

Focus Contrast in Web Harvested Data

Mats Rooth

The talk presents a method for collecting data that bear on the phonetics-phonology and semantics-pragmatics of intonation from web sources. The targets are phrases like “than I did”, where intonation varies in a way which correlates with grammatical and pragmatic context, and common phrases with inherent contrastive potential such as “in my opinion” and “on the tenor saxophone”. Audio files and time offsets in them are identified using a search engine based on automatic speech recognition, retrieved, and cut to a relevant segment under program control. An SVM machine learning classifier is trained to make a grammatically determined distinction in intonation based on purely acoustic cues.