VAST: A Corpus of Video Annotation for Speech Technologies

Jennifer Tracey, Stephanie Strassel
Linguistic Data Consortium, University of Pennsylvania
3600 Market Street, Suite 810, Philadelphia, PA. 19104 USA
{garjen, strassel}@ldc.upenn.edu

Abstract

The Video Annotation for Speech Technologies (VAST) corpus contains approximately 2900 hours of video data collected and labeled to support the development of speech technologies such as speech activity detection, language identification, speaker identification, and speech recognition. The bulk of the data comes from amateur video content harvested from the web. Collection was designed to ensure that the videos cover a diverse range of communication domains, data sources and video resolutions and to include three primary languages (English, Mandarin Chinese and Arabic) plus supplemental data in 7 additional languages/dialects to support language recognition research. Portions of the collected data were annotated for speech activity, speaker identity, speaker sex, language identification, diarization, and transcription. A description of the data collection and each of the annotation types is presented in this paper. The corpus represents a challenging data set for language technology development due to the informal nature of the majority of the data, as well as the variety of languages, noise conditions, topics, and speakers present in the collection.

Keywords: speech corpora, video corpora, multilingual resources

1. Introduction

The Video Annotation for Speech Technologies (VAST) corpus is the result of an effort to collect 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. The corpus comprises approximately 2900 hours of data, targeting three primary languages (English, Mandarin Chinese and Arabic) plus supplemental data in 7 additional languages/dialects to support language recognition research. Portions of the collected data were annotated for speech activity, speaker identity, speaker sex and language, diarization, and transcription. The collection of English, Mandarin, and Arabic data is referred to here as the “main corpus,” and the data from supplemental languages as well as the annotations are referred to as the “sub-corpora.”

2. Main corpus

Videos in the main corpus contain speech in English (including US, UK, and other varieties), Mandarin Chinese, and Arabic. Videos are assigned a primary language designation, but some videos may contain more than one language due to codeswitching in naturally occurring data. The majority of data in the VAST corpus consists of amateur videos harvested from the Internet. In addition, a small amount of audio from broadcast television news and/or informal talk shows is included in the main corpus. No annotations were performed on the broadcast data, and the discussion of the data in this paper therefore focuses primarily on the amateur video. Criteria for inclusion of videos in the main corpus are as follows:

- 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

Annotators hired and trained by LDC performed a "data scouting" task, in which they searched the web for appropriate content. During a given work session data scouts were instructed to search for videos appropriate for inclusion in the main corpus, or else they were instructed to do more focused searching for videos suitable for the SID or LID sub-corpora. Data scouting was conducted using a customized user interface developed for VAST, known as VScout. The VScout toolkit is a Firefox add-on consisting of an annotation form displayed on the left side of the browser window. Data scouts use the browser in the usual way to search, navigate video websites, and watch videos. When they find a suitable video they fill out the VScout webform, and the results are logged to a database. The data scouting process results in the following information about each video: page URL, number of speakers (1, 2, or 3+), sound conditions (background noise/speech, outdoors/indoors), speaker overlap, and language.

The average duration of amateur video clips is approximately three minutes, and Table 1 below provides a summary of some of the other features of the data. The features summarized are those that were noted during the data scouting process. Note that a video may contain both indoor and outdoor settings, so the percentages sum to greater than 100%. As can be seen in the summary, the majority of videos have some amount of background noise, and just over half have three or more speakers, making this a challenging dataset for annotation and transcription.

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

Table 1: Summary of Data Features

Language	Source Data	SAD	SID	LID	Diarization	Transcription
Arabic	818	197	0	91	0	91
English	768	187	99	32	43	32
Mandarin	720	280	0	20	0	29
Min Nan	11	0	0	11	0	0
Spanish	63	0	0	63	0	0
Portuguese	21	0	0	21	0	0
Russian	23	0	0	23	0	0
Polish	20	0	0	20	0	0
Total	**2444**	**664**	**99**	**281**	**43**	**152**

Table 2: Hours of amateur video in VAST corpus by subcorpus

3. Sub-corpora

Portions of the corpus were selected for inclusion in one or more sub-corpora: language ID (LID), speaker ID (SID), speech activity detection (SAD), diarization, and transcription. Except for the LID sub-corpus, all videos included in the sub-corpora were selected from the main corpus; the LID sub-corpus consists of some files selected from the main corpus as well as some additional files collected in languages not included in the main corpus. Videos that are extremely difficult for humans to annotate (due to noise conditions, number of overlapping speakers, etc.) were avoided when possible; however, limited availability of videos in some languages/dialects required inclusion of some of these more challenging videos in order to meet data volume targets. For each task, annotators were given guidelines and training, and were not permitted to perform the task in the production pipeline until task coordinators were satisfied with their level of competence. All tasks were performed by LDC annotators, with the exception of transcription, some of which was performed by LDC annotators and some of which was performed by external transcription vendors. Table 2 shows the data volume (hours of amateur video) in each subset of the corpus. As discussed in the sections that follow, some data appears in multiple sub-corpora.

3.1 SAD Sub-corpus

A portion of the data from the main corpus was annotated for speech activity detection (SAD), which included distinguishing speech from music, as well as labeling speech segments for language and speaker sex. Annotators review the output of an automatic SAD system or created segments from scratch (correction of automatic output proved less efficient than fully manual segmentation in many cases, so annotators were permitted to ignore automatic output). During this manual SAD correction and segmentation pass, annotators distinguished three categories:

- Non-speech. This category includes silence and non-vocal background noise. It may also include very short duration filled pauses or other speaker vocalizations that occur in isolation during a lengthy period of non-speech or music and were not detected by the automatic SAD pass.
- Speech. This category includes all speech including discernable background speech, backchannels, filled pauses, non-lexemes, laughter, coughing and all other speaker vocalizations.
- Music. This category includes vocal and non-vocal music: sung, instrumental, or rapped music, as well as rhythmic or chanted slogans. Music is sound which is intentionally produced to create melody (carrying a tune) and/or rhythm.

Non-speech was not explicitly segmented or labeled; any audio that is not contained within a segment is considered non-speech.

All speech segments are contained within a single "speech track"; during the SAD task annotators did not create separate tracks for each speaker's speech. Similarly, all music is contained within a single "music track". Speech segments and music segments may overlap and are maximally long and maximally inclusive. In other words, if the recording contains continuous speech for 15 seconds with no noticeable pauses, annotators created a segment on the speech track whose duration is 15 seconds – even if the number of speakers or the languages spoken changed during that segment.

While SAD annotation primarily relies on the audio recording, annotators were permitted to consult the video for help in disambiguating difficult cases. If the video and audio presented conflicting information, the annotator relied primarily on the audio. SAD annotation was performed on a total of 187 hours of English, 197 hours of Arabic, and 280 hours of Chinese.

3.1.1 Speaker Sex Labeling

Concurrent with the manual SAD correction and segmentation task, annotators also labeled each speech segment with respect to speaker sex. Segments were labeled as male (one or more speakers, all male), female (one or more speakers, all female), mixed (both male and female speakers), or unknown (speaker's sex cannot be determined).

3.1.2 Light LID Labeling

In addition to speaker sex, annotators also labeled each speech segment with respect to language. Labels are target (segment contains only target language), non-target (segment does not contain target language), mixed (segment contains target language plus one or more other languages), or unknown (language of the segment cannot be determined). For this task, "target language" is defined as the expected predominant language of the recording based on the source data auditing task.

3.2 Speaker Diarization Sub-corpus

To help address interest in speech processing in multi-speaker data, LDC performed speaker diarization on a subset of 43 hours from the English portion of the main corpus. This annotation results in distinct SAD segments for each individual speaker within this subset of data. Diarization was performed on speech segments only (i.e., voices of singers in music segments were not included), and the speaker sex and light LID labels from the SAD task were not applied to the diarized segments.

3.3 Transcription Sub-corpus

A portion of the main corpus was selected for transcription in each of the primary languages The Transcription sub-corpus includes 30 hours of English, 40 hours of Iraqi Arabic, 50 hours of Egyptian Arabic, and 29 hours of Mandarin Chinese, transcribed in a Quick Rich transcription style (Glenn et al., 2010; Bendahman et al., 2008). Selection of videos for transcription favored data with a high ratio of speech to non-speech. We also tried to maximize the number of distinct speakers in this sub-corpus. The English files for transcription were selected from the Speaker Diarization sub-corpus described above. For the Egyptian and Iraqi transcription efforts, guidelines for standardized spelling of the dialectal varieties were taken from previous transcription projects involving these dialects undertaken at LDC (Maamouri et al., 2004; Habash et al., 2012). The transcription was carried out in at least two passes, with the first pass focusing on correct verbatim transcription and speaker identification. The second pass focused especially on correcting use of transcription conventions for special categories like proper names, as well as adherence to the standardized orthography for the Egyptian and Iraqi Arabic transcription.

3.4 SID Sub-corpus

The SID sub-corpus contains multiple videos from each of 300 English speakers. For each speaker, 2-10 videos containing that speaker’s voice were collected, with the aim of including a variety of interlocutors, speaking styles and acoustic/physical environments in the cluster for each speaker. The SID judgment was applied at the file level; that is, no segment-level annotation of speaker ID was applied. Each cluster of videos was audited to verify that the target speaker’s voice occurs in each of the videos and that there is sufficient diversity of interlocutors and environments in the cluster. In the delivered corpus, videos in a cluster are marked "yes" or "no" for inclusion as the "core" cluster. Files included in the core are considered sufficiently diverse to form a valid cluster. Files marked "no" for inclusion in the core cluster are additional files containing the target speaker, but were found to be redundant with other videos in the cluster with respect to either interlocutors or environment. All clusters have a minimum of two videos in the "core". An effort was made to include 1-2 videos from each SID cluster in the Speaker Diarization corpus, but some clusters were collected too late in the project to be included in the Diarization annotation.

3.5 LID Sub-corpus

The LID sub-corpus contains approximately 11-23 hours of data for each of 15 languages (or language varieties), for a total of roughly 280 hours.

Languages were selected to match the language clusters used in the NIST LRE 2015 evaluation. In general, the VAST LID languages are a subset of the languages used in LRE 2015, with the exception of Gulf Arabic, which was added in VAST due to its substantial presence in the Arabic data. We included all languages from the LRE list where at least 10 hours of video were collected, plus Gulf

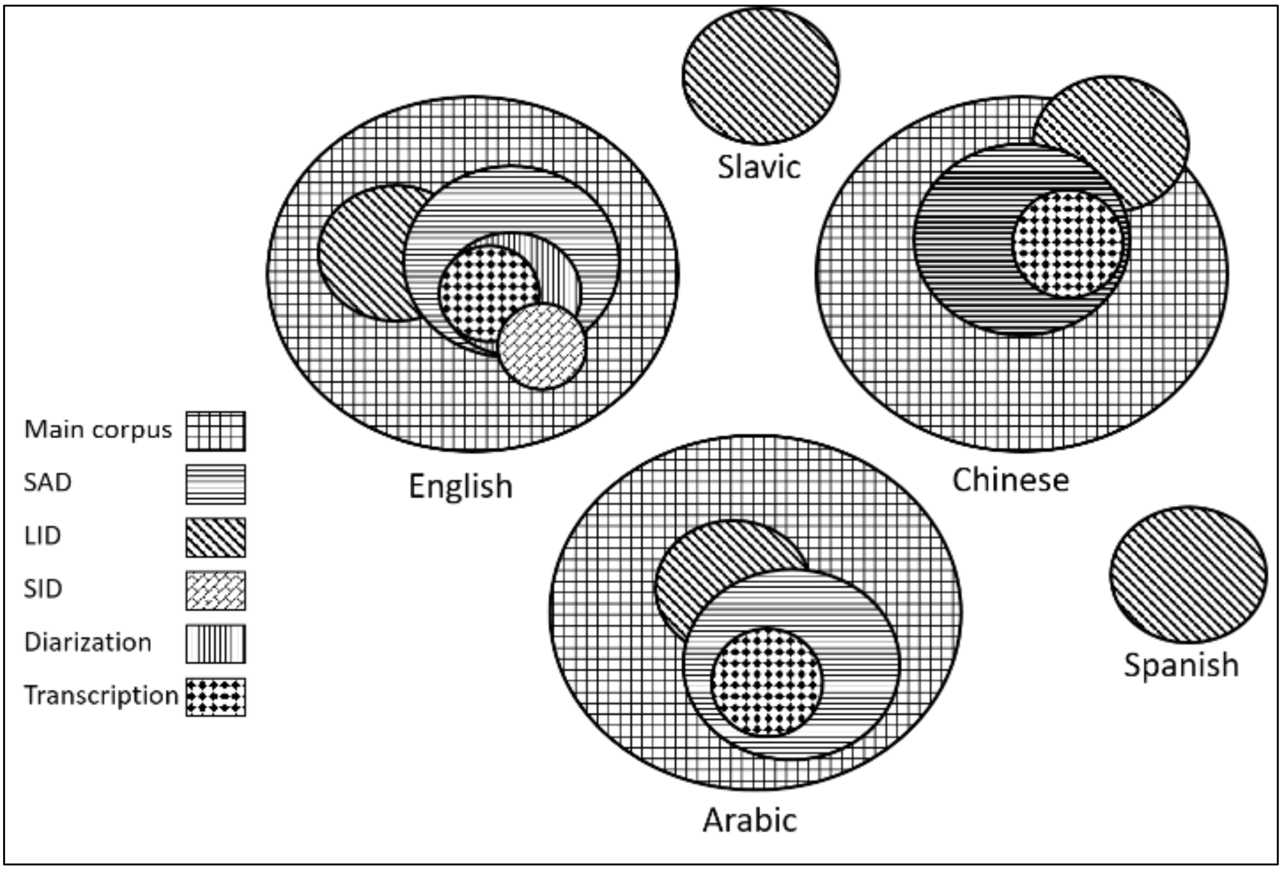

Figure 1: Relationship among VAST Sub-Corpora

Arabic. The final clusters for VAST are shown in Table 2. All included languages have at least 20 hours of data except for Maghrebi and Gulf Arabic (15 hours each), Min Nan (11 hours), and British English (12 hours). Videos designated as part of the LID corpus contain some speech in the target variety, preferably more than 50% of the speech in the video; however, some videos included may have lower proportions of target language variety speech.

<table>
<tr><th>Arabic Cluster</th><th>Slavic Cluster</th><th>Chinese Cluster</th></tr>
<tr><td>Egyptian Arabic
Iraqi Arabic
Levantine Arabic
Maghrebi Arabic
Gulf Arabic</td><td>Polish
Russian</td><td>Mandarin
Min Nan</td></tr>
<tr><th>English Cluster</th><th colspan="2">Spanish Cluster</th></tr>
<tr><td>British English
Gen. American English</td><td colspan="2">Caribbean Spanish
European Spanish
Latin American Spanish
Brazilian Portuguese</td></tr>
</table>

Table 3: VAST LID Language Clusters

3.6 Relationship among the Sub-corpora

SAD, Diarization, Transcription, and Speaker ID sub-corpora are all subsets of the main corpus, where Diarization and Speaker ID come from the English portion of the main corpus only, and SAD and transcription files are taken from all three primary languages in the main corpus. The LID sub-corpus contains some files from the main corpus (all varieties in the Arabic and English clusters, and Mandarin in the Chinese cluster), along with some files that are not part of the main corpus (Min Nan from the Chinese cluster, as well as the Slavic and Spanish clusters). With a very few exceptions, all Diarization and Transcription files are a subset of the SAD sub-corpus; English Transcription files are a subset of the files in the Diarization sub-corpus. In addition, there is partial overlap between the SID sub-corpus and the Diarization and English Transcription sub-corpora, as well as between the English, Chinese, and Arabic clusters of the LID corpus and the SAD and Transcription Corpora. Figure 1 illustrates the interactions between the various sub-corpora.

4. Availability of VAST Data and Future Work

To date, 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). Additional annotation of the VAST corpus is current in progress, including additional SAD annotation on data that may be used in future test sets, as well as some annotation of video features for a subset of the data. While some portions of the VAST corpus are being withheld for use in future Open SAT, SRE, and LRE evaluations, specific sub-corpora will appear in LDC's public catalog starting in 2018. The first releases will include approximately 2000 files per language (Arabic, Chinese, and English) with SAD annotation, as well as the 29 hours of Chinese transcription data.

5. 6. Bibliographical References

Bendahman, C., Glenn, M., Mostefa, D., Paulsson, N., Strassel, S. (2008). Quick Rich Transcriptions of Arabic Broadcast News Speech Data. In Nicoletta Calzolari (Conference Chair), et al., editors, Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08), pages 3605–3608, Marrakech, Morocco, May 28-30. European Language Resource Association (ELRA).

Glenn, M., Strassel, S., Lee, H., Maeda, K., Zakhary, R., Li, X. (2010). Transcription Methods for Consistency, Volume and Efficiency. In Nicoletta Calzolari (Conference Chair), et al., editors, Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10), pages 2915–2920, Valletta, Malta, May 17-23. Workshop on Language Resources and Human Language Technologies for Semitic Languages. European Language Resource Association (ELRA).

Habash, N., Diab, M., Rambow, O. Conventional Orthography for Dialectal Arabic. In Nicoletta Calzolari (Conference Chair), et al., editors, Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'12), pages 711–718, Istanbul, Turkey, May 21-27. European Language Resource Association (ELRA).

Maamouri, M., Buckwalter, T., Cieri, C. Dialectal Arabic Telephone Speech Corpus: Principles, Tool Design, and Transcription Conventions. In Proceedings of NEMLAR International Conference on Arabic Language Resources and Tools, Cairo, Egypt, September 22-23.

NIST. (2017a). 2017 (Pilot) Speech Analytic Technologies Evaluation. https://www.nist.gov/itl/iad/mig/nist-2017-pilot-speech-analytic-technologies-evaluation. Accessed February, 2018.

NIST. (2017b). 2017 NIST Language Recognition Evaluation. https://lre.nist.gov/. Accessed February, 2018.