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, USA

Linguistic Data Consortium

Outline

- Data Requirements
- Language Selection
- Broadcast Collection
- Telephone Collection
- Segment Selection
- Auditing
- Auditor Agreement
- Data Distribution and Corpus Summary
- Conclusions

Data Requirements for LRE 2011

- Distribution of previous LRE data to new participants
  - Previous test sets
  - LRE 2009 training data, including large broadcast news corpus
- New resources for LRE 2011
  - As in LRE 2009, includes both conversational telephone speech (CTS) and broadcast narrowband speech (BNBS)
  - Both genres for most but not all languages
    - Arabic varieties limited to broadcast-only (MSA) or telephone-only (Iraqi, Levantine, Maghrebi)
  - Target 24 languages/dialects, some of which may be mutually intelligible to some extent by humans
    - 400 segments per language
    - At least 2 unique sources per language
      - Broadcast source is provider-program (so CNN Larry King is different source than CNN Headline News)

Linguistic Data Consortium

Language Selection

- Reviewed information sources like Ethnologue
- Compiled list of candidates plus 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; at least some evidence that (some) humans may find the varieties mutually intelligible to some extent
- Candidate set of 38 languages whittled down to 24 with NIST and sponsor input, and considering
  - Availability of broadcast sources
  - Availability of claques/auditors

*Throughout we use ***language*** as shorthand for a linguistic variety that may be referred to by different sources as a language or dialect

Linguistic Data Consortium

Potential Confusability for LRE 2011 Languages

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

Language	ISO 639-3 or LDC 3-letter code	Confusablity Score	Language(s) of Possible Confusion
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

Broadcast Collection

- Multiple broadcast sources
  - Existing, unexposed VOA1 data
  - New and unexposed archival data from local satellite collections in Philadelphia, Tunis and Hong Kong
  - New collection from cable, satellite and off-the air sources via portable collection platform installed in New Delhi
  - New collection from streaming web radio sources
- Variety of formats
  - Satellite data is MPEG1 Audio Layer II (.mp2)
    - 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 sources mp3, 128kbps bitrate, 44.1kHz sample rate

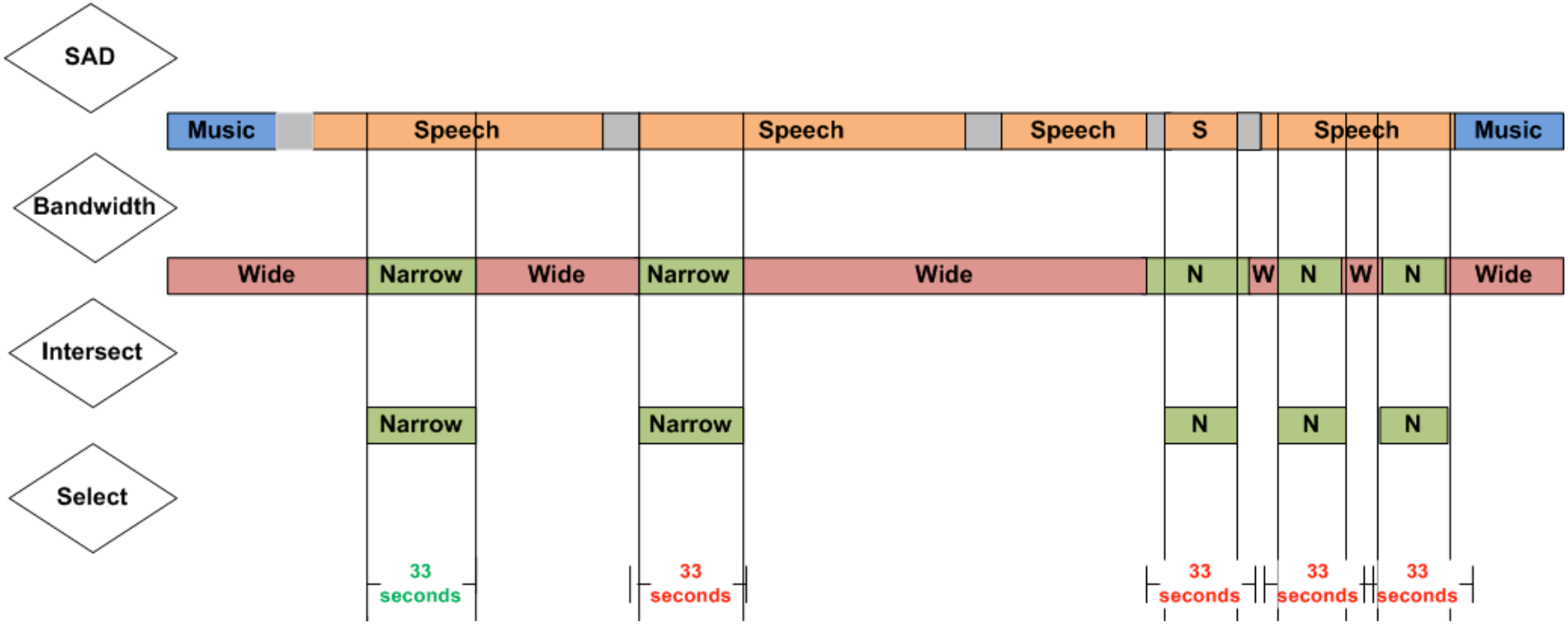
Telephone Collection

- Claque-based collection model
  - Claque is a native speaker informant
  - Eases recruitment burden
  - Claques later serve as auditors
- 2-5 claques recruited per language
- Each claque makes a single call to each of 15-30 individuals in their existing social network
  - Callee hears pre-recorded message and provides consent prior to call being recorded
  - Steps taken to ensure different claques' callees did not overlap
  - Claque call sides excluded from corpus
  - Require at least some calls within US to avoid bi-uniqueness of channel/language conditions
- Calls collected on LDC's CTS platform in 8kHz, 8-bit μlaw

Selection of Segments for Auditing

- Full recordings passed through SAD system to distinguish speech vs. silence, music, other non-speech
- For CTS data we extract 2 segments per call, 30-35 seconds each
- For BNBS, additional bandwidth filter prior to selection
  - From the intersection of speech and bandwidth filters, continuous regions of 33+ seconds selected
    - For regions > 33 sec, single 33-sec segment chosen from center
      - No selection of multiple segments from single stretch of speech
    - When necessary to get a sufficient number of auditable segments for a given language, shorter continuous segments (down to a minimum of 10 sec) were selected
      - No concatenation of separate, short BN segments

Broadcast Segment Selection



Although many speech segments are large enough to yield multiple 30 second sub-segments we do not further segment them in order to maximize the number of potential speakers in the corpus

Data Preparation for Auditing

- Extracted segments converted to auditor format
  - BNBS: 16 KHz, 16 bit
  - CTS: 8 KHz single-channel
  - Converted to pcm, ms-wav file format for browser compatibility
- Auditor presented with entire segment
  - BNBS: typically 33 sec long, but possibly as little as 10 sec
  - CTS: entire 30-35 second segment

Auditing Kit Construction

- Baseline: segments expected to be in the auditor's language
  - For BNBS, proportional selection of all available segments
  - For CTS, other claques' callee sides
- Up to 10% **distractor** segments from non-confusable language
  - Presented to the auditor at random, to keep them attentive
- Up to 10% **dual** segments, also assigned to other auditor(s)
- For languages with confusable/buddy languages, also include **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 non-linear nature of the collection, actual kit makeup varies
  - So one kit may be predominately CTS, or have < 10% dual segments

Auditor Screening and Training

- Preliminary online screening of potential auditors for language skills
  - Questions about language background, education
  - Listening test comprising 10 segments including target language, distractor and potentially confusable language segments
  - Screening results also helped point out areas where auditor training was required, e.g. to clarify language labels
- Of ~130 who took screening, 84 passed and were hired and given additional training (typically in-person)
  - Telephone calling instructions
  - Bandwidth detection training via “Signal Quality Perception Test”
    - Train, then test on ability to distinguish wide- and narrow-band segments (described as “phone-like” quality and “studio-like” quality)
    - Auditors could revisit test anytime to refresh their memory

Linguistic Data Consortium

Auditing Questions

- Goal of auditing is to ensure that segments
  - Contain only speech
  - Are in the target language variety
  - Are narrowband
  - Contain only one speaker
  - Audio quality is acceptable
  - In contrast to previous LRE, no question about speaker uniqueness
- Auditors judge each segment via a web-based interface
  - Required to listen to entire segment
  - Instructed to use good-quality headphones
  - Formal auditing instructions explain how to answer each question

Linguistic Data Consortium

Auditing Interface

Click here to review **INSTRUCTIONS** and sample audio

Welcome, Karen Jones. Your target language is **Arabic, MSA**.

Segments completed: 20. Segments remaining: 24. **Current audit ID# is 247** 247

ara | Arabic, MSA

Audio Quality

Is the segment all speech (no music or sound effects)? yes no

Is it all "telephone-like" in quality (not studio quality)? yes no

clear and easy to understand

somewhat unclear

very unclear, hard to understand

Check all that apply: distortion noise drop-outs interference other

Comment (optional):

Language

Is all of the speech in Arabic, MSA? yes no

Click here if the content is offensive:

Comment (optional):

Speaker

Is all the speech from a single speaker? yes no

What is the speaker's sex? male female unsure

What is the speaker's dialect/accent? speaker uses the expected dialect/accent

speaker uses a different dialect/accent

not a native speaker

Comment (optional):

Submit your answers

Auditing Consistency - Context

- 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 delivered to NIST
- Segments that showed discrepant auditor judgments or indeterminacy in manual language labeling were excluded from delivery
- Numbers reported here are from segments assigned during normal auditing process, not during a post-hoc consistency analysis task

Linguistic Data Consortium

Within-Language Agreement

- Comparing multiple judgments

  - *where the expected language of the segment was the language of the auditors*

  - *what is language-label agreement?*

- e.g. two Bengali speakers judge clips that are purported to be Bengali (e.g. because of the collection source)

Within Language (1)

Arabic varieties

Auditor Language	Expected Language	Count Total	% Agreement
ara	ara	77	42.86%
arm	arm	41	85.37%
acm	acm	39	92.31%
alv	alv	54	98.15%

Farsi/Persian, Dari

Auditor Language	Expected Language	Count Total	% Agreement
fas	fas	291	98.63%
prs	prs	107	98.13%

Within Language (2)

Slavic varieties

Auditor Language	Expected Language	Count Total	% Agreement
ces	ces	48	100.00%
pol	pol	52	98.08%
slk	slk	72	98.61%
ukr	ukr	205	99.51%
rus	rus	44	100.00%

South Asian varieties

Auditor Language	Expected Language	Count Total	% Agreement
pnb	pnb	119	98.32%
ben	ben	49	97.96%
tam	tam	63	100.00%
hin	hin	148	89.19%
urd	urd	44	90.91%

Within Language (3)

Other languages

Auditor Language	Expected Language	Count Total	% Agreement
cmn	cmn	44	100.00%
spa	spa	147	97.28%
tur	tur	60	98.33%
tha	tha	83	93.98%

No dual annotation for Lao due to lack of auditors

Auditor Language	Expected Language	Count Total	% Agreement
eng	eng	57	94.74%
eni	eni	123	98.37%

?

Dual Annotation Results

Includes all multiply judged segments by auditors of same language, regardless of expected segment language

Overall Language Agreement

general	N	Y
N	951	
Y	214	1499

Disagreement mostly from 3 languages

lang	disagr	total	% disagr
ara	14	52	26.92%
hin	31	252	12.30%
urd	116	268	43.28%

Dual Annotation Confusion Matrices

All Speech?

All Speech	N	Y
N	309	
Y	**451**	1904

Single Speaker?

SnglSpkr	N	Y
N	206	
Y	**220**	2238

All Narrowband?

AllNB	N	Y
N	67	
Y	**118**	2479

Speaker Sex

SpkrSex	N	Y	U
N	1087		
Y	**58**	1383	
U	**24**	**28**	20

Signal Quality Judgment

SigQual	diffic	easy	medium
difficult	19		
easy	**51**	2029	
medium	**43**	**337**	92

Dialect Judgment

Dialect	marked	non-ntv	normal
marked	30		
non-ntv	**11**	4	
normal	**165**	**20**	1420

Includes all multiply judged segments by auditors of same language, regardless of expected language of segment

Cross-Language Agreement

- Reporting judgments where

  - a segment was confirmed by an annotator to be in their language

  - that language was the expected language

  - independently judged by an annotator of another language to also be in that language

- E.g. a Hindi speaker verifies an expected Hindi segment to be Hindi, and an Urdu speaker judges the same segment to be Urdu

Linguistic Data Consortium

Cross-Language

Arabic varieties

Expected Language	Auditor Language	Count Total	% Confusion
alv	acm	108	4.63%
ara	acm	108	4.63%
ara	alv	121	19.01%
ara	arm	104	18.27%
arm	acm	111	0.90%
arm	alv	120	0.83%

Cross-Language

Czech - Slovak

Expected Language	Auditor Language	Count Total	% Confusion
ces	slk	179	1.12%
slk	ces	120	0.83%

Thai - Lao

Expected Language	Auditor Language	Count Total	% Confusion
lao	tha	140	10.71%
tha	lao	73	6.85%

American English – Indian English

Expected Language	Auditor Language	Count Total	% Confusion
eng	eni	160	28.75%
eni	eng	154	0.65%

Linguistic Data Consortium

Cross-Language

Hindi - Urdu

Expected Language	Auditor Language	Count Total	% Confusion
hin	urd	496	25.40%
urd	hin	786	53.31%

Dari – Farsi/Persian

Expected Language	Auditor Language	Count Total	% Confusion
fas	prs	307	0.33%
prs	fas	18	77.78%

Data Distribution

- Data and audit results distributed to NIST in 6 incremental releases
- Packages contain
  - Full source audio recordings from which segments extracted, in original format
  - Auditor-versions of extracted segments
  - Audit results for segments that meet these criteria
    - Is the segment in the target language? (YES only)
    - Does the segment contain only speech? (YES only)
    - Is all the speech from one speaker? (YES or NO)
    - Does the entire segment sound like narrow-band signal? (YES or NO)
  - Segment metadata table

audid	numeric ID of audit submission in lb_aud_ann table		**noise**	cmt free-text auditor comment on signal quality
segid	numeric ID of audited segment		**spkr**	cmt free-text auditor comment on speaker
lngid	3-letter language ID as confirmed by auditor		**lng_cmt**	free-text auditor comment on language
result	concatenation of responses to yes/no questions		**ref**	reference status
sex	speaker gender (M/F)		**auditor**	numeric ID of auditor
spkr	typ speaker's dialect category (native/non-native/etc)		**src**	path/name of source audio file
noise	amt auditor's judgment of noise level (easy/hard/etc)		**duration**	length in seconds of the audio segment
noise	typ auditor's list of noise conditions (distortion/etc)			

Linguistic Data Consortium

Data Summary

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

Language	Broadcast Sources	Broadcast NB Segments	CTS Segments	Total Useable Segments in LRE-11
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

Linguistic Data Consortium

Conclusions

- Significant volumes of new telephone and broadcast data collection for 24 languages which include several confusable varieties
  - New collection strategies needed to support corpus requirements
- 84 auditors made 22,561 audit judgments yielding 9889 LRE segments
- Auditing kits constructed to support consistency analysis
  - Within-language agreement over 95% except for
    - Most Arabic varieties
    - Hindi, Urdu
    - Thai (but not Lao)
  - Cross-language confusion for
    - Some Arabic pairs, especially involving Modern Standard Arabic
    - Thai/Lao (asymmetry)
    - American English/Indian English (strong asymmetry)
    - Hindi/Urdu (asymmetry)
    - Farsi/Dari (asymmetry; small sample size)
- Corpus supported LRE 2011 evaluation and will be published in LDC catalog pending authorization by sponsors

Acknowledgements

- Thanks to 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
- Thanks to Alvin Martin and Craig Greenberg for their ongoing support
- Thanks to our collection partners in Tunis, New Delhi and Hong Kong