annotator expertise, improving SER's ability to handle subjective emotion perception.

Through extensive evaluations using different standard emotional speech datasets and features, this thesis demonstrates the effectiveness of the proposed methods in addressing SER-specific challenges. These contributions advance the development of SER systems that are more robust, flexible, and capable of understanding complex human emotions, facilitating their integration into practical applications.

Keywords: Speech emotion recognition, duration-dependent bias, mixed-emotion model, spontaneous elicitation, annotator subjectivity.