A brief Introduction to Language Recognition

Alberto Abad

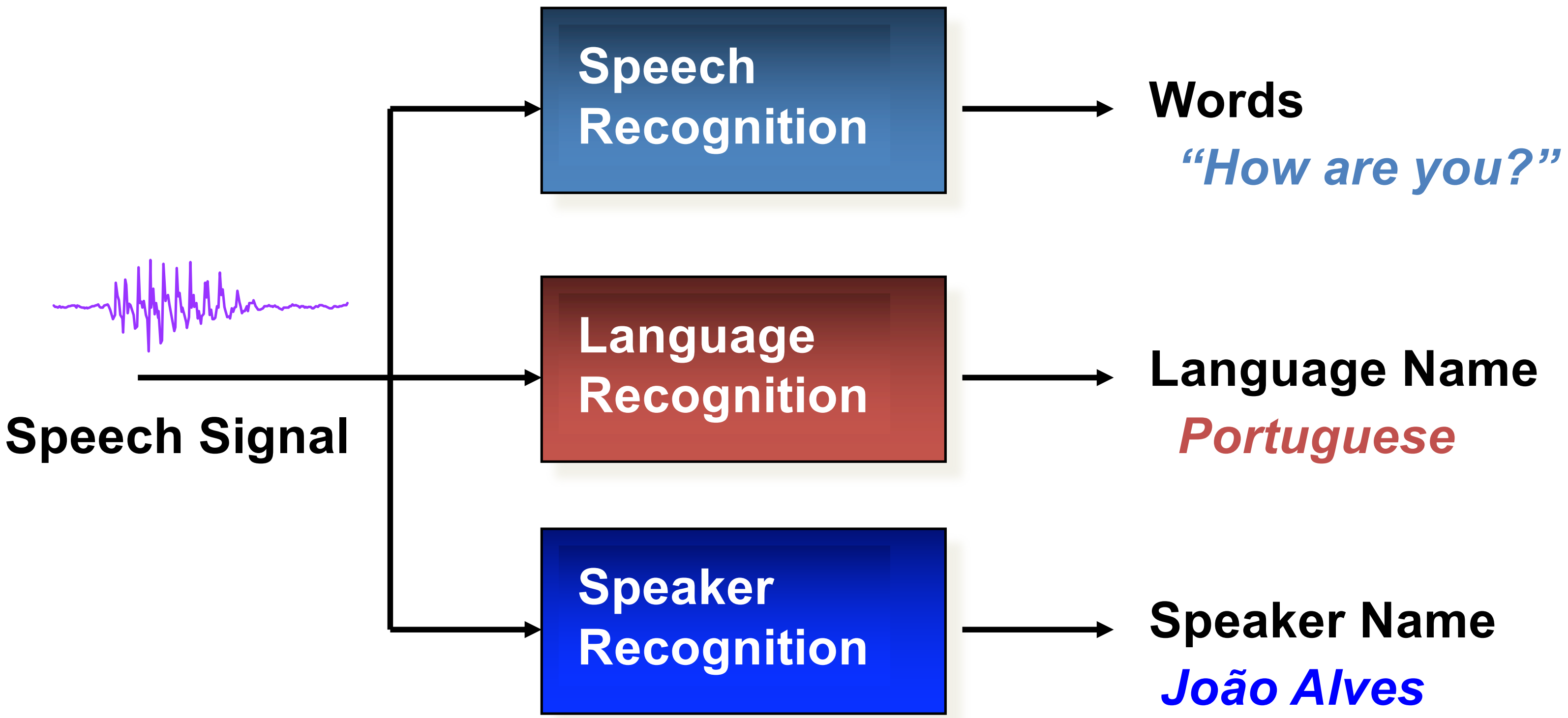
IST/INESC-ID Lisboa, Portugal

alberto.abad@tecnico.ulisboa.pt

Speech Processing - **IST** Lisboa, May, 2019

Intro: Speech processing

- Large area includes: analysis/synthesis, coding, **recognition**.



- Some commonalities
- Also many particulates → We will see some of them today!!

Intro: LR vs SR

- Speaker Recognition (SR) and Language Recognition (LR) are closely related topics that share some techniques/methods:
  - Similar feature extraction.
  - GMM short-term acoustics modeling.
  - SVM modelling (instead of GMMs) methods.

  ... but also have some particularities, ie:
  - LR: Phonotactic approaches, many samples for training, etc.
  - SR: Inter-session variability, 1-few samples for training, etc.
- SR and LR have seen great recent improvements (partially) motivated by NIST SRE and LRE competitive evalauation workshops.

Outline

- LR application and approaches
  - Acoustic approaches
  - Phonotactic approaches
- Evaluation and performance
- Other topics:
  - Variety identification
  - Native language (L1) identification

Language recognition applications

- **Language Recognition** has the potential of being of great utility in the Broadcast News processing chain:
  - Select right ASR (and other language-dependent modules)
  - Reject segments for processing in case of not-covered (or unknown) languages
  - Enrich transcription of spoken documents
  - Select/purify material for unsupervised training
- Variety or **dialect** recognition poses similar (more challenging) problems
- Recent and current work at L^2F:
  - LRE evaluation campaigns: ALBAYZIN-2008, NIST LRE 2009, ALBAYZIN-2010 (CTS & BN), LRE 201, ALBAYZIN-2012, ComParE2015 (Nativeness degree), ComParE2016 (Native language detection)
  - Portuguese variety identification: EP, BP & AP

Language recognition approaches

What does people do for Language Verification (LV)?

- Different LV approaches classified according to the kind of source of information they rely on:
  - **Acoustic phonetics:**
    - Short-term modelling with GMM, NN, SVM, i-vectors…
  - **Phonotactics:**
    - Model rules that govern phoneme combinations.
  - **Others** less common:
    - prosody, morphology, syntax…

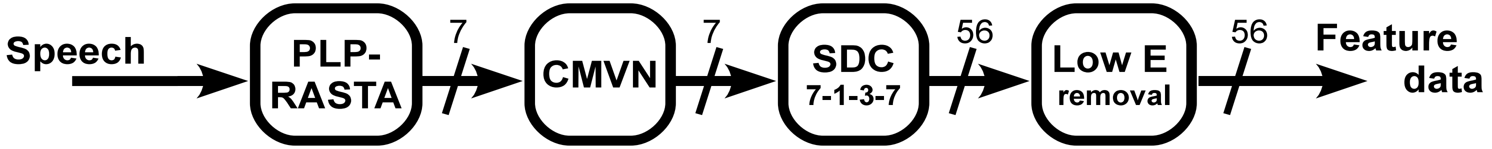
LR acoustic approaches

- LR acoustic methods are very similar to the methods used for SR, including:
  - Cepstral-based features
  - GMM-UBM
  - Gaussian supervectors
  - Factor analysis methods
  - Feature normalization, channel compensation, etc…

- Some of the differences are:
  - **Features** Use features that try to incorporate speech evolution information
  - **Models** In LR we have large amounts of samples of the target classes in contrast to SR where we usually have few utterances
    - Channel compensation is important, but it is not as dramatic as in SR
  - **Back-end** Language scores are used as a kind features for a back-end

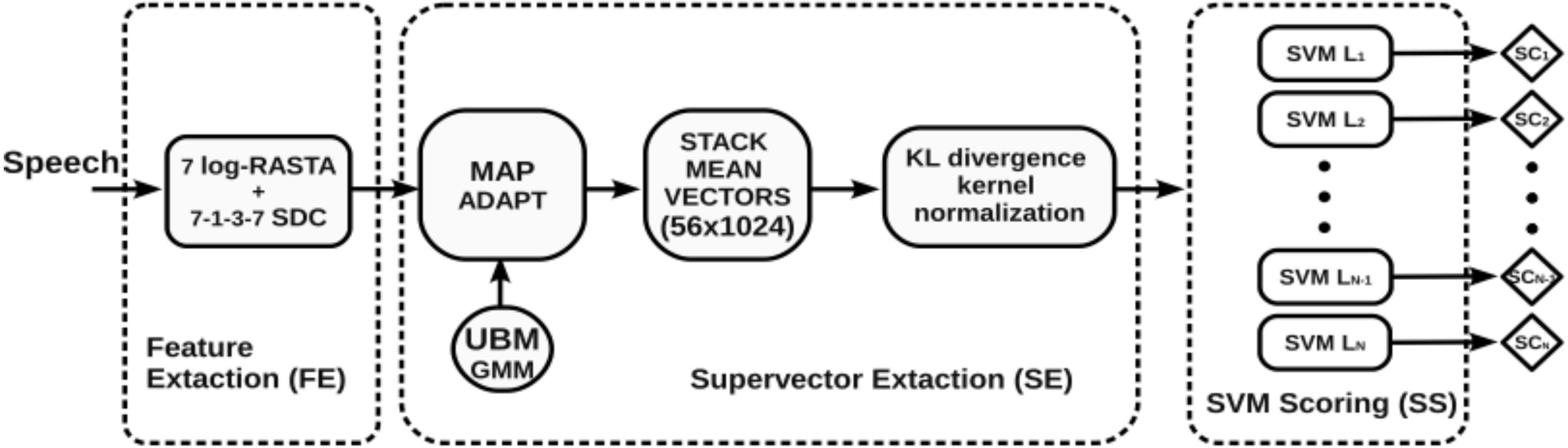
LR acoustic approaches: SDC

- Shifted-delta cepstrum (**SDC**) features are standard for acoustic based LR
  - Concatenate delta frames
  - Typical configuration 7-1-3-7
- Example of front-end (used by us in our systems):



LR acoustic approaches: GSV

- 1 LR Gaussian super-vector system (as used for NIST LRE 2011)

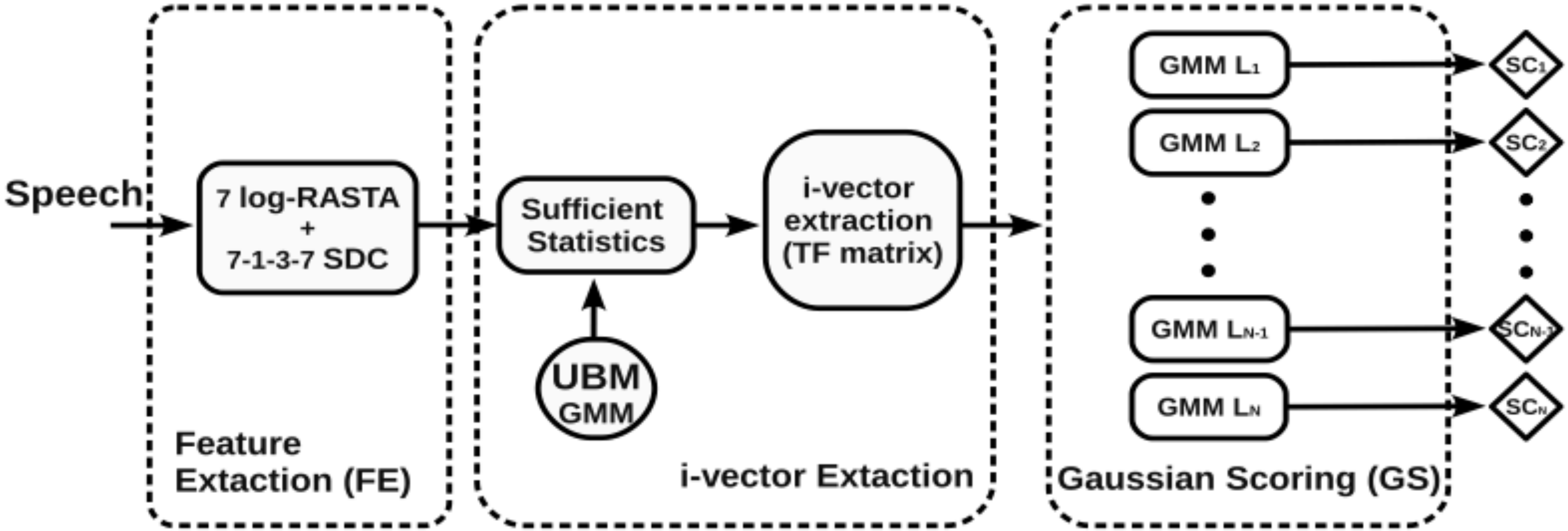


LR acoustic approaches: i-vectors

- The availability of large number of target examples has an impact on techniques.
- For instance in the i-vectors approach:
  - In SR, model and test i-vectors are extracted and cosine score is used.
  - In LR, i-vectors are used as features to train Language Models (GMM, SVM, etc…)

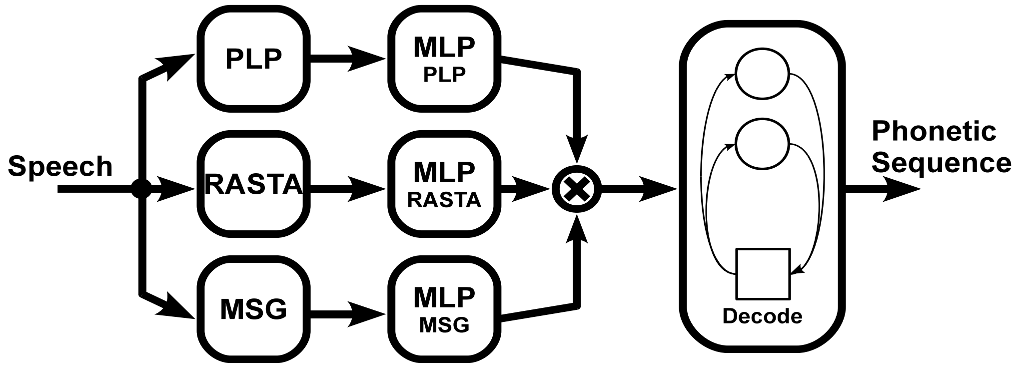
LR acoustic approaches: i-vectors

- i-vector based LR system (as used for NIST LRE2011)



LR: Phonotactics basics (PRLM)

- Use a phonetic tokenizer (of any language) to extract phonetic sequences of every speech segment:



Train For every target language, train an n-gram model with all the training sequences of this language

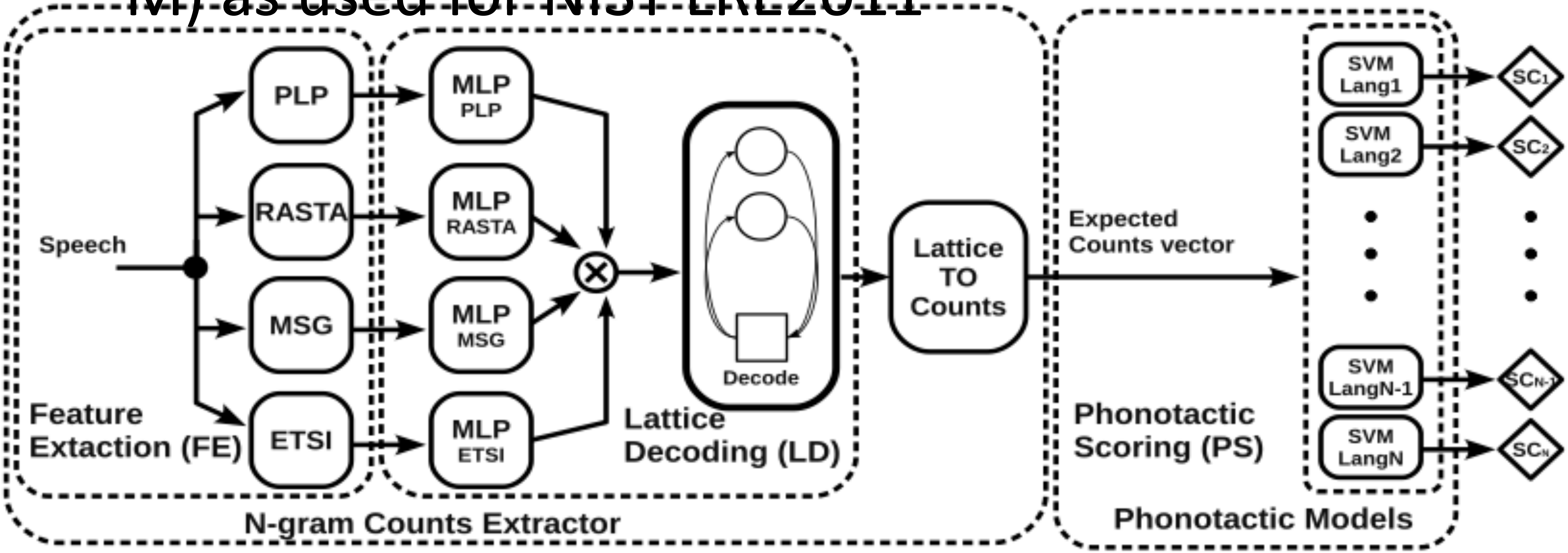
Test Tokenize test segment and compute likelihood for every target language n-gram model

LR: PRLM improvements

- PRLM methods work extremely well for LR
- Some common approaches to improve PRLM methods include:
  - Parallel systems (PPRLM)
  - Model vector of counts with SVM (instead of n-grams)
    - Use expected n-gram counts
    - Higher orders
    - Dimensionallity reduction

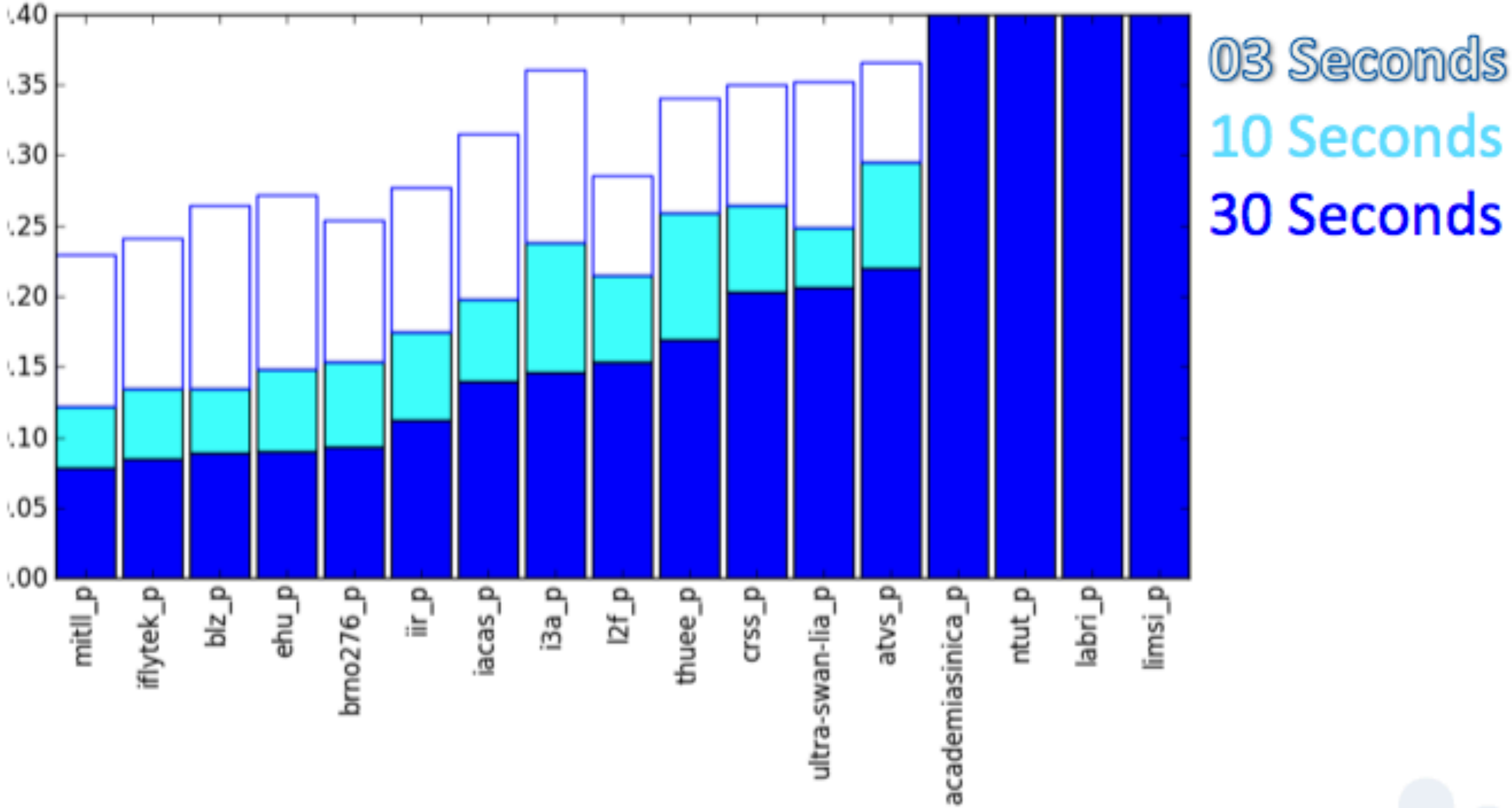
LR: PRLM improvements

- 4 Phone-recognizers followed by SVM modelling (PRSVM) as used for NIST LRE2011



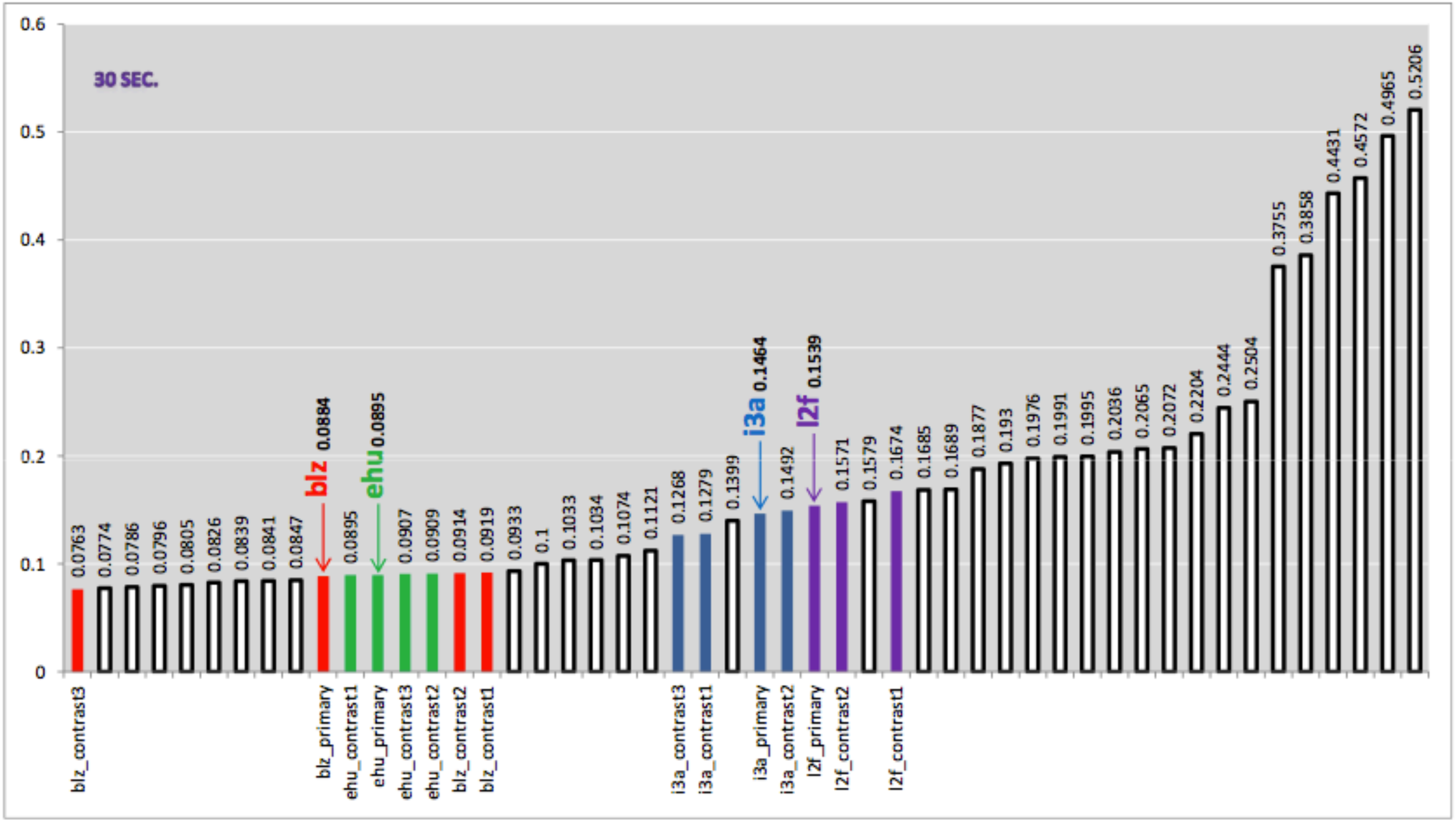
LR evaluation: NIST LRE2011

- Addressed to CTS and telephone BN data
- 24 target highly-confusable languages → Language pair detection task
- Cost → Average of the 24 more confusable pairs (worse)



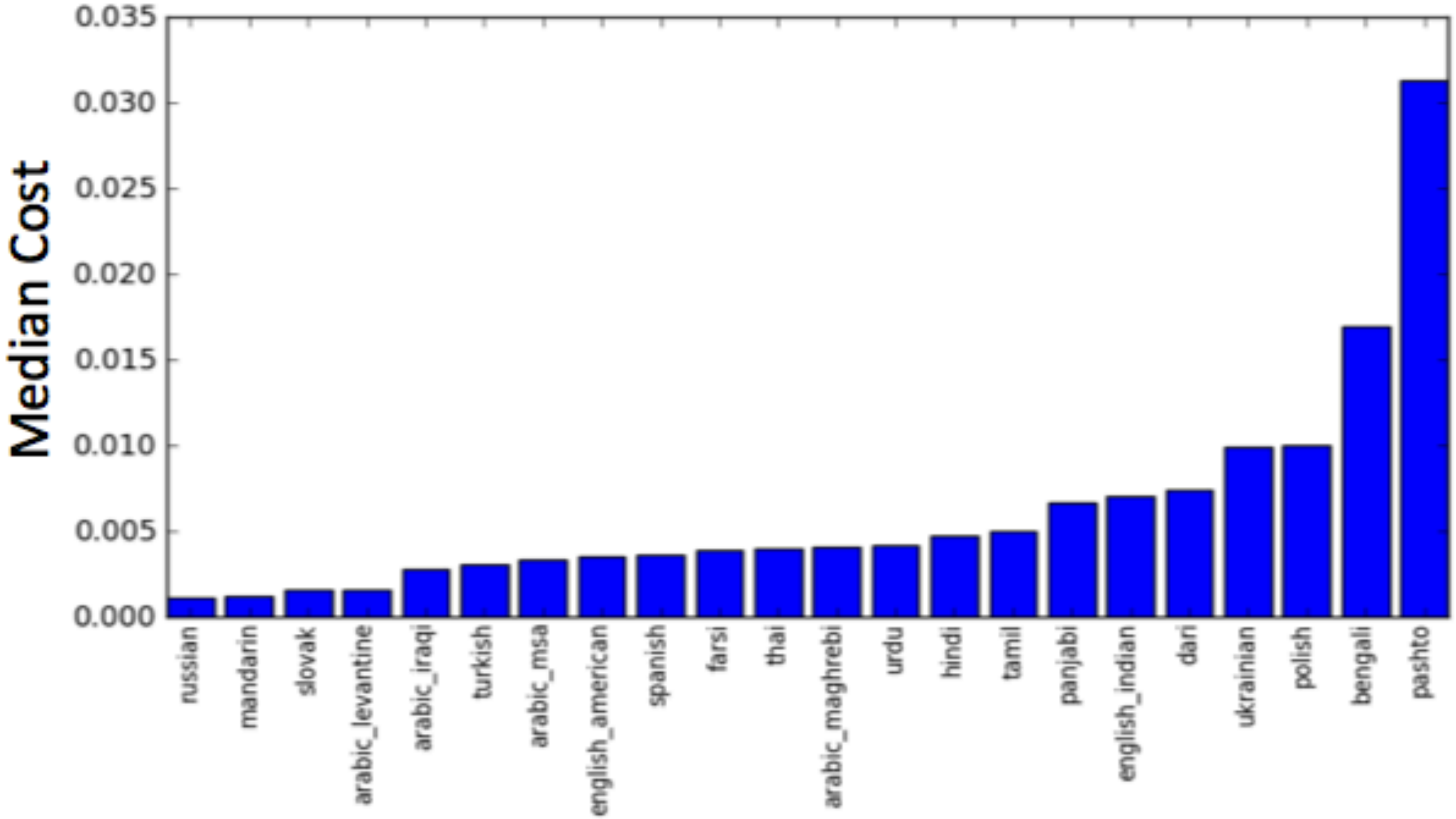
LR evaluation: NIST LRE2011

- 30 seconds (all submitted systems)



LR evaluation: NIST LRE2011

- Most confusable languages



LR evaluation: NIST LRE2011

- Most confusable pairs

Pair	min-C_{avg}	act-C_{avg}	act - min	%
Hindi-Urdu	0.2775	0.2952	0.0177	6.38
Lao-Thai	0.1232	0.1724	0.0492	39.94
Panjabi-Urdu	0.1163	0.1446	0.0283	24.33
Hindi-Panjabi	0.0656	0.0761	0.0105	16.01
Czech-Slovak	0.0578	0.0762	0.0184	31.83
Arabic_Maghrebi-Pashto	0.0484	0.0996	0.0512	105.79
Arabic_Iraqi-Arabic_Levantine	0.0471	0.0519	0.0048	10.19
Panjabi-Pashto	0.0413	0.0641	0.0228	55.21
Polish-Ukrainian	0.0401	0.0564	0.0163	40.65
Russian-Ukrainian	0.0344	0.0953	0.0609	177.03
Arabic_Levantine-Arabic_MSA	0.0342	0.0358	0.0016	4.68
Arabic_Levantine-Pashto	0.0316	0.0721	0.0405	128.16
Slovak-Ukrainian	0.0314	0.0413	0.0099	31.53
Arabic_Maghrebi-Panjabi	0.0299	0.0368	0.0069	23.08
Arabic_Iraqi-Pashto	0.0293	0.0643	0.0350	119.45
Dari-Farsi	0.0261	0.0668	0.0407	155.94
Arabic_Levantine-Arabic_Maghrebi	0.0260	0.0309	0.0049	18.85
Panjabi-Tamil	0.0254	0.0797	0.0543	213.78
Pashto-Tamil	0.0249	0.0344	0.0095	38.15
Dari-Pashto	0.0236	0.0966	0.0730	309.32
Bengali-Pashto	0.0235	0.0274	0.0039	16.60
Bengali-Panjabi	0.0229	0.0587	0.0358	156.33
English_Indian-Hindi	0.0216	0.0355	0.0139	64.35
Arabic_Maghrebi-Arabic_MSA	0.0196	0.0213	0.0017	8.67

LR other topics: Variety identification

- Use automatic variety identification to get more material for unsupervised training (BN shows with mixed AP / EP)

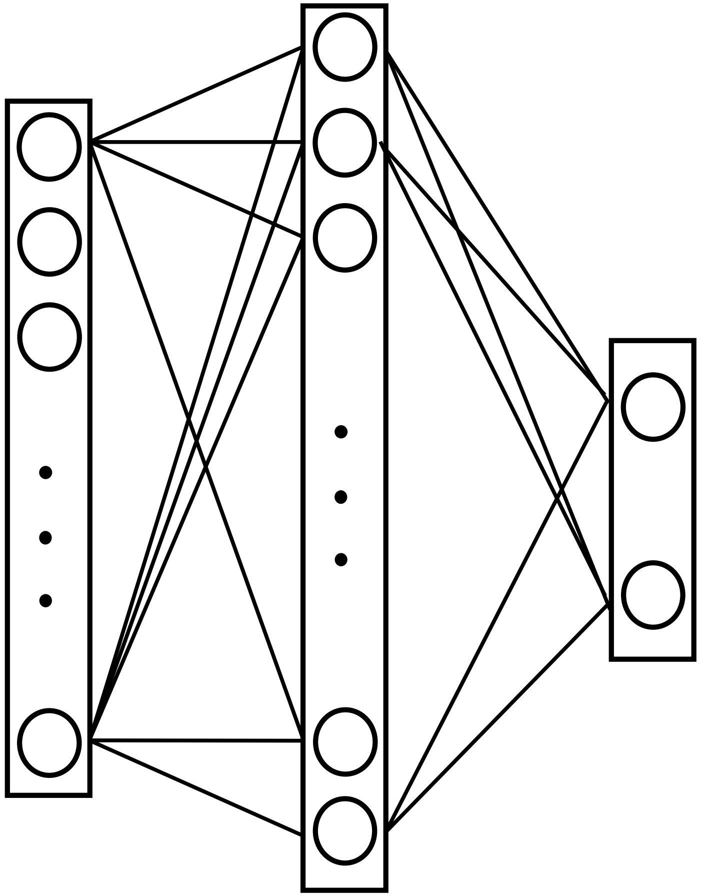
- Based on the combination (LLR fusion) of:
  - Conventional PPRLM
  - Conventional GSV
  - **NEW** PRLM mono-phonemic approach
    - Phones that appear in a single variety

1. determine mono-phones 2. train phone recognizer 3. train prlm with new phone recognizer

Train Binary MLPs for each pair (AP+EP) of phones

AP Training Corpus

EP Training Corpus



AP Transcription

... al6~j~ d$@_{AP}$ d$@_{AP}$Skr$@_{AP}$r...

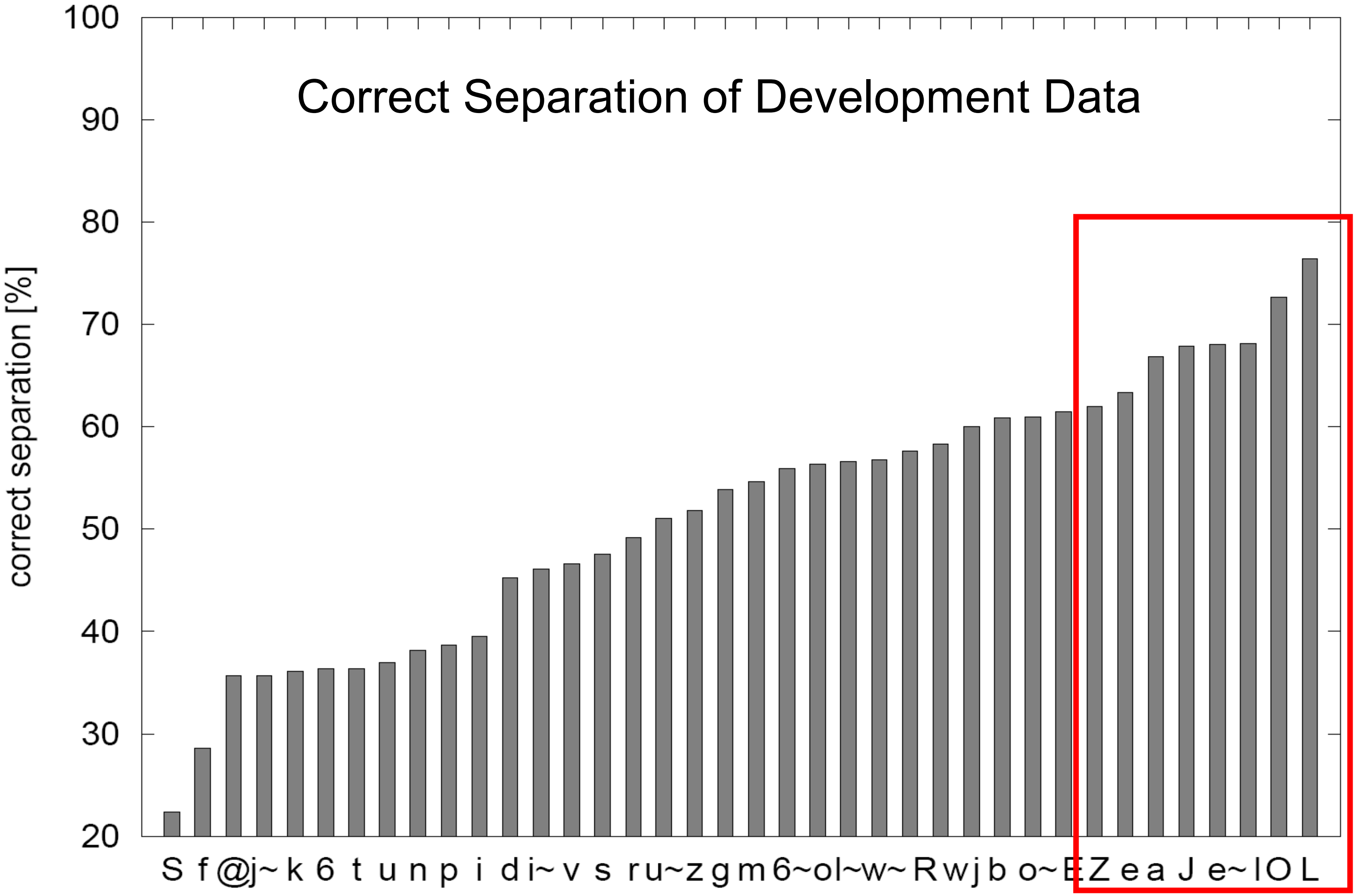
EP Transcription

... sEri$@_{EP}$ $@_{EP}$ f$@_{EP}$nO$@_{EP}$ ...

LR other topics: Variety identification

1. determine mono-phones 2. train phone recognizer 3. train prlm with new phone recognizer

/L/ /O/
/l/ /e~/
/J/ /a/
/e/ /Z/



LR other topics: Variety identification

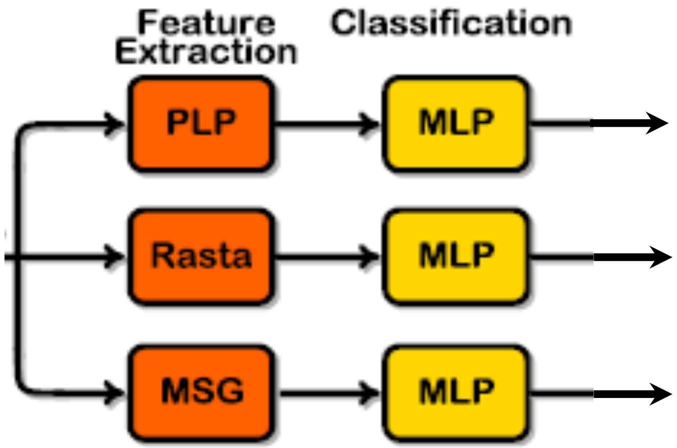
1. determine mono-phones **2. train phone recognizer** 3. train prlm with new phone recognizer

EP/AP	EP/BP	AP/BP
/L/ /O/	/o~/ /j/	/o~/
/l/ /e~/	/R/	/u~/
/J/ /a/	/u~/	/R/ /j/
/e/ /Z/	/6~/	/6~/

AP +EP

Training Corpus

(450 min each)



Mono-Phonetic Alignment of AP + EP

Outputs:

30 poly-phones+

2*8 mono-phones

+silence

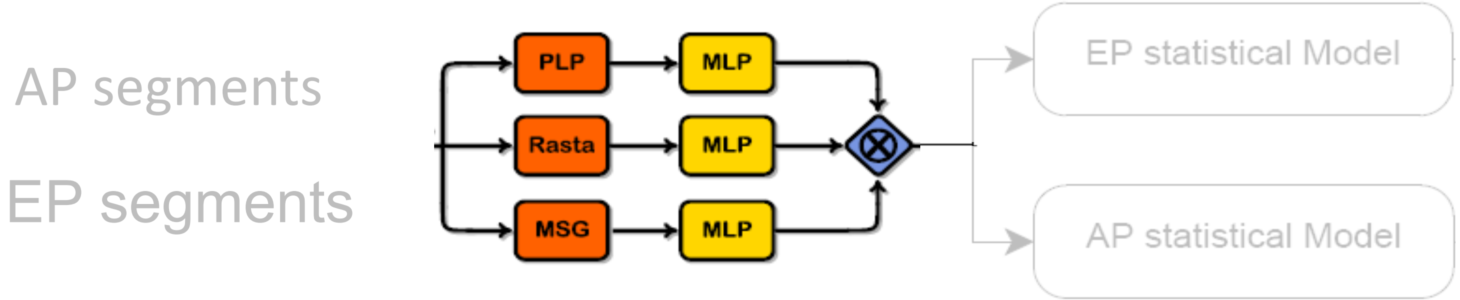
LR other topics: Variety identification

1. determine mono-phones 2. train phone recognizer **3. train prlm with new phone recognizer**

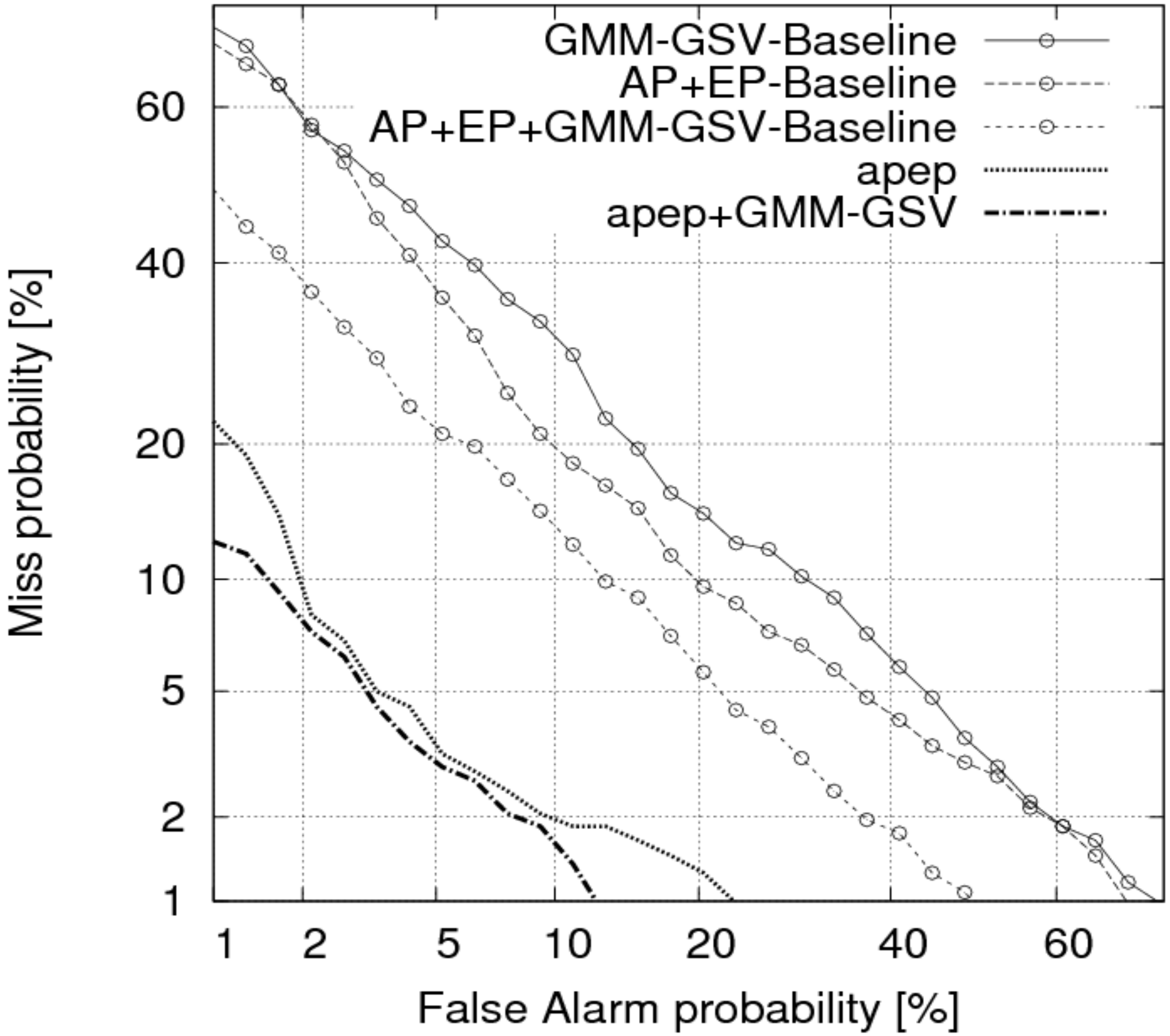
Train Data	AP	EP
duration [min.]	238.8	279.1
segments	1424	1283
∅ dur./segm. [s]	10.1	13.1
<3s [%]	16.9	0.1
3-10s [%]	42.3	49.6
10-30s [%]	38.7	44.1
>30s [%]	2.2	6.3

3-gram, Witten-Bell discounting
SRILM Toolkit [Stolcke, 2002]

1-best filtered sequence



LR other topics: Variety identification



- AP & EP varieties are the most difficult to distinguish (BP is more different)
- Nice improvements thanks to mono-phonemes
- **Our experience** it helps a lot in highly confusable pairs

LR other topics: Multi-variety ASR

	AP	BP	EP	all
AP-ASR	**24.5**	49.0	22.4	32.4
BP-ASR	52.2	**22.1**	62.1	44.8
EP-ASR	27.2	57.0	**16.7**	34.2

- Best in the diagonal (matched variety)
  - Average WER 21.1% for oracle system
  - Best individual in the complete set → AP with WER 32.4%
- Cross-variety observations
  - AP and EP closer among them than BP (in terms of ASR)
    - ASR systems are more similar
  - AP set is more challenging
  - BP most distant, but seems closer to AP?

LR other topics: Multi-variety ASR

- **MV ASR results**

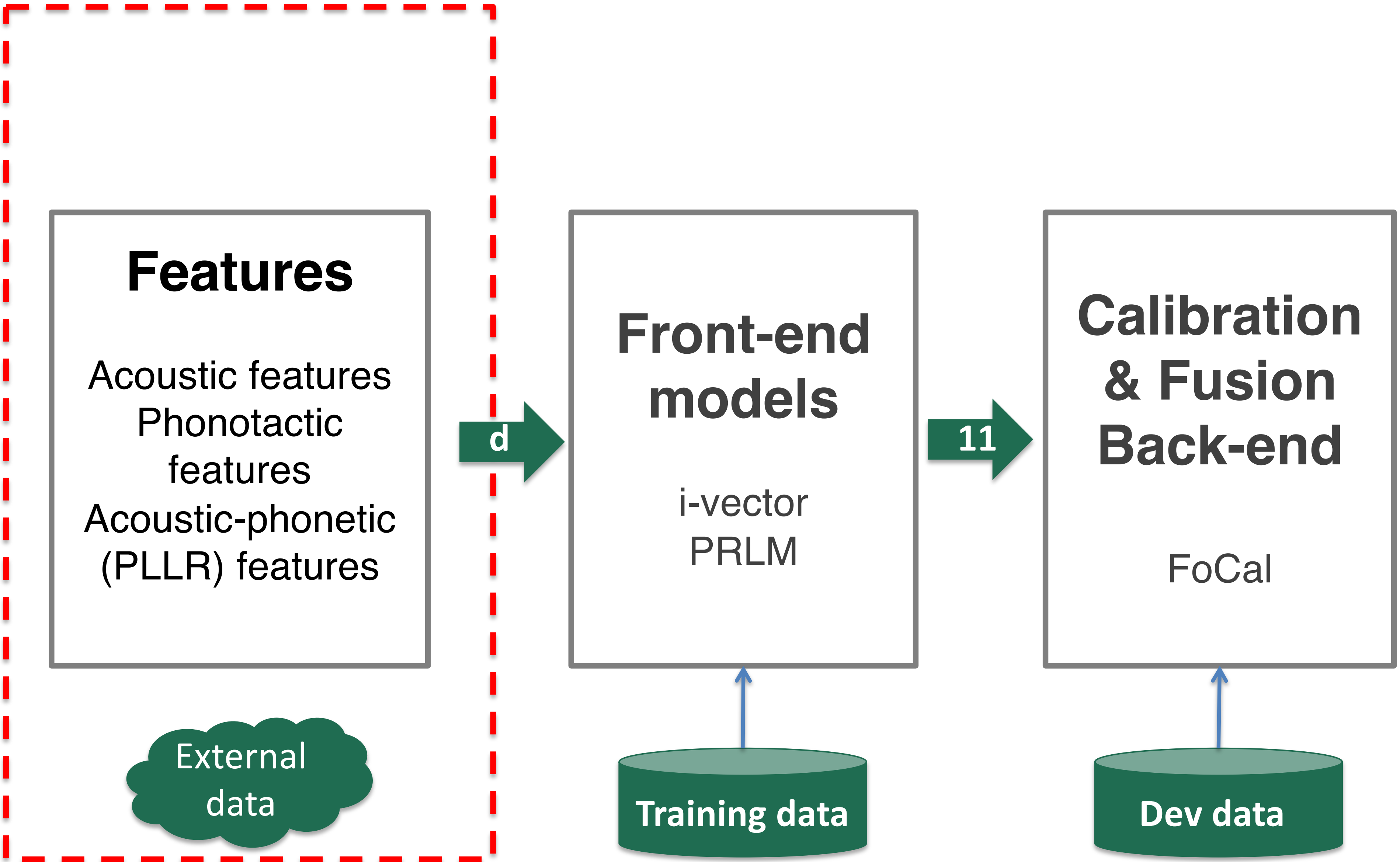
	AP	BP	EP	all
oracle ASR	24.5	22.1	16.7	$\boxed{\mathbf{21.1}}$
multi-variety ASR	24.5	22.6	21.0	$\boxed{\mathbf{22.7}}$

  - AP and BP almost equivalent to oracle
  - Significant (but not dramatic) drop in EP

LR other topics: Native Language (L1) identification

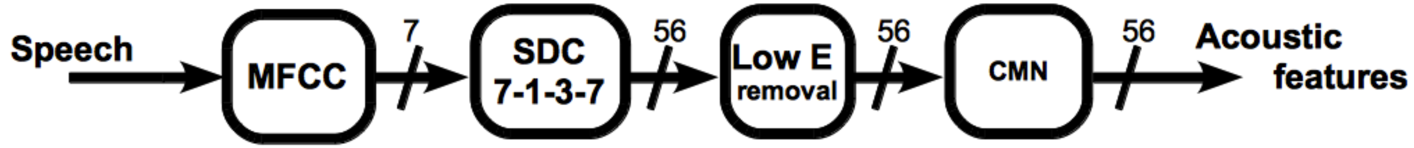
- The **ComParE 2016 Native Language** task aims at identifying L1 of non-native English speakers:
  - Similar to language, accent, and dialect ID in Spoken Language Recognition (SLR)
    - Most successful systems are based on acoustic or phonotactic information
    - Combination tends to provide increased performance
    - **Phone Log-Likelihood Ratio (PLLR)** features convey frame-by-frame acoustic-phonetic information:
      - Can be used in conventional Total Variability Factor Analysis (i-vector)
      - One of the best individual system results on relevant benchmarks

- The main **objective** is to explore PLLR features in the L1 detection task, and also:
  - Comparison of PLLR with acoustic and phonotactic approaches
  - Use of (as much as possible) in-house already available technology
  - Explore NN strategies on the top of features and i-vectors
  - Develop a (hopefully) good performing system and have fun!!

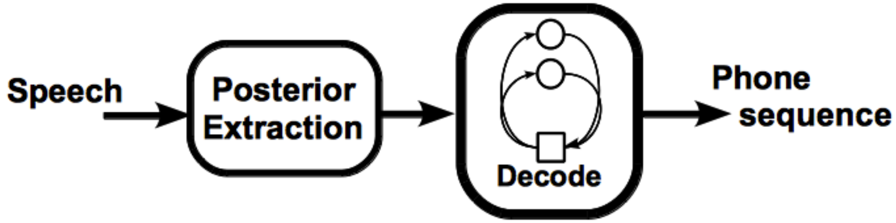

Features
Acoustic features
Phonotactic features
Acoustic-phonetic (PLLR) features
External data
d
Front-end models
i-vector
PRLM
Training data
11
Calibration & Fusion Back-end
FoCal
Dev data

Acoustic, phonotactic & acoustic-phonetic features

1. **Acoustic features**



2. **Phonotactic features**



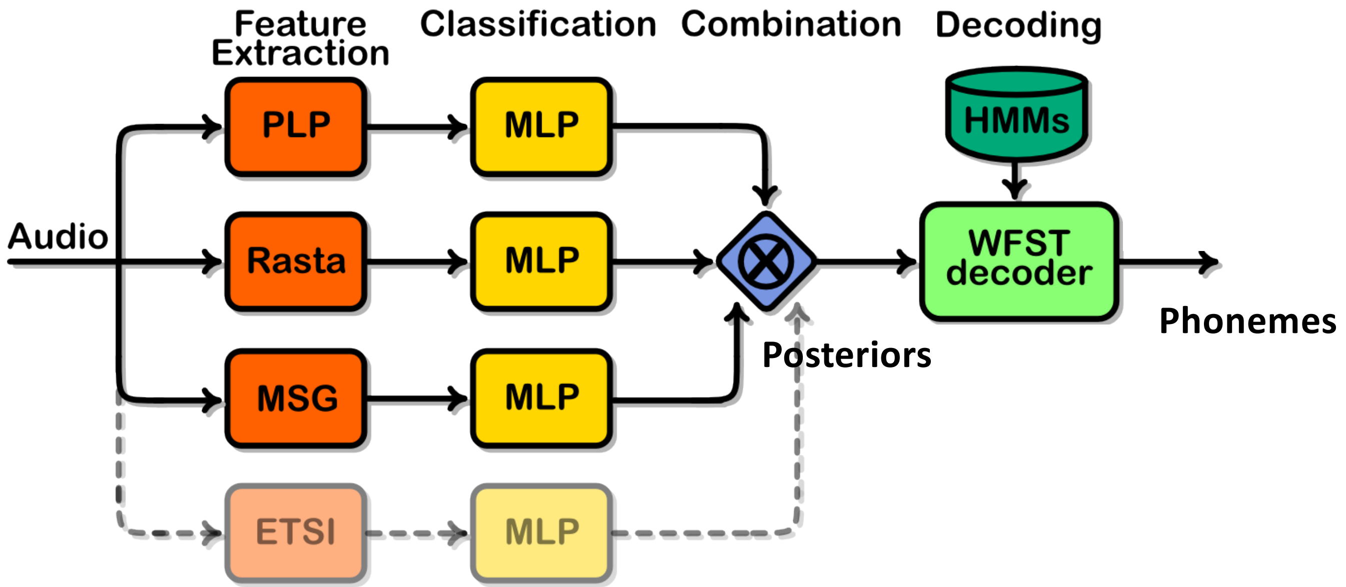
3. **Acoustic-phonetic features (PLLR)**

- Considering a phone decoder that provides frame-by-frame phone posteriors p_i, the PLLR features are computed as follows:

$$\boxed{r_i = \text{logit}(p_i) = \log \frac{p_i}{(1-p_i)} \qquad i = 1, \ldots, N.}$$

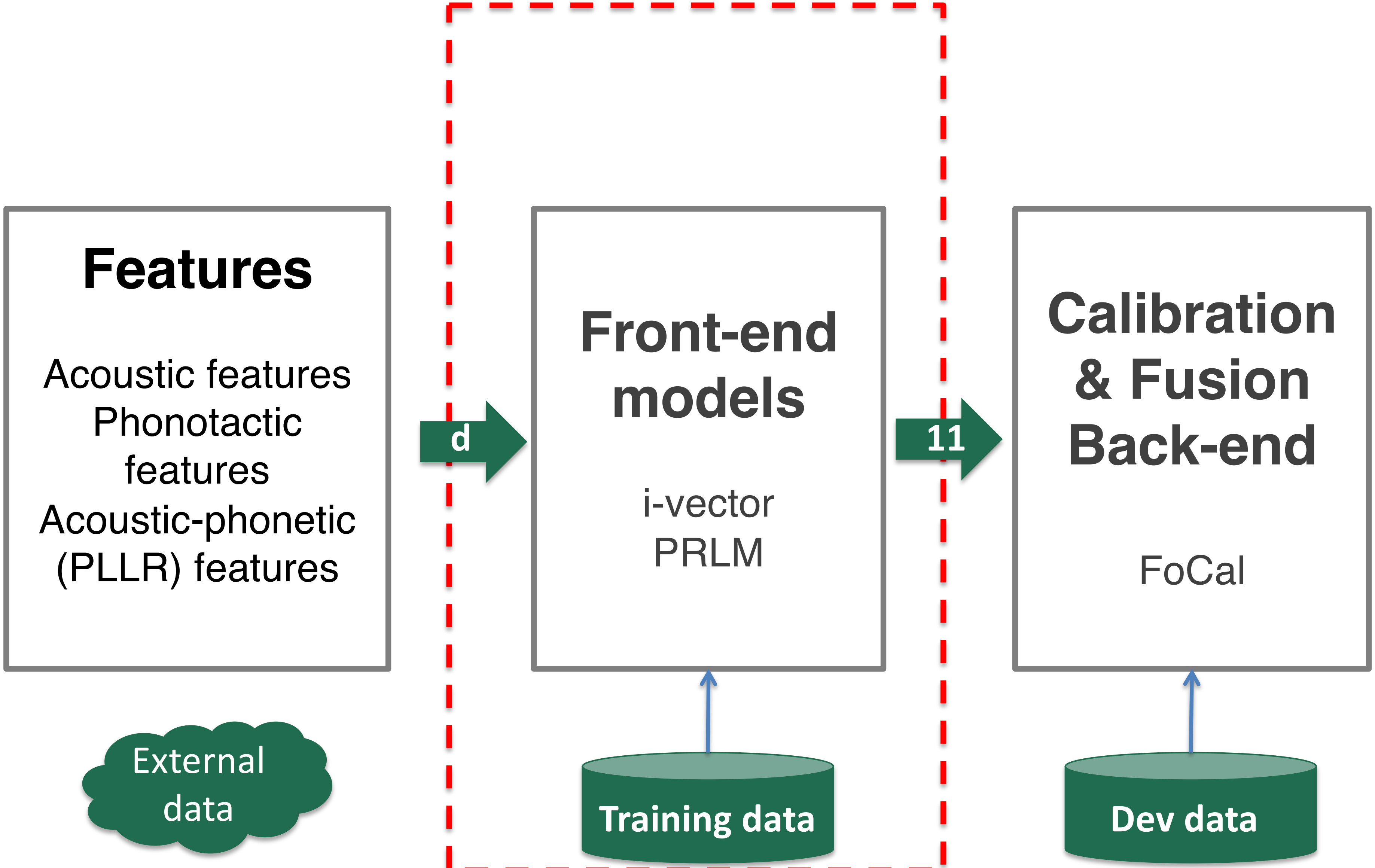
Phonetic Classifiers

- Phonetic classifiers based in in-house MLP networks are used for:
  1. Posterior probability extraction for **PLLR** feature computation
  2. Phoneme tokenization used for **phonotactic** systems



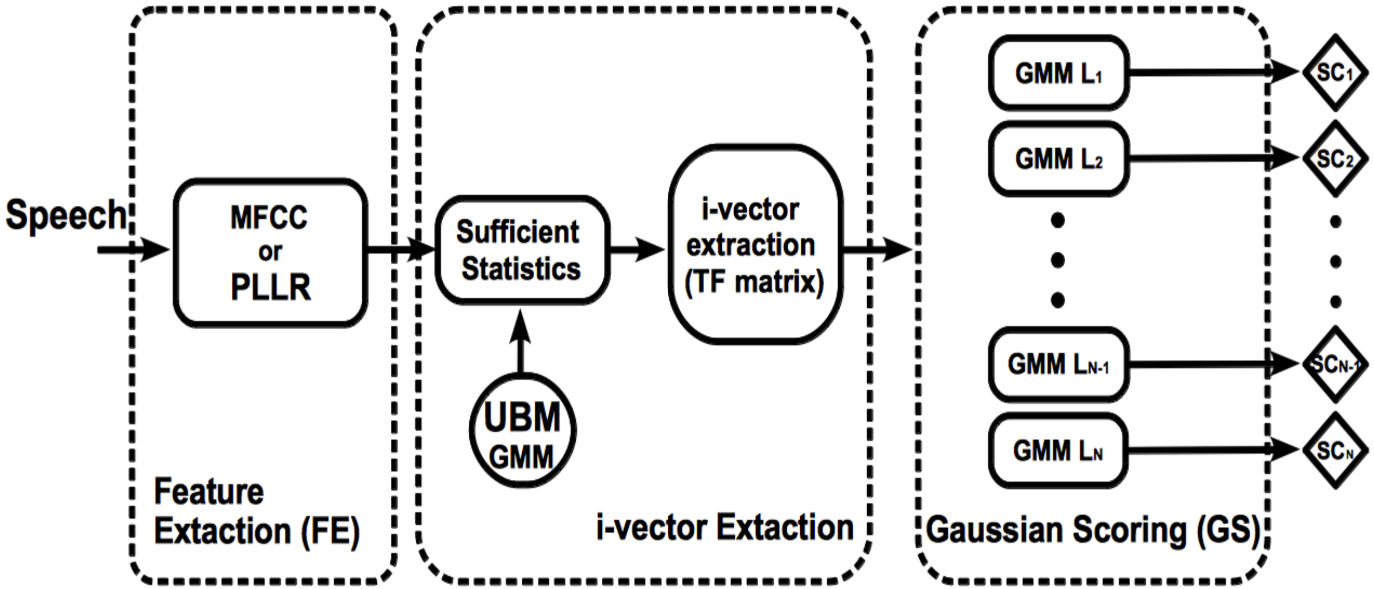
- **Feature extraction** Multi-stream 26 PLP, 26 logRASTA-PLP, 28 MSG and 39 ETSI
- **MLP** Several context input frames (13-15), 2 hidden-layers (500 units) and 1 output layer
  - <u>Output layer size</u> 39 for *pt*, 40 for *br*, 30 for *es*, 41 for *en*
- **Data** *pt* 115 hours (57 BN+58 tel); *br* 13 hours of BN data; *es* 57hours (36 BN+21 tel); en 142 hours (HUB4 96 & 97)

INESC-ID approaches for L1 identification



i-vector sub-systems

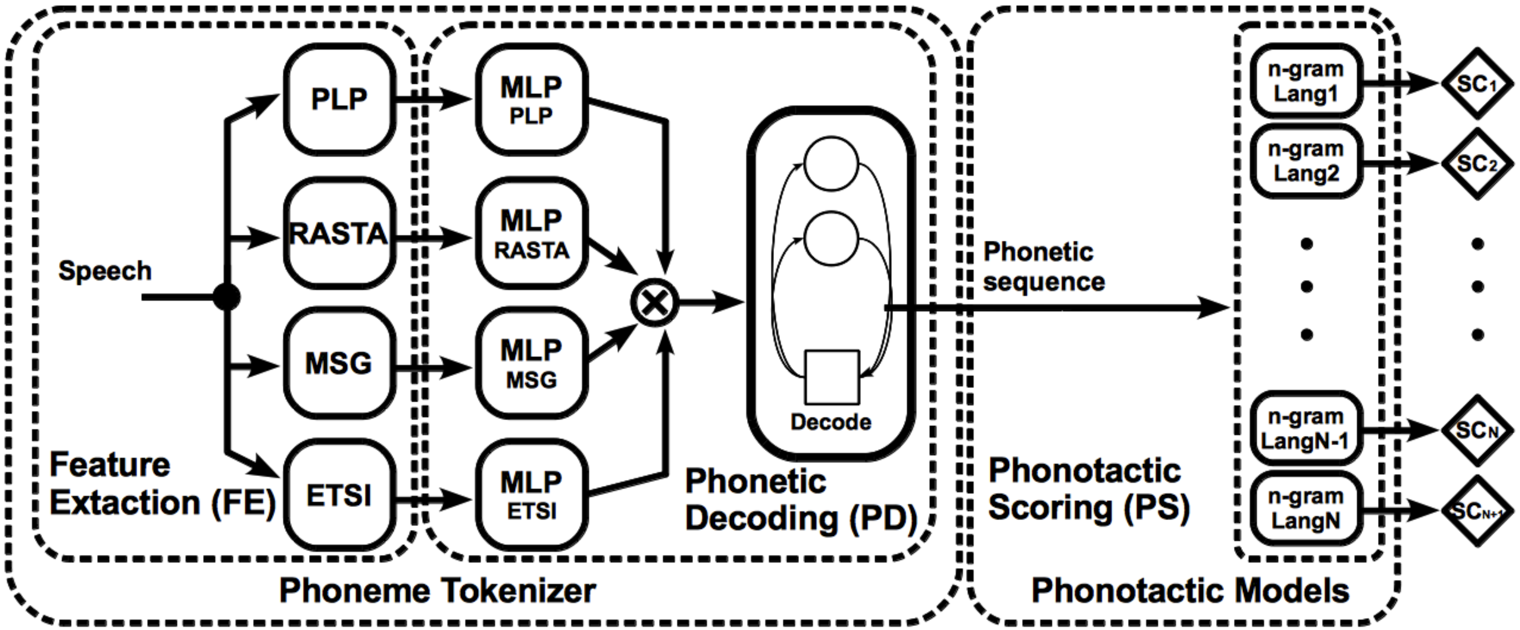
5 i-vector systems: **1 acoustic (MFCC) & 4 acoustic-phonetic (PLLR-{*en,es,pt,br*})**



- TV modelling and i-vector extraction:
  - GMM-UBM of 1024 mixtures
  - T-matrix sub-space of 400 dimensions
  - Centering + whitening + unit length norm.
- Language modelling and scoring
  - Single Gaussian with shared full-covariance
  - Log-likelihood scoring
- All the challenge training data used for UBM, T-matrix, and Gaussian modeling (no partitions on data)

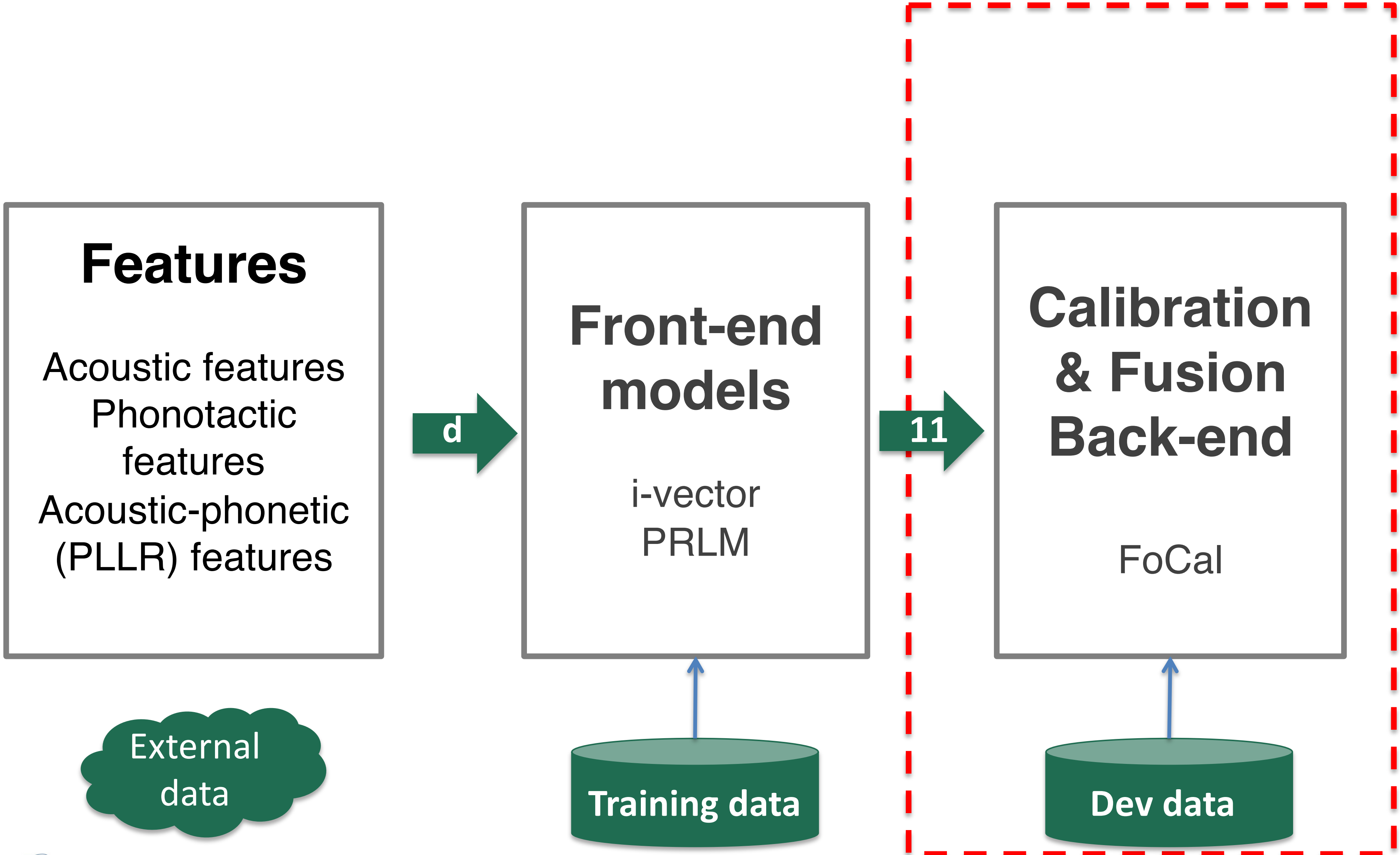
Phonotactic sub-systems

4 PRLM systems: PRLM-{*en,es,pt,br*}



- 3-gram phonotactic models trained for each L1 target language
- The 11 likelihoods of the L1 target languages form the vector of scores

INESC-ID approaches for L1 identification



Calibration & Fusion Back-end

- Linear Gaussian Back-End for each sub-system

$$\boxed{\mathbf{s}_i = \mathbf{A}_i \mathbf{x}_i + \mathbf{o}_i}$$

- Fusion of sub-systems linear logistic regression fusion

$$\boxed{\mathbf{l} = \sum_i \alpha_i \mathbf{s}_i + \mathbf{b}}$$

  - During development, the back-end parameters were trained and evaluated on the development set (kind of 2-fold cross-validation)
  - For the submissions, all the DEV data was used for fusion and calibration:
    - Possible over-fitting to DEV set
  - Calibration was carried out using the **FoCal Multi-class Toolkit**

Results in the DEV set

	UAR [%]	Acc [%]
Baseline	45.1	44.9
Phonotactic (BR)	46.4	46.2
Phonotactic (EN)	51.4	51.4
Phonotactic (ES)	50.0	49.8
Phonotactic (PT)	53.1	53.1
Phonotactic (ALL) **(I)**	63.3	63.2
i-vectors (MFCC) **(II)**	76.2	76.3
i-vectors (BR-PLLR)	76.9	76.9
i-vectors (EN-PLLR)	79.2	79.2
i-vectors (ES-PLLR)	77.6	77.4
i-vectors (PT-PLLR)	80.6	80.5
i-vectors (ALL PLLR) **(III)**	83.0	82.9
(I) + (II)	78.6	78.7
(II) + (III)	**84.6**	**84.6**

Comparison with the baseline
Results in the DEV set

ComPaRe 2016 Official Baseline

	ARA	CHI	FRE	GER	HIN	ITA	JPN	KOR	SPA	TEL	TUR
ARA	**29**	3	5	7	5	5	6	6	7	6	7
CHI	4	**38**	5	4	5	2	5	10	6	4	1
FRE	11	7	**29**	8	0	4	3	1	11	0	6
GER	5	3	5	**55**	1	7	1	2	5	1	0
HIN	4	1	1	0	**47**	2	2	2	2	21	1
ITA	6	2	9	6	6	**46**	0	4	10	1	4
JPN	4	13	4	2	2	1	**36**	11	10	1	1
KOR	4	19	1	2	2	3	14	**32**	5	3	5
SPA	6	11	15	6	2	4	9	9	**32**	1	5
TEL	2	0	2	2	24	2	2	2	2	**43**	2
TUR	6	5	5	5	2	6	7	8	5	0	**46**

INESC-ID ComPaRe 2016 system

	ARA	CHI	FRE	GER	HIN	ITA	JPN	KOR	SPA	TEL	TUR
ARA	**77**	0	3	1	0	1	1	0	1	0	2
CHI	0	**78**	0	1	0	1	2	0	1	1	0
FRE	3	0	**64**	2	0	2	2	0	5	0	2
GER	2	1	2	**78**	0	0	0	1	0	0	1
HIN	0	0	0	0	**67**	0	0	0	0	16	0
ITA	1	0	5	2	0	**79**	1	1	3	0	2
JPN	1	1	1	0	0	0	**70**	8	4	0	0
KOR	2	4	1	1	0	0	5	**77**	1	0	0
SPA	2	1	2	1	0	5	4	5	**77**	1	2
TEL	0	0	0	0	18	0	0	0	0	**65**	0
TUR	0	1	1	3	1	2	0	2	1	0	**84**

Final results in the TEST set

	DEV [UAR %]	TEST [UAR %]
ComPaRe 2016 Official Baseline	45.1%	47.5%
INESC-ID ComPaRe 2016 system	**84.6%**	**81.3%**

COMPARE 2016 - Quiz

1 2 3 4 5 6 7

Arabic | French | German | Italian | Japanese | Mandarin Chinese | Spanish

Conclusions

- Language recognition is a very active research field in the area of speech processing:
  - Some overlap in techniques (and community) with speaker recognition
- Recent advances (fostered by International evaluations) have led the technology to:
  - High performances for certain tasks.
    - In some cases, better than humans
  - Exploring more challenging tasks:
    - similar language pairs, variety/accent, L1, etc.
- Most common approaches are based on modelling of short term acoustics:
  - Current leading methods are based on factor analysis (JFA, i-vectors…).
  - Recent: Short-time feature extraction based on NN (posteriors, bottle-neck, etc.)
  - However, best systems are based on the combination of several sub-systems.


technology
from seed
inesc id
lisboa
L^{2}F - Spoken Language Systems Laboratory