Harvesting Speech Datasets for Linguistic Research on the Web

Mats Rooth (Cornell University)

Spoken language is growing explosively on the web. Though generic spoken language search is hardly available, some sites index spoken language using speech recognition, and it is possible to run off-the-shelf speech recognition in a laboratory setting. This talk will present methodology for creating single-phrase web datasets, which consist of multiple utterances of a single short word sequence. A workflow created at the Cornell Computational Linguistics Lab and the McGill Prosody Lab consists of web harvest of underlying data, identification of true tokens and transcription in a web database interface, phoneme alignment using an HMM aligner, and subsequent acoustic and linguistic analysis. The procedure results in large, diverse, and naturalistic datasets that make it possible to re-examine issues such as the acoustic form and contextual conditioning of prosody.