New Resources for Recognition of Confusable Linguistic Varieties: The LRE11 Corpus

Stephanie Strassel, Kevin Walker, Karen Jones, Dave Graff, Christopher Cieri

Linguistic Data Consortium
University of Pennsylvania, Philadelphia, PA USA
{strassel|walkerk|karj|graff|cieri}@ldc.upenn.edu

Abstract

The NIST 2011 Language Recognition Evaluation focuses on language pair discrimination for 24 languages/dialects, some of which may be considered mutually intelligible or closely related. The LRE11 evaluation required new data for all languages, comprising both conversational telephone speech and broadcast narrowband speech from multiple sources in each language. Given the potential confusion among varieties in the collection, manual language auditing required special care including the assessment of inter-auditor consistency. We report on collection methods, auditing approaches, and results.

1. Data Requirements

The NIST Language Recognition (LRE) campaigns began in 1996 to with the goal of evaluating performance on language recognition in narrowband speech. The most recent campaign, LRE11, targets language pair discrimination for 24 languages/dialects, some of which may be mutually intelligible to some extent by humans [1]. Data requirements for LRE11 demanded collection of speech sufficient to yield at least 400 narrowband segments for each language. Traditionally LRE evaluations have utilized large collections of conversational telephone speech (CTS). The 2009 LRE corpus represented the first departure from the standard approach in its reliance on narrowband segments embedded in broadcast, typically coming from listener call-ins, phone interviews of pundits and some correspondent reports and man on the street interviews. LRE11 targets collection of both CTS and broadcast narrowband speech (BNBS) for each language, with a few exceptions. Modern Standard Arabic (ara) is a formal variety that wouldn't typically be spoken during spontaneous conversation and was excluded as a CTS collection target. Conversely, the dialectal Arabic varieties of Iraqi, Levantine and Maghrebi were not expected to appear in formal broadcast news programs and were therefore excluded as a BNBS target. Collection also targeted multiple broadcast sources, where "source" is a provider-program (so Larry King Live is different from CNN Headline News).

To satisfy the need for data in languages that might exhibit a high degree of confusability (whether for humans or systems), we reviewed sources including Ethnologue [2] and compiled a preliminary list of candidate languages[1]. Each language was assigned a confusability index score:

- 0 - Not likely to be confusable with another candidate language
- 1 - Possibly confusable with another candidate language; languages are related and may be confused by (some) systems if not by (most) humans
- 2 - Likely confusable with another candidate language; evidence that (some) humans may find the varieties mutually intelligible to some extent

Language	ISO 639-3 or LDC 3-letter code	Confusablity Score	Language(s) of Possible Confusion
Arabic Iraqi	acm	2	other Arabic
Arabic Levantine	alv	2	other Arabic
Arabic Maghrebi	arm	2	other Arabic
Arabic MSA	ara	2	other Arabic
Bengali	ben	1	other Indic
Czech	ces	1	slk
Dari	prs	2	fas
English (American)	eng	1	emi
English (Indian)	emi	1	eng
Farsi/Persian	fas	2	prs
Hindi	hin	2	urd
Lao	lao	2	tha
Mandarin	cmn	0	
Pashto	pus	0	
Polish	pol	1	other Slavic
Punjabi, Western	pnb	1	other Indic
Russian	rus	1	other Slavic
Slovak	slk	1	ces
Spanish	spa	0	
Tamil	tam	0	
Thai	tha	1	lao
Turkish	tur	0	
Ukrainian	ukr	1	other Slavic
Urdu	urd	2	hin

Table 1: Target Languages in the LRE11 Evaluation

When evaluating confusability we took care to distinguish those varieties with multiple names which are generally recognized as the same language (e.g. Persian/Farsi); those which are mutually intelligible varieties but given different language names for historical, social or political reasons (e.g. Hindi and Urdu); and those which are really different languages (or mutually unintelligible, e.g. Mandarin Chinese and Cantonese). From this exercise a set of 38 candidate languages was identified; that list was ultimately whittled down to 24 after researching the availability of broadcast sources for the language and considering the availability of employable native speakers. The final set of LRE11 languages, along with any confusable language varieties for each target, are presented in Table 1.

[1] Throughout the paper we use *language* as shorthand for a linguistic variety that may be referred to by different sources as a language or dialect.

2. Speaker/Auditor Recruitment and Screening

Speaker and auditor recruitment for LRE11 was particularly challenging given the short timeline for data collection and the large number of languages being targeted. The collection model for the CTS component of the corpus was similar to that used in the LDC's first LRE CTS collection (CallFriend, LDC96S46 - LDC96S60), but with two notable differences. The original Callfriend protocol was designed to yield exactly one call per speaker: in order to collect 200 speakers per language, we recruited 100 people, and provided incentives to each one in return for making a single phone call to another speaker of their language who was in the U.S. For the LRE11 collection, we recruited fewer individuals per language, and gave them incentives to call as many other speakers of their language as they could; if necessary, they could call to acquaintances outside the U.S. This small core of recruited callers, or "claques", would be present in all recorded calls, so the yield of unique-speaker call sides would be lower than in CallFriend, but this would be offset by the relative efficiency of recruiting. The allowance of overseas calls raised concerns about possible correlations between particular regional telephone networks and particular languages, so we sought to enforce guidelines to ensure that each language would be represented by calls to multiple geographic regions, with a strong preference to have as many callees as possible within the U.S.

This methodology added another dimension to the standard set of recruitment challenges: recruits not only had to possess the right combination of language and professional skills, but also had to be socially well connected. Recruitment materials underscored this requirement, stating that it would be necessary to "Contact up to 30 people you know who are fluent speakers of your target language and are willing to have their voices recorded for research purposes." We targeted a minimum of 3 recruits per language; this number was established to ensure the required CTS collection volume, to permit some amount of dual-auditing for purposes of establishing inter-auditor consistency rates, and to avoid the conflict of interest that would be created by having an individual audit segments from calls where he also acted as a claque.

Given the large number of recruits targeted and the short timeline for project completion, it was critical to have an efficient and effective recruitment strategy. Recruitment was broad, targeting local and regional community organizations as well as online user communities. Initial candidate assessment was achieved by means of a multi-part online screening process. The initial screening was administered to any applicant who expressed legitimate interest in the study, and was designed to assess a candidate's availability and employability, social network density, and competence in the target language. The language skills portion of the screening test addressed many dimensions of competence including how the candidate learned the language and how often the language was used for common tasks like reading the news or conversing with friends and family.

Candidates who passed the initial screening were then subject to a secondary test that required them to listen to ten segments of speech and identify those that were in their target language. This language ID test was specifically designed to include segments in languages considered mutually intelligible and/or in the same language family, and as such emulated the actual auditing task required to support LRE11. The test also provided a valuable opportunity both to determine the conceptions of particular languages a candidate might possess and also for us to explain how language categories were being used for the purpose of LRE11. This was especially useful in the case of a language with multiple labels. Dari, for example, is frequently called Farsi by its speakers, but for the purpose of LRE11 needed to be distinguished from Farsi (Persian) as spoken in Iran. Within each LRE11 language category auditors were not expected to be experts on the various dialects of their language. It was accepted that auditors came with their own intuitions about specific languages that may or may not have been in line with the LRE11 categories.

Although many applicants were multilingual and were interested in making calls and auditing more than one target language, each recruit was assigned to a single target language (the one for which they demonstrated the highest degree of nativeness). One reason for this restriction was to maximize speaker variety in the collection as a whole; another was to reduce the chances of an applicant overstating their language skills in an attempt to procure more work and therefore greater compensation. Of approximately 130 candidates who took the initial screening test, 84 were ultimately employed as claques and/or auditors.

3. Collection

3.1. Telephone Speech Collection

The CTS portion of the LRE11 corpus was collected using LDC's existing collection infrastructure. LDC operates three computer telephony systems specifically for collecting speech from the telephone network. Each system is connected to a dedicated T-1 line, which provides 24 audio channels and has toll-free service enabled. The systems incorporate Dialogic telephony hardware; specifically, each system houses a Dialogic D/480JCT-2T1 telephony board which can perform interactive voice response functions and call logging functions. In addition, one of the systems incorporates an AudioCodes DP6409 Passive-Tap call logging board. 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.

For LRE11 the call platform software was configured to support a CallFriend-style protocol, which requires the claque to dial in to the designated toll-free number and enter their unique PIN. They then key in the number of their call partner and the system's Robot Operator places the call. When the call partner answers, the Robot Operator plays a pre-recorded prompt announcing the purpose of the call and requesting permission to record the conversation. Collected calls are initially written to disc on the platform itself in 8kHz, 8-bit μlaw; supporting software handles transfer of recordings to the main LDC network and updates to the associated call and speaker databases.

Special efforts were taken in LRE11 to avoid bi-uniqueness of channel conditions and language in the telephone collection. Claques for each language were encouraged to make calls to multiple countries, and claques from all languages were required to make calls within US.

3.2. Broadcast Collection

The LRE09 BNBS corpus utilized an existing archive of several thousand hours of Voice of America broadcasts previously collected by LDC. Most of that material was exposed during LRE09, making it unsuitable for use as test data in LRE11. Moreover, LRE11 required data from a large number of languages with multiple sources for each language. These factors necessitated new data collection and from a wide range of broadcast sources, including multiple satellite feeds and over the air broadcasters from three locations around the world. Limited collection from streaming web sources was also targeted.

Locally, LDC operates an extensive collection system dedicated to the capture and processing of broadcast content from a wide range of sources; this system is depicted in Figure 1. The system is able to collect audio and video from satellite, cable (CATV) and terrestrial TV. The satellite reception facilities allow us to address up to three simultaneous C-Band and Ku-Band satellite downlinks, as well as Dish Network and DirecTV satellite downlinks.

In addition to the primary broadcast collection system we maintain portable broadcast collection platforms in Tunis and Hong Kong; the system in Tunis supplied multiple broadcast sources for LRE11. Each portable platform is a TiVO style digital video recording (DVR) system capable of recording two streams of A/V material simultaneously. The platform includes integrated analog and digital Satellite DVB-S reception components; it supports international specifications and is capable of recording programming outside of the United States. The system has a very small footprint and is suitable for transportation as a piece of carry-on luggage.

The portable platforms and the main LDC collection system share the same code base and rely on a modular, unified hardware specification. Improvements in the main collection platform therefore translate into benefits for both platforms. The systems run Ubuntu Linux, and are equipped with two *TechnoTrend S-1500* DVB-S PCI receiver/decoder boards capable of processing one satellite transponder. Each capture card has an associated, dedicated capture drive. We used a combination of Open Source utilities and supporting scripts written in Python and Perl to capture a predefined set of PIDs (program IDs) from a given transponder at a scheduled time. For the purposes of LRE, we focused on transponder PIDs associated with audio streams; because the bandwidth of the audio streams was relatively compact, we were able to capture multiple PIDs from a given transponder in parallel. Equipping each system with two receiver/decoders allowed us to capture from two separate transponders simultaneously. In all cases, the audio streams that we selected were unencrypted, MPEG-1 Audio Layer II audio in an MPEG transport stream. We used TS Tools version 1.11 to convert the audio from transport stream to elementary stream. Each capture system included a table of sources of interest, along with corresponding transmission parameters including transponder frequency, polarization, symbol rate, source PID, and language.

All collection activity on the platforms is driven by a supervisor computer with a customized scheduling database. The supervisor computer is responsible for controlling receivers, audio video matrix routing, and recording job initialization. The system also incorporates 8TB of local storage, dedicated automatic speech recognition systems, dedicated multimedia transcoding systems, a 24TB LTO4 tape backup system, and two experimental logging systems which can be used to capture entire transponder transport streams from satellite downlinks. The platform deployed in Tunisia is maintained remotely by personnel at LDC, with recordings scheduled from LDC and automatically downloaded into LDC's collections server.

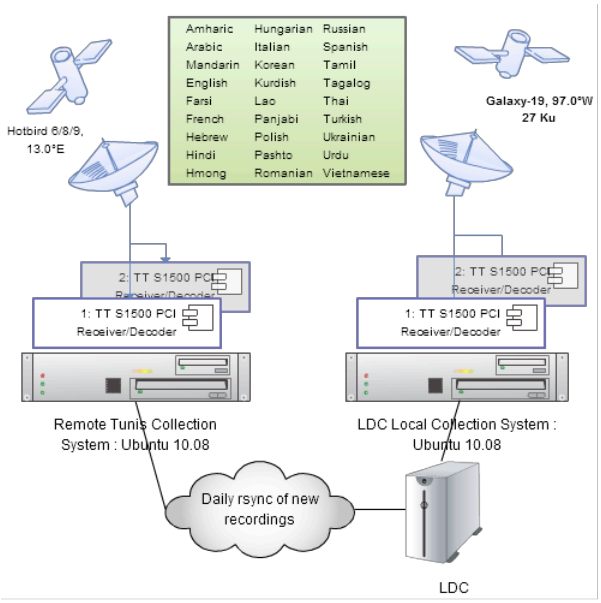

Figure 1: Satellite Collection Diagram

Given the language varieties targeted for LRE11 and the difficulty of finding appropriate sources in Philadelphia or Tunisia, it soon became necessary to deploy an additional remote collection platform in a new location. After surveying potential data sources and evaluating several collection sites, a new platform was deployed in New Delhi, and this site contributed a substantial portion of the recordings from southeast and central Asian languages. In contrast to the Tunis platform, source research and recording schedules for the New Delhi system were performed onsite. New recordings were downloaded from a local FTP server and reviewed on a rolling basis by LDC collections staff. Collection targets were refined over time based on this iterative feedback loop.

A limited supplemental collection of streaming radio sources was undertaken in Philadelphia to augment the satellite collection, particularly to address low-yield languages and/or those with a limited number of sources available via the primary collections. Promising channels were identified using off-the-shelf streaming radio database software, and we used Audials RadioTracker to drive a Perl application that captured multiple streams in parallel. The captured audio was written directly to disc in its native format.

Regardless of the collection method an initial effort was required to identify potential broadcast sources that produce programming in multiple languages. This is not only efficient but it also reduces the connection between linguistic and channel features that may hinder the development and evaluation of language recognition technologies. In the process of developing the source databases, we relied heavily on information from the online site *http://www.lyngsat.com*. In the case of the Philadelphia collection system the greatest yield came from Galaxy-25 (used by GlobeCast to deliver

broadcast content in many languages) and from Galaxy-IIIC (for Chinese programming and SCOLA). The Tunis platform primarily utilized Hotbird-6, comprising a large number of transponders, each of which carried between 5 and 15 independent streams. Hotbird coverage is very wide, including North Africa, Europe, and parts of the Middle East. Many of the New Delhi site broadcast sources were recorded from local over-the-air radio and television transmissions, while the remainder were collected from a set top box fed by a local Satellite TV provider.

During collection we record entire programs in languages of interest before extracting narrowband regions within those segments. This approach resulted in very large volumes of unaudited audio, with only a fraction of the collected speech yielding narrowband speech in the target language. To maximize yield we attempted to tailor the collection schedule to target each transponder of interest at the ideal time. This was an iterative process, due to the fact that for most of the broadcasters we had little or no scheduling information. Our strategy was to make survey recordings over a 24-hour period, review those recordings and identify timeslots with high potential, make longer test recordings based on that information, review the results, and iterate as needed.

Another challenge was the range of audio formats that appear in naturally occurring data from a variety of sources. Channels, sampling rate, sample size, compression and audio file headers vary independently in complex ways. The data collected via satellite is all MPEG1 Audio Layer II (.mp2), 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

All streaming data was mp3, 128kbps bitrate, yielding 44.1kHz sample rate. Prior to distribution all data is normalized to flac compressed, linear sampled audio at 8kHz sampling rate with 8 bit samples and NIST SPHERE file headers.

4. Auditing

4.1. Segment Preparation

Selection of potential LRE segments from the much larger inventory of collected data is a multi-stage process. First, the audio is passed through a speech activity detection system to eliminate from further consideration any silence, music or other non-speech. For CTS data this yields segments of 30-35 seconds in duration. For BNBS data a second filter is required to distinguish any narrowband signal. From the intersection of speech and bandwidth filters, continuous regions of 33 seconds duration or more are selected. For regions longer than 33 seconds, a 33-second segment is chosen from its center. This selection process is illustrated in Figure 2.

Given the difficulty of finding sufficient BNBS segments for all LRE11 languages, in some cases it was necessary to reduce the 33-second duration requirement and instead select shorter continuous segments, down to a minimum of 10 seconds.

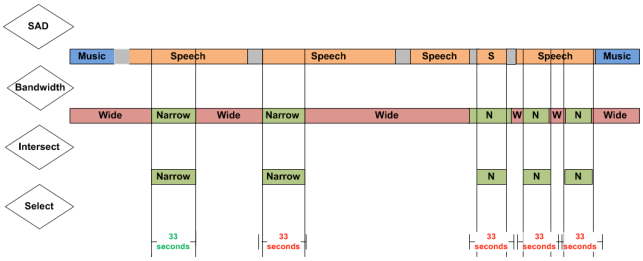

Figure 2: Selecting BNBS Segments

We do not concatenate smaller, separate (possibly non-contiguous) segments to build the 33-second segments. And although many speech segments are large enough to yield *multiple* 33-second sub-segments, we do not further segment them because we wish to maximize the number of potential speakers in the corpus; taking multiple cuts from a very long segment is likely to yield multiple segments from the same speaker. Selected BNBS segments were saved as 16 kHz, 16 bit while CTS segments are 8 kHz single-channel, converted to pcm, ms-wav file format for compatibility with the web-based auditing toolkit.

4.2. Auditing Task

Auditing was performed using a web-based GUI developed by LDC, shown in Figure 3. The goals of auditing were to determine that

- the audio segment contained only speech
- all the speech was in the target language
- the speech was clear
- the signal was narrowband
- there was just one speaker
- the quality of the audio was acceptable

Auditors received hands-on training for using the auditing GUI, and were given written and verbal instructions for how to answer each question in the auditing task. Auditors also completed a short online training exercise to establish an understanding of the difference between narrowband and wideband audio. In this exercise auditors listened to several examples of both kinds of bandwidth before listening to unlabeled audio samples and making a judgment on whether each segment's sound was telephone quality (narrowband) or studio quality (wideband). Auditors could repeat the exercise until confidence in their ability to make this distinction was achieved. They could also revisit the test at any point during the auditing process to refresh their memory about this distinction.

Trained auditors accessed their assignments via a secure LDC server. They were instructed to use good quality headphones and listen to each segment of speech in its entirety before making their judgments on three aspects of the speech segment: audio quality, language and speaker. The sections below describe the auditing process in detail.

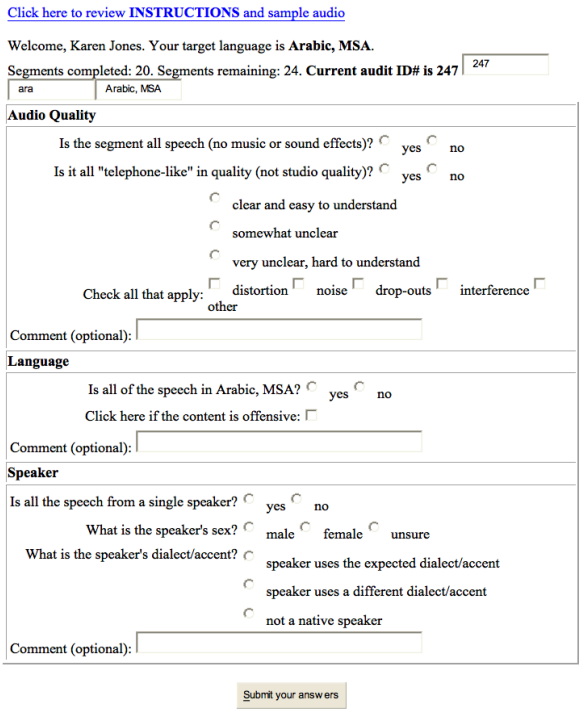

Figure 3: LRE Auditing Interface

4.2.1. Audio Quality

First, auditors were required to answer a series of questions about the quality of the audio recording itself, to establish that the segment was appropriate for inclusion in LRE. There were three mandatory questions:

- Is the segment all speech?
- Is it all "telephone-like" in quality (not studio quality?)
- How clear is the audio?

Auditors were also required to indicate whether the recording contained any distortion, interference or other recording problems. If the call contained background noise (dogs barking, music) auditors were asked to note that in the comments.

4.2.2. Language

Next, auditors were instructed to make a yes/no decision about whether all of the speech was in their target language. Since many of the CTS segments originated from bilingual speakers auditors were told to listen carefully for speech streams in which the speaker had switched language and comment on segments that displayed this characteristic.

Auditors were also strongly encouraged to write in the comment box if they heard a language or variety that was not the target but was one they thought they could name. Although auditors were not assumed to possess any specialized language expertise beyond fluency in the target language, encouraging them to leave detailed comments about segments not in their own language proved beneficial. For instance, auditor comments helped reveal a large number of narrowband broadcast segments that were believed to be Pashto based on the source from which they were collected, but in fact proved to be Dari. Based on auditor comments it was possible to re-assign these segments to a Dari auditor for confirmation and eventual inclusion in the final corpus.

4.2.3. Speaker

Finally, auditors were required to answer questions about the speaker heard in the recording:

- Is there only one speaker?
- What is the speaker's sex?
- What is the speaker's dialect/accent?
  - a native speaker using a standard accent or dialect of the target language,
  - a native speaker using a non-standard accent or dialect
  - a non-native speaker of the target language.

Again, auditors were encouraged to leave comments about their judgments. In contrast to previous LRE evaluation corpora, auditors in LRE11 were not asked to indicate whether a given speaker was unique in the corpus; this question wasn't feasible given the large number of segments being judged and the existence of multiple auditors for each language.

4.3. Auditing Kit Construction

Given the high degree of mutual intelligibility for some languages in the LRE11 collection it was especially important to collect judgments from auditors about languages that might be confused with their own. Additionally, to establish inter-auditor agreement rates we also targeted some amount of dual annotation in which multiple auditors independently judged the same segments. Segments were compiled into "kits" which were assigned to each auditor. The baseline component of each kit was a set of segments expected to be in the auditor's language, comprising a proportional selection of all available BNBS segments plus CTS segments from other claques' call partners.

In addition to the baseline, each kit contained up to 10% "distractor" segments, which were drawn from any other LRE11 language. These distractors were randomly interspersed with the baseline segments, primarily as a way to keep auditors attentive to the task. Kits also contained up to 10% "dual" segments, which are also independently assigned to one or more of the other auditors for that language; these segments are used to calculate within-language inter-auditor consistency. For those languages potentially confusable with other LRE11 languages, kits also included "confusable" segments comprising 10%, 25% or 100% of the baseline amount, as follows:

- 10% for related/possibly confusable varieties (e.g. Polish/Slovak)
- 25% for likely confusable varieties (e.g. Lao/Thai)
- 100% for known confusable varieties (e.g. Hindi/Urdu)

Given the non-linear nature of the data collection, actual kit makeup varied, so that a given kit could be predominately CTS, or have fewer than 10% dual segments. Crucially, auditors did not know anything about the expected makeup of their kits and were simply instructed to judge each segment using the same standards.

4.4. Inter-Auditor Consistency

We assessed inter-auditor agreement on several dimensions. The first of these is within-language agreement, in which we

compare multiple judgments where the expected language of the segment was the language of both auditors; for instance, we compare the judgments from two Bengali auditors judging segments expected to be Bengali. As expected, within-language agreement approaches 100% for most languages. There are some exceptions, most notably involving Modern Standard Arabic (42.86% agreement) and several of the dialectal Arabic varieties (85.37% for Maghrebi; 92.31% for Iraqi). These results are not terribly surprising given the diglossic situation for Arabic. Hindi and Urdu also show relatively low within-language agreement (89.19% and 90.91% respectively); it's worth noting that these varieties are considered highly mutually intelligible. Thai also showed a lower-than-expected agreement rate (93.98%) which may be attributable to some confusion with Lao.

Looking at all multiply-labeled segments, for the corpus as a whole we observe 214 cases of disagreement (out of a possible 2664) where one auditor labeled the segment as being in the target language while another rejected the same segment as not being in the target language. The vast majority of these segments are attributable to confusion involving Hindi and Urdu, with some residual confusion related to Modern Standard Arabic and also Pashto.

Focusing only on cross-language agreement, we measure agreement for cases where a segment was confirmed by an annotator to be in their language when that language was the expected language, but the segment was independently judged by an annotator of another language to be in that second auditor's language. For instance, a Hindi speaker verifies an expected Hindi segment to be Hindi, and an Urdu speaker judges the same segment to be Urdu.

Segment Expected Language	Auditor Language	Total Count	Disagreement Count	% Disagreement
alv	acm	108	5	4.63%
ara	acm	108	5	4.63%
ara	alv	121	23	19.01%
ara	arm	104	19	18.27%
arm	acm	111	1	0.90%
arm	alv	120	1	0.83%
eng	eni	160	46	28.75%
eni	eng	154	1	0.65%
fas	prs	307	1	0.33%
prs	fas	18	14	77.78%
hin	urd	496	126	25.40%
urd	hin	786	419	53.31%
lao	tha	140	15	10.71%
tha	lao	73	5	6.85%
ces	slk	179	2	1.12%
slk	ces	120	1	0.83%

Table 2: Cross-language agreement among LRE11 auditors

This analysis, summarized in Table 2, indicates confusability among some language pairs and clusters. The most noticeable and least surprising is the cluster of Arabic varieties, particularly when the segment's purported language is Modern Standard Arabic. There are several cases of strong asymmetry between language pairs; for instance, in the case of American English vs. Indian English, Indian English auditors are quite likely to judge purported American English segments as actually being Indian English, while American English auditors do not show this tendency for purported Indian English segments. Similar asymmetry also exists to some extent for Lao~Thai, Hindi~Urdu and Farsi~ Dari (though note the small sample size for purported Dari segments, which reflects the overall difficulty of obtaining Dari data for LRE11). It's worth noting that the Hindi~Urdu pair was found to be highly confusable for auditors in previous LRE evaluations as well.

5. Corpus Distribution

LRE11 data was distributed to NIST in six incremental releases. Each package contained full source audio recordings from which audited segments had been extracted, in their original format; the extracted segments as presented to the auditor; and auditing results for each segment. Only "useable" segments were included in the data release. To be useable a segment had to have been judged as being in the target language and containing only speech. Each release was accompanied by a set of segment metadata specifying the following:

- audid: numeric ID of audit submission in annotation table
- segid: numeric ID of audited segment
- lngid: 3-letter language ID as confirmed by auditor
- result: concatenation of responses to yes/no questions regarding segment "useability"
- sex: speaker sex (M/F)
- spkr_typ: speaker's dialect category (native, non-native, etc)
- noise_amt: auditor's judgment of noise level (easy, hard, etc)
- noise_typ: auditor's list of noise conditions (distortion, etc)
- noise_cmt: free-text auditor comment on signal quality
- spkr_cmt: free-text auditor comment on speaker
- lng_cmt: free-text auditor comment on language
- ref: reference status
- auditor: numeric ID of auditor
- src: path/name of source audio file
- duration: length in seconds of the audio segment

The various comment fields are typically empty for usable segments, since auditors usually make comments only when there is something "wrong" with a segment. The "reference status" is primarily for LDC-internal use; when two or more auditors judged a single segment and gave different decisions, the segment was withheld until the difference could be adjudicated. In that case, the "ref" field is used to isolate the adjudicated result for the segment. Most segments were judged by only one auditor, so this field was typically empty. The "src" field was again intended primarily for LDC-internal use; it shows when different audited segments come from the same source recording, and also shows which collection platform was used to record the original audio.

The audited segments delivered for LRE11 were limited to just those where (a) we had only one auditor judgment on record, or (b) the two or more auditor judgments were in agreement. When one of those was true, and the judgment indicated a usable segment (in the auditor's target language, and all speech), the segment was included in the corpus delivery. Segments that showed discrepant auditor judgments or indeterminacy in manual language labeling were excluded from delivery. The resulting LRE11 corpus comprises 9889 useable segments. Table 3 summarizes the inventory of

segments for each language along with the number of unique broadcast sources for each language that included narrowband broadcast segments.

The linguistic resources described in this paper have been distributed to LRE11 performers as training, development and evaluation data.

Language	Broadcast Sources	Broadcast NB Segments	CTS Segments	Total Useable Segments in LRE-11
acm	0	0	408	408
alv	0	0	408	408
ara	22	406	0	406
arm	0	0	405	405
ben	17	227	220	447
ces	4	279	179	458
cmn	9	173	259	432
emi	42	366	50	416
eng	8	331	121	452
fas	27	208	197	405
hin	34	348	70	418
lao	1	125	126	251
pnb	15	11	397	408
pol	1	239	242	481
prs	20	374	25	399
pus	12	257	155	412
rus	3	302	139	441
slk	4	242	172	414
spa	10	188	231	419
tam	11	214	200	414
tha	5	338	65	403
tur	8	305	167	472
ukr	8	67	175	242
urd	8	256	222	478

Table 3: Useable segments and source variety by language

LDC will wherever possible distribute the data more broadly, for example to its members and licensees, through the usual mechanisms, for instance via publication in the LDC catalog. General catalog releases will include complete audit results rather than the filtered versions delivered within the program. Upon sponsor request some subsets of data may be reserved for use during a specified time period within LRE only.

6. References

NIST 2011. *The 2011 NIST Language Recognition Evaluation Plan (LRE11)*. (http://nist.gov/itl/iad/mig/upload/LRE11_EvalPlan_releasev1.pdf)

Ethnologue, Languages of the World. (http://www.ethnologue.com/)

LyngSat (http://www.lyngsat.com/)

7. Acknowledgements

We thank the Speech@FIT group in the Faculty of Information Technology at Brno University of Technology (BUT) in Czech Republic (Brno) for providing speech and bandwidth detection technologies.