The Broadcast Narrow Band Speech Corpus: A New Resource Type for Large Scale Language Recognition

Christopher Cieri, Linda Brandschain, Abby Neely, David Graff, Kevin Walker, Chris Caruso¹, Alvin Martin, Craig Greenberg²

¹Linguistic Data Consortium, University of Pennsylvania
²National Institute of Standards and Technology, Gaithersburg, MD.

{ccieri,brandschain,neeley,graff,walker,carusoc}@ldc.upenn.edu,
{alvin.martin,craig.greenberg}@nist.gov

Abstract

This paper describes a new resource type, broadcast narrow band speech for use in large scale language recognition research and technology development. After providing the rational for this new resource type, the paper describes the collection, segmentation, auditing procedures and data formats used. Along the way, it addresses issues of defining language and dialect in found data and how ground truth is established for this corpus. Index Terms: multilingual speech corpora, language recognition, language identification, language detection, language, dialect, mutual intelligibility, broadcast news, conversational speech

1. Introduction

The NIST Language Recognition (LRE) campaigns began in 1996 to establish a baseline for technology performance where the task was to determine whether a target language was spoken in a sample of conversational telephone speech (CTS). After a hiatus, the campaigns restarted in 2003 and have continued biennially since then. Until recently, LRE has relied upon collections of CTS, originally the 15 CALLFRIEND corpora published by LDC (LDC 1996). However, over the years CTS has proven expensive to collect given the need for a many speakers in many languages who each contribute relatively little speech, parameters that poorly match the traditional compensation schemes for human subjects. This paper describes the creation of a multilingual corpus of broadcast narrow band speech (BNBS), typically telephone speech embedded in news broadcast, as a cost-effective alternative to CTS to support language recognition technology development and evaluation in particular in the 2009 NIST Language Recognition Evaluations (LRE), the relative costs of collection and the breadth of languages for which BNBS is available made it worthwhile to test the hypotheses that this data will be useful for LRE and early results from LRE 2009 have been encouraging.

The Voice of America “is a multimedia international broadcasting service funded by the U.S. Government through the Broadcasting Board of Governors. VOA broadcasts approximately 1,500 hours of news, information, educational, and cultural programming every week” (VOA 2009) via satellite and AM, FM and shortwave radio frequencies and also distributes programming via the Internet in streaming media and downloadable formats. VOA currently broadcasts in more than 45 languages or linguistic varieties that the service labels: Afan Oromo, Albanian, Amharic, Armenian, Azeri, Bangla, Bosnian, Burmese, Cantonese, Chinese, Creole, Croatian, Dari, English, French, Georgian, Greek, Hausa, Hindi, Indonesian, Khmer, Kinyarwanda, Kirundi, Korean, Kurdi, Kurdish, Lao, Macedonian, Mandarin, Ndebele, Pashto, Persian, Portuguese, Russian, Serbian, Shona, Somali, Spanish, Swahili, Thai, Tibetan, Tigrigna, Turkish, Ukrainian, Urdu, Uzbek, Vietnamese. The languages and number of hours of broadcast per language, vary according to the priorities of the United States government.

In 1996, United States Public Law 104-269 authorized Voice of America to cooperate with LDC in sharing programming for use in research, education and technology development. LDC began collecting VOA directly from the service’s internal broadcast distribution network in 1998 and has, to date, collected more than 17,000 hours in more than 70 languages. The VOA data in this corpus have been collected at LDC in three distinct efforts designated VOA1, VOA2 and VOA3 and described below. Since the BNBS Corpus is planned to support blind language recognition technology

10.21437/Interspeech.2009-732
evaluations, some details regarding the languages included and the collection epoch are withheld.

During the VOA1 epoch, multi-channel MPEG-1 level II audio streams were captured via dedicated satellite dish, de-multiplexed, and converted to digital audio by dedicated hardware designed to match other VOA downlink and broadcasting sites around the world. Pairs of digital audio channels were fed into Townsend DatLINK A/D converters, connected via SCSI to Sun Sparc workstations where scripts controlled the scheduling of audio capture on up to 12 channels at a time, following broadcast time-tables provided by VOA. Audio was digitized at 16 kHz, into 16-bit PCM samples, and stored to disk in NIST SPHERE file format.

After initial capture, VOA1 files were audited manually to confirm the success of the recordings and to determine the start and end points of individual broadcasts. Based on the audits, segments of various lengths (typically 15, 30 or 60 minutes) were extracted from the initial capture files, and stored as separate, single-channel, single-language programs; these files were then compressed using the lossless shorten (Wikipedia 2009) algorithm as implemented in the NIST SPHERE utility w_encode (NIST 2009b), and archived. During the VOA2 epoch, the broadcasts were recorded directly from the VOA network at 48kHz. The original files are MPEG-2 audio in transport stream container. The stream is made up of ~0.024 second audio frames.

The majority of VOA3 data were collected initially from the VOA website by the Speech@FIT group in the Faculty of Information Technology at the Brno University of Technology in the Czech Republic (hereafter, Brno). A small number of more recent broadcast recordings were subsequently added by LDC. Recent VOA recordings are publicly available on VOA’s website. However, VOA removes older broadcasts from its site making it impossible to completely recreate VOA3 simply by re-harvesting that site. There is some variation in the audio files types on the VOA site. Most files were 11K, stereo files with an mpeg2.5 audio ID; a smaller number were 22K mono files with an mpeg2.0 audio ID. About 1% of the files were of other types including sampling rates up to 48K and mpeg1.0 audio IDs.

3. Segment Definition and Selection

In general, radio broadcasts may contain music, brief periods of silence, commercials, news and features. The news related speech may include studio reporting, interviews and group discussions, correspondent reports, call-ins and previously recorded speech. Where there is reasonable expectation that the audience is multilingual the broadcast may also be multilingual. Broadcasts may even include speech in languages not known to the target audience, for example when interpretation is provided. In addition to these general features, many VOA broadcasts begin with an announcements in English stating that the programming is VOA and indicating the language to follow. For these reasons, raw broadcast speech must be processed, segmented and audited (annotated) before it can provide reliable data for system training and ground truth for system evaluation.

Segments are defined using a multistep process. First the audio files are processed by the Brno speech activity detection engine that distinguishes speech from music and silence. The same files are then processed by software, also provided by Brno, that distinguishes narrow band from wide band signal. A third process then consolidates the output of these two processes to select the largest sub-segments wherever a segment of speech overlaps with a segment of narrow band signal. Where these new segments are less than 33 seconds they are discarded. Specifically we attempt no concatenation of smaller, perhaps non-contiguous, sub-segments in order to build up 33 second segments of narrowband speech though such a process is conceivable. Although many of these segments are large enough to yield multiple 30 second sub-segments we do not further segment them because taking multiple cuts from a very long segment is likely to yield multiple instances of the same speaker. The fourth process extracts a new sub-segment exactly 30 seconds long from the center of each 33 second segment thus leaving out initial and final margins of at least 1.5 seconds to minimize wideband speech at the transitions. Figure 1 illustrates the segmentation process.

4. Auditing Format

Once the narrow band speech segments have been identified, the next step is to audit them to assure they contained exclusively speech in the target language or linguistic variety. Given the very large number of languages in the BNBS Corpus, it was not possible to have speakers of all of these languages on staff or physically located at LDC. Thus, we had to develop an auditing interface that could be used by remote auditors with minimal computing skills to play audio clips and enter judgments easily, on a variety of computing platforms via the Internet.

For VOA1 and VOA3 data sets, the audio format provided to auditors is perceptually equivalent to the archival format; each 30-sec audit segment is extracted from the original file and converted to mp3 before transfer to the auditor's browser. The VOA2 data was converted into an archival format of flac compressed RIFF before audio was extracted, converted to 16-kHz, 16-bit, single-channel PCM with lossless flac compression and converted to mp3 for transfer.

5. Language versus Dialect

The data for this effort come from BNBS, specifically VOA. Different sources use the terms language and dialect differently according to the relative importance those sources assign to social, historical, political and scientific factors. Some sources may refer to two linguistic varieties as separate languages because they are spoken in two different countries while other sources may refer to them as dialects of the same language because they are mutually intelligible. In dealing with this variability we respect the names of linguistic varieties asserted by their broadcast sources and use the term variety to express agnosticism as to their status.

VOA broadcasts may contain some mix of dialects both across and even within broadcasts asserted to be in a single language. We do not expect the VOA data selected for training and development material to differ significantly from that selected for evaluation material with respect to the inventory and relatively frequency of speech samples in different languages and dialects. However, the audit described below does not attempt to control for these effects either. All
Auditors are asked to indicate if any segment they review contains more than one speaker or any speech in a language other than the target.

6. Audit

The goal of the audit is to determine whether the segments include only speech in the target language, to confirm that they are narrow band and to identify the number of speakers and whether these speakers have been heard previously. Auditors are native speakers of the target language selected for their reliability, speed, ability to follow directions given in English and eligibility to work. Auditors are not necessarily trained as linguists. Given that news broadcasts routinely include speakers of multiple mutually intelligible dialects, particularly among the correspondents, pundits and listeners calling in, auditors are, perforce, not native speakers of each dialect. Auditors access their individual tasking via the Internet using authenticated login to a secure LDC server.

Before beginning the task, auditors undergo training and a test in narrow band detection. During training, auditors are presented with 5 ten-second, wide band segments described as studio quality and 5 ten-second, narrow band segments described as phone-like quality. Auditors listen to the segments repeating if necessary, until they are comfortable that they can discern the difference. They are then presented with 20, ten-second segments, which they must listen to and label. All segments contain speech entirely in English. The decision process is two-fold, auditors must choose whether the segment is phone-like or studio quality and indicate their confidence in their decision. Confidence levels are certain, pretty sure and unsure.

The results of this test are tabulated to determine the reliability of human judgments of narrow band signals. Auditors who perform significantly below average receive additional coaching to determine that they understand the distinction and are using high quality headphones and are invited to train a second time to improve performance. However, auditors are not accepted or rejected on the basis of their performance on this task alone. The average time required to train a single auditor, including this training and test, was two to three hours.

Next, auditors are presented with one 30-second segment at a time and are required to listen to it in its entirety using good quality headphones. The segment itself is stored in a web-accessible directory on the LDC server, and a link to the file is included in an "<EMBED>" tag in a web form presented to the auditor, producing an audio playback widget compatible with the installed browser and including the usual play and stop buttons.

The form collects audit judgments for the embedded segment. After listening to the entire segment auditors indicate whether: 1) The entire segment consists only of speech (no music or sound effects). 2) The speech is entirely in the target language. 3) The entire segment is "telephone-like" in quality (not studio quality). 4) All the speech is from a single speaker. 5) The current speaker is different from those audited previously. Auditors must also select one label from among the three sets below: 6) The speaker is: Male | Female | Unsure 7) The speech is: Native | Regional | Non-native 8) The audio is: Fairly clear/easily understood | Somewhat noisy/some parts muddled | Heavy noise or distortion/hard to understand.

Instructions, always available from each audit form, remind auditors to:
- Use high quality headphones or ear buds and avoid using desk-top or built-in speakers
- Answer the first five questions for every segment and, if the segment is all speech in the target language, answer all remaining questions as well
- If the segment contains speech in a non-target language the auditor recognizes, indicate that language in the comment box.
- If the segment contains two speakers one of whom is new, select 'yes' for 'different speaker', and 'no' for 'single speaker'.

The segments used to train auditors to detect wide band and narrow band speech are also always available from the audit form for each segment.

Once auditors submit their decisions for a given segment, that segment is no longer available and the next segment appears. An auditor’s workload consists of a fixed number of segments to be audited in their entirety. A tally at the top of the form shows the number of segments completed by that auditor and remaining. Auditors are asked to report to LDC immediately if many of the segments are not in the target language or contain the same speaker, silence, wideband speech or music.

Auditing yielded 18,721 database records, one per segment per auditor, of which 17,985 were used. The total number of auditing decisions counting every question answered and every comment submitted was 154,467. There were 658 instances in which an auditor reviewed the same segment. The average time required to audit a 30-second segment was 42 seconds. In 358 cases, the multiple judgments occurred within seconds to months of each other. In 300 cases, the judgments were separated by two years. Auditors agreed with their prior decisions concerning the language of the segment 99.4% of the time. Segments from two languages were also audited independently by two different auditors. The French auditors agreed with each other’s language labels 99.5% of the time. The Bosnian auditors agreed 89.6% of the time, the lower agreement possibly due to differences in experience, the one auditor being a Bosnian dominant Bosnian-Croatian bilingual, the other being Croatian dominant.

To explore the degree to which auditors can actually distinguish mutually intelligible varieties, auditors of a linguistic variety, y, audited segments that an auditor of mutually intelligible variety, x, had labeled as belonging to x. Table 1, shows the y varieties in the left column, x varieties across the top and the percent of segments distinguished by auditor y as not belonging to variety y.

<table>
<thead>
<tr>
<th>dist.</th>
<th>Dari</th>
<th>Persian</th>
<th>Hindi</th>
<th>Urdu</th>
<th>Creole</th>
<th>Croatian</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dari</td>
<td>97.6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Persian</td>
<td>89.6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hindi</td>
<td>92.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Urdu</td>
<td>91.9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>French</td>
<td>100</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bosnian</td>
<td>98.9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1: Percentage of segments of a variety x distinguished by an auditor of a mutually intelligible variety y.
7. Distribution Format

The completed corpus as used in LRE 2009 includes all: audio, segment boundaries and accompanying labels (music/speech/silence, bandwidth, language) for all audited languages across all three VOA collections. Prior to distribution to LRE09 participants, all audio was converted to single channel, 8 bit, 8-KHz μ-law to provide continuity with previous evaluations.

The automatic labels of signal content were delivered as a set of four text files for each data set: VOA1, VOA2, VOA3. The first file contains a mapping of filename to language label where the language label is a list of three candidates in the case of VOA2 file id lang1 (lang2 lang3) The second label file lists the time stamps for bandwidth labels: file id start time end time {phone|wideband|other} The third label file lists the time stamps for speech detection: file id start time end time {speech|other|unknown} The fourth label file lists the intersection of speech and bandwidth labels, in the form of groups of lines beginning with filename and language label followed by time stamps and combined speech and bandwidth labels: file id lang1 (lang2 lang3) start time end time {phone|wideband|other} {speech|other|unknown}. Time boundaries describe the time elapsed from the start of the radio program audio file, inclusive, and are given in the greatest precision measured for that particular metadata. "lang" is a four-letter abbreviation.

The metadata for each audited segment includes the following fields of information:

- p_seg_id: unique identifier for a segment
- filepath: relative path to the corresponding audio file
- dataset: corpus from which the audio is drawn
- aud_start: offset in seconds from beginning of audio corresponding to where segment audit starts
- aud_end: offset in seconds from beginning of audio corresponding to where segment audit ends
- lang_id: language spoken in segment
- lang_comment: comments the auditor made regarding the language of the segment
- spkr seks: sex of speaker in segment
- spkr_dialect: {native | non-native | dialect}, “Native” means speaker is a native speaker and dialect is standard. "non-native" means speaker is non-native speaker of the language. "dialect" means speaker is native speaker and dialect is not standard.
- all_1_spkr: true means the segment contains speech from a single speaker; false denotes multiple speakers.
- uniq_spkr: True means auditor has not been heard speaker in previous segments; false denotes a speaker that has been heard before.
- spkr_comment: free response auditor comments regarding the speaker of the segment.
- sig_quality: audit assessed quality of signal
- sign_comment: auditor comments regarding signal
- all_phbnd: True means entire segment was perceived to be phone band limited; false means portions of segment were perceived to be outside phone bandwidth.
- all_speech: True means that segment is composed entirely of speech; false means that some non-speech is present.

8. Discussion

The minimum requirements for including a language in the LRE 2009 main task were 80 training segments, 80 development segments and between 80 and 400 evaluation segments. Where at least 160 segments of previously exposed CTS data were available, they were used in place of BNBS data. In addition, LRE2009 will include a number of “out of set” languages not previously exposed. The data collected, segmented and audited by LDC have yielded adequate segments to support a NIST LRE evaluation for 23 languages and 16 out of set languages. LDC plans to release the VOA2 and VOA3 portions and the exposed subset of VOA1 after its use in the LRE evaluation campaign.

9. Disclaimer

Commercial equipment and materials are identified in order to adequately specify certain procedures. In no case does such identification imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that the materials or equipment identified are necessarily the best available for the purpose.

10. Conclusions

Broadcast news is a rich source of BNBS that can be used to support Language Recognition technology development and evaluation which has previously relied upon CTS collected to specification at considerably higher cost. The long term usefulness of the BNBS corpus for this purpose is not yet known but the presence of a new corpus of more than 11,000 hours of speech in nearly 70 languages partially audited to provide ground truth should support a variety of HLTs.

11. Acknowledgements

LDC would like to thank the researchers at the Speech@FIT group in the Faculty of Information Technology at the Brno University of Technology (BUT) in the Czech Republic (Brno) for providing the technologies to detect speech and narrow band signals, technologies that were absolutely necessary in the creation of this corpus.

12. References