Multi-language Speech Collection for NIST LRE

Karen Jones, Stephanie Strassel, Kevin Walker, David Graff, Jonathan Wright

Linguistic Data Consortium, University of Pennsylvania
Philadelphia, PA 19104 USA
Email:{karj,strassel,walkerk,graff,jdwright}@ldc.upenn.edu

Abstract

The Multi-language Speech (MLS) Corpus supports NIST's Language Recognition Evaluation series by providing new conversational telephone speech and broadcast narrowband data in 20 languages/dialects. The corpus was built with the intention of testing system performance in the matter of distinguishing closely related or confusable linguistic varieties, and careful manual auditing of collected data was an important aspect of this work. This paper lists the specific data requirements for the collection and provides both a commentary on the rationale for those requirements as well as an outline of the various steps taken to ensure all goals were met as specified. LDC conducted a large-scale recruitment effort involving the implementation of candidate assessment and interview techniques suitable for hiring a large contingent of telecommuting workers, and this recruitment effort is discussed in detail. We also describe the telephone and broadcast collection infrastructure and protocols, and provide details of the steps taken to pre-process collected data prior to auditing. Finally, annotation training, procedures and outcomes are presented in detail.

Keywords: language recognition; speech; telephone; broadcast

1. Introduction

The Multi-Language Speech (MLS) Corpus was created by Linguistic Data Consortium (LDC) to support the ongoing Language Recognition Evaluation (LRE) campaign conducted by the National Institute of Standards and Technology (NIST). With speech data originating from over 9000 speakers, the MLS Corpus is roughly double the size of the last corpus developed by LDC for this evaluation series, the LRE11 Corpus (Strassel 2012). The current MLS Corpus is large enough to provide data for multiple LRE evaluation cycles. The MLS Corpus was most recently used in support of the LRE15 evaluation (NIST 2015), providing speech data in twenty linguistic varieties including 6600 segments manually verified for language.

As with the LRE11 Corpus, the MLS Corpus was designed to support the evaluation of system capabilities for distinguishing closely related or confusable linguistic varieties. To support this goal we collected and audited audio for twenty distinct linguistic varieties across six defined "language clusters", where the varieties within a given cluster can be considered mutually intelligible and/or typologically related to some degree. The MLS Corpus required collection of new conversational telephone speech (CTS) and/or broadcast narrowband speech (BNBS) for each language, as well as updates to the procedures for selecting and preparing test segments for inclusion in the LRE15 evaluation. In the sections that follow we describe construction of the MLS Corpus in detail.

2. Data Requirements

The goal behind construction of the MLS Corpus was to collect narrowband speech from 400 unique speakers in each of 20 languages. Collection included two genres: conversational telephone speech (CTS) conversations between people who know one another, and broadcast narrowband speech (BNBS) taken from listener call-ins, person-on-the-street interviews or other instances where telephone data is embedded in a broadcast recording

The twenty languages selected for inclusion in the MLS corpus were chosen from an original list of 78 candidate languages. Final selection criteria included several considerations including:

- Sponsor interest
- Confusability with other linguistic varieties
- Availability of speakers to make calls and perform auditing work
- Availability of broadcast sources
- Availability of existing LRE training data

The varieties finally chosen for the MLS Corpus were categorized into one of six clusters of confusable varieties, shown in Table 1.

1. ARABIC	**2. SPANISH**	**3. ENGLISH**
Egyptian Arabic Iraqi Arabic Levantine Arabic Maghrebi Arabic Modern Standard Arabic	Caribbean Spanish European Spanish Latin American Spanish Brazilian Portuguese	British English Indian English General American English
4. CHINESE	**5. SLAVIC**	**6. FRENCH**
Cantonese Mandarin Min Nan Wu	Polish Russian	West African French Haitian Creole

Table 1: MLS Corpus Language Clusters

It should be noted that the degree to which clustered linguistic varieties might be considered mutually intelligible varies considerably. For example, while the varieties in the English cluster are certainly mutually intelligible, and even some pairings of the Arabic dialects are to some extent mutually intelligible, the four Chinese languages are generally considered to be mutually unintelligible languages despite being commonly referred to as “dialects” of Chinese.

The ratio of CTS data to BNBS data in the corpus varies by language. For Modern Standard Arabic (MSA), for example, 100% of speech segments came from BNBS, since MSA would not ordinarily be used in telephone conversations between friends or relatives. Also, for some languages collection and auditing of BNBS significantly outpaced CTS speaker recruitment, such that the size of the CTS collection could be reduced considerably. Conversely, for other languages very little broadcast collection was possible and we relied largely or entirely on CTS collection to yield the required number of segments.

Given the reliance on multiple collection strategies and the variable level of difficulty in either recruiting speakers or identifying suitable broadcast sources for collection, the time to complete each language varied considerably, as shown in Figure 1. Languages with no CTS collection are excluded from the figure.

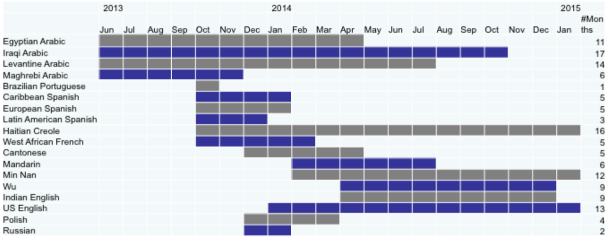


Figure 1: Variable time for completion per language

For each CTS call the goal was to extract two 30-second segments for use as training or test data. For each BNBS recording, only one segment was extracted. Where meeting collection goals proved challenging for a particular language, it was permissible to extract two segments of speech from a single broadcast recording. In such cases it was important that the two segments were taken from maximally distal portions of the source recording to reduce the chances that the segments contained speech from the same speaker.

For broadcast data there was an additional requirement that for any language there should be at least two sources, where “source” is a feature of channel and is defined as a combination of a show producer and a service provider on that channel. By this definition, CNN received via Comcast is a different source from CNN received via Verizon FiOS, since the channel characteristics may differ in the two cases.

3. Speaker Recruitment

To create the CTS portion of the corpus, we recruited native speakers following the same approach used for LRE11, in which recruited callers (known as “claques”) were required to make single calls to multiple unique individuals within their established social networks (friends, family, acquaintances). The primary challenge in building the MLS Corpus centered around the requirement that speech from 400 speakers was required for each language. For language varieties for which there was no available broadcast data this meant recruiting double the number of speakers compared to LRE11. In addition to possessing native or near-native fluency in the target language, potential recruits also needed to possess basic English skills (so that they could be effectively managed by English speaking staff at LDC). More importantly, they had to be socially well-connected, such that they would have no difficulty in making calls to dozens of different friends and family members who also spoke their language. Claques could reside anywhere in the US, and advertising was targeted toward places where we expected to find large concentrations of speakers of a given language; we used US Census data to inform our advertising strategies and also reached out to organizations that catered to the population in question.

Assessment of potential claques included several stages:

- A short questionnaire designed to gather information about language background and skills
- A brief screening assessment in which candidates were asked to identify audio recordings containing speech in their own stated dialect from a set of audio files in confusable dialects
- An in-person or phone interview to provide more information about the project and make a final assessment about a candidate’s suitability.

A total of 456 claque candidates were assessed, 219 were offered work on the project, and 190 actually completed calls.

4. Collection

4.1 CTS Collection

The CTS portion of the MLS Corpus was collected using LDC’s existing collection infrastructure. LDC operates a computer telephony system for specifically collecting speech from the telephone network. The system consists of a T-1 line, which provides 24 audio channels and operates in toll-free mode. A Dialogic D/480JCT-2T1 telephony board performs interactive voice response functions and call logging functions. In addition, an AudioCodes DP6409 Passive-Tap call logging board reduces the risk of losing data in the event of a failure of the primary collection platform. The telephony hardware provides the ability to record up to 12 two-person conversations simultaneously. Customized IVR software is installed on each system; the telephony application handles all interactions with callers, connects callers to one another, and starts/stops recordings. All collected calls were recorded directly to disk in 8kHz, 8-bit, µ-law format.

All claques were provided with a personal identification number (PIN) and a printable set of guidelines outlining the telephone call collection task. Claques were responsible for recruiting several dozen unique callees from among their friends and family, explaining that the calls would be recorded for research purposes, and arranging times to make each call.

For each call, claques followed the steps listed below:

- Dial a toll-free number provided by LDC
- Enter their own PIN
- Enter the phone number of their call partner

While recruited claques provided their names and contact information in order to be compensated, callees were entirely anonymous and were not directly compensated (though some claques choose to share their own compensation with their callees). Callees were not assigned a PIN and provided no demographic information or contact information. Prior to the start of recording, the telephone platform's Robot Operator played a pre-recorded message announcing the purpose of the call and requesting permission to record the conversation as soon as the call partner answered the phone.

To ensure that conversations were as natural as possible, claques and callees were free to talk about any subject of their own choosing with the caveats that sensitive or personal issues should be avoided and also that personal identifying information such as full names and addresses should not be revealed during the call.

To reduce the potential for an observable correlation between the acoustic properties of any particular telephony channel and a particular language, claques for every language were instructed to call speakers both within the US and outside of the US. In this way, all linguistic varieties in the corpus have some calls using the same US-based telephony network, thus eliminating the risk of channel-language bi-uniqueness.

Only the callee side was used when extracting deliverable speech segments for LRE evaluations; we wanted to reduce the presence of repeat speakers in the collection, and by design the claque call side consisted of the same speaker making multiple calls. Further, callees were asked to assert that they had not participated in prior LRE calls (whether in the current collection or in previous LRE collections). Claques were prohibited from calling the same individual callee more than once. While it was expected that a claque may wish to speak to different members of the same household who happened to share the same phone number, the telephone platform was also configured to prevent a claque calling the same number more than three times. In addition, given the primary focus on the callee side as the corpus deliverable, claques were required to encourage their call partner to talk as much as possible.

To meet the target number of speakers for each language, roughly 200 claques were recruited across the various languages, with the requirement that each claque make calls to a minimum of 15 different acquaintances. Claques made 23 calls on average, with a few claques managing just a single call and one claque making as many as 133 calls for the study.

4.2 BNBS Collection

The BNBS component of the corpus utilized LDC's existing collection infrastructure, with new data from both satellite networks and web sources. For some languages it was also possible to utilize previously unexposed broadcast recordings from LDC's existing holdings.

Galaxy 19 free-to-air programming was collected via LDC's rooftop dish for multiple MLS languages, and additional SCOLA programming was collected the Spanish language varieties. LDC's satellite collection system utilizes Ubuntu linux and incorporates TechnoTrend S-1500 DVB-S PCI receiver/decoder boards for processing one satellite transponder. Pre-identified Program IDs (PIDs) were captured from the transponder at scheduled times with the use of Python/Perl scripts and open source utilities. Collected audio streams were unencrypted MPEG-1 Audio Layer II (.mp2) that came in a variety of formats including:

- MPEG ADTS, layer II, v1, 128 kbps, 48 kHz, Stereo
- MPEG ADTS, layer II, v1, 160 kbps, 48 kHz, Stereo
- MPEG ADTS, layer II, v1, 192 kbps, 48 kHz, Stereo
- MPEG ADTS, layer II, v1, 64 kbps, 44.1 kHz, Monaural
- MPEG ADTS, layer II, v1, 64 kbps, 48 kHz, Stereo

Web radio sources were also collected to augment the satellite collection, particularly to address low-yield languages. Potential broadcast sources were identified and reviewed by claques (to determine appropriateness of language) and technical staff. A web downloader process utilizing open source software ran 24/7 and checked lists of sources that were flagged for download, then collected data streams as 30-minute captures.

Additionally, unexposed broadcast data from prior LDC collection efforts contributed additional recordings for Indian English, Mandarin, Modern Standard Arabic and US English.

In contrast to CTS, identification of individual speakers in the broadcast data is unfeasible. Nevertheless, two steps were taken to reduce and/or identify cases of speaker recurrence in this data. First, broadcast schedules were analyzed to determine the optimal interval that should occur between recordings of successive broadcasts of a given program, in order to minimize speaker repeats due to rebroadcasts. Second, we utilized speaker identification software developed by Phonexia to produce an exhaustive set of speaker trials on all collected BNBS data. In these trials every segment within a language group was treated as a "model", and then measured against every other segment as a "test signal", with the goal of helping assess the likelihood of two segments having the same speaker. The results of the speaker trials were delivered to NIST along with the corpus, for possible use in segment selection for evaluation.

5. Auditing

5.1 Segment Preparation

Careful manual auditing of collected data was an essential procedure for verifying the accuracy of language labels and ensuring that all data satisfied quality requirements. Because exhaustive auditing of complete recordings would be cost-prohibitive, manual auditing was limited to small segments extracted from each recording.

For broadcast recordings, the first step in segment preparation was to identify the areas of the recording that contained narrowband signal. Bandwidth detection technology developed by Phonexia was utilized for this purpose. For both BNBS and CTS segments a speech activity detection (SAD) tool developed by LDC was used to identify areas of silence, music and other non-speech.

A segment was deemed eligible for auditing if 33 seconds of speech was detected within a 33-90 second window. In a change from LRE11 where discontinuous segments of speech were concatenated to produce CTS segments of the requisite duration, segments in the LRE14 collection were used with their internal silence intervals kept intact. This means that the segment presented for manual auditing could be up to 90 seconds in duration, with 33 seconds of speech present somewhere within the segment. Figure 2 provides an illustration of the pre-processing steps leading to the generation of audit segments.

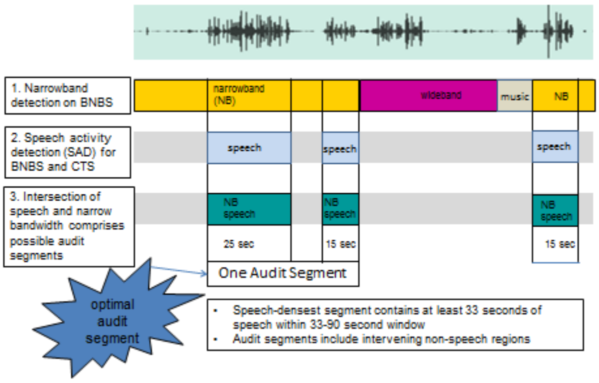

1. Narrowband detection on BNBS
narrowband (NB)
wideband
music
NB
2. Speech activity detection (SAD) for BNBS and CTS
speech
speech
speech
3. Intersection of speech and narrow bandwidth comprises possible audit segments
NB speech
NB speech
NB speech
25 sec
15 sec
15 sec
One Audit Segment
optimal audit segment
Speech-densest segment contains at least 33 seconds of speech within 33-90 second window
Audit segments include intervening non-speech regions

Figure 2. Preparing Audio Segments for Audit

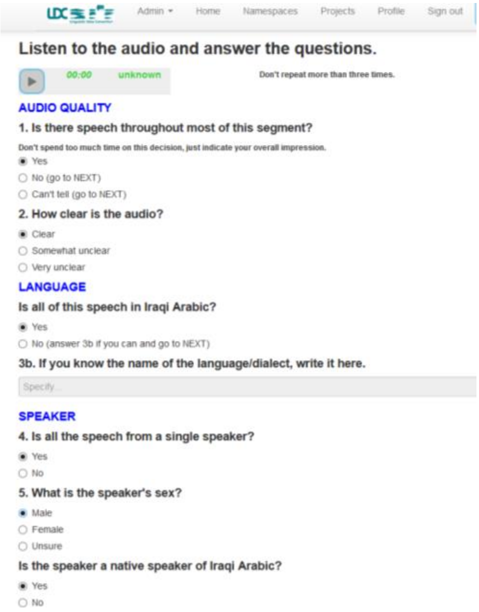
5.2 Manual Auditing

Auditors were native speakers of the target language and were trained in the specifics of the auditing task and were required to successfully complete a practice assignment before working on corpus data. Auditors used a web-based GUI, shown in Figure 3, which presented one segment at a time for review. The GUI was configured to require complete playback of each segment before any questions could be answered about the segment. For each segment, auditors answered the following questions:

- Is there speech throughout most of this segment?
- How clear is the audio?
- Is all of the speech in [expected language]?
- Is all of the speech from a single speaker?
- Is the speaker a native speaker?
- What is the speaker’s sex?

A number of logical constraints were built into the auditing procedure. For example, if an auditor answered “No” to the question “Is there speech throughout most of this segment?”, then all subsequent questions would be hidden from view and the auditor would immediately be forced to move on to the next segment. Constraints of this type were added to the audit logic to ensure that the auditing task was performed efficiently, without spending unnecessary time on segments that did not meet requirements.

In many cases, the best claques from data collection were retained to act as auditors; to avoid bias auditors were barred from auditing their own calls. Further constraints, which were implemented via automated audit assignment logic, included prioritizing CTS segments for auditing over broadcast segments, and also ensuring that all segments extracted from a given telephone call were audited before moving on to segments from any other call.



Listen to the audio and answer the questions.

00:00 unknown

Don't repeat more than three times.

AUDIO QUALITY

1. Is there speech throughout most of this segment?

Don't spend too much time on this decision, just indicate your overall impression.

● Yes
○ No (go to NEXT)
○ Can't tell (go to NEXT)

2. How clear is the audio?

● Clear
○ Somewhat unclear
○ Very unclear

LANGUAGE

Is all of this speech in Iraqi Arabic?

● Yes
○ No (answer 3b if you can and go to NEXT)

3b. If you know the name of the language/dialect, write it here.

Specify...

SPEAKER

4. Is all the speech from a single speaker?

● Yes
○ No

5. What is the speaker's sex?

● Male
○ Female
○ Unsure

Is the speaker a native speaker of Iraqi Arabic?

● Yes
○ No

Figure 3. Auditing Interface

Following a procedure first developed in LRE11, auditing assignments consisted of batches of segments called “kits”, where each kit was comprised of target segments (segments believed to be in the auditor’s own language) plus three other types of segments:

- 10% cross-audit segments (segments from another dialect in the same cluster included for the purpose of assessing language/dialect confusability)
- 5% dual segments (where an in-language segment is audited by another speaker of the same dialect with the purpose of measuring inter-annotator agreement)
- 5% distractor segments (segments from a completely different cluster of languages, which were included to keep auditors alert)

Because the MLS Corpus is being used in multiple LRE evaluations, corpus-wide statistics including inter-annotator agreement numbers are potentially evaluation sensitive and cannot be reported until the full corpus has been exposed.

6. Results

Complete audit segments along with full CTS calls, audit results, and call and segment metadata for the complete MLS corpus were delivered to NIST in seven incremental deliveries, for selection of train and test segments to be used in LRE15 and subsequent evaluations. Audio was delivered in its original format as initially captured by LDC during collection.

LDC worked closely with NIST to develop corpus metadata requirements, definitions and values. Tables 2-3 list the metadata fields along with their descriptions, for audio and annotation metadata respectively. LDC did not pre-filter audited segments to select “good” segments to deliver to NIST. Instead, all audited speech segments were delivered along with their associated metadata and annotations, which NIST could use to inform selection of train and test segments for LRE. In addition to LDC’s own pre-delivery checks of all data and metadata, NIST performed a series of independent checks.

Field Name	Field Description
audio_id	6- or 7-digit numeric ID for the audio segment
datetime	yyyy-mm-dd hr:mn:sc = date of audio recording
btime	offset (seconds) from start of recording to start of segment
duration	segment duration (seconds)
file_size	byte count of segment file
file_type	file format (flac)
md5_checksum	checksum of segment file
source_duration	source recording duration (seconds)
source_file	full file name of source recording
source	call-id and channel (CTS) or producer/provider (BNBS)
segment_type	either "CTS" or "BNBS"
origin_info	Anonymized phone "number" (CTS) or country/broadcaster (BNBS)

Table 2: MLS Corpus Audio Metadata

Field Name	Field Description
auditor_id	numeric ID of auditor
audit_type	'target', 'distractor', 'confusable'
auditor_lang	language that the auditor is listening for
audio_id	6- or 7-digit numeric ID for the audio segment
language_code	assumed language of the audio segment
all_target_lang	Is all of the speech in [language]? (Yes/No/ NO RESPONSE)
off_target_lang	Auditor's comment if segment is not in their language
mostly_speech	Is there speech throughout most of this segment? (Yes/No)
speech_clarity	How clear is the audio? (clear/some unclear/very unclear/ NO RESPONSE)
single_speaker	Is all of the speech from a single speaker? (yes/no/unsure/NO RESPONSE)
native_speaker	Is the speaker a native speaker? (yes/no/unsure/NO RESPONSE)
speaker_sex	What is the speaker's sex? (male/female/unsure/NO RESPONSE)

Table 3: MLS Corpus Annotation Metadata

Table 4 shows the number of BNBS and CTS segments from the MLS Corpus selected for use in the LRE15 evaluation, with a total of 6600 segments selected across the two genres.

Language	# BNBS Segments	# CTS Segments
MSA	130	0
Iraqi Arabic	0	478
Maghrebi Arabic	0	440
Levantine Arabic	0	364
Egyptian Arabic	0	426
Caribbean Spanish	25	99
Latin American Spanish	95	274
European Spanish	103	204
Brazilian Portuguese	232	13
Indian English	7	361
US English	55	313
British English	421	0
West African French	37	332
Haitian Creole	0	484
Min Nan	0	452
Wu	0	397
Mandarin	78	240
Cantonese	38	86
Russian	118	44
Polish	182	72
TOTAL	1521	5079

Table 4: MLS Segments in LRE15

7. Conclusions

The MLS corpus consists of carefully labeled and annotated conversational telephone and broadcast narrowband speech data in twenty languages, and the use of this corpus for segment selection for the NIST LRE15 evaluation is a testament to its importance to human language technology researchers. Once the resources described in this paper are no longer sequestered for use in ongoing NIST Language Recognition Evaluation campaigns, they will be published in the LDC catalog making them available to the research community at large.

8. Acknowledgements

Thanks are due to Craig Greenberg and John Howard at NIST, and also to Doug Reynolds and Elliot Singer at Lincoln Laboratories, MIT. We are also grateful to Phonexia for use of two types applications: Phonexia's Speaker identification software was used post-collection to provide NIST with tables of possible speaker matches across the full CTS/broadcast datasets for each completed language. Phonexia's bandwidth detection technology was used as a pre-processing aid to identify regions of narrowband speech in collected broadcast data.

9. References

NIST (2015). The 2015 NIST Language Recognition Evaluation Plan (LRE15). http://www.nist.gov/itl/iad/mig/upload/LRE15_EvalPlan_v23.pdf. Retrieved March 17, 2016.

Ryant, N. (2013). LDC HMM Speech Activity Detector (v.1.0.3a). LDC, University of Pennsylvania.

Strassel, S., Walker, K., Jones, K., Graff, D., Cieri, C. (2012). New Resources for Recognition of Confusable Linguistic Varieties: The LRE11 Corpus. In Proceedings of Odyssey 2012 - The Speaker and Language Recognition Workshop, Singapore, June 25-28, pp.202-208.