様式８

論　文　内　容　要　旨

報告番号	甲　先　第　458　号	氏　名	刘 铮
学位論文題目	Research on Speech Emotion Recognition Based on Machine Learning Approaches (機械学習手法に基づく音声感情認識に関する研究)		

内容要旨

With the continuous development of deep learning technology, research on human-machine emotional interaction has made significant progress. Speech is one of the critical means of information transmission, playing an indispensable role in human life by conveying not only semantic but also emotional information. In recent years, emotion recognition based on speech has received much attention due to the widespread application of deep learning technology and the rise of affective computing. Accurately extracting emotional information from speech signals is a vital issue in this field.

However, one of the significant challenges in developing high-performance speech emotion recognition systems is the lack of sufficient data. Constructing a high-quality emotional corpus requires a considerable investment of time and resources, including professional actors to perform voices with various emotions in specific scenes, and efficient data labeling to overcome the issue of imbalanced samples. Meanwhile, in order to improve the accuracy of speech emotion recognition, traditional methods only extract features from local datasets, which leads to overfitting of the models due to the weak robustness of the features. Therefore, constructing effective speech features is crucial to improving the accuracy of speech emotion recognition systems.

In this study, we propose an integrated active learning sampling strategy and an efficient framework to construct a speech emotional corpus effectively. Our method outperforms other active learning algorithms by improving sampling efficiency and selecting small category samples to be labeled with preference in imbalanced datasets. In actual corpus construction experiments, our method can prioritize selecting small class emotion samples, achieving an accuracy rate of 90% even with less than 50% of labeled data. This greatly enhances the efficiency of constructing the speech emotion corpus.

Additionally, we enhance the robustness of speech features using self-supervised learning and propose a feature fusion model (called Dual-TBNet) that consists of two 1D convolutional layers, two Transformer modules, and two bidirectional long short-term memory (BiLSTM) modules. Our model fuses five pre-trained features and acoustic features using the attention mechanism to capture the correspondence between the two features and enhance the contextual information of the fused features. In the comparison experiments, the Dual-TBNet model achieved a recognition accuracy and F1 score of 95.7% and 95.8% on the CASIA dataset, 66.7% and 65.6% on the eNTERFACE05 dataset, 64.8% and 64.9% on the IEMOCAP dataset, 84.1% and 84.3% on the EMO-DB dataset and 83.3% and 82.1% on the SAVEE dataset. The Dual-TBNet model effectively fuses acoustic features of different lengths and dimensions with pre-trained features, enhancing the robustness of the features, and achieved the best performance.