SPEAKER SOURCE LOCALIZATION USING AUDIO-VISUAL DATA AND ARRAY PROCESSING BASED SPEECH ENHANCEMENT FOR IN-VEHICLE ENVIRONMENTS

Xianxian Zhang, John H. L. Hansen,
Kazuya Takeda, Toshiki Maeno, and Kathryn Arehart

ABSTRACT

Human-Computer interaction for in-vehicle systems requires effective audio capture, tracking of who is speaking, environmental noise suppression, and robust processing for applications such as route navigation, hands-free mobile communications, and human-to-human communications for hearing impaired subjects. In this paper, we consider two interactive speech processing frameworks for in-vehicle systems. First, we consider integrating audio-visual processing for localization the primary speech for a driver using a route navigation system. Integrating both visual and audio content allows us to reject unintended speech to be submitted for speech recognition within the route dialog system. Second, we consider a combined multi-channel array processing scheme based on a combined fixed and adaptive array processing scheme (CFA-BF) with a spectral constrained iterative Auto-LSP and auditory masked GMMSE-AMT-ERB processing for speech enhancement. The combined scheme takes advantage of the strengths offered by array processing methods in noisy environments, as well as speed and efficiency for single channel methods. We evaluate the audio-visual localization scheme for route navigation dialogs and show improved speech accuracy by up to 40% using the CIAIR in-vehicle data corpus from Nagoya, Japan. For the combine array processing and speech enhancement methods, we demonstrate consistent levels of noise suppression and voice communication quality improvement using a subset of the TIMIT corpus with four real noise sources, with an overall average 26dB increase in SegSNR from the original degraded audio corpus.