# University of the Basque Country (EHU) Systems for the NIST 2011 LRE

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

GTTS, Dept. Electricity and Electronics
University of the Basque Country (EHU)
Leioa, Spain

mikel.penagarikano@ehu.es

**NIST 2011 LRE Workshop**
Atlanta (Georgia), USA
December 6-7, 2011

# Outline

1. Train and development data
   - New target languages
   - Data partitioning
2. System description
   - Short description
   - Phonotactic subsystems
   - Acoustic subsystems
   - Backend & Fusion
   - Submission
3. Analysis of the results
   - Subsystem comparison
   - Post-eval analisys
4. Conclusions

GTTS Tecnologías Software

# New target languages

- 9 new target languages: Arabic Iraqi, Arabic Levantine, Arabic Maghrebi, Arabic MSA, Czech, Lao, Panjabi, Polish, Slovak.
- NIST data: 100 30-second segments per new language. Randomly split in two halves:
  - *lre11-train*, for training
  - *lre11-dev*, for development/test
- Aditional data used by BLZ consortium (*BLZ-train*)[1]:
  - Arabic Iraqi: CTS from LDC2006S45
  - Arabic Levantine: CTS from LDC2006S29
  - Arabic Maghrebi: BN speech from Arrabia TV (Morocco)
  - Arabic MSA: BN speech from Kalaka-2 (Al Jazeera)
  - Czech:
    - BN speech from the COST278 BN database
    - Telephone speech from LDC2000S89 and LDC2009S02
  - Lao: Telephone speech from VOA3 (LRE09)
  - Panjabi: no data
  - Polish: BN speech from Telewizja Polska
  - Slovak: BN speech from the COST278 BN database

[1] Broadcast news speech was downsampled to 8 kHz and applied the *Filtering and Noise Adding Tool* (FANT) to simulate a telephone channel.

# Data partitioning

- Development: restricted to segments audited by NIST.
  - The evaluation set of NIST 2007 LRE
  - The evaluation set of NIST 2009 LRE
  - lre11-dev
  - 8500 30-second segments
- Train: 66 training subsets, including target and non-target languages:
  - CTS from previous LREs (18 subsets)
  - Narrow-band speech (telephone speech?) from VOA/LRE2009 (30 subsets)
  - lre11-train (9 subsets)
  - BLZ-train (9 subsets)
  - 35000 long (>30-second) segments

# Short description

- High-level subsystems (phonotactic):
  - Czech phone-lattice phonotactic SVM
  - Hungarian phone-lattice phonotactic SVM
  - Russian phone-lattice phonotactic SVM
- Low-level subsystems (acoustics):
  - Linearized Eigenchannel GMM (Dot-Scoring) with channel compensated statistics
  - Generative iVectors
- Optional ZT-norm
- Generative backend
- Multiclass linear logistic regression
- Minimum expected cost Bayes decision

# Disk failure

- Two weeks before the submission deadline, and due to a mechanical failure of a disk we lost the LRE11 data:
  - Indexes (VOA time marks)
  - Speech wave files
  - Baum-Welch statistics
  - Expected counts of *n*-grams (up to 4-grams)
- No time to start again (nor money for professional data recovery)
- We found partial copies of:
  - Channel-compensated Baum-Welch statistics
  - Expected counts of 3-grams
- The submission was adapted to use the available data (speech signals, statistics, etc.)
  - Phonotactic subsystem was limited to 3-grams.
  - iVectors were computed on the compensated sufficient statistics space
- See: Stuck inside of a disk failure

GTTS Tecnologías Software

# Phonotactic subsystems

**Common approach to SVM-based phonotactic language recognition**

Phone Decoder → Phone-state Posteriors → Lattice → Expected counts of *n*-grams → SVM-based Language Models

Freely available software was used in all the stages:

- **Phone Decoders:** TRAPS/NN phone decoders developed by BUT for Czech (CZ), Hungarian (HU) and Russian (RU).
- **Phone-state Posteriors & Lattice:** HTK along with the BUT recipe
- **Expected counts of *n*-grams:** The *lattice-tool* from *SRILM*
- **SVM modeling:** *LIBLINEAR* (a fast linear-only version of libSVM). Modified by adding some lines of code to get the regression values (instead of class labels).

# Experimental setup

- An energy-based voice activity detector is applied to split and remove long-duration non-speech segments from signals.
- Non-phonetic units: *int* (intermittent noise), *pau* (short pause) and *spk* (non-speech speaker noise) are mapped to a single non-phonetic unit.
- A ranked (frequency-based) sparse representation, which involved only the $M$ most frequent features (unigrams + bigrams + ... + $n$-grams) is used
- SVM vectors consist of expected counts of phone $n$-grams extracted from the lattices, converted to frequencies and weighted with regard to their background probabilities as:

$$w_i = \frac{1}{\sqrt{p(d_i|background)}}$$

- The SVM language models are trained using a L2-regularized L1-loss support vector classification solver.

# Acoustic subsystems

Both systems have in common the acoustic parameters:

- 7MFCC + SDC (7-2-3-7) & gender independent 1024 mixture GMM

## Dot-Scoring

Statistics extraction → Channel compensation → Dot-Scoring

Channel matrix:

- estimated using only target languages data
- 500 channels
- 10 ML-MD iterations

## Generative iVector subsystem

iVector extraction → Generative Gaussian Language Models

Total variability matrix:

- estimated using only target languages data
- 500 dimensions
- 10 ML-MD iterations

# Backend & Fusion

- An independent backend and fusion was estimated for each nominal duration (3, 10 and 30 sec). Both the backend and the fusion were estimated with the FoCal toolkit.
- A ZT-norm was optionally applied to the scores prior to the backend
- Each subsystem produced 66 scores that were mapped to 24 target languages by means of a generative Gaussian backend
  - Discriminative Gaussian backends were tried but showed no improvement at development.
- Multiclass linear logistic regression based fusion was applied
  - Pairwise and language family-wise regressions were tried but showed no improvement at development.
- Minimum expected cost Bayes decisions were made

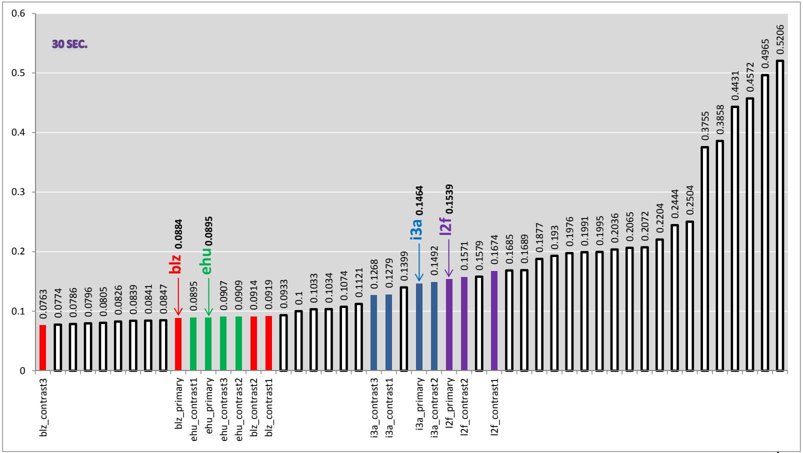
# Submission

- One primary and three contrastive systems were submitted.
- The 5 subsystems were included in each submission.
- Submissions differ in the use of ZT-norm and the development subsets used for the estimation of fusion and calibration parameters of test signals with nominal duration of 10 and 3 seconds.

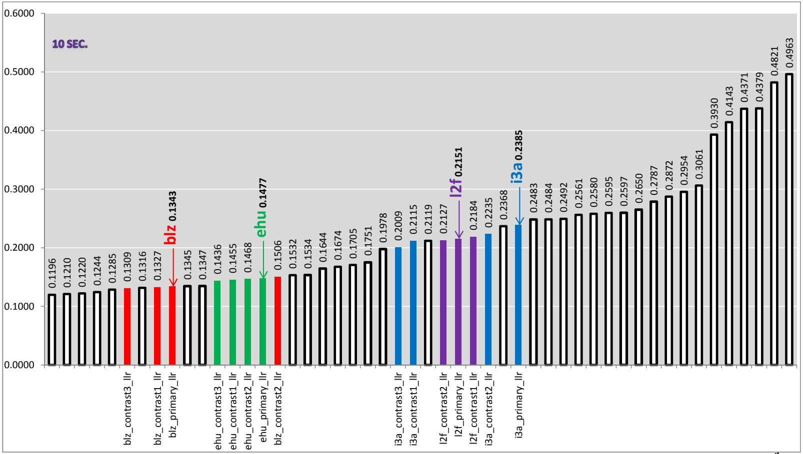
Table: Main features of the EHU primary and contrastive systems.

| System | zt-norm | Backend & Fusion Train Dataset | | |
|---|---|---|---|---|
| | | 30s | 10s | 3s |
| **Primary** | No | dev30 | dev10 | dev03 |
| **Contrastive 1** | No | dev30 | dev10+dev30 | dev03+dev10+dev30 |
| **Contrastive 2** | Yes | dev30 | dev10 | dev03 |
| **Contrastive 3** | Yes | dev30 | dev10+dev30 | dev03+dev10+dev30 |

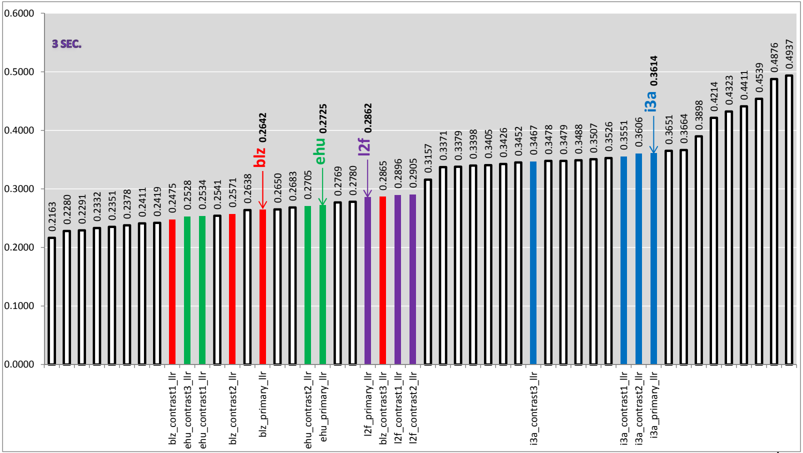
# Subsystem comparison - 30 seconds



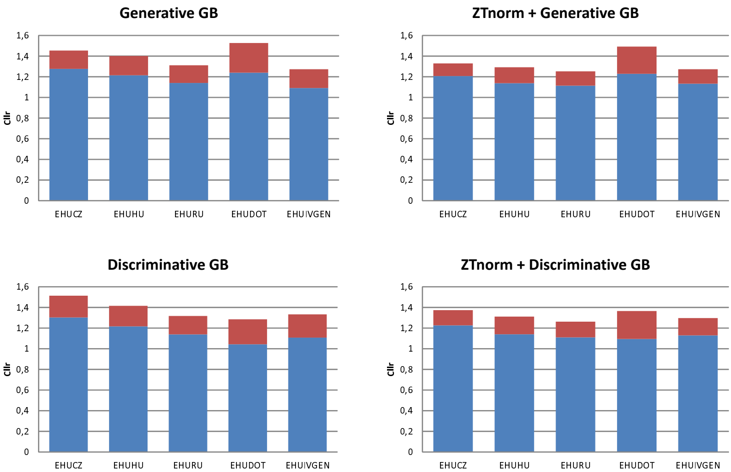
# Subsystem comparison - 10 seconds



# Subsystem comparison - 3 seconds



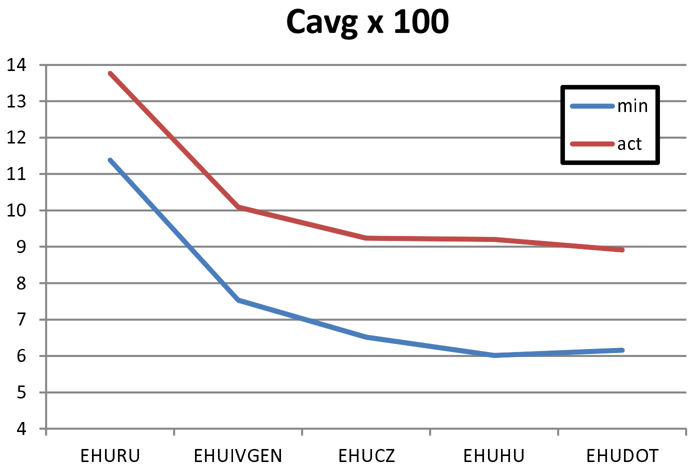
# ZT-norm & generative/discriminative backend - 30 seconds



# Phonotactic vs. Acoustic - 30 seconds

| | new Cavg x 100 | | full Cavg x 100 | |
|---|---|---|---|---|
| | min | act | min | act |
| EHUCZ | 12,15 | 14,02 | 2,97 | 3,76 |
| EHUHU | 11,96 | 14,28 | 2,71 | 3,62 |
| EHURU | 11,38 | 13,76 | 2,57 | 3,46 |
| **Phonotactic** | 7,73 | 10,13 | 1,47 | 2,28 |
| EHUDOT | 11,62 | 14,18 | 2,19 | 3,17 |
| EHUIVGEN | 11,58 | 14,15 | 2,60 | 3,50 |
| **Acoustic** | 11,18 | 13,30 | 2,00 | 2,85 |
| **ALL** | 6,16 | 8,92 | 0,94 | 1,69 |

# Greedy selection - 30 seconds



# Conclusions

- A very competitive submission was obtained based on state of the art language recognition technology.
- Data collection may have been the key.
- For 3-second tests, using a larger development set (3, 10 and 30-second segments) increased the robustness of the system.
- Unlike the BLZ submision, the ZT-norm didn't provide any improvement.
- The discriminative backend improved only the Dot-Scoring system.
- Third participation, with a great performance improvement. In 2007, avgCost was around 0,30 and in 2009 it was around 0,07.

# Thank you!