IARPA Babel Program

Stavros Tsakalidis, Frantisek Grezl

SLTU 2016

Activities on Under-resourced Languages

Motivation:

- Speech recognition technology has limited coverage of languages
- Much of the Speech Technology Research had been conducted on English with substantial amounts of data collected
- Need capability to quickly train speech models in any language with potentially noisy, limited resources

Activities:

- Collected and transcribed about 100 hours of speech from 27 under-resourced languages (e.g. Guarani, Igbo, Dholuo, Tok Pisin)
- Developed methods for building high-performing ASR and KWS systems for any new language in one week, given limited data and linguistic resources in that language

Lessons Learned

Obstacles

- No major obstacles mainly due to efficient Program Management by the Babel PM, Mary Harper
- Untransribed audio size was small (40-60 hrs) to have an impact on semi-supervised training methods

Recipe for success

- Data for each year was available at the beginning of each year
- Work with diverse languages (27 total)
- Real recording conditions
  - Acquire speech data in-country
  - Acquire data in a variety of conditions (e.g., in café, on the street) with different recording devices (e.g., cell phone, hands free microphone) from a variety of dialects
- Rigorous evaluation
  - Use a “surprise” language for evaluation each period
  - Study multiple languages each program period

Collaboration

- Offer to the community ability to extract multilingual bottleneck features
  - Allows to train robust models for a new language with much less or even zero acoustic training data
- Make the data available to the community and pool resources across under-resource project (similar to the OPUS project for parallel corpora for MT)
- Offer Speech-to-Speech Translators to organizations involved with relief efforts (e.g. Haiti earthquake, Syrian refuge crisis)