Evaluation of Spoken Language Recognition Technology Using Broadcast Speech: Performance and Challenges

Luis J. Rodríguez-Fuentes, Amparo Varona, Mireia Diez,
Mikel Penagarikano, Germán Bordel

Software Technologies Working Group (http://gtts.ehu.es)
Department of Electricity and Electronics, University of the Basque Country
Barrio Sarriena s/n, 48940 Leioa, Spain
email: mikel.penagarikano@ehu.es

Odyssey 2012, Singapore
June 27, 2012

Contents

- Introduction
  - Context
  - Motivation
- An Overview of the Albayzin LREs
- Albayzin LRE datasets
- SLR system
- Performance analysis
  - Closed-set Clean-speech (CC)
  - Open-set Clean-speech (OC)
  - Noisy speech (Albayzin 2010 LRE)
- Conclusions and future work

Context (I)

- Spoken Language Recognition (SLR) technology advancements largely fostered by NIST LREs
- NIST providing data + researchers providing the algorithms
- NIST LRE datasets: 8 kHz, conversational telephone speech (CTS) + narrow-band broadcast news (NBBN)
- Up to 24 target languages (including variants of the same language)
- Issues:
  (1) focus on telephone speech and large-scale verification applications
  (2) lack of resources to objectively assess technology improvements on wide-band speech
  (3) challenges specific to other kind of data (e.g. wide-band broadcast speech) not addressed

Context (II)

- Albayzin 2008 and 2010 LRE aimed to expand the scope of SLR technology assessment
- Inspired by NIST 2007 LRE: same task, test procedures, performance measures, file formats, etc.
- Differences:
  (1) speech signals from wide-band TV broadcasts involving multiple speakers
  (2) small set of target languages, but potentially challenging due to acoustic, phonetic and lexical similarities
  (3) target application: Spoken Document Retrieval (SDR)

Motivation

To identify the most challenging conditions in SLR tasks, which may eventually guide the design of future evaluations

To that end...

- SLR system based on SoA approaches developed and evaluated on the Albayzin 2008 and 2010 LRE datasets
- System performance analysed with regard to:
  - the set of target languages
  - the amount of training data
  - background noise (clean vs. noisy speech)

Albayzin LRE: common features

- Task: language detection
  - trial = target language (L) + test segment (X)
  - deciding (by computational means) whether or not L was spoken in X
  - providing a likelihood score (which is assumed to support the decision)
- System performance measured on a set of trials, by comparing system decisions with reference labels stored in a keyfile
- Each test segment featuring a single language: target language or an Out-Of-Set (OOS) language (for open-set verification trials)
- Following NIST LRE, test segments of three different nominal durations (3, 10 and 30 seconds) evaluated separately
- Performance measures:
  - Average cost C_{avg} (pooled across target languages), with the same priors and costs used in NIST 2007 and 2009 LRE
  - Detection Error Tradeoff (DET) curves: to compare the global performance of different systems for a given test condition

Albayzin LRE: things that were different

Albayzin 2008 LRE

- Target languages: Basque, Catalan, Galician, Spanish
- Two separate tracks depending on the data used to build systems:
  - restricted (only train and dev data provided for the evaluation)
  - free (any available data)
- Only clean speech

Albayzin LRE: things that were different

Albayzin 2008 LRE

- Target languages: Basque, Catalan, Galician, Spanish
- Two separate tracks depending on the data used to build systems:
  - restricted (only train and dev data provided for the evaluation)
  - free (any available data)
- Only clean speech

Albayzin 2010 LRE

- Target languages: Basque, Catalan, Galician, Spanish, Portuguese, English
- Free development
- Two separate tracks depending on the background noise:
  - clean: only clean-speech test segments were considered
  - noisy: all the test segments (containing either clean or noisy/overlapped speech) were considered
- Separate sets of clean and noisy/overlapped speech segments provided for training

Albayzin LRE datasets: shared features

- Speech segments are continuous excerpts from TV broadcast shows involving one or more speakers
- Recording setup: Roland Edirol R-09 digital recorder (directly connected to cable TV)
- Audio signals stored in WAV files: uncompressed PCM, 16 kHz, single channel, 16 bits/sample
- Disjoint sets of TV shows posted to training, development and evaluation, as an attempt to achieve speaker independence

Albayzin 2008 LRE: KALAKA

- Segments containing background noise, music, speech overlaps, etc. filtered out
- OOS languages: French, Portuguese, English, German
- Training: more than 8 hours per target language

	Spanish	**Catalan**	**Basque**	**Galician**
#segments	282	278	342	401
time (minutes)	529	538	531	532

- Development and evaluation: 1800 segments each (600 per nominal duration, 120 per target language and 120 containing OOS languages)
- More than 50 hours of speech: 36 hours for training + 7.7 hours for development + 7.7 hours for evaluation

GTTS
Tecnologías Software

Albayzin 2010 LRE: KALAKA-2

- KALAKA fully recycled for KALAKA-2
- New recordings, specially for Portuguese, English and OOS languages
- Noisy segments collected from existing and newly recorded materials
- Evaluation dataset completely new and independent of KALAKA
- OOS languages: Arabic, French, German, Romanian
- Training: more than 10 hours of clean speech and more than 2 hours of noisy speech per target language

	Clean speech		Noisy speech	
	#segments	time (minutes)	#segments	time (minutes)
Basque	406	644	112	135
Catalan	341	687	107	131
English	249	731	136	152
Galician	464	644	125	134
Portuguese	387	665	160	197
Spanish	342	625	133	222

- Development and evaluation: more than 150 segments per target language and nominal duration (4950 and 4992 segments, respectively)
- 125 hours of speech: 82 hours for training + 21.24 hours for development + 21.43 hours for evaluation

SLR system: acoustic subsystems

- SLR system identical to that developed for NIST 2011 LRE, with very competitive performance
- Fusion of 2 acoustic and 3 phonotactic subsystems

SLR system: acoustic subsystems

- SLR system identical to that developed for NIST 2011 LRE, with very competitive performance
- Fusion of 2 acoustic and 3 phonotactic subsystems
- Acoustic subsystems
  - Acoustic features: MFCC-SDC (7-2-3-7)
  - UBM: gender-independent 1024-mixture GMM
  - High-dimensional representation: zero-order + centered and normalized first-order Baum-Welch statistics
  - Subsystem 1 - Linearized Eigenchannel GMM: channel matrix estimated only on data from target languages
  - Subsystem 2 - Generative iVector: total variability matrix estimated only on data from target languages

SLR system: phonotactic subsystems + backend/fusion

- Phonotactic subsystems
  - Phone-Lattice SVM approach
  - BUT TRAPs/NN phone decoders for Czech, Hungarian and Russian providing phone posteriors
  - Phone lattices built on posteriors by means of HTK (BUT recipe)
  - Expected counts of phone *n*-grams computed by means of SRILM (up to 3-grams, weighted counts)
  - L2-regularized L1-loss SVM classification by means of LIBLINEAR

SLR system: phonotactic subsystems + backend/fusion

- Phonotactic subsystems
  - Phone-Lattice SVM approach
  - BUT TRAPs/NN phone decoders for Czech, Hungarian and Russian providing phone posteriors
  - Phone lattices built on posteriors by means of HTK (BUT recipe)
  - Expected counts of phone *n*-grams computed by means of SRILM (up to 3-grams, weighted counts)
  - L2-regularized L1-loss SVM classification by means of LIBLINEAR
- Backend and Fusion
  - Parameters optimized on the development set of Albayzin 2010 LRE and then applied to both 2008 and 2010 evaluation sets
  - Gaussian backend applied only in the open-set condition
  - Fusion/Calibration parameters estimated by linear logistic regression under a multiclass paradigm
  - Minimum expected cost Bayes decisions based on the calibrated scores
  - FoCal toolkit by Niko Brümmer

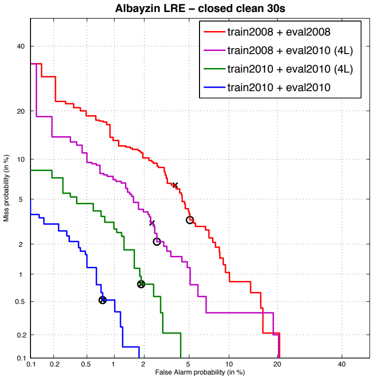
Performance analysis

Outline

- **Clean speech (closed-set and open-set):**
  - Comparison across Albayzin 2008 and 2010 LRE
  - Confusion of languages with each other
- **Noisy speech (only Albayzin 2010 LRE):**
  - Degradation compared to clean speech

Closed-set Clean-speech (CC): comparison across evaluations

Performance on the 2008 LRE dataset much worse than on the 2010 LRE dataset (red vs. blue) - see details here



(1) Different amount of training data to estimate models (purple vs. green)

(2) Portuguese and English (2010 LRE) less confused with the other languages than the average (green vs. blue)

(3) Task intrinsically more difficult in 2008 than in 2010, probably due to higher acoustic variability related to background noise (red vs. purple)

Closed-set Clean-speech (CC): confusion of languages with each other

Miss probabilities (diagonal) and false alarm probabilities (out of the diagonal) on the CC-3s condition of the Albayzin 2010 LRE

		Target					
		eu	**ca**	**en**	**gl**	**pt**	**es**
Segment	**eu**	0.054	0.046	0.015	0.139	0.000	0.162
	ca	0.107	0.060	0.013	0.181	0.107	0.195
	en	0.015	0.037	0.015	0.000	0.052	0.022
	gl	0.099	0.198	0.033	0.207	0.083	**0.397**
	pt	0.027	0.075	0.034	0.055	0.027	0.055
	es	0.112	0.152	0.024	**0.336**	0.016	0.144

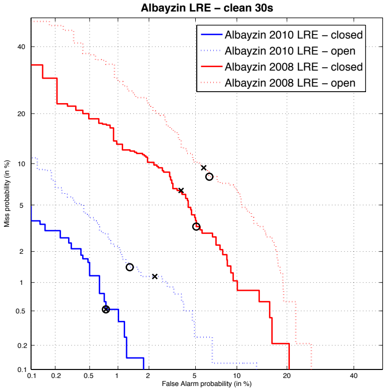
(1) Romance languages in Spain feature high error rates, remarkably Spanish and Galician: many Galician speakers having Spanish as first (mother) language

(2) Lowest error rates for English and Portuguese (and then Basque, which is confused mostly with Spanish)

(3) Low confusion rates for Portuguese: comparatively little contact with Romance languages in Spain (except for Galician, see (1))

Open-set Clean-speech (OC): comparison across evaluations

Again, performance on the 2008 LRE dataset much worse than on the 2010 LRE dataset (red dotted vs. blue dotted) - see details here



(1) Difference in performance for equivalent tasks (clean-speech, 30s) in 2008 and 2010 LRE: around 5 points in terms of EER

(2) Albayzin 2010 LRE: larger training dataset, less confusable languages (on average)...

(3) Similar differences in performance between open-set and closed-set for both datasets (dotted vs. continuous)

GTTS
Tecnologías Software

Open-set Clean-speech (OC): confusion of languages with each other

Miss probabilities (diagonal) and false alarm probabilities (out of the diagonal) on the OC-3s condition of the Albayzin 2010 LRE (including OOS segments)

		Target					
		eu	**ca**	**en**	**gl**	**pt**	**es**
Segment	**eu**	0.062	0.062	0.000	0.146	0.000	0.231
	ca	0.094	0.107	0.000	0.201	0.074	0.201
	en	0.000	0.007	0.052	0.000	0.007	0.000
	gl	0.116	0.223	0.000	0.141	0.074	**0.587**
	pt	0.000	0.027	0.014	0.048	0.041	0.041
	es	0.136	0.208	0.000	**0.616**	0.008	0.112
	OOS	0.149	**0.304**	0.123	0.113	0.159	0.210

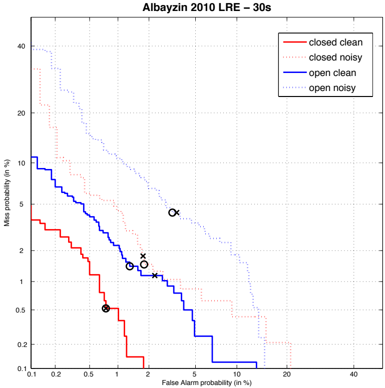
(1) OOS segments had a strong impact on false alarm rates for all the target languages:

- Strongest relative impact for Portuguese and English
- Strongest absolute impact for Catalan and Spanish

(2) Overall, best performance for English and Portuguese

(3) Highest confusion (by far) between Galician and Spanish

Performance on noisy speech (Albayzin 2010 LRE)



(1) SLR system built on clean and noisy speech signals: not specially optimized to deal with noisy speech

(2) Performance on the noisy-speech condition far worse than on the clean-speech condition (dotted vs. continuous) - see details here

(3) Moving from clean to noisy (continuous red to dotted red) produced higher degradation than moving from closed-set to open-set (continuous red to continuous blue)

(4) Performance on the Open-set Noisy-speech (ON) condition: between 2 and 6 times worse than in the Closed-set Clean-speech (CC) condition, depending on the nominal duration (the shorter the segments the smaller the differences in performance)

Conclusions (I)

- Tasks defined for Albayzin 2008 LRE more challenging than those defined for Albayzin 2010 LRE, due to:
  1. Amount of training and development data
  2. Average confusability of languages with each other
  3. Intrinsic features of the evaluation datasets (acoustic variability)
- Closely related languages (e.g. Romance languages in Spain) the most confused
- OOS segments producing a strong impact on false alarm rates for all the target languages
- Highest degradation found when dealing with noisy speech

Conclusions (II)

Most challenging conditions:

- Background noise, conversations, etc. (outdoor environments)
- Similarity of target languages (dialects)
- Amount of speech available to make decisions (short segments)
- Lack of training/development data (low-resource target languages)

Three possible setups proposed for future evaluations:

(1) **Dialect recognition:** intrinsically difficult, already addressed in NIST LRE

(2) **Large-scale European language recognition:** many closely related languages, collaboration of research groups throughout Europe required for data collection

(3) **Language recognition in the wild:** uncontrolled resources in the internet, small set of target languages, many/few/no training data

Future work (actually, current work)

Albayzin 2012 Language Recognition Evaluation

- **New KALAKA-3 database**
  - Includes all the materials of KALAKA-2 for training
  - Development and evaluation data: **any kind of speech found in the Internet**
  - Two tasks: **Plenty-of-Training** (Basque, Catalan, English, Galician, Portuguese, Spanish) and **Empty-Training** (French, German, Greek, Italian)
  - Many new OOS languages

Future work (actually, current work)

Albayzin 2012 Language Recognition Evaluation

- **New KALAKA-3 database**
  - Includes all the materials of KALAKA-2 for training
  - Development and evaluation data: **any kind of speech found in the Internet**
  - Two tasks: **Plenty-of-Training** (Basque, Catalan, English, Galician, Portuguese, Spanish) and **Empty-Training** (French, German, Greek, Italian)
  - Many new OOS languages
- **Schedule:**
  - **July 16:** registration deadline (training and development data released via web)
  - **September 3:** evaluation data released via web
  - **September 24:** deadline for submitting system results
  - **October 15:** keyfile and preliminary results released to participants
  - **November 21-23:** Evaluation Workshop, at **IberSpeech 2012**, Madrid (Spain)

Future work (actually, current work)

Albayzin 2012 Language Recognition Evaluation

- **New KALAKA-3 database**
  - Includes all the materials of KALAKA-2 for training
  - Development and evaluation data: **any kind of speech found in the Internet**
  - Two tasks: **Plenty-of-Training** (Basque, Catalan, English, Galician, Portuguese, Spanish) and **Empty-Training** (French, German, Greek, Italian)
  - Many new OOS languages
- **Schedule:**
  - **July 16:** registration deadline (training and development data released via web)
  - **September 3:** evaluation data released via web
  - **September 24:** deadline for submitting system results
  - **October 15:** keyfile and preliminary results released to participants
  - **November 21-23:** Evaluation Workshop, at **IberSpeech 2012**, Madrid (Spain)

More info at http://iberspeech2012.ii.uam.es/ (under Albayzin Evaluations)

Future work (actually, current work)

Albayzin 2012 Language Recognition Evaluation

- **New KALAKA-3 database**
  - Includes all the materials of KALAKA-2 for training
  - Development and evaluation data: **any kind of speech found in the Internet**
  - Two tasks: **Plenty-of-Training** (Basque, Catalan, English, Galician, Portuguese, Spanish) and **Empty-Training** (French, German, Greek, Italian)
  - Many new OOS languages
- **Schedule:**
  - **July 16:** registration deadline (training and development data released via web)
  - **September 3:** evaluation data released via web
  - **September 24:** deadline for submitting system results
  - **October 15:** keyfile and preliminary results released to participants
  - **November 21-23:** Evaluation Workshop, at **IberSpeech 2012**, Madrid (Spain)

More info at http://iberspeech2012.ii.uam.es/ (under Albayzin Evaluations)

You are all invited to participate !!!

GTTS Tecnologías Software

Performance (C_{avg}) on the closed-set clean-speech condition

	CC-30s	**CC-10s**	**CC-3s**
train2008 + eval2008	0.0514	0.0761	0.1722
train2008 + eval2010 (4L)	0.0275	0.0552	0.1535
train2010 + eval2010 (4L)	0.0133	0.0506	0.1466
train2010 + eval2010	0.0063	0.0263	0.0888

Back to performance on CC-30s

Performance (C_{avg}) on the open-set clean-speech condition

	OC-30s	**OC-10s**	**OC-3s**
Albayzin 2008 LRE	0.0759	0.1211	0.2004
Albayzin 2010 LRE	0.0171	0.0437	0.1094

Back to performance on OC-30s

Performance (C_{avg}) on the noisy-speech condition (Albayzin 2010 LRE)

	CN-30s	**CN-10s**	**CN-3s**
Albayzin 2010 LRE	0.0177	0.0599	0.1476
	ON-30s	**ON-10s**	**ON-3s**
	0.0390	0.0867	0.1740

Back to performance on the noisy-speech 30s condition