VAST: A Corpus of Video Annotation for Speech Technologies

Jennifer Tracey and Stephanie Strassel
{garjen,strassel}@ldc.upenn.edu

Linguistic Data Consortium

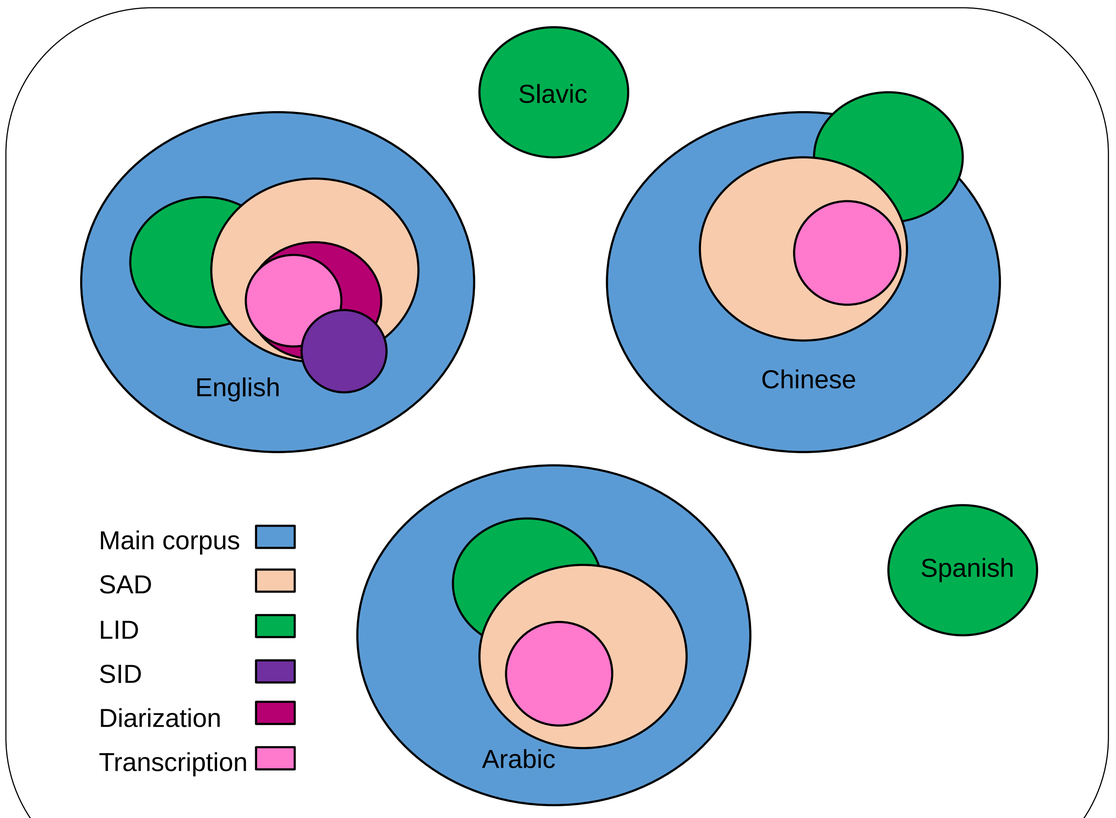
Video Annotation for Speech Technologies (VAST)

- Video content covering a diverse range of communication domains, data sources and video resolutions
- For use in training, development and testing of multiple speech technologies
- ~2900 hours of data in three primary languages (English, Mandarin Chinese and Arabic) plus 7 additional languages/dialects
- Portions of the collected data were annotated for speech activity, speaker identity, speaker sex and language, diarization, and transcription

Features of the corpus

- Videos must contain speech in one of the three primary languages
- Any variety/dialect of Arabic or English is acceptable for the main corpus, while Chinese videos must contain Mandarin
- Multi-party, informal speech is preferred over monologs, telephone-style dialogs or interviews
- There is no restriction on topic (variety of topics preferred)
- Speaker(s) are not required to appear on camera

Feature	Percent of files
Indoor setting	59%
Outdoor setting	48%
Single speaker	25%
Two speakers	24%
Three or more speakers	51%
Background noise	67%



Relationship among corpus components

Speech Activity Annotation

- Speech and music segments for a subset of Arabic, Chinese, and English data
- Speech segments labeled for speaker sex and language
- 187 hours of English, 197 hours of Arabic, and 280 hours of Chinese

Diarization

- Distinct SAD segments for each individual speaker in 43 hours of English data

Speech Segments → English Speaker Segments

Transcription

- 30 hours of English, 40 hours of Iraqi Arabic, 50 hours of Egyptian Arabic, and 29 hours of Chinese
- For the Egyptian and Iraqi transcription, guidelines provided for standardized spelling of the dialectal varieties

Language Identification

- 11-23 hours of data for each language/variety
- May include codeswitching

English cluster	Slavic cluster	Chinese cluster
British English Gen. American English	Polish Russian	Mandarin Min Nan
Arabic cluster	**Spanish cluster**	
Egyptian Arabic Iraqi Arabic Levantine Arabic Maghrebi Arabic Gulf Arabic	Caribbean Spanish European Spanish Latin American Spanish Brazilian Portuguese	

Speaker Identification

- 2-10 videos from each of 300 English speakers from amateur videos
- Aim for a variety of interlocutors, speaking styles and acoustic/physical environments in the cluster for each speaker
- Speaker information applied at the file level (no segment-level speaker ID)
- Effort was made to include 1-2 files per speaker in the diarization set

Data Availability

- Portions of the VAST data have been used in the NIST 2017 (Pilot) Speech Analytic Technologies Evaluation (NIST 2017a) and in the 2017 NIST Language Recognition Evaluation (NIST 2017b)
- First public data releases: subset of Arabic, Chinese, and English SAD annotation, Chinese transcription