Frame-Level Features Conveying Phonetic Information for Language and Speaker Recognition

Mireia Diez

GTTS, Department of Electricity and Electronics.
University of the Basque Country, UPV/EHU
mireia.diez@ehu.es

Leioa, September 2015

GTTS Tecnologías Software

GTTS Tecnologías Software

The Task

Spoken language recognition is a pattern recognition task that consists of recognizing the language spoken in an utterance by computational means

The main **complexity** of language recognition tasks comes from dealing with undesired variabilities present in the utterances due to several factors: **channel variabilities**

Extracting informative features robust against those variabilities or designing modeling techniques capable of patterning after the desired variabilities, while discarding the noisy ones, are the main focus of research in the task

GTTS Tecnologías Software

Applications

The possible applications include, among others:

- Phone service automation
- Phone call filtering
- Search and indexing of audiovisual resources
- Product customization
- Intelligence
- Forensics
- Speech preprocessing in multilingual dialog systems
- Speech preprocessing for suitable model/dictionary selection in data recovery systems
- Security

GTTS Tecnologías Software

Evolution

Applications becoming more common in the technology surrounding us, thanks to the great human effort put on it in the last decades which has promoted the evolution of SLR systems.

Two main factors have enabled this evolution:

- Growth in computational power
- The increasing amount of available data

GTTS Tecnologías Software

Datasets

NIST Datasets Starting on 1996 and held every two years from 2003 to 2011: Narrow band (8kHz) signals. Mainly conversational telephone speech. Evaluations for 3s, 10s and **30s** speech segments

- **NIST 2007 LRE:** 14 target languages
- **NIST 2009 LRE:** 23 target languages, radio broadcasts included as resources (only telephone bandwidth speech)
- **NIST 2011 LRE:** Focused on pairwise language detection task, 24 target languages

GTTS Tecnologías Software

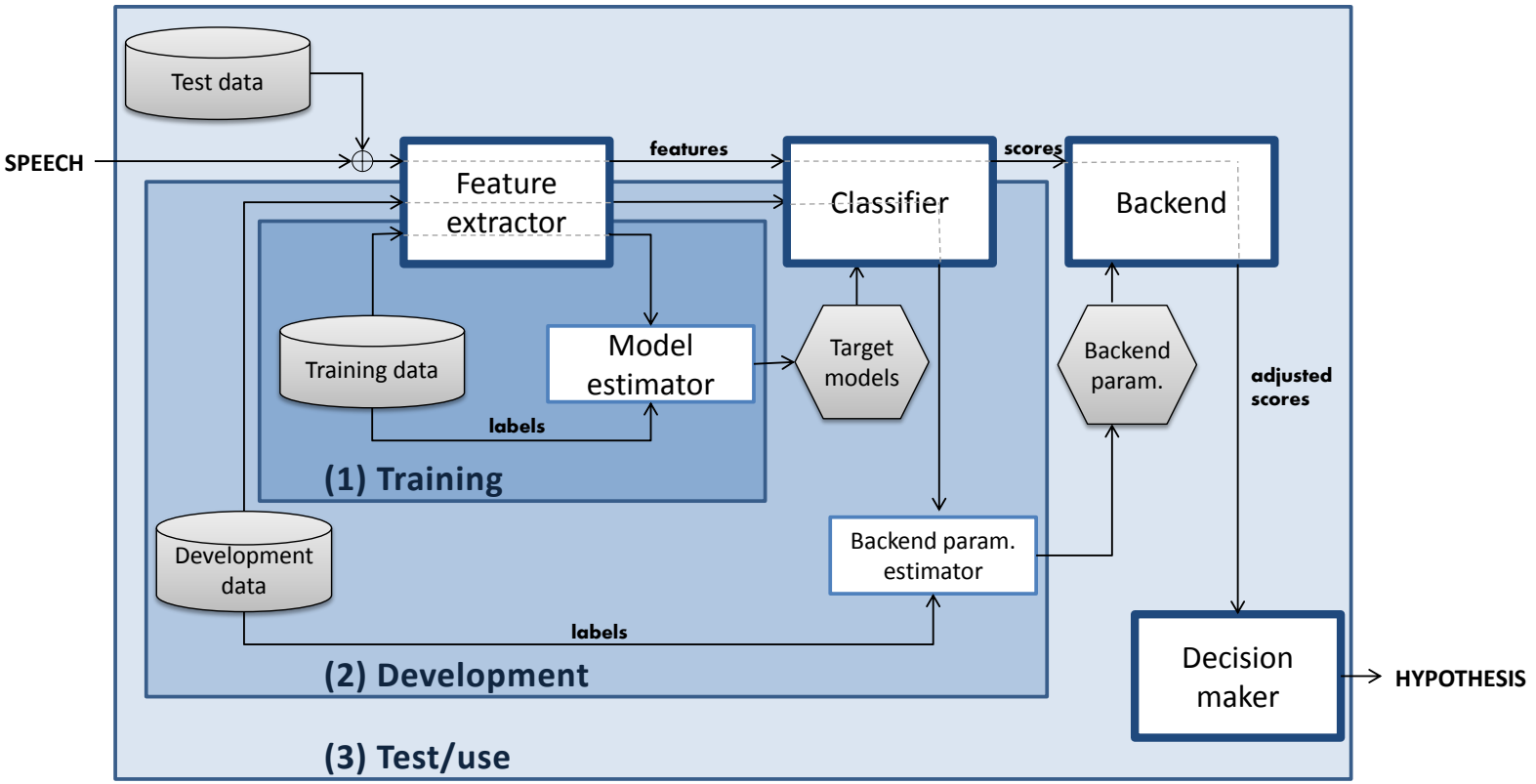
Datasets

KALAKA-2: Extension of KALAKA. Wide-band TV broadcast speech signals downsampled to (single channel) 16kHz. 6 target languages. Two main evaluation tasks: clean and noisy speech

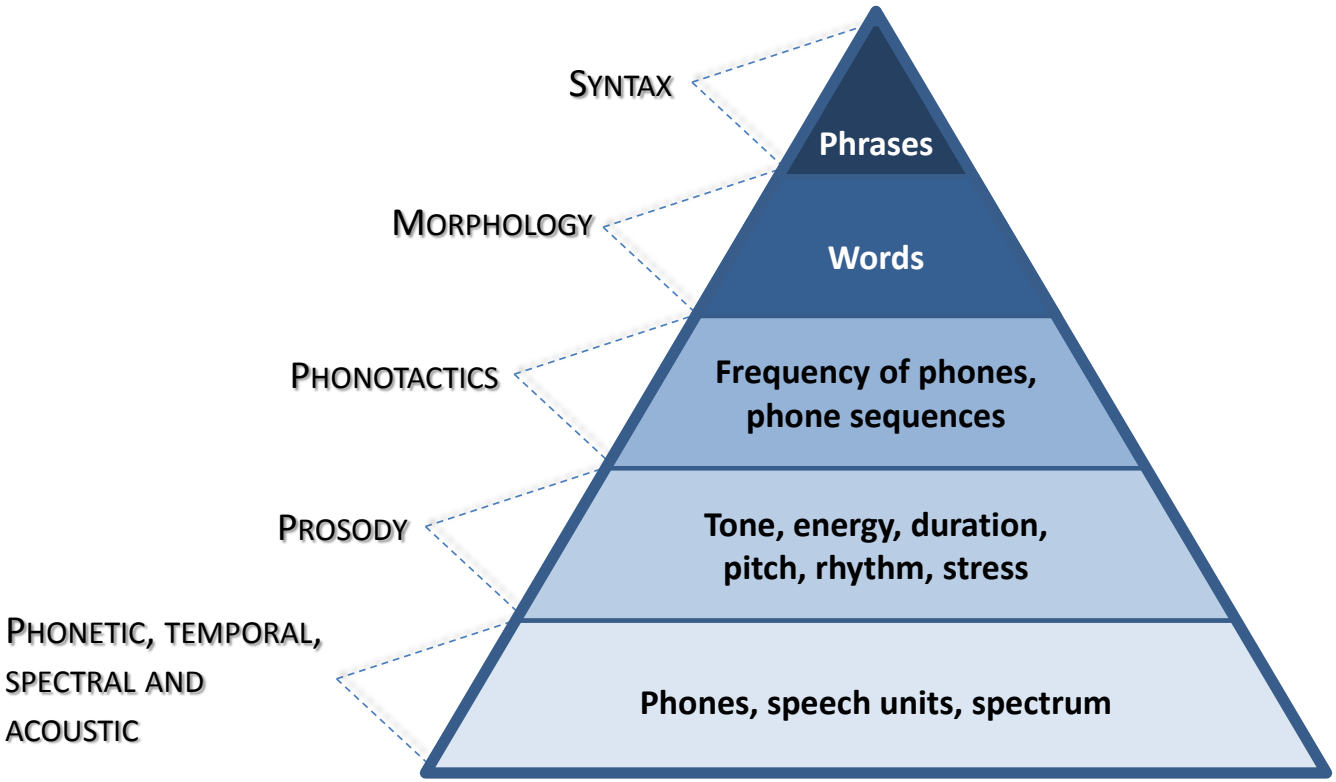
RATS: Designed to perform SLR in challenging scenarios, focusing on noisy environments. Data for 5 target languages, retransmitted through 8 different communication channels

GTTS Tecnologías Software

Structure of Recognition Systems

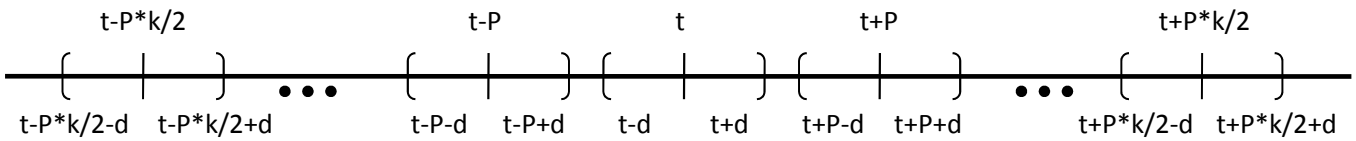


Feature Extraction



Acoustic Features

- **Mel Frequency Cepstral Coefficients (MFCC)**: frame level features based on Mel scale of frequency
- **Shifted Delta Cepstrum (SDC)**: characterize the language by the evolution of local variations of the spectrum around the analysis window, are specified by four parameters: N-d-P-k



Baseline acoustic system

- Features are used to train a **Gaussian Mixture Model** (**GMM**), that is used as **Universal Background Model** (**UBM**)
- Acoustic approaches: MAP-GMM, eigenchannel compensation
- **i-vectors**: The i-vector approach maps high-dimensional input data to a low-dimensional feature vector, hypothetically maintaining most of the relevant information

GTTS Tecnologías Software

i-vector system

Under the i-vector modeling assumption, an utterance GMM supervector (stacking the means of a GMM which is estimated by MAP adaptation of the UBM to the input utterance) is defined as:

$$\mathbf{M} = \mathbf{m} + \mathbf{T}\mathbf{w} \tag{1}$$

- $\mathbf{m}$ is the utterance independent mean supervector
- $\mathbf{T}$ is the total variability matrix (a low-rank rectangular matrix)
- $\mathbf{w}$ is the so called *i-vector*, a normally distributed low-dimensional latent vector
- $\mathbf{M}$ is assumed to be normally distributed with mean $\mathbf{m}$ and covariance $\mathbf{T}\mathbf{T}'$
- The latent vector $\mathbf{w}$ can be estimated from its posterior distribution conditioned to the Baum-Welch statistics

GTTS Tecnologías Software

- **Generative modeling approach** for i-vectors (each language modeled by a single Gaussian distribution)
- i-vectors classified using **Logistic Regression**
- i-vectors used to train **Neural Network** Classifiers

High level features

Most common representations are based on the information provided by **phone decoders**

- Large amount of labeled data to train the models for each of the phones of their **phonetic inventory**
- Rely on different **features**: MFCCs, PLPs, etc. Usually augmented to obtain information for larger temporal contexts
- Input frames are scored with regard to the trained phonetic units of the phonetic inventory. Different **modelings**: GMMs, HMMs, NNs, Hybrid models

GTTS Tecnologías Software

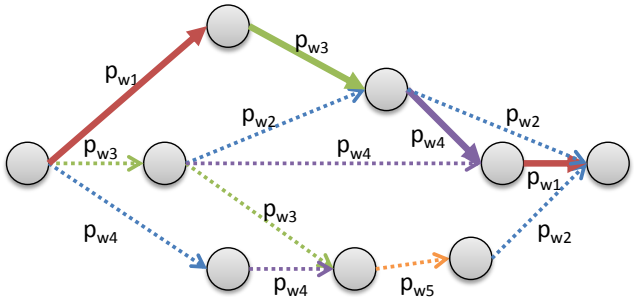
Phone decoders

- Phonemes usually span several frames. **Phone-state posteriors** represent shorter units and allow a **better acoustic modeling** of the speech

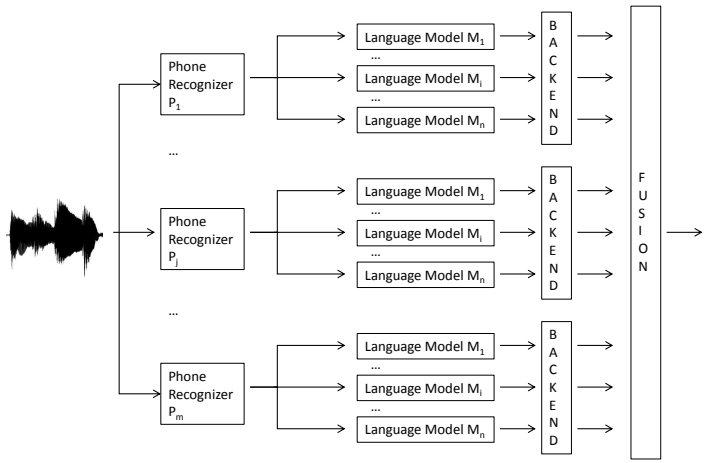
Given an input sequence of acoustic observations X, they provide an acoustic posterior probability of each state s $(1 < s < S)$ of each phone model i $(1 < i < N)$ at each frame t, $p(i|s;t)$

Phonotactic approaches

1-best decoding: the most likely phone sequence
Phone lattices: perform a summation over phone sequences



PRLM: phone recognition language modeling
PPRLM: parallel phone recognition language modeling, different phonotactic scores combined



Phone-lattice-SVM approach

- **Support Vector Machines**, are **discriminative models** that try to find the boundaries between two classes
- The phone lattice produced by a decoder i is stored for each target language j, then **feature vectors** are built from **expected counts of phone n-grams**
- A **SVM model is estimated on the outputs of the phone decoder** i for the training dataset, taking j as the target language

GTTS Tecnologías Software

Calibration and Fusion

- **Score Normalization** Helps removing the environmental effects on the score space: Z-Norm, T-Norm, ZT-Norm
- **Backend** Is a calibration stage that transforms the space of scores to get reliable estimates of the class probabilities, and can map scores to the space of target languages: Generative/Discriminative Gaussian backends, Logistic Regression, etc.
- **Fusion** Combines scores of several systems to give a final set of calibrated and fused scores

Evaluation Measures

Average Cost (C_{avg}): A combination of P_{miss} and P_{fa} pooled across target languages. For closed-set evaluation tasks:

$$C_{\text{avg}} = \frac{1}{L}\sum_{i=1}^{L}\left\{C_{\text{miss}}P_{\text{T}}P_{\text{miss}}(i) + \frac{1}{L-1}\sum_{l_N} C_{\text{fa}}(1-P_{\text{T}})P_{\text{fa}}\left(l_T, l_N\right)\right\} \quad (2)$$

Average Cost (C_{avg}^{24}): The primary measure for NIST 2011 LRE, averages the actual cost for the 24 pairs with the highest minimum cost

C_{LLR}: Evaluates system performance globally by means of a single numerical value. It only depends on the scores (not on application dependent parameters), on their ability to discriminate amongst target languages and on how well they are calibrated

GTTS Tecnologías Software

GTTS Tecnologías Software

Phone Log-Likelihood Ratio Features

- **Phone Log-Likelihood Ratios (PLLR)**: features for **language** and **speaker** recognition
- PLLRs designed to provide **acoustic-phonetic information** in a sequence of **frame-level feature vectors** that can be easily plugged into acoustic systems by simply replacing the MFCC-based features

GTTS Tecnologías Software

Phone Log-Likelihood Ratio Features

- Phone decoder including: N phone units, represented by a model of S states. Given an input sequence of acoustic observations X, we assume that the acoustic posterior probability of each state s $(1 \leq s \leq S)$ of each phone model i $(1 \leq i \leq N)$ at each frame t, $p(i|s,t)$, is output by the phone decoder
- The acoustic posterior probability of a phone unit i at each frame t can be computed by:

$$p(i|t) = \sum_{\forall s} p(i|s,t) \tag{3}$$

GTTS Tecnologías Software

Phone Log-Likelihood Ratio Features

In this way, the decoder outputs an n-dimensional vector of phone posteriors at each frame t: $\mathbf{p}(t) = (p(1|t), p(2|t), \ldots, p(n|t))'$

$$\sum_{i=1}^{n} p(i|t) = 1 \tag{4}$$

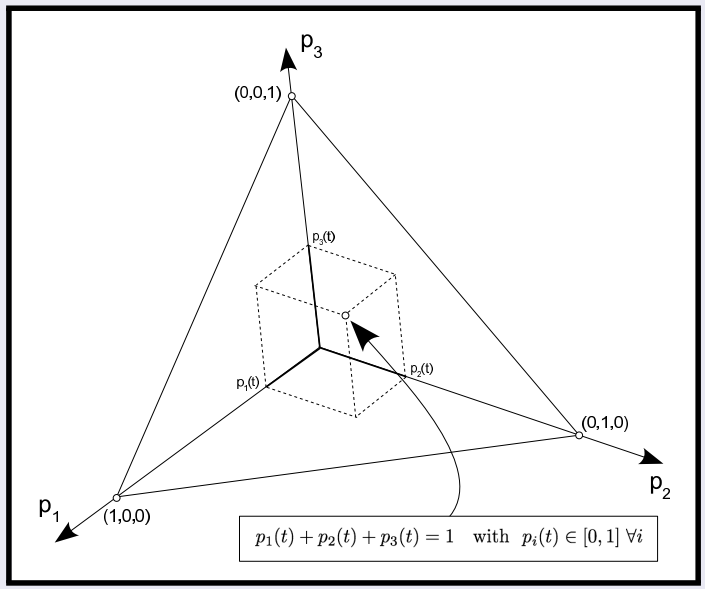
$$p(i) \in [0, 1] \qquad i = 1, 2, \ldots, n. \tag{5}$$

The vector $\mathbf{p}(t)$ defines a certain mixture of phones, the one that, according to the parameters of the phone decoder, best describes the spectral content of the analysis window.

Phone Log-Likelihood Ratio Features

Geometrically, the vector $\mathbf{p}(t)$ can be also interpreted as a point inside an $(N-1)$-dimensional region known as *standard* $(N-1)$-*simplex*

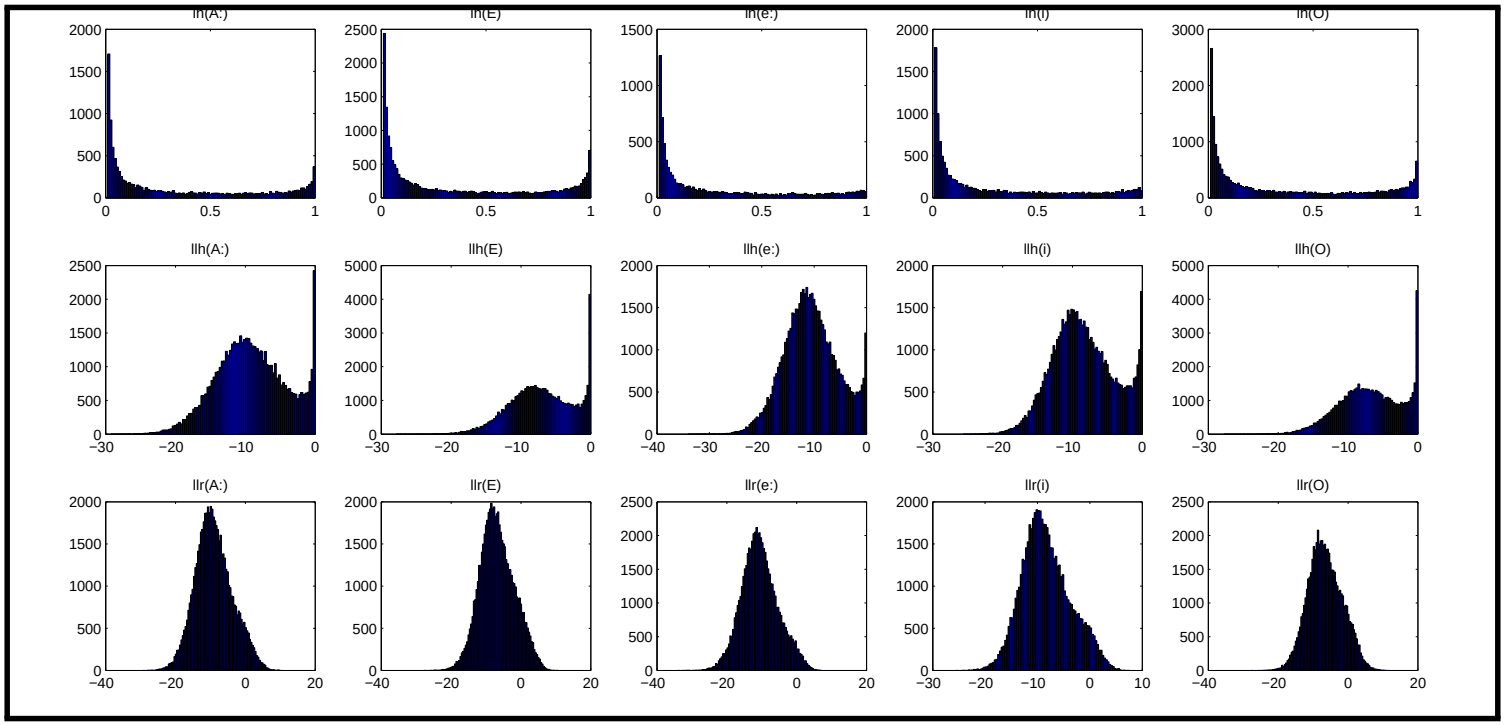
The standard $(N-1)$-simplex $\Delta^{(N-1)}$ is the subset of points in $\mathbb{R}^N$ given by:



Standard 2-simplex defined by phone posteriors in the case of a phone decoder with 3 phonetic units

$$\Delta^{(N-1)} = \{(x_0, \ldots, x_{N-1}) \in \mathbb{R}^N \mid \sum_{i=0}^{N-1} x_i = 1 \ \wedge \ x_i \geq 0 \ \forall i\}$$

Phone Log-Likelihood Ratio Features



Distributions of frame-level likelihoods (first row), log-likelihoods (second row) and log-likelihood ratios (third row) for the Hungarian phones (A:, E, e:, i and O).

Back

- Assuming a classification task with flat priors, the log-likelihood ratios at each frame t can be computed from posterior probabilities as follows:

$$PLLR(i|t) = \log \frac{p(i|t)}{\frac{1}{(N-1)}(1 - p(i|t))} \qquad i = 1, ..., N \qquad (6)$$

- In this way, n log-likelihood ratios are computed at each frame t, carrying the same information as the n phone posteriors, but seemingly featuring Gaussian distributions
- These are the Phone Log-Likelihood Ratio (PLLR) features

GTTS Tecnologías Software

PLLR computation

- **Open-software Temporal Patterns Neural Network** phone decoders, developed by **Brno University of Technology** (BUT)
- Decoders for **Czech**, **Hungarian** and **Russian** which include 42, 58 and 49 phonetic units, respectively, plus 3 non-phonetic units
- Non-phonetic units integrated into a **single non-phonetic unit**: 43 (CZ), 59 (HU) and 50 (RU) PLLR features per frame
- Most results in this presentation using **HU phone decoder**

GTTS Tecnologías Software

System Configuration

- **Voice activity detection** performed by removing the feature vectors whose **highest PLLR value** corresponds to the integrated **non-phonetic unit**
- Gender independent **1024**-mixture GMM **UBM**
- i-vector approach, 500 dimensional i-vectors

GTTS Tecnologías Software

Backend and Fusion

- The backend setup was **separately optimized for each dataset**
  - **NIST 2007, 2009 LRE**: ZT-Norm followed by a discriminative Gaussian backend
  - **NIST 2011 LRE**: no normalization, generative Gaussian backend
  - **KALAKA-2**: no normalization, no backend
- Discriminative multiclass calibration/fusion estimated on the development set and applied to evaluation scores using the **FoCal toolkit**

GTTS Tecnologías Software

PLLR Extraction Configuration

PLLR extraction configuration optimized on the NIST 2007 LRE dataset

Dynamic coefficients

System	$C_{avg} \times 100$	C_{LLR}
PLLR	3.45	0.564
PLLR+Δ	**2.66**	**0.382**
PLLR+Δ+ΔΔ	3.60	0.506

Static + first order dynamic coefficients: PLLR+Δ → 118 features (HU)

Feature level channel compensation: Feature normalization, feature warping, RASTA filtering.

System	$C_{avg} \times 100$	C_{LLR}
PLLR	**2.66**	**0.382**
PLLR +FN	2.95	0.436
PLLR +FW	3.21	0.435
PLLR +RASTA	8.67	1.149

GTTS Tecnologías Software

Contrastive Systems

- **Acoustic feature-based i-vector** system
  - Concatenation of **MFCC** and **SDC** coefficients under a 7-2-3-7 configuration, 56 dimensional feature vector
  - **Same configuration** for the i-vector system as in the PLLR feature-based approach: Voice activity detection, GMM estimation, total variability matrix training, scoring

GTTS Tecnologías Software

Contrastive systems

- **Phone-lattice-SVM system**
  - Three phonotactic systems using **BUT TRAPs/NN phone decoders** for Czech, Hungarian and Russian
  - Phone posteriors converted to **phone lattices** by **HTK**
  - Expected counts of **phone n-grams** computed using the **lattice tool of SRILM**
  - **SVM LIBLINEAR classifier**, vectors consisting of expected frequencies of phone n-grams (up to n=3)

Overall Results on NIST 2007 LRE

System			$C_{avg} \times 100$	C_{LLR}
MFCC-SDC i-vector		(a)	2.85	0.407
HU	**Phonotactic**	**(b)**	**2.08**	**0.310**
	PLLR i-vector	(c)	2.66	0.382
Fusion	(a)+(b)		1.08	0.152
	(a)+(c)		1.40	0.215
	(b)+(c)		1.20	0.166
	(a)+(b)+(c)		**0.82**	**0.124**

▸ Reported results on NIST 2007 LRE

GTTS Tecnologías Software

Overall Results on NIST 2009 LRE

System			$C_{avg} \times 100$	C_{LLR}
MFCC-SDC i-vector		(a)	2.70	0.535
HU	Phonotactic	(b)	2.49	0.502
	PLLR i-vector	**(c)**	**2.42**	**0.505**
Fusion	(a)+(b)		1.67	0.346
	(a)+(c)		1.79	0.392
	(b)+(c)		1.69	0.357
	(a)+(b)+(c)		**1.48**	**0.321**

▸ Reported results on NIST 2009 LRE

GTTS Tecnologías Software

Overall Results on NIST 2011 LRE

System			$C_{avg} \times 100$	C_{LLR}
MFCC-SDC i-vector		(a)	5.96	1.088
HU	Phonotactic	(b)	7.15	1.280
	PLLR i-vector	**(c)**	**5.18**	**0.982**
Fusion	(a)+(b)		4.34	0.823
	(a)+(c)		4.00	0.789
	(b)+(c)		4.39	0.829
	(a)+(b)+(c)		**3.63**	**0.714**

▸ Reported results on NIST 2011 LRE

GTTS Tecnologías Software

Results on Albayzin 2010 LRE

System			Clean		Noisy	
			$C_{avg} \times 100$	C_{LLR}	$C_{avg} \times 100$	C_{LLR}
MFCC-SDC i-vector		(a)	2.12	0.176	3.95	0.325
HU	Phonotactic	(b)	2.35	0.218	7.28	0.621
	PLLR i-vector	**(c)**	**1.41**	**0.127**	**3.17**	**0.308**
Fusion	(a)+(b)		1.10	0.106	2.43	0.211
	(a)+(c)		1.20	0.109	2.65	0.227
	(b)+(c)		1.09	0.092	2.65	0.228
	(a)+(b)+(c)		**0.97**	**0.086**	**1.86**	**0.168**

▸ Reported results on Albayzin 2010 LRE

GTTS Tecnologías Software

GTTS Tecnologías Software

Dimensionality Reduction

- Static + first order dynamic coefficients: PLLR+Δ $\rightarrow$ 86 features (CZ), 118 features (HU), 100 features (RU)
- Computational problem (for some approaches): PLLR representation larger than common acoustic representations
- **Goal:** To reduce the set of phone units in the PLLR representation
- Several supervised and unsupervised techniques tested

Supervised Techniques

For each family, the posterior of each phonetic class was computed by adding the posteriors of the phones included in it, then the log-likelihood ratios were computed

- *Family-R*: The reduced (R) set of phones used by Soufifar et al.[1] to limit the number of n-gram counts (33 phone classes)
- *Family-SL*: this set is defined by merging **Short and Long (SL) phones** (31 phone classes)

[1] M. Soufifar et al. **"Discriminative Classifiers for Phonotactic Language Recognition with i-vectors"**, *Proc. ICASSP*, Japan, 2012

Supervised Techniques

- *Family-MP*: A set of phonemes defined according to phonetic categories following IPA charts, (23 phone classes):
  - Consonants produced with the same **Manner and Place (MP)** of articulation were merged

		Place of articulation					
		Labial	Alveolar	Post-alveolar	Palatal	Velar	Glottal
Manner of articulation	Nasal	m	n		ɳ		
	Stop	p b	t d		c ɟ	k g	
	Affricate		t͡s d͡z	t͡ʃ d͡ʒ	c͡ç ɟ͡ʝ*		
	Fricative	f v	s z	ʃ ʒ			h
	Trill		r				
	Approximant		l		j		

  - **Vowels belonging to the same regions** in the IPA charts were also merged

GTTS Tecnologías Software

Supervised Techniques

- *Family-M*: More generic (14 phone classes)
  - Consonants produced with the same **Manner (M)** of articulation were merged

		Place of articulation					
		Labial	Alveolar	Post-alveolar	Palatal	Velar	Glottal
Manner of articulation	Nasal	m	n		ɲ		
	Stop	p b	t d		c ɟ	k g	
	Affricate		t͡s d͡z	t͡ʃ d͡ʒ	c͡ç ɟ͡ʝ*		
	Fricative	f v	s z	ʃ ʒ			h
	Trill		r				
	Approximant		l		j		

  - **Vowels belonging to the same regions** in the IPA charts were also merged

Unsupervised Techniques

- *Correlation*: An iterative clustering algorithm is used. In each step, the algorithm **merges** the closest pair of phones according to the **correlation among the phone posterior probabilities**
- *Frequency*: The N phones with the **highest posterior probabilities** overall in the training set are **selected** and used as (reduced) phone set
- PCA was also tested. Since PCA is an orthogonal transformation that is assumed to deal with normally distributed data ranging in $(-\infty, \infty)$, it is not a suitable transformation to be applied on the phone posterior probability space. Instead, PCA is directly applied on the normally distributed PLLR space

GTTS Tecnologías Software

Comparison of dimensionality Reduction Techniques

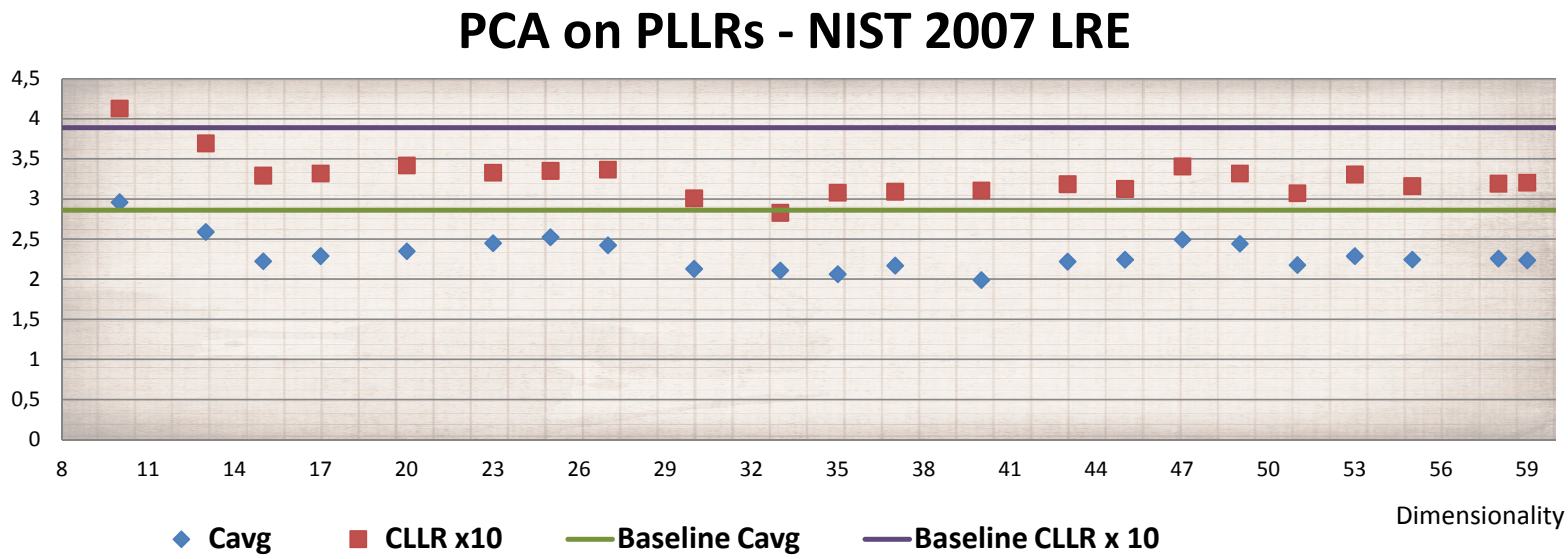
NIST 2007 LRE

HU PLLR System				Dim	$\%C_{avg}$	C_{LLR}
Baseline				59+Δ	**2.86**	**0.389**
Supervised	Merge Phones	Family	R	33+Δ	3.07	0.422
			SL	31+Δ	3.46	0.467
			MP	23+Δ	**2.98**	**0.426**
			M	14+Δ	4.22	0.580
Unsupervised	Merge Phones	Correlation		23+Δ	3.76	0.523
	Select Phones	Frequency		23+Δ	3.56	0.480
	PLLR Projection	PCA		23+Δ	**2.45**	**0.333**

PCA Dimensionality Study

- PCA projection of the features decorrelates the feature space, making the features more suitable for the diagonal covariance GMMs used in the approach
- Search for an **optimal compromise** between the feature vector **size** and the **performance** of the system
- Find the **minimum feature vector size** without high degradation to test the effect of the **Shifted-Delta** transformation over the reduced set of features

PCA Dimensionality Study



Degradation of the systems starts when reducing the features to around 13 dimensions

Shifted Delta PLLRs

SLR performance of an i-vector system based on SD-PLLR features was optimized using different N, P and d values, on the NIST 2007 LRE 30s test set

System	C_{avg}	C_{LLR}
Baseline	2.66	0.382
SD-PLLR 13-2-3-7	1.71	0.260

Performance on the NIST 2011 LRE 30s test set

System	C_{avg}	C_{LLR}	$\%C^{24}_{\text{avg}}$
Baseline	5.18	0.982	12.12
SD-PLLR 13-2-3-7	4.10	0.826	10.48

GTTS Tecnologías Software

GTTS Tecnologías Software

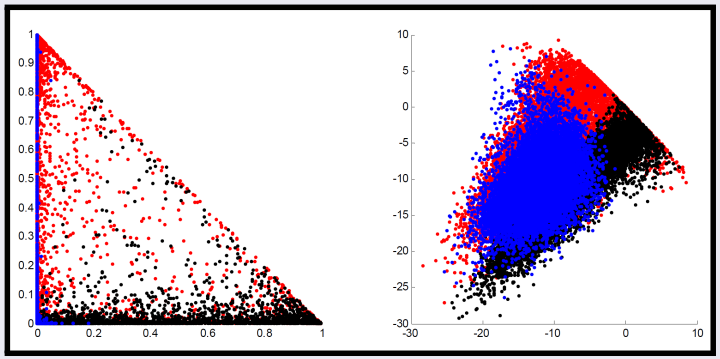
PLLR Feature Space

PLLRs address the non-Gaussian nature of phone posteriors for each (one-dimensional) individual phone model

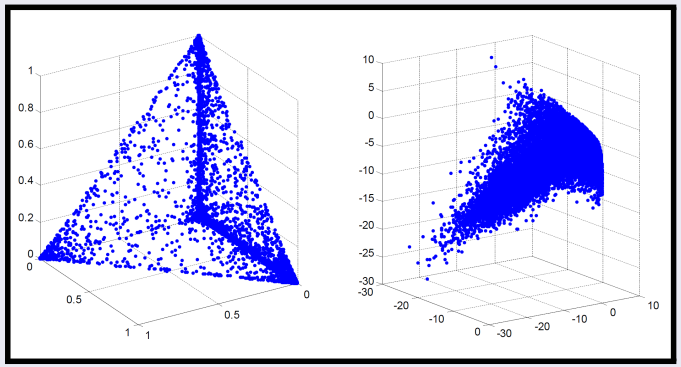
▸ One dimensional distributions

When further analyzing the multidimensional distribution of PLLRs, the seemingly Gaussian distributed features show a strongly bounded distribution

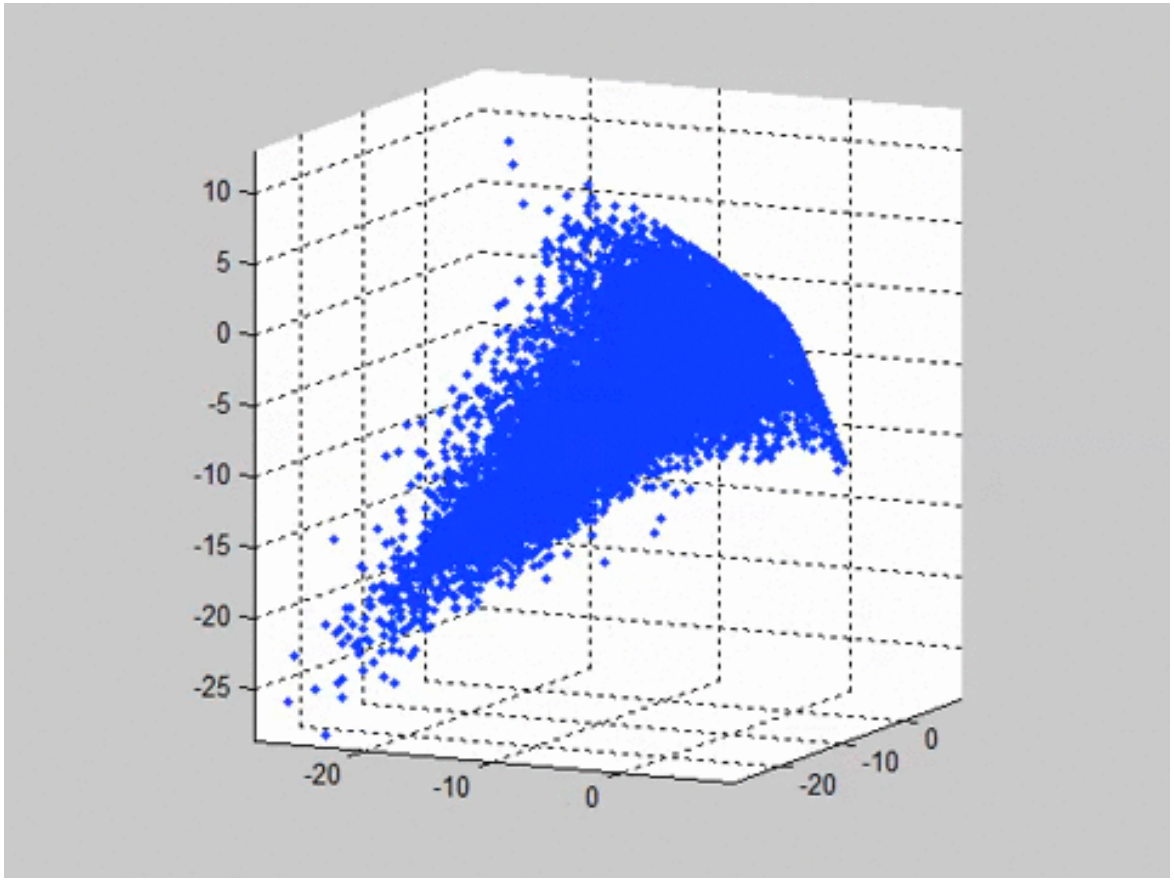
PLLR Feature Space



Distributions of (a) phone posteriors and (b) PLLRs, for three pairs of phones, *a:* vs *E* (red), *i* vs *i:* (black) and *dz* vs *h* (blue).



Distributions of (a) phone posteriors and (b) PLLRs, for the set of phones (*a:*, *E*, *O*).


10
5
0
-5
-10
-15
-20
-25
-20
-10
0
-20
-10
0

PLLR Feature Space

Phone posterior vector $\mathbf{p}(t) = (p(1|t), p(2|t), \ldots, p(N|t))'$

$$PLLR(i|t) = \log \frac{p(i|t)}{(1 - p(i|t))} \qquad i = 1, ..., N \qquad (7)$$

As phone posteriors range in [0,1], PLLRs would seemingly range in $(-\infty, \infty)$

But the constraint among phone posteriors ($\sum_{i=1}^{N} p(i|t) = 1$) is transferred into the PLLR space.

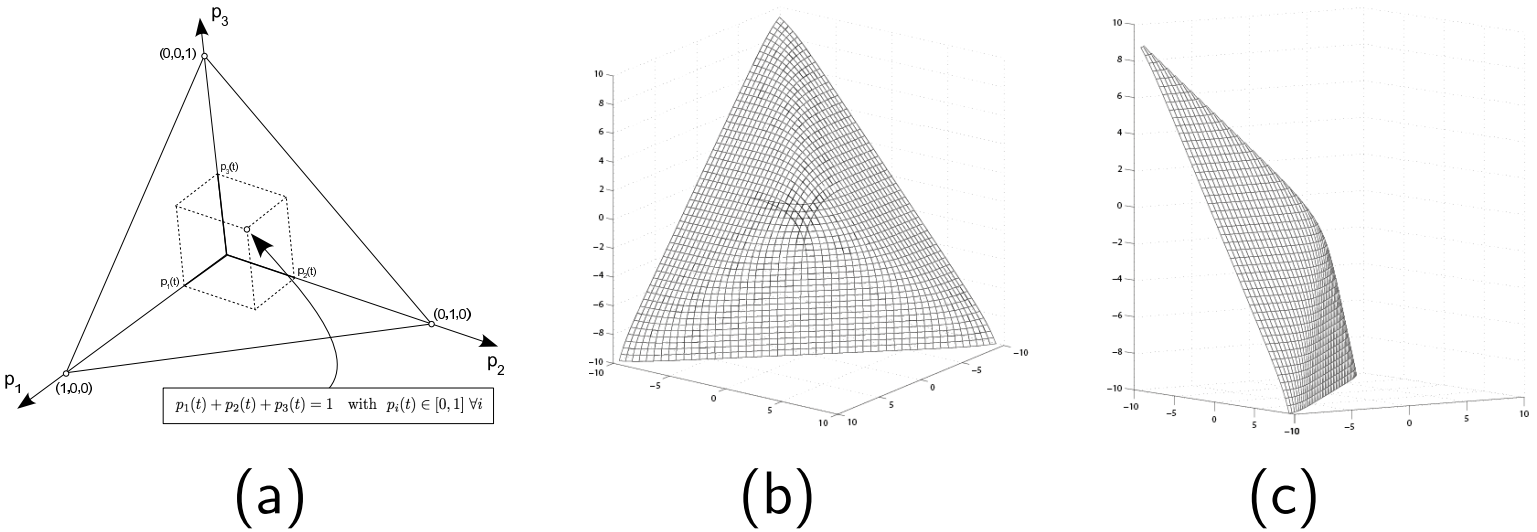
PLLR Feature Space

From the simplex (space where the phone posteriors lay) and the PLLR definition, we derive the hyper-surface $\mathcal{S}$ where the PLLRs lay:

$$\mathcal{S}^{(N-1)} = \left\{ \mathbf{r} \in \mathbb{R}^N \;\middle|\; \mathcal{G}(\mathbf{r}) = \sum_{i=1}^{N} \frac{1}{1+e^{-r_i}} - 1 = 0 \right\} \qquad (8)$$

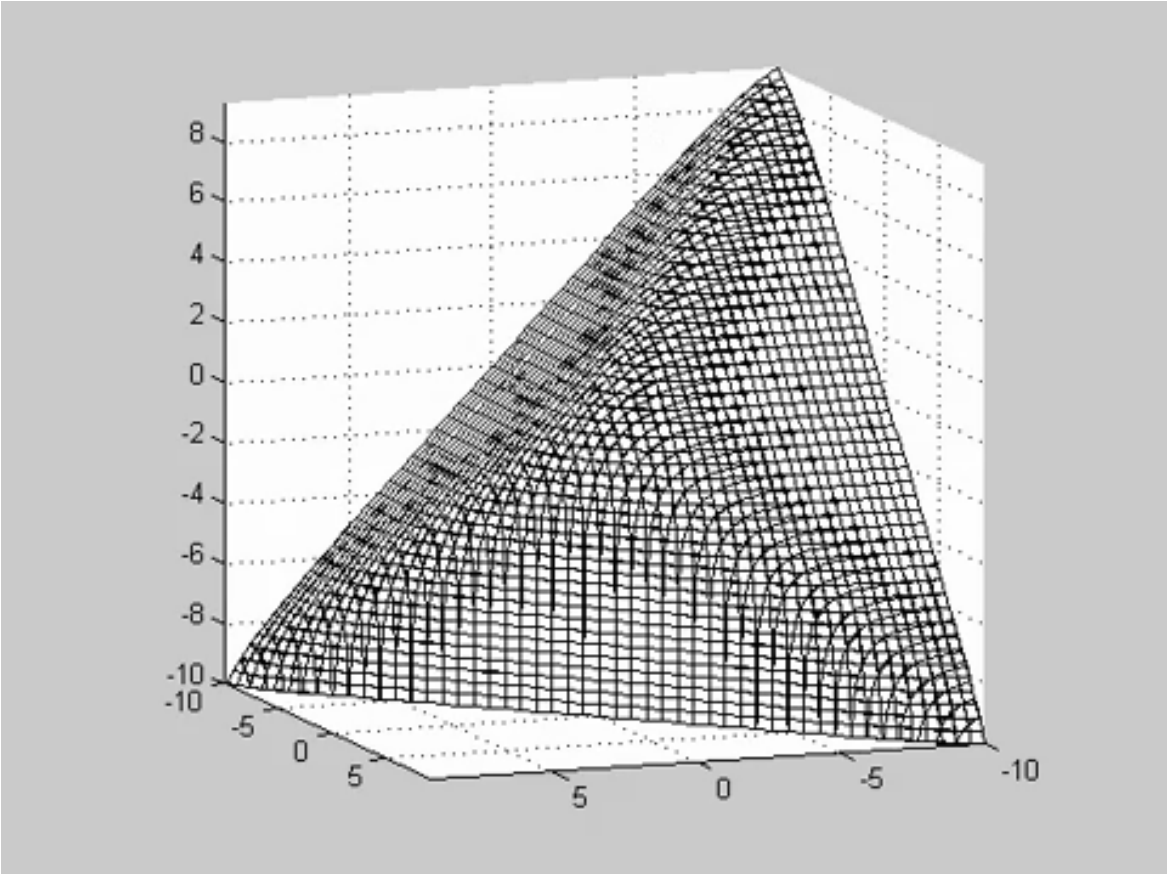
where $\mathcal{G}(\mathbf{r})$ is the implicit hyper-surface function in the PLLR space.

PLLR Feature Space



(a) (b) (c)

(a) Standard 2-simplex defined by phone posteriors in the case of a phone decoder with 3 phonetic units. Graphs (b) and (c) show the hyper-surface where PLLRs lie for the case of a phone decoder with 3 phonetic units.


8
6
4
2
0
-2
-4
-6
-8
-10
-10
-5
0
5
5
0
-5
-10

PLLR Projection

The hyper-surface $\mathcal{S}$ is asymptotically perpendicular to the basis of PLLRs, which explains the bounded distributions shown.

▸ Demonstration

To avoid this bounding effect, we propose to project PLLRs into the hyper-plane tangential to the surface at the point where all the posteriors take the same value $p_i = \frac{1}{N}$, that is, the top of the convex surface, where the normal vector is:

$$\mathbf{n}|_{r_i=-\log(N-1)} = \frac{(N-1)}{N\sqrt{N}} \cdot \hat{\mathbf{1}} \tag{9}$$

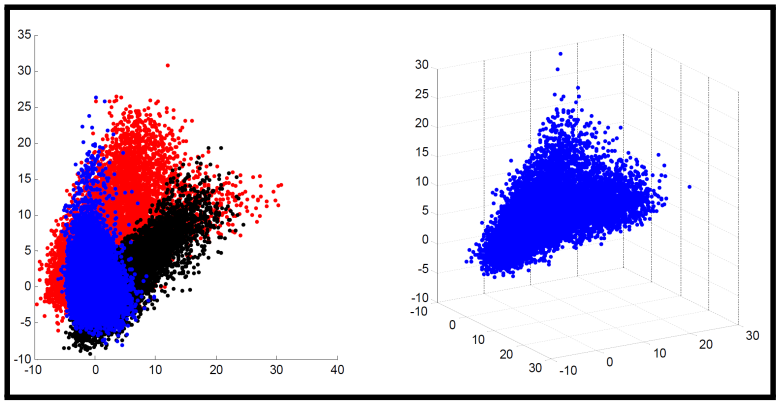
where $\hat{\mathbf{1}} = \frac{1}{\sqrt{N}}[1_1, 1_2, ..., 1_N]$.

GTTS Tecnologías Software

PLLR Projection

In the general case (N dimensions), the kernel (null space) of the desired projection is $\hat{\mathbf{1}}$, then the projection matrix P is given by:

$$P = \mathbb{I} - \hat{\mathbf{1}}' * \hat{\mathbf{1}} \tag{10}$$



Distribution of the PLLRs shown in previous figures after projecting them into the defined hyper-plane.

Finally, in order to decorrelate the parameters, we apply PCA on the transformed PLLRs

Results

PLLR System	$\%C_{\text{avg}}$ (r.i.)	C_{LLR} (r.i.)
Baseline 59	2.86	0.389
Projection 59	2.31 (19%)	0.320 (18%)
Projection+PCA 58	**2.10 (27%)**	**0.310 (20%)**
PCA 59	2.24 (22%)	0.321 (17%)
PCA 58	2.26 (21%)	0.319 (18%)

$\%C_{\text{avg}}$ and C_{LLR} performance (and relative improvements) for the PLLR i-vector baseline system, and systems using PLLR features on the NIST 2007 LRE primary evaluation task.

GTTS Tecnologías Software

Results

Dataset	System		$\%C_{\text{avg}}$	C_{LLR}	$\%C^{24}_{\text{avg}}$
2009 LRE	PLLR	(a)	2.42	0.505	-
	PLLR+Projection+PCA	(b)	**2.19**	**0.443**	-
	Acoustic MFCC	(c)	2.70	0.535	-
	Phonotactic	(d)	2.49	0.502	-
	(c)+(d)		1.67	0.346	-
	(a)+(c)+(d)		1.48	0.321	-
	(b)+(c)+(d)		**1.42**	**0.307**	-
2011 LRE	PLLR	(a)	5.18	0.981	12.12
	PLLR+Projection+PCA	(b)	**4.30**	**0.824**	**11.33**
	Acoustic MFCC	(c)	5.95	1.088	13.56
	Phonotactic	(d)	7.15	1.280	14.28
	(c)+(d)		4.34	0.823	10.43
	(a)+(c)+(d)		3.63	0.714	9.14
	(b)+(c)+(d)		**3.33**	**0.667**	**8.91**

GTTS Tecnologías Software

PLLRs on RATS

- Data of the RATS program 5 target languages and 10 non-target languages
- Levantine Arabic and Czech phone recognizers, based on a hybrid NN/HMM approach
- Projected PLLRs+*Delta* used to compute 600 dimensional i-vectors
- Two PLLR approaches based on different classifiers
  - Multi-class regularized Logistic Regression
  - Three-layer NN classifiers with 300 neurons in the hidden layer and 6 outputs
- Contrastive Perceptual Linear Prediction and Subspace n-gram Model feature-based systems

GTTS Tecnologías Software

PLLRs on RATS

Results with LR - C_{avg} [%]

System	120s	30s	10s	3s
PLP2	7.72	11.69	16.39	23.04
SnGM-LE	5.86	12.28	18.53	26.45
SnGM-CZ	8.59	15.76	20.89	27.95
PLLR-LE	**4.56**	**7.98**	**12.61**	21.48
PLLR-CZ	6.95	10.76	15.13	**21.32**

GTTS Tecnologías Software

PLLRs on RATS

Results with NN - C_{avg} [%]

System		120s	30s	10s	3s
1	PLP2	7.21	9.21	12.43	18.58
2	SnGM-LE	5.53	9.34	15.61	22.76
3	SnGM-CZ	7.23	10.46	15.38	24.05
4	PLLR-LE	**5.37**	**7.31**	**11.46**	**17.63**
5	PLLR-CZ	5.81	8.83	12.30	19.52

Fusions	120s	30s	10s	3s
4+5	5.19	6.79	10.14	16.04
1+4	5.80	6.43	8.69	15.37
2+4	**5.12**	6.61	10.48	16.93
3+5	5.74	8.33	11.28	18.11
1+2+4	5.38	6.31	**8.53**	14.90
1+4+5	5.29	6.43	8.71	14.65
1+2+3+4+5	5.59	**6.21**	8.75	**14.37**

GTTS Tecnologías Software

Shifted Delta Projected PLLRs

Results on the NIST 2007 LRE dataset

PLLR System	$\%C_{\text{avg}}$	C_{LLR}
Projection + PCA 58	2.10	0.310
Projection + PCA 13	2.43	0.330
Projection + PCA 13 + SD	1.52	0.225

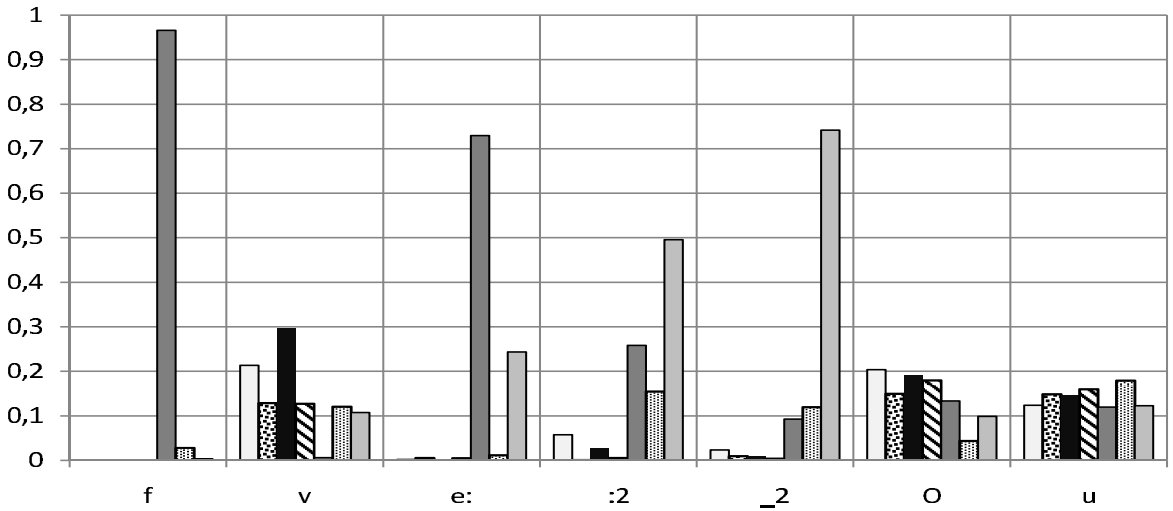
Results on the NIST 2011 LRE dataset

PLLR System	$\%C_{\text{avg}}$	C_{LLR}	$\%C^{24}_{\text{avg}}$
Projection + PCA 58	4.30	0.824	11.33
Projection + PCA 13 + SD	4.84	0.916	12.49
Projection + PCA 15 + SD	4.39	0.833	11.26
Projection + PCA 17 + SD	3.80	0.756	9.52

GTTS Tecnologías Software

GTTS Tecnologías Software

PLLRs on Speaker Recognition



Normalized average posteriors $\hat{p}(i|s)$ of seven Hungarian phones on the same utterance of the TIMIT dataset for 7 different speakers. The subset of phones represents fricative labiodental consonants (f and v) and a subset of vowels (e:, :2, _2, O, u), as defined in the International Phonetic Alphabet.

GTTS Tecnologías Software

PLLRs on Speaker Recognition

- **PLLR+Δ** configuration found also the best parameterization
- **i-vector - Gaussian PLDA Approach**
- Gender dependent **1024**-mixture **UBMs**
- Gender dependent Total Variability matrices
- **i-vector** dimensionality set to **500**

Acoustic Feature-Based System

- **13 MFCC coefficients**, including the zero (energy) coefficient
- **Cepstral Mean Subtraction** and **Feature Warping**
- Feature vector augmented with **first and second order deltas**: 39-dimensional feature vector

Datasets

NIST SRE Datasets

- Annual evaluations from 1996 to 2006, biannual ever since. Conversational speech data from the Mixer Corpora. Gender dependent, text independent datasets
- Training resources cover several conditions: telephone conversation excerpts, two-channel conversations, interview segments, summed-channel conversations
- Test conditions also covered several conditions

Evaluation Measures

Detection Cost Function (DCF)

$$DCF = \frac{C_{\text{miss}} P_{\text{T}} P_{\text{miss(i)}} + \sum C_{\text{fa}} (1 - P_{\text{T}}) P_{\text{fa}} (L_T, L_N)}{min\{C_{miss} P_T, C_{fa}(1 - P_T)\}} \quad (11)$$

EER

This measure reports system performance at the operation point for which the false alarm error rate (P_{fa}) is equal to the miss error rate (P_{miss})

Results on the NIST 2010 SRE

	Condition	System	EER	MinDCF	ActDCF
(1)	Interview same microphone in training and test	MFCC	1.86	0.417	0.439
		PLLR+Δ	4.05	0.653	0.854
		Fusion	**1.40**	**0.363**	**0.367**
(2)	Interview different microphone in training and test	MFCC	2.99	0.562	0.633
		PLLR+Δ	6.39	0.804	0.819
		Fusion	**2.36**	**0.492**	**0.553**
(3)	Interview training telephone test	MFCC	3.63	0.625	**0.848**
		PLLR+Δ	9.20	0.862	0.978
		Fusion	**3.25**	**0.522**	0.874
(4)	Interview training telephone test rec. over microphone	MFCC	1.71	0.443	0.475
		PLLR+Δ	5.52	0.690	0.703
		Fusion	**1.69**	**0.372**	**0.406**
(5)	Telephone in training and test	MFCC	4.64	0.600	0.712
		PLLR+Δ	8.41	0.848	0.869
		Fusion	**4.29**	**0.560**	**0.688**

GTTS Tecnologías Software

Results on the NIST 2012 SRE

Condition		System	EER	MinDCF	ActDCF
(2)	Telephone with No Added Noise	MFCC	1.77	0.272	0.290
		PLLR+Δ	3.12	0.419	0.440
		Fusion	**1.39**	**0.215**	**0.246**
(5)	Telephone Recorded in Noise	MFCC	1.93	0.260	0.294
		PLLR+Δ	3.72	0.449	0.481
		Fusion	**1.64**	**0.219**	**0.283**

GTTS Tecnologías Software

GTTS Tecnologías Software

Summary

- **PLLR** feature-based systems improve Spoken Language Recognition performance under the i-vector approach, even outperforming acoustic and/or phonotactic approaches
- The system **contributes in fusions**, providing complementary information to both acoustic and phonotactic systems
- **PLLR systems** built on different decoders can be **fused** to get significant **gains in performance**
- Good performance of the system under **noisy conditions**
- **PLLR feature vector dimensionality** can be **significantly reduced** attaining almost **the same performance**

GTTS Tecnologías Software

Summary

- Applying **PCA** in the PLLR feature space not only **reduces the computational cost**, but also **improves system performance**
- **PLLRs** can be **projected to get unbounded distributions** which **enhances the information retrieved** by the system
- On Speaker Recognition, **PLLR** feature-based system provides **significant improvements in fusions**
- PLLRs provide a suitable way of conveying complementary acoustic-phonetic information for language and speaker recognition

GTTS Tecnologías Software

Current & Future Work

- Test the approach under different modeling approaches
- Further analyze performance of PLLRs on short signals and noisy environments
- Use other phone decoders or combinations of their outputs to compute PLLRs
- Study and optimize PLLRs for speaker recognition
- PLLRs for text-dependent speaker recognition

GTTS Tecnologías Software

Thank you!

Reported results on the primary task of the NIST 2007 LRE

Approach	Model	EER	$C_{avg} \times 100$
Acoustic	GMM-MMI [Torres et.al. 08]	–	2.10
	GSV-SVM [Torres et.al. 08]	–	1.92
	Discriminative GMM-MAP [Brummer et.al. 09]	–	1.74
Phonotactic	HU, Phone-SVM, lattices [Tong et.al. 10]	1.84	–
	HU, Phone-SVM, lattices [Richardson et.al. 08]	2.40	–
	EN, Phone-SVM, lattices [Richardson et.al. 08]	1.80	–
PLLR	**HU, i-vector, generative**	2.69	2.66
Fusions	2 acoustic subsystems [Torres et.al. 08]	–	1.55
	2 phonotactic subsystems [Torres et.al. 08]	–	1.55
	3 phonotactic subsystems [BenZeghiba et.al. 12]	–	0.90
	4 subsystems [Torres et.al. 08]	0.93	0.97
	acoustic+phonotactic+PLLR	0.80	0.82

Back

Reported results on the primary task of the NIST 2009 LRE

Approach	Model	EER	$C_{avg} \times 100$
Acoustic	GMM-MMI [DehakN et. al. 11]	2.30	–
	SVM-GSV [Torres et. al. 09]	–	2.30
	JFA [Jancik et. al. 10]	–	2.02
	i-vector (LDA+WCCN) [DehakN et. al. 11]	2.40	–
	i-vector, generative [Martinez et. al. 12]	–	3.09
	i-vector (Logistic reg.) [Plchot et. al. 12]	–	2.35
Phonotactic	EN, Phone-SVM [Torres et. al. 09]	–	2.34
	HU, Phone-SVM [Mikolov et. al. 10]	–	3.85
	RU, Phone-SVM [Mikolov et. al. 10]	–	3.03
PLLR	**HU, i-vector, generative**	2.43	2.42
Fusions	2 acoustic subsystems [Torres et. al. 09]	–	2.00
	2 acoustic subsystems [Plchot et. al. 12]	–	1.78
	3 phonotactic subsystems [Mikolov10]	–	2.39
	3 phonotactic subsystems [BenZeghiba et. al. 12]	–	1.99
	3 subsystems [Torres et. al. 09]	–	1.64
	36 subsystems [Castaldo et. al. 10]	–	1.16
	acoustic+phonotactic+PLLR	1.47	1.48

Back

Reported results on the primary task of the NIST 2011 LRE

Approach	Model	$C_{avg} \times 100$	$C^{24}_{avg} \times 100$ min	$C^{24}_{avg} \times 100$ actual
Acoustic	i-vector (LDA) [Singer et. al. 12]	4.15	–	8.90
	i-vector (HLDA) [Brummer et. al. 12]	–	–	10.35
	GMM-SVM [HuaiYou et. al. 12]	–	10.41	–
Phonotactic	RU, PCA [Brummer et. al. 12]	–	–	14.32
	HU, *n*-gram i-vector [Brummer et. al. 12]	–	–	15.42
PLLR	**HU, i-vector, generative**	5.18	9.83	12.12
Fusions	5 subsystems [Singer et. al. 12]	3.30	–	7.00
	8 subsystems [Rodriguez et. al. 12]	3.35	5.09	7.64
	3 subsystems [Brummer et. al. 12]	–	–	8.47
	3 subsystems [HuaiYou et. al. 12]	–	9.02	–
	acoustic+phonotactic+PLLR	3.63	6.68	9.14

Back

GTTS Tecnologías Software

Reported results on the primary task: Albayzin 2010 LRE

Condition	Approach	Model	$C_{avg} \times 100$
Clean	Acoustic	JFA [Martinez et.al. 11]	1.86
		GMM-MMI [Martinez et.al. 11]	4.33
	Phonotactic	HU, Phone-ML [Martinez et.al. 11]	4.41
		EN, Phone-ML [Rodriguez et.al. 11]	3.26
	PLLR	**HU, i-vector, generative**	1.41
	Fusions	8 subsystems [Martinez et.al. 11]	1.84
		5 subsystems [Rodriguez et.al. 12]	1.77
		2 subsystems [Abad et.al. 10]	1.81
		acoustic+phonotactic+PLLR	0.97
Noisy	Fusions	5 subsystems [Rodriguez et.al. 12]	3.90
		2 subsystems [Abad et.al. 10]	2.53
		acoustic+phonotactic+PLLR	1.86

Back

GTTS Tecnologías Software

Results Using Multiple Decoders

NIST 2011 LRE

PLLR System		$\%C_{\text{avg}}$	C_{LLR}	$\%C_{avg}^{24}$
Baseline	CZ (43+Δ)	5.31	0.978	12.46
	HU (59+Δ)	5.18	0.982	12.12
	RU (50+Δ)	4.70	0.898	11.27
	CZ+HU+RU	**3.79**	**0.720**	**9.10**
Family-MP	CZ (25+Δ)	5.53	1.054	13.62
	HU (23+Δ)	5.40	1.015	12.64
	RU (21+Δ)	5.13	0.961	11.57
	CZ+HU+RU	**3.82**	**0.693**	**9.79**
PCA	CZ (25+Δ)	4.46	0.855	11.20
	HU (23+Δ)	4.48	0.877	10.88
	RU (21+Δ)	4.20	0.803	11.01
	CZ+HU+RU	**3.21**	**0.634**	**8.45**

GTTS Tecnologías Software

PLLR feature space

The hyper-surface $\mathcal{S}$ is asymptotically perpendicular to the basis of PLLRs, which explains the bounded distributions shown.
The normal vector to the hyper-surface $\mathcal{S}$ is:

$$\mathbf{n} = \nabla \mathcal{G}(\mathbf{r}) \tag{12}$$

where each component n_i of $\mathbf{n}$ is given by:

$$n_i = \frac{e^{r_i}}{(1 + e^{r_i})^2} \tag{13}$$

PLLR feature space

Let us consider the case in which a subset of phones $\mathcal{I} \subset \{1, 2, ..., N\}$ accounts for most of the probability mass, that is, $\sum_{\forall i \in \mathcal{I}} p_i = 1 - \epsilon$. As these phones tend to take all the probability mass, it follows that $\sum_{\forall i \in \mathcal{I}} p_i \to 1$ and $\epsilon = \sum_{\forall i \notin \mathcal{I}} p_i \to 0$ (therefore, $r_i \to -\infty \ \forall i \notin \mathcal{I}$). Accordingly, for the normal vector $\mathbf{n}$ it holds:

$$\lim_{\epsilon \to 0} n_i = 0 \qquad \forall i \notin \mathcal{I} \tag{14}$$

That is, the normal vector tends to lie in the subspace $\mathcal{Q}$ where the set of phones $\mathcal{I}$ are confined. Hence, the surface is asymptotically perpendicular to any basis defined on $\mathcal{Q}$.

Back