World Academy of Science, Engineering and Technology
International Journal of Electronics and Communication Engineering
Vol:18, No:10, 2024

Cross Attention Fusion for Dual-Stream Speech Emotion Recognition

Authors : Shaode Yu, Jiajian Meng, Bing Zhu, Hang Yu, Qiurui Sun

Abstract : Speech emotion recognition (SER) is for recognizing human subjective emotions through audio data in-depth analysis. From speech audios, how to comprehensively extract emotional information and how to effectively fuse extracted features remain challenging. This paper presents a dual-stream SER framework that embraces both full training and transfer learning of different networks for thorough feature encoding. Besides, a plug-and-play cross-attention fusion (CAF) module is implemented for the valid integration of the dual-stream encoder output. The effectiveness of the proposed CAF module is compared to the other three fusion modules (feature summation, feature concatenation, and feature-wise linear modulation) on two databases (RAVDESS and IEMO-CAP) using different dual-stream encoders (full training network, DPCNN or TextRCNN; transfer learning network, HuBERT or Wav2Vec2). Experimental results suggest that the CAF module can effectively reconcile conflicts between features from different encoders and outperform the other three feature fusion modules on the SER task. In the future, the plug-and-play CAF module can be extended for multi-branch feature fusion, and the dual-stream SER framework can be widened for multi-stream data representation to improve the recognition performance and generalization capacity.

Keywords : speech emotion recognition, cross-attention fusion, dual-stream, pre-trained
Conference Title : ICISPCS 2024 : International Conference on Intelligent Signal Processing and Communication Systems
Conference Location : London, United Kingdom
Conference Dates : October 17-18, 2024