Topic: District Names Speech Corpus for Urdu ASR

Presenter: Sahar Rauf

Presentation Date: 15th March

Venue: KICS Seminar Hall

Abstract:

Speech corpus is a collection of audio recordings that is a necessary element to build the ASR systems. ASR is a promising field of research and a part of services related to healthcare, agriculture, weather forecasting and mobile applications. The speech corpus can be a good source of capturing variability occurred due to age, gender, dialect, background noise and language of a speaker. The purpose is to present a speech corpus that is developed for Urdu automatic speech recognition (ASR) system. The corpus comprises of single word utterances fixed vocabulary consisting of district names of Pakistan. The data is recorded over a telephone channel from all over Pakistan to cover six major accents; Punjabi, Urdu, Saraiki, Pashto, Sindhi, and Balochi. The data was collected in challenging acoustic environments; the major issues were silence, background noise and alternate pronunciations, which can affect the performance of the system. In order to address these issues, comprehensive data verification and cleaning guidelines are proposed. The proposed process serves as a data pre-processing step for the development of ASR, which is successfully integrated in an Urdu dialog system to provide weather information of Pakistan. A Pakistan map is also presented on the Center for Language Engineering (CLE) website that presents the color coded information of different districts from where the data was collected.