Audio-visual automatic speech recognition and related bimodal speech technologies: A review of the state-of-the-art and open problems

Gerasimos Potamianos

Institute of Informatics and Telecommunications
National Centre for Scientific Research "Demokritos"
GR-15310 Athens, Greece

gpotam@iit.demokritos.gr

Abstract

The presentation will provide an overview of the main research achievements and the state-of-the-art in the area of audio-visual speech processing, mainly focusing in the area of audio-visual automatic speech recognition. The topic has been of interest in the speech research community due to the potential of increased robustness to acoustic noise that the visual modality holds. Nevertheless, significant challenges remain that have hindered practical applications of the technology most notably difficulties with visual speech information extraction and audio-visual fusion algorithms that remain robust to the audio-visual environment variability inherent in practical, unconstrained interaction scenarios and audio-visual data sources, for example multi-party interaction in smart spaces, broadcast news, etc. These challenges are also shared across a number of interesting audio-visual speech technologies beyond the core speech recognition problem, where the visual modality has the potential to resolve ambiguity inherent in the audio signal alone; for example, speech activity detection, speaker diarization, and source separation.