THE USE OF LIP INFORMATION FOR ROBUST SPEAKER IDENTIFICATION AND SPEECH RECOGNITION

Ertan Cetingul, Engin Erzin, Yucel Yemez and A. Murat Tekalp

ABSTRACT

This study investigates the benefits of multimodal fusion of audio, lip motion and lip texture modalities for speaker and speech recognition. The audio modality is represented by the well-known mel-frequency cepstral coefficients (MFCC) along with the first and second derivatives, whereas lip texture modality is represented by the 2D-DCT coefficients of the luminance component within a bounding box about the lip region. A new lip motion modality representation based on discriminative analysis of the dense motion vectors within the same bounding box is employed for speaker/speech recognition. The fusion of audio, lip texture and lip motion modalities is performed by the so-called Reliability Weighted Summation (RWS) decision rule. Experimental results show that inclusion of lip motion and lip texture modalities provides further performance gains in both speaker identification and speech recognition scenarios.