World Academy of Science, Engineering and Technology
International Journal of Cognitive and Language Sciences
Vol:18, No:11, 2024

Multi-Modal Feature Fusion Network for Speaker Recognition Task

Authors : Xiang Shijie, Zhou Dong, Tian Dan

Abstract : Speaker recognition is a crucial task in the field of speech processing, aimed at identifying individuals based on their vocal characteristics. However, existing speaker recognition methods face numerous challenges. Traditional methods primarily rely on audio signals, which often suffer from limitations in noisy environments, variations in speaking style, and insufficient sample sizes. Additionally, relying solely on audio features can sometimes fail to capture the unique identity of the speaker comprehensively, impacting recognition accuracy. To address these issues, we propose a multi-modal network architecture that simultaneously processes both audio and text signals. By gradually integrating audio and text features, we leverage the strengths of both modalities to enhance the robustness and accuracy of speaker recognition. Our experiments demonstrate significant improvements with this multi-modal approach, particularly in complex environments, where recognition performance has been notably enhanced. Our research not only highlights the limitations of current speaker recognition methods but also showcases the effectiveness of multi-modal fusion techniques in overcoming these limitations, providing valuable insights for future research.

Keywords : feature fusion, memory network, multimodal input, speaker recognition
Conference Title : ICSLP 2024 : International Conference on Spoken Language Processing
Conference Location : Bangkok, Thailand
Conference Dates : November 25-26, 2024