This is the revised version (after referee comments) of an article submitted to IEEE TRANSACTIONS ON AUDIO, SPEECH AND LANGUAGE PROCESSING on July 19, 2012. On December 2012, we received the reviewers' comments and, after including most of the suggested changes, we produced this revised version, which was re-submitted to IEEE-TASLP on January 2013. After 18 months (by July 2014), the Associate Editor made a final decision and the paper was rejected. We decided to make it available on Internet because it can be considered a seminal reference to our PLLR features. If you are so kind to cite this paper, please use the URL http://gtts.ehu.es/gtts/NT/fulltext/DiezPLLR2013.pdf.

Using Phone Log-Likelihood Ratios as Features in an iVector Approach to Language Recognition

Mireia Diez, *Student Member, IEEE,* Amparo Varona, *Member, IEEE,* Mikel Penagarikano, *Member, IEEE,* Luis Javier Rodriguez-Fuentes, *Member, IEEE,* and German Bordel, *Member, IEEE*

***Abstract*—In acoustic approaches to Spoken Language Recognition (SLR), short-term spectral features such as MFCC-SDC are commonly used under different modeling techniques, including the highly competitive total variability (iVector) approach. On the other hand, phonotactic SLR systems are built on segmental-level phone n-gram statistics computed on phone lattices/sequences produced by phone decoders, some of which use frame-level phone posterior probabilities as intermediate features. In this paper, we propose the use of log-likelihood ratios of frame-level phone posterior probabilities, the so called Phone Log-Likelihood Ratios (PLLR), as features under an iVector approach. A PLLR-based iVector system is fully described, evaluated and compared to: (1) an acoustic iVector system (trained on MFCC-SDC features); and (2) a phonotactic (Phone-lattice-SVM) system. To evaluate the goodness of the proposal, experiments on four different benchmarks have been carried out: the NIST 2007, 2009 and 2011 LRE databases (which include narrow band, mostly single-speaker telephone conversational speech) and the Albayzin 2010 LRE database (which includes wide-band, multi-speaker broadcast speech). In these experiments, the iVector system trained on PLLR features outperformed the iVector system trained on MFCC-SDC. The fusion of the PLLR-based iVector system —either with the acoustic iVector system or with the phonotactic system— provided significant improvements in all benchmarks, which may indicate that PLLR features are complementary to both acoustic and phonotactic features. Finally, the fusion of the three approaches yielded the best or among the best performances reported so far on the considered benchmarks[1].**

***Index Terms*—Spoken Language Recognition, Phone Posterior Probabilities, Log-Likelihood Ratios, iVectors**

I. Introduction

SPOKEN Language Recognition (SLR) refers to the task of recognizing by computational means the language spoken in an utterance. Two main types of systems are commonly applied in SLR tasks [1]: low-level acoustic systems and high-level phonotactic systems. Due to the complementarity of these approaches, best results are usually obtained by discriminatively fusing several acoustic and phonotactic systems [2] [3].

This work has been supported by the University of the Basque Country, under grant GIU10/18 and project US11/06, by the Government of the Basque Country, under program SAIOTEK (project S-PE12UN055), and the Spanish MICINN, under *Plan Nacional de I+D+i* (project TIN2009-07446, partially finance by FEDER funds). Mireia Diez is supported by a 4-year research fellowship from the Department of Education, University and Research of the Basque Country.

M. Diez, A. Varona*, M. Penagarikano, L.J. Rodriguez-Fuentes and G. Bordel are with the Department of Electricity and Electronics, University of the Basque Country, UPV/EHU, 48940, Leioa, Spain. e-mail: mireia.diez@ehu.es, amparo.varona@ehu.es (corresponding author, Phone +34 946015540 Fax +34 946013071), mikel.penagarikano@ehu.es, luisjavier.rodriguez@ehu.es, german.bordel@ehu.es.

[1]EDICS: Multilingual Recognition and Identification

In low-level acoustic systems, the target language is modeled with information taken from the spectral characteristics of the audio signal. The approach known as Gaussian Mixture Model / Universal Background Model (GMM-UBM) was successfully applied to language recognition by using the concatenation of Mel-Frequency Cepstral Coefficient (MFCC) and Shifted Delta Cepstrum (SDC) features [4]. In the following years, besides introducing GMM supervectors for Support Vector Machine (SVM) classifier [5], acoustic systems improved due to discriminative GMM training [6] and acoustic adaptation (CMLLR) [7]. Also, Joint Factor Analysis (JFA) [8], previously applied to speaker recognition, was successfully applied to spoken language recognition [9].

Recently, an approach derived from JFA, known as *Total Variability Factor Analysis*, for which low-dimensional features known as *iVectors* are extracted and then processed under different modeling approaches, has become state-of-the-art in the speaker and language recognition fields due to its excellent performance, low complexity and low dimensionality [10] [11] [12].

New approaches combining the iVector technology with other complementary features are emerging, as in [13], where prosodic features (pitch, energy and duration) were introduced, or in [14], where speaker vectors from subspace GMM were used as features. It has been reported that these systems alone do not yield outstanding results, but performance improves significantl when fusing them with a system based on spectral features.

Among high-level approaches, nowadays the most common is the so called *Phone-lattice-SVM* approach, which combines phonotactic features (expected counts of phone n-grams computed on phone lattices provided by phone decoders) with SVM classifier [15]. Variations of this approach have been recently proposed leading to improved performance [16] [17]. Also, remarkable efforts are being devoted to deal with high-dimensionality representations, by applying either feature selection [18] or dimensionality reduction [19] techniques. Recently, high-level phonotactic features (n-gram counts) have been applied to SLR tasks under an iVector approach [20].

In this paper, we explore the use of frame-level log-likelihood ratios of phone posterior probabilities produced by a phone decoder —hereafter called *Phone Log-Likelihood Ratios* (PLLR)— as features under an iVector approach. We provide some insight into PLLR features and discuss why using them could be useful for characterizing speech sounds in different languages. We also provide quantitative evidence of the effectiveness of an iVector system trained on PLLR features.

The PLLR-based iVector system is compared to and fused with two acoustic and phonotactic baseline systems, revealing that PLLR features provide complementary information, since performance increases significantl in all cases after fusing the PLLR-based system.

This complementarity can be explained in terms of the time span of the events implicitly described by each feature set. The acoustic baseline (an iVector system) is based on MFCC-SDC features, which convey static and dynamic short-term spectral information, in contrast to PLLR features, which convey long-term spectral (i.e. phonetic-level) information. On the other hand, the phonotactic baseline (a Phone-Lattice-SVM system) takes expected counts of phone n-grams (conveying segmental-level information) as features. These features are typically computed from frame-level phone posteriors, but a sizeable amount of low-level phonetic information is lost when phone posteriors are collapsed into few likely sequences of phones (the phone lattices), on which phone n-gram counts are computed.

The competitiveness and complementarity of the PLLR-based iVector system has been experimentally tested on four challenging benchmarks: the datasets used for the National Institute of Standards and Technology (NIST) Language Recognition Evaluations (LRE) in 2007, 2009 and 2011 (which include narrow-band, mostly single-speaker conversational telephone speech) and the Albayzin 2010 LRE dataset (which includes wide-band multi-speaker broadcast speech).

The rest of the paper is organized as follows. Section II reviews previous work that motivated the proposed approach. In Section III, an interpretation of phone posteriors as a reference system to characterize speech sounds is briefl discussed and the PLLR features are formally defined Section IV describes the PLLR-based iVector system along with the two baseline systems and the backend and fusion models applied to system scores. Section V describes the datasets used in the experiments and Section VI formally define the evaluation measures applied in this work. Section VII presents a series of experiments carried out to fin the best configuratio of the PLLR system, using NIST 2007 LRE as development set. Section VIII presents the results obtained on the three remaining benchmarks, compares the performance of the proposed approach to baseline systems, and for a more complete perspective, provides results reported by other authors on the same datasets. The contribution and potential usefulness of PLLR features in the fiel of spoken language recognition is discussed in Section IX. Finally, conclusions are given in Section X.

II. Motivation

It is widely accepted that the spectral and phonotactic features on which most SLR systems rely provide complementary information. However, it is not common practice to combine them into a single feature set, mainly because spectral features are computed on a frame-by-frame basis, whereas phonotactic features provide segmental-level information, and thus there is no clear way to mix them. Most authors build separate acoustic and phonotactic systems and fuse them at the score level to get best SLR performance [3] [21] [22] [23].

Previous works aiming to combine acoustic and phonetic/phonotactic information in SLR systems have focused on searching acoustic differences conditioned on the phone sequence, based on the observation that certain phones are realized in different ways across languages/dialects.

In [7], acoustic likelihoods computed on dialect-adapted phone lattices were used to score input utterances in dialect recognition tasks. During training, dialect-specifi phonetic models were created by unsupervised MAP adaptation of dialect-neutral models. During testing, for each input utterance, a single phone lattice was computed using the phone decoder with dialect-neutral models. This lattice was then rescored using dialect-adapted phonetic models to generate dialect-specifi lattices. Finally, these latter were used to compute the acoustic likelihood of the input utterance for each dialect. The same idea was further developed in [24], by selecting those biphones that best discriminated between dialects and training dialect-specifi models for them. This approach fused well with classic phonotactic systems, probably because it reflecte acoustic differences between dialects not present in phone sequences/lattices.

A more recent dialect recognition approach scored the target dialects based on differences between dialect-specifi acoustic models corresponding to the same phone type [25]. First, a phone recognizer was applied to get the most likely phone sequence and segmentation. Then, a phone-GMM supervector was estimated for each phone type, based on the feature vectors aligned to the instances of that phone appearing in the input utterance. The phonetic characteristics of each input utterance were thus summarized in a single vector of phone-type supervectors. Binary SVM classifier were then trained for each pair of dialects, taking phone-GMM supervectors as input (thus accounting only for acoustic differences in the realizations of phones), and applied to estimate posterior probabilities (used as scores) for each input utterance.

In the approaches reported above, dialect-specifi sets of acoustic-phonetic models are estimated and a scoring function is provided that allows to decide what set of models (i.e. what dialect) better fit an input utterance. In both cases, however, systems rely on spectral features and can be considered, in this regard, as acoustic approaches.

Partly inspired by [25], our firs thought was to apply frame-level phone posteriors as weights in the estimation of phone-type specifi models, using MFCC-SDC as features. After realizing that frame-level phone posteriors *already* conveyed acoustic-phonetic information, we concluded that they may be characterizing speech sounds spanning the duration of a phone and centered at the given frame. This led us to the idea of using phone posteriors alone as features. Since phone posteriors are assumed to characterize speech at the phonetic level, which is between short-term spectral and phonotactic levels, it was very likely that the proposed approach contributed complementary information to existing approaches. Finally, from a practical point of view, frame-level phone posteriors would just replace spectral features in acoustic SLR systems, keeping the remaining components exactly the same, which makes the proposed approach very easy to implement.

III. Using Phone Posteriors as Features

In this study, a specifi type of phone decoders is used, consisting of a set of phonetic models (each featuring a linear structure of states) and a neural network that provides estimates of phone state posteriors. The neural network takes as input a relatively long (around 300 ms) window, consisting of a sequence of acoustic feature vectors, and outputs the phone state posteriors. Assuming a frame step of 10 ms, two consecutive analysis windows will be mostly overlapped and the sequence of posteriors will thus evolve more smoothly than spectral features (for which analysis windows are much shorter, typically 20-30 ms long). Besides, since the neural network is trained on a large and diverse dataset, the posteriors are expected to be robust to speaker, channel and other sources of variability.

Let us consider one of such decoders, with n phonetic units, each of them represented by a left-right model of S states. At each frame t, the posterior probability of each state s ($1 \leq s \leq S$) of each phone model i ($1 \leq i \leq n$), $p_{i,s}(t)$, is output, so that the posterior probability of a phonetic unit i can be computed by adding the posteriors of its states:

$$p_i(t) = \sum_{\forall s} p_{i,s}(t) \tag{1}$$

In this way, the decoder outputs an n-dimensional vector of phone posteriors at each frame t: $\mathbf{p}(t) = (p_1(t), p_2(t), \ldots, p_n(t))'$ (the symbol $'$ denoting the transpose), such that $\sum_{i=1}^{n} p_i(t) = 1$ and $p_i \in [0,1]$ for $i = 1, 2, \ldots, n$. The vector $\mathbf{p}(t)$ define a certain mixture of phones, the one that, according to the neural network parameters, best describes the spectral content of the analysis window. Geometrically, the vector of posteriors 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:

$$\Delta^{(n-1)} = \{(x_0, \ldots, x_{n-1}) \in \mathbb{R}^n \mid \sum_{i=0}^{n-1} x_i = 1 \ \wedge \ x_i \geq 0 \ \forall i\} \tag{2}$$

Figure 1 shows the case of a phone decoder with 3 phonetic units: each analysis window is assigned a point inside the standard 2-simplex (a triangle). Note that vertices represent *pure* phonetic units, whereas edges represent different mixtures of the two phonetic units connected by them.

Under this geometric interpretation, a phone decoder can be seen as a reference system for representing speech sounds in any language, its phonetic units playing the role of bases. This seemingly simple representation involves a complex model, since each phonetic unit gets activated for specifi sequences of spectral features, which include both static and dynamic information. Some units could be strongly correlated among each other, so they would get activated at the same time (e.g. all the nasal consonants would activate when a nasal sound appeared in the input), but each phonetic unit will also provide specific information that will help catching the subtle differences between sounds. This specifi nature of posteriors is expected to provide enough degrees of freedom to represent speech sounds in different languages. In fact, the goodness of this representation will depend on the richness of the inventory of phonetic units (i.e. how well they cover the sounds appearing in other languages) and the optimality of the classification/mixin process (i.e. how well the neural network estimates the mix of phonetic units to best represent any sound).

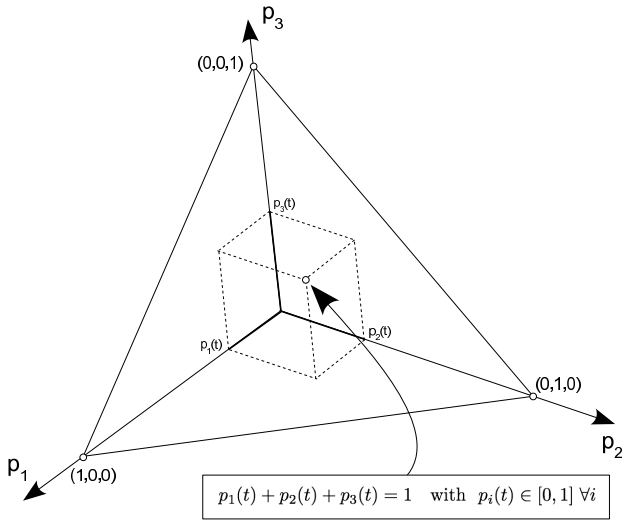

Fig. 1. Standard 2-simplex define by phone posteriors in the case of a phone decoder with 3 phonetic units.

The phone posteriors computed according to Eq. 1 exhibit extremely skewed and sparse (non-Gaussian) distributions (see Figure 2, firs row). This is not suitable for the kind of models we apply in SLR systems, since they typically assume that features are Gaussian-distributed. To address the non-Gaussian nature of phone posteriors, we can transform them into log-posteriors, obtaining more suitable but still non-Gaussian distributions (see Figure 2, second row). If we go further and take the logarithm of the likelihood ratio (using phone posteriors as likelihoods), the obtained distributions are nearly Gaussian (see Figure 2, third row).

At each frame t, log-likelihood ratios are computed in the classical way for a binary verificatio task (using fla priors), as follows:

$$LLR_i(t) = \log \frac{p_i(t)}{\frac{1}{(n-1)} \sum_{\forall j \neq i} p_j(t)} \qquad i = 1, ..., n \tag{3}$$

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

IV. System Configuration

A. Baseline MFCC-SDC iVector System

Under the total variability modeling approach [10], an utterance dependent GMM supervector $\mathbf{M}$ (stacking GMM mean vectors) is decomposed as follows:

$$\mathbf{M} = \mathbf{m} + \mathbf{T}\mathbf{w} \tag{4}$$

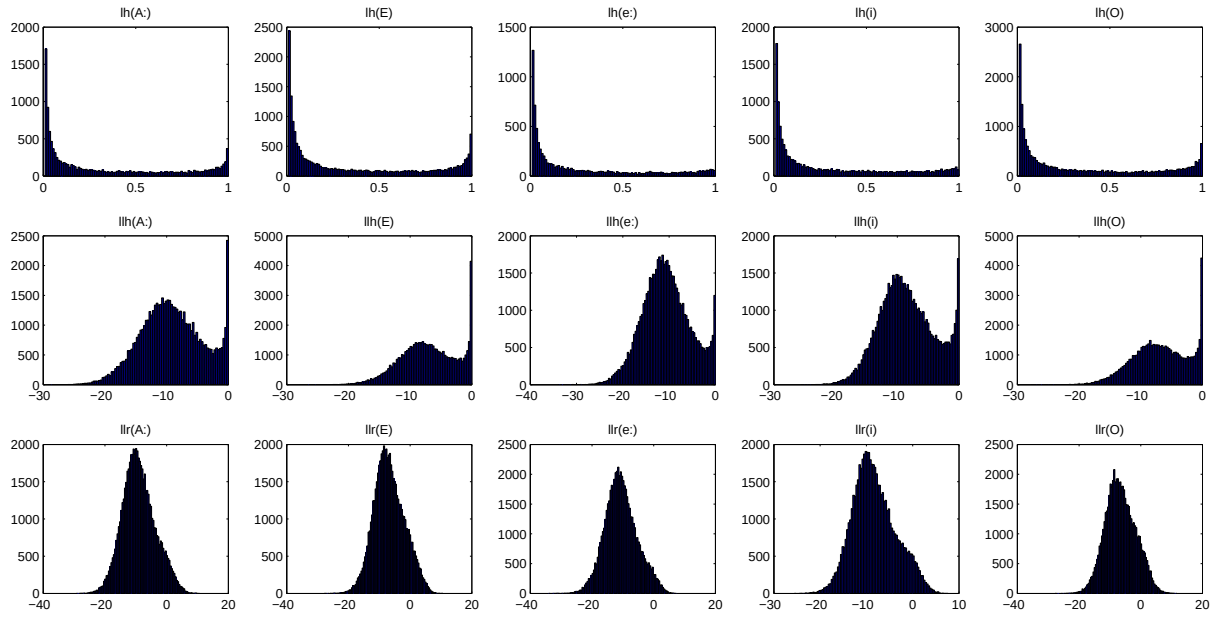

Fig. 2. Distributions of frame-level phone likelihoods (lh, first row), phone log-likelihoods (llh, second row) and phone log-likelihood ratios (llr, third row) for 5 phonetic units (A:, E, e:, i, O) of the Brno University of Technology decoder for Hungarian, computed on a subset of the NIST 2007 LRE test set.

where $\mathbf{m}$ is the utterance independent mean supervector, $\mathbf{T}$ is the total variability matrix (a low-rank rectangular matrix) and $\mathbf{w}$ is the so called *iVector* (a normally distributed low-dimensional latent vector). That is, $\mathbf{M}$ is assumed to be normally distributed with mean $\mathbf{m}$ and covariance $\mathbf{TT}'$. The latent vector $\mathbf{w}$ can be estimated from its posterior distribution conditioned to the Baum-Welch statistics extracted from the utterance and using a Universal Background Model (UBM). The iVector approach maps high-dimensional input data (a GMM supervector) to a low-dimensional feature vector (an iVector), hypothetically maintaining most of the relevant information.

As in previous works [26] [27], the concatenation of MFCC and SDC coefficient under a 7-2-3-7 configuratio was used as acoustic representation for the baseline acoustic iVector system. Voice Activity Detection (VAD) can be performed by applying a phone decoder to detect and discard non-speech segments. In [28] [29], the Brno University of Technology (BUT) phone decoder for Hungarian [30] was successfully applied as a VAD system. We performed VAD in a similar way, by removing the feature vectors whose highest PLLR value corresponded to the non-phonetic unit, using the BUT phone decoder for Hungarian (see Section IV-C for details on the computation of PLLRs).

A gender independent 1024-mixture UBM was estimated by the Maximum Likelihood criterion on the training dataset, using binary mixture splitting, orphan mixture discarding and variance flooring The total variability matrix $\mathbf{T}$ was estimated according to the procedure define in [10], but using only data from target languages, as in [12].

A generative modeling approach was applied in the iVector feature space [12], the set of iVectors of each language being modeled by a single Gaussian distribution. Thus, the iVector scores were computed as follows:

$$score(f, l) = \mathcal{N}(\mathbf{w}_f; \mu_l, \mathbf{\Sigma}) \tag{5}$$

where $\mathbf{w}_f$ is the iVector for target signal f, μ_l is the mean iVector for language l and $\mathbf{\Sigma}$ is a common (shared by all languages) within-class covariance matrix.

B. Baseline Phonotactic Systems

The three phonotactic systems applied in this work were developed under the Phone-lattice-SVM approach [31] [19]. Given an input signal, an energy-based voice activity detector was applied in firs place, which split and removed long-duration non-speech segments. Then, the open software Temporal Patterns Neural Network (TRAPs/NN) phone decoders, developed by the Brno University of Technology (BUT) for Czech (CZ), Hungarian (HU) and Russian (RU) [30] were applied. The main features of these decoders are:

- Czech decoder - 8 kHz, trained on the Czech SpeechDat(E) database, containing 12 hours of speech from 1052 Czech (526 male, 526 female) speakers, recorded over the Czech fi ed telephone network.
- Hungarian decoder - 8 kHz, trained on the Hungarian SpeechDat(E) database, containing 10 hours of speech from 1000 Hungarian (511 male, 489 female) speakers, recorded over the Hungarian fi ed telephone network.
- Russian decoder - 8 kHz, trained on the Russian SpeechDat(E) database, containing 18 hours of speech from 2500 Russian (1242 male, 1258 female) speakers, recorded over the Russian fi ed telephone network.

Regarding channel compensation, noise reduction, etc. the three systems relied on the acoustic front-end provided by BUT decoders.

BUT decoders were configure to produce phone posteriors that were converted to phone lattices by means of HTK [32] along with the BUT recipe [30]. Then, expected counts of phone n-grams were computed using the *lattice-tool* of SRILM [33]. Finally, a Support Vector Machine (SVM) classifie was applied, SVM vectors consisting of expected frequencies of phone n-grams (up to $n = 3$), weighted as in [18]. A sparse representation was used, which involved only the most frequent features according to a greedy feature selection algorithm [19]. L2-regularized L1-loss support vector regression was applied, by means of LIBLINEAR [34].

C. PLLR iVector System

As a firs step to get the PLLR features, the BUT TRAPs/NN phone decoders for Czech, Hungarian and Russian [30] were applied. These decoders feature 45, 61 and 52 phonetic units, respectively, each unit being represented by a three-state model.

For each input utterance, BUT decoders produce a sequence of numbers $x_{i,s}(t)$ encoding the posterior probabilities $p_{i,s}(t)$, for each state s of each phone model i at each frame t, in the following way:

$$x_{i,s}(t) = \sqrt{-2 \log p_{i,s}(t)} \tag{6}$$

Thus, the posterior probability $p_{i,s}(t)$ can be obtained as follows:

$$p_{i,s}(t) = e^{-\frac{(x_{i,s}(t))^2}{2}} \tag{7}$$

For each decoder, the non-phonetic units *int* (intermittent noise), *pau* (short pause) and *spk* (non-speech speaker noise) were integrated into a single non-phonetic unit. Then, according to Equation 1, a single posterior probability was computed for each model i $(1 \leq i \leq n)$ by adding the posterior probabilities of its states. Finally, log-likelihood ratios were computed according to Equation 3. In this way, PLLR vectors of size 43 (CZ), 59 (HU) and 50 (RU) were obtained at each frame t.

Once the PLLR features were computed, the remaining elements of the iVector system: voice activity detection, GMM, total variability matrix and iVector scoring, were designed and applied in the same way as for the acoustic iVector system (see Section IV-A).

D. Backend Approaches and Setup

The backend serves as a precalibration stage that transforms the space of scores to get reliable estimates of the true class probabilities. Besides, when the set of languages for which models have been trained does not match the set of target languages, the backend maps the available scores to the space of target languages. In the case of NIST LRE datasets, separate models can be trained for different dialects of a target language or for different data sources (telephone conversational speech, radio broadcast speech, etc.), and non-target languages can be modeled as well. Different backends can be applied [26]:

- *Generative Gaussian Backend*: In a generative Gaussian backend, the distribution of language scores is modeled by a multivariate normal distribution $\mathcal{N}(\mu_t, \mathbf{\Sigma})$ for each target language t, where the full covariance matrix $\mathbf{\Sigma}$ is shared across all target languages. Maximum Likelihood (ML) estimates of the means and the covariance matrix are computed.
  Given a score vector $\mathbf{s}$ of size K, the output (calibrated) log-likelihood vector $\hat{\mathbf{s}}$ is obtained by:
  $$\hat{\mathbf{s}} = \mathbf{A}\mathbf{s} + \mathbf{b} + \mathbf{c} \tag{8}$$
  where the rows of $\mathbf{A}$ are:
  $$\mathbf{a}_t = \mu_t' \mathbf{\Sigma}^{-1} \tag{9}$$
  and the elements of $\mathbf{b}$ and $\mathbf{c}$ are (note that $\mathbf{c}$ is a constant vector):
  $$b_t = -\frac{1}{2} \mu_t' \mathbf{\Sigma}^{-1} \mu_t \tag{10}$$
  $$c_t = -\frac{K}{2} \log(2\pi) - \frac{1}{2} \log|\mathbf{\Sigma}| - \frac{1}{2} \mathbf{s}' \mathbf{\Sigma}^{-1} \mathbf{s} \tag{11}$$
- *Discriminative Gaussian Backend*: In this case, ML estimates of the means and the common covariance matrix are used initially, but further reestimates of the means are iteratively computed in order to maximize the Maximum Mutual Information (MMI) criterion:
  $$F_{\mathbf{MMI}}(\lambda) = \sum_{\forall \mathbf{s}} \log \frac{p_\lambda(\mathbf{s}|l(\mathbf{s}))^C}{\sum_{\forall l} p_\lambda(\mathbf{s}|l)^C p(l)} \tag{12}$$
  where $p_\lambda(\mathbf{s}|l(\mathbf{s}))$ is the likelihood of the score vector $\mathbf{s}$ given the true target language $l(\mathbf{s})$ and model parameters λ, $p(l)$ is the probability of language l and C is an heuristic factor. In this work, 20 MMI iterations were performed and C was set to 10.
- *ZT-Norm*: Score normalization techniques such as Z-norm and T-norm [35] can help removing the environmental effects on the score space. Nevertheless, they are rarely applied alone in SLR systems. Instead, they are usually applied before some other backend. The Z-norm aims to compensate for deviations related to the target language, thus Z-norm parameters are estimated using target scores on a set of development signals containing non-target languages. The T-norm aims to compensate for deviations related to the test signal, thus T-norm parameters are estimated using non-target scores on the test signal.

The backend setup was separately optimized for each dataset. A ZT-norm followed by a discriminative Gaussian backend was applied in experiments on the NIST 2007 and 2009 LRE datasets, whereas a generative Gaussian backend was applied in experiments on the NIST 2011 LRE dataset. In experiments on the Albayzin 2010 LRE dataset, no backend was applied, since none of them improved performance. This latter result could be due to unreliable estimates of backend parameters, since the development set of the Albayzin 2010 LRE was small compared to those of NIST LRE.

E. Calibration and Fusion

The *FoCal* multiclass toolkit was applied to perform the calibration and fusion of SLR systems. The FoCal toolkit provides a single framework to calibrate/fuse the scores of

several systems, assuming that they represent (or can be interpreted) as log-likelihoods. The principles underlying this approach have been extensively described in [36] [2] [37].

Let L be the number of target languages and let $s_j(X,t)$ be the score corresponding to system j for the input utterance X and target language t. The calibration and fusion of k (including $k=1$) systems can be simultaneously performed by computing the fused scores $s_f(X,t)$ as a linear combination of the system scores, as follows [37]:

$$s_f(X,t) = \sum_{j=1}^{k} \alpha_j s_j(X,t) + \beta_t \qquad (13)$$

The weights α_j ($j \in [1,k]$) and β_t ($t \in [1,L]$) are estimated in a discriminative way, by applying linear logistic regression to minimize the so called *Log-Likelihood Ratio Cost* C_{LLR}, which measures the information provided by the scores in order to take *good* decisions (see Section VI for a formal definitio of C_{LLR}).

V. Datasets

A. NIST 2007 LRE

The NIST 2007 LRE [38] define a spoken language recognition task for conversational speech across telephone channels, involving 14 target languages.

Training and development data used in this work were limited to those distributed by NIST to all 2007 LRE participants: (1) the Call-Friend Corpus[2]; (2) the OHSU Corpus provided by NIST for the 2005 LRE[3]; and (3) the development corpus provided by NIST for the 2007 LRE[4]. A set of 23 languages/dialects was define for training, including target and non-target[5] languages. For development purposes, 10 conversations per language were randomly selected, and the remaining conversations (amounting to around 968 hours) were used for training. Development conversations were further divided into 30-second speech segments. The total number of 30-second segments was 3073 (see Table I for more details). Results reported in this paper have been computed on the subset of 30-second speech segments of the test set for the closed-set condition (2158 segments), which was the primary task in the NIST 2007 LRE.

B. NIST 2009 LRE

The NIST 2009 LRE featured 23 target languages [39], involving 11 target languages for which Conversational Telephone Speech (CTS) was available in the NIST 2007 LRE dataset, plus 12 new target languages for which Broadcast Narrow-Band Speech was provided. Most of the speech provided for the latter consisted of telephone calls included in Voice of America (VOA) broadcasts. For the 12 new target languages, NIST distributed between 141 and 199 30-second audited VOA segments per language. Additional non-audited materials were provided for the 23 target languages and for several non-target languages (see Table II).

[2]See http://www.ldc.upenn.edu/.

[3]OHSU Corpora, http://www.ohsu.edu/.

[4]See http://www.itl.nist.gov/iad/mig/tests/lre/2007/.

[5]French was the only non-target language used for NIST 2007 LRE.

TABLE I
2007 NIST LRE CORE CONDITION: TRAINING DATA (HOURS), DEVELOPMENT AND EVALUATION DATA (# 30S SEGMENTS), DISAGGREGATED FOR TARGET AND NON-TARGET LANGUAGES.

	Hours	**# 30s cuts**	
Language	Train	Devel	Eval
Arabic	52.59	179	80
Bengali	5.0	76	80
Chinese	166.12	567	398
English	143.7	288	240
Farsi	46.22	225	80
German	57.03	173	80
Industani	64.35	243	240
Japanese	79.11	141	80
Korean	72.86	150	80
Russian	5.0	66	160
Spanish	117.35	531	240
Tamil	58.18	165	160
Thai	5.0	64	80
Vietnamese	46.7	205	160
Non-Target	48.74	-	-
TOTAL	967.95	3073	2158

TABLE II
2009 NIST LRE CORE CONDITION: TRAINING DATA (HOURS), DEVELOPMENT AND EVALUATION DATA (# 30S SEGMENTS), DISAGGREGATED FOR TARGET AND NON-TARGET LANGUAGES.

	Hours		**# 30s cuts**			
	Train		Devel			Eval
	2007	2009	2007 (CTS)		2009	2009
Language	(CTS)	(VOA)	devel	eval	(VOA)	
Amharic	-	58.31	-	-	262	398
Bosnian	-	5.63	-	-	259	355
Cantonese	5.0	2.45	83	80	104	378
Creole	-	7.21	-	-	256	323
Croatian	-	6.45	-	-	190	376
Dari	-	69.05	-	-	276	389
EngAmerican	130.7	9.01	204	80	230	896
EngIndian	13.0	-	84	160	-	574
Farsi/Persian	46.22	25.16	225	80	294	390
French	48.74	67.91	222	80	293	395
Georgian	-	4.32	-	-	166	399
Hausa	-	48.31	-	-	274	389
Hindi	59.35	10.06	174	160	178	667
Korean	72.86	5.70	150	80	250	463
Mandarin	151.1	32.4	331	158	230	1015
Pashto	-	184.3	-	-	281	395
Portuguese	-	25.74	-	-	240	397
Russian	5.0	147.76	66	160	299	511
Spanish	117.35	45.44	531	240	242	385
Turkish	-	6.67	-	-	289	394
Ukrainian	-	5.59	-	-	281	388
Urdu	5.0	36.60	69	80	299	379
Vietnamese	46.7	9.5	205	160	240	315
Non-Target	266.92	408.82	-	-	-	-
TOTAL	967.95	1222.39	2344	1518	5433	10571

Training and development data used in this work were limited to those distributed by NIST to all 2009 LRE participants. A set of 64 languages/dialects was define for training models. Each of them was mapped either to a target language or to non-target languages[6]. For example, Mainland and Taiwan from NIST 2007 LRE and Mandarin from VOA were all mapped to Mandarin, whereas Arabic was mapped to non-target languages. Persian and Farsi were mapped to the same language, as was properly pointed out in [28].

For languages appearing in VOA recordings, the longest speech segments out of each fil were posted to the training

[6]The set of non-target languages define for the NIST 2009 LRE includes: Arabic, Bengali, German, Japanese, Tamil and Thai from CTS recordings, and Albanian, Azerbaijani, Bangla, Burmese, Greek, Indonesian, Khmer, Kinyarwanda/Kirundi, Kurdish, Macedonian, Ndebele, Oromo, Serbian, Shona, Somali, Swahili, Tibetan, Tigrigna, and Uzbek from VOA broadcasts.

TABLE III
NIST 2011 LRE CORE CONDITION: TRAINING DATA (HOURS) AND DEVELOPMENT AND EVALUATION DATA (# 30S SEGMENTS), DISAGGREGATED FOR TARGET AND NON-TARGET LANGUAGES

	Hours				**# 30s cuts**					
	Train				Devel					Eval
Language	2007 (CTS)	2009 (VOA)	2011 (30s audit)	Other sources	2007 (CTS)	2007 (eval)	2009 (VOA)	2009 (eval)	2011 (audit)	2011
Arabic Iraqi	-	-	0.48	20.34	-	-	-	-	48	308
Arabic Levantine	-	-	0.47	27.56	-	-	-	-	49	308
Arabic Maghrebi	-	-	0.41	1.79	-	-	-	-	54	305
Arabic MSA	-	-	0.47	1.87	-	-	-	-	51	306
Bengali	5.0	54.40	-	-	76	80	296	43	-	412
Czech	-	-	0.41	4.19	-	-	-	-	56	261
Dari	-	69.05	-	-	-	-	276	389	-	267
English American	130.7	9.01	-	-	204	80	230	896	-	221
English Indian	13.0	-	-	-	84	160	-	574	-	387
Farsi/Persian	46.22	25.16	-	-	225	80	294	390	-	404
Hindi	59.35	10.06	-	-	174	160	178	667	-	213
Lao	-	-	0.50	2.22	-	-	-	-	41	62
Mandarin	151.1	32.40	-	-	331	158	230	1015	-	360
Panjabi	-	-	0.50	-	-	32	-	9	45	299
Pashto	-	184.38	-	-	-	-	281	395	-	383
Polish	-	-	0.51	1.79	-	-	-	-	46	267
Russian	5.0	147.76	-	-	66	160	299	511	-	441
Slovak	-	-	0.41	1.69	-	-	-	-	56	280
Spanish	117.3	45.44	-	-	531	240	242	385	-	419
Tamil	58.18	-	-	-	165	160	-	-	-	414
Thai	5.0	-	-	-	64	80	-	188	-	375
Turkish	-	6.67	-	-	-	-	289	394	-	276
Ukrainian	-	5.59	-	-	-	-	281	388	-	170
Urdu	5.0	36.60	-	-	69	80	299	379	-	478
Non-Target	257.76	406.93	-	-	-	-	-	-	-	-
TOTAL	853.61	1033.45	4.16	61.45	1989	1470	3135	6623	446	7616

dataset, using no more than 2 segments per file and a minimum of 225 segments per language. The number of segments extracted per fil was relaxed (augmented) for those languages with few file in VOA.

The whole training dataset (CTS from NIST 2007 LRE and VOA broadcast speech from NIST 2009 LRE) amounted to 2190 hours. For development, some materials taken from the development and evaluation datasets of the NIST 2007 LRE were used (see Table I). For languages appearing in VOA, besides the audited segments provided by NIST, additional randomly extracted speech segments, each around 30 seconds long (specificall , between 25 and 35 seconds long), were used. The whole development dataset consisted of 9295 segments. Results reported in this paper were computed on the NIST 2009 LRE evaluation corpus, specificall on the 30-second, closed-set condition (primary evaluation task).

C. NIST 2011 LRE

In the NIST 2011 LRE, 24 target languages were considered (see Table III). Among them, 9 languages had never been used before in NIST LRE. Development data specificall collected for these 9 languages were sent to participants, including 100 30-second segments per language. For a better coverage, we randomly split those subsets into two disjoint subsets (each having approximately half the segments for each language/dialect): the firs half was used to train specifi models for the new languages, and the second half was used to estimate backend and fusion parameters [19].

To train more robust models for the target languages, we added data from databases distributed by the Linguistic Data Consortium (LDC), some of them containing conversational telephone speech (LDC2006S45 for Arabic Iraqi, LDC2006S29 for Arabic Levantine) and others containing broadcast speech (LDC2000S89 and LDC2009S02 for Czech). For these latter, only automatically detected telephone-speech segments were used.

The remaining materials were extracted from wide-band broadcast news recordings, dowsampling them to 8 kHz and applying the Filtering and Noise Adding Tool[7] (FANT) to simulate a telephone channel. The COST278 Broadcast News database [40] was used to get speech segments for Czech and Slovak. Arabic MSA was extracted from Al Jazeera broadcasts included in the KALAKA-2 database created for the Albayzin 2010 LRE [41]. Finally, broadcasts were also *captured* from video archives in TV websites to get speech segments in Arabic Maghrebi (Arrabia TV, http://www.arrabia.ma) and Polish (Telewizja Polska, TVP INFO, http://tvp.info). TV broadcasts were fully audited, so that only reasonably clean speech segments were selected for training. We couldn't collect additional training materials for Panjabi by any means. Therefore, a single model (trained on just 55 segments) was used for this language.

A set of 66 languages/dialects was define for training. Each of them was mapped either to a target language or to non-target languages[8]. The training dataset includes the data mentioned above (one-half of the audited segments plus other sources) plus 2007 CTS and 2009 VOA signals (see Tables I and II). The whole training dataset for the NIST 2011 LRE benchmark amounts to 1953 hours.

[7]Available online: http://dnt.kr.hsnr.de

[8]The set of non-target languages define for the NIST 2011 LRE includes: French, German, Japanese, Korean and Vietnamese from CTS recordings, and Albanian, Amharic, Creole, French, Georgian, Greek, Hausa, Indonesian, Kinyarwanda/Kirundi, Korean, Ndebele, Oromo, Shona, Somali, Swahili, Tibetan and Tigrigna from VOA broadcasts.

For development purposes, the second half of the audited segments provided for new target languages, along with the NIST 2007 and 2009 evaluation datasets, and 30-second signals used for development in 2007 and 2009 (see Tables I and II) were used. The whole development dataset consists of 13663 segments. Results reported in this paper were computed on the NIST 2011 LRE evaluation corpus, specificall on the 30-second, closed-set condition (primary evaluation task) (see Table III for more details).

D. Albayzin 2010 LRE (KALAKA-2)

The Albayzin 2010 LRE dataset (KALAKA-2) contains wide-band 16 kHz TV broadcast speech signals for six target languages (see Table IV). The Albayzin 2010 LRE [41] featured two main evaluation tasks, on clean and noisy speech, respectively. In this work, acoustic processing involved downsampling signals to 8 kHz, since all the systems were designed to deal with narrow-band signals.

The training, development and evaluation datasets used for this benchmark match exactly those define for the Albayzin 2010 LRE. For the primary clean-speech language recognition task, more than 10 hours of clean speech per target language were used for training. For the noisy-speech language recognition task, besides the clean speech subset, more than 2 hours of noisy/overlapped speech segments were used for each target language. The distribution of training data, which amounts to around 82 hours, is shown in Table IV. Only 30-second segments were used for development purposes. The development dataset used in this work consists of 1192 segments, amounting to more than 10 hours of speech. Results reported in this paper were computed on the Albayzin 2010 LRE evaluation corpus, specificall on the 30-second, closed set condition (for both clean speech and noisy speech conditions). The distribution of segments in the development and evaluation datasets is shown in Table IV. For further details, see [42].

TABLE IV
ALBAYZIN 2010 LRE: DISTRIBUTION OF TRAINING DATA (HOURS) AND DEVELOPMENT AND EVALUATION DATA (# 30S SEGMENTS)

	Clean Speech			Noisy Speech		
	Hours	# 30s cuts		Hours	# 30s cuts	
Language	Train	Devel	Eval	Train	Devel	Eval
Basque	10.73	146	130	2.25	29	74
Catalan	11.45	120	149	2.18	47	55
English	12.18	133	135	2.53	60	69
Galician	10.74	137	121	2.23	60	83
Portuguese	11.08	164	146	3.28	77	58
Spanish	10.41	136	125	3.70	83	79
TOTAL	66.59	836	806	16.17	356	418

VI. EVALUATION MEASURES

The four benchmarks considered in this work define the same type of SLR task, which is known as *spoken language verification*. Given a trial, consisting of a test segment and a target language, the system must decide whether or not the target language is spoken in the test segment (Accept/Reject).

In spoken language verificatio tasks, two types of errors are considered: (1) *misses*, those for which the correct answer is *Accept* (target trials) but the system says *Reject*; and (2) *false alarms*, those for which the correct answer is *Reject* (impostor trials) but the system says *Accept*. Therefore, for any test condition the corresponding error rates can be computed as the fraction of target trials that are rejected (*miss error rate*, P_{miss}) and the fraction of impostor trials that are accepted (*false alarm error rate*, P_{fa}), and suitable cost functions can be define as combinations of these basic error rates. Note that the decision of the system may vary according to application dependent parameters of the cost function (typically, the prior probabilities of the target languages and the costs of misses and false alarms). A well calibrated system should be able to automatically adapt the decisions to each particular application.

A. Equal Error Rate (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}). EER is a very simple measure, useful in many contexts, but it does not allow us to measure the global performance of a system (i.e. for a wide range of operation points). Moreover, it does not take into account the ability of the system to be positioned at the EER operation point (i.e. the performance loss due to bad calibration), since the threshold value is implicitly chosen *a posteriori* by the evaluator.

B. Average Cost (C_{avg})

This measure is a combination of P_{miss} and P_{fa} pooled across target languages. For closed-set evaluation tasks, it is computed as follows:

$$C_{avg} = \frac{1}{L}\sum_{i=1}^{L}\{C_{miss}P_{target}P_{miss(i)} + \frac{1}{L-1}\sum_{\substack{j=1\\ j\neq i}}^{L} C_{fa}(1-P_{target})P_{fa}(i,j)\} \quad (14)$$

where L is the number of target languages, and P_{target} (the target prior), C_{miss} (the miss error cost) and C_{fa} (the false alarm error cost) are application-dependent parameters. In this work, $P_{target} = 0.5$ and $C_{miss} = C_{fa} = 1$.

Unlike EER, the C_{avg} accounts for the calibration loss, but it is still limited to a single operation point. The C_{avg} is chosen as the primary evaluation measure in this work due to historical reasons: C_{avg} was the primary measure in NIST evaluations until 2011, and many authors have reported system performance in terms of C_{avg}.

C. NIST 2011 LRE metric, C_{avg}^{24}

In the NIST 2011 LRE an alternative metric was used to evaluate system performance, based on a pairwise cost function define by:

$$C(L1,L2) = C_{L1}P_{L1}P_{miss}(L1) + C_{L2}(1-P_{L1})P_{miss}(L2) \quad (15)$$

where $L1$ and $L2$ denote languages 1 and 2, respectively, $C_{L1} = C_{L2} = 1$ and $P_{L1} = 0.5$.

The overall C_{avg}^{24} measure was define as the mean of the $C(L1, L2)$ values over the 24 language pairs for which the C_{min} values were greatest[9]. For further details, see [43].

D. Log-Likelihood Ratio Cost (C_{LLR})

When the scores represent (or can be interpreted as) log-likelihoods, systems can be evaluated in terms of the so called C_{LLR} [36], which has been used as alternative performance measure in some NIST evaluations. C_{LLR} allows us to evaluate the system performance globally by means of a single numerical value. It only depends on the scores (it does not depend on application dependent parameters), on their ability to discriminate amongst target languages each other and on how well they are calibrated, the two key features of a SLR system. On the other hand, it has higher statistical significanc than EER or C_{avg}, since it is computed from verificatio scores (in contrast to EER or C_{avg}, which depend only on Accept/Reject decisions). Let us now recall how C_{LLR} is computed.

Let $LR(X, i)$ be the *likelihood ratio* corresponding to segment X and target language i. The likelihood ratio can be expressed in terms of the conditional probabilities of X with regard to the alternative target and non-target hypotheses, as follows:

$$LR(X, i) = \frac{prob(X|i)}{prob(X|\neg i)} \tag{16}$$

Let E be an evaluation dataset, consisting of the union of L disjoint subsets: E_j ($j \in [1, L]$) containing speech segments in the target language j. Pairwise costs $C_{LLR}(i, j)$, for $i, j \in [1, L]$, are define as follows:

$$C_{LLR}(i,j) = \begin{cases} \frac{1}{|E_i|} \sum\limits_{X \in E_i} \log_2(1 + LR(X,i)^{-1}) & j = i \\ \frac{1}{|E_j|} \sum\limits_{X \in E_j} \log_2(1 + LR(X,i)) & j \neq i \end{cases} \tag{17}$$

Finally, the average C_{LLR} is computed by adding the pairwise costs for all the combinations of target and non-target languages, as follows:

$$C_{LLR} = \frac{1}{L} \sum_{i=1}^{L} \{P_t \cdot C_{LLR}(i,i) + \sum_{\substack{j=1 \\ j \neq i}}^{L} P_{nt} \cdot C_{LLR}(i,j)\} \tag{18}$$

where P_t is the prior probability of target languages and $P_{nt} = (1 - P_t)/(L - 1)$ is the prior probability of non-target languages.

The C_{LLR} takes unbounded non-negative values expressed in information units (bits), with lower values representing better performance, the value 0 corresponding to a perfect system and the value $\log_2(L)$ corresponding to a system which just relies on priors, thus providing no information to decide a trial. In this work, the C_{LLR} has been computed by means of the FoCal toolkit [44]. Further details about the reasons for using this measure and its interpretation can be found in [36] [2].

[9] $C_{min}(L1, L2)$: minimum of the pairwise cost function, found for the optimal operation point (threshold).

VII. Development Experiments

A detailed study was carried out on the NIST 2007 LRE dataset, with the aim to fin the optimal configuratio of the iVector system trained on PLLR features. The NIST 2007 LRE dataset was selected as our primary benchmark for two main reasons: (1) this database is well-known in the SLR community, and has been used as benchmark for many techniques and applications; and (2) the size of the database makes it suitable for extensive experimentation.

A. Dynamic coefficients

Dynamic coefficient of acoustic representations have proven to be effective in speech, speaker and language recognition, so we firs evaluated the usefulness of this technique when applied to PLLR features. Three different parameterizations based on the PLLR features were tested. The firs one comprised only the PLLR static features. In the second one, the feature vector was augmented with first-orde dynamic coefficient (PLLR+Δ), as define in [32]:

$$\Delta f(t) = \frac{\sum_{d=1}^{D} d\,[f(t+d) - f(t-d)]}{2\sum_{d=1}^{D} d^2} \tag{19}$$

where $f(t)$ is a PLLR feature at time t, and $2D + 1$ is the size of the regression window. First-order dynamic coefficient were computed using Eq. 19 with $D = 2$. The third parameterization comprised static features plus first and second-order (acceleration) dynamic coefficient (PLLR + Δ + $\Delta\Delta$). Second-order dynamic coefficient were computed using Eq. 19 on first-orde dynamic coefficients with $D = 1$.

Tests were carried out using the BUT decoders for Czech, Hungarian and Russian. Similar results and conclusions can be obtained with any of them. Therefore, for the sake of clarity, Table V shows only results for one of them (the BUT decoder for Hungarian).

TABLE V
EER, $C_{avg} \times 100$ AND C_{LLR} PERFORMANCE FOR IVECTOR SYSTEMS USING PLLR, PLLR+Δ AND PLLR+Δ+$\Delta\Delta$ FEATURES COMPUTED WITH THE HU BUT DECODER, ON THE LRE07 PRIMARY EVALUATION TASK.

System	EER	$C_{avg} \times 100$	C_{LLR}
PLLR	3.77	3.45	0.564
PLLR+Δ	2.69	2.66	0.382
PLLR+Δ+$\Delta\Delta$	3.58	3.60	0.506

Clearly, the use of first-orde dynamic coefficient improved significantl the performance of the system. However, adding second-order dynamic coefficient did not improve but instead degraded performance. This could be reflectin a dimensionality issue: when using the BUT decoder for Hungarian, PLLR+Δ+$\Delta\Delta$ feature vectors reach 177 dimensions, thus producing supervectors of 181248 dimensions for a 1024-mixture GMM. Given such a large dimensionality, probably we don't have enough data to get reliable estimates of all the parameters. Therefore, in the experiments reported in the following sections, PLLR+Δ will be used as features.

B. Variability compensation

SLR performance degradation is related in most cases to channel/noise variability. Several techniques can be applied to

minimize the effect of this variability. Among them, Feature Normalization [45] and Feature Warping [46] are two of the most effective (the latter, commonly used in speaker recognition). However, as shown in Table VI, no improvement in performance was attained when applying any of these techniques under the PLLR-iVector approach.

TABLE VI
EER, $C_{avg} \times 100$ AND C_{LLR} PERFORMANCE FOR IVECTOR SYSTEMS USING PLLR FEATURES COMPUTED WITH THE BUT HU DECODER, WITH: (A) NO NOISE REDUCTION TECHNIQUE, (B) FEATURE NORMALIZATION (FN) AND (C) FEATURE WARPING (FW), ON THE LRE07 PRIMARY EVALUATION TASK.

System	EER	$C_{avg} \times 100$	C_{LLR}
PLLR	2.69	2.66	0.382
PLLR +FN	3.04	2.95	0.436
PLLR +FW	3.18	3.21	0.435

C. Overall results on NIST 2007 LRE

Results obtained under the PLLR-iVector approach with the three BUT decoders, using PLLR features (with firs order dynamic coefficient and without noise reduction/compensation) are shown in Table VII. The performance of acoustic and phonotactic baseline systems is also presented and compared to that of the PLLR-iVector systems, along with all the possible fusions, to study their complementarity.

TABLE VII
EER, $C_{avg} \times 100$ AND C_{LLR} PERFORMANCE FOR THE MFCC-SDC IVECTOR BASELINE SYSTEM, IVECTOR SYSTEMS USING PLLR FEATURES, PHONOTACTIC BASELINE SYSTEMS AND THE FUSION OF THEM, FOR EACH OF THE BUT DECODERS, ON THE LRE07 PRIMARY EVALUATION TASK.

System			EER	$C_{avg} \times 100$	C_{LLR}
MFCC-SDC iVector		(a)	2.75	2.85	0.407
CZ	Phonotactic	(b1)	2.97	2.94	0.440
	PLLR iVector	(c1)	4.29	4.18	0.550
Fusion	(a)+(b1)		1.24	1.22	0.189
	(a)+(c1)		1.93	1.95	0.280
	(b1)+(c1)		1.74	1.79	0.257
	(a)+(b1)+(c1)		1.20	1.24	0.176
HU	**Phonotactic**	**(b2)**	**1.97**	**2.08**	**0.310**
	PLLR iVector	(c2)	2.69	2.66	0.382
Fusion	(a)+(b2)		1.02	1.08	0.152
	(a)+(c2)		1.40	1.40	0.215
	(b2)+(c2)		1.10	1.20	0.166
	(a)+(b2)+(c2)		**0.80**	**0.82**	**0.124**
RU	Phonotactic	(b3)	2.70	2.69	0.383
	PLLR iVector	(c3)	4.15	4.08	0.549
Fusion	(a)+(b3)		1.25	1.13	0.182
	(a)+(c3)		1.69	1.72	0.265
	(b3)+(c3)		1.75	1.76	0.240
	(a)+(b3)+(c3)		1.03	1.10	0.163

If we focus on single-systems, the best performance was yielded by the HU phonotactic system, which attained 2.08 $C_{avg} \times 100$, followed by the HU PLLR iVector system, which outperformed the acoustic MFCC-SDC iVector system. The RU and CZ phonotactic systems performed worse that the HU phonotactic system. The performance of the RU and CZ PLLR iVector systems degraded, with regard to the HU PLLR iVector system, in the same proportion as the corresponding phonotactic systems.

Regarding the fused systems, similar conclusions can be drawn from any of the CZ/RU/HU sets of results. The fusion of the acoustic iVector system and the PLLR iVector system yielded very good results. As may be expected, the fusion of the phonotactic and the acoustic iVector systems provided very competitive performance, but was closely followed by the fusion of the phonotactic and PLLR systems. This latter result is specially relevant, since phonotactic and PLLR features share a common source (the phone posterior probabilities provided by BUT decoders) but systems proved to be complementary. This supports our claim that PLLR features convey acoustic-phonetic information that is partly lost when collapsing phone posteriors to get phone lattices. Finally, the best performance was always yielded by the fusion of the three systems (remarkably, for HU: 0.80 EER, 0.82 $C_{avg} \times 100$ and 0.124 C_{LLR}). These results support our claim that PLLR features could be complementary to both short-term spectral and phonotactic features.

D. Statistical Significance

In this Section, we study the statistical significanc of the relative performance improvements attained when adding the PLLR-based system to baseline systems. First, we provide a quick overview of such improvements in Figure 3, in terms of the absolute number of miss and false alarm errors for the baseline systems alone (dark gray columns) and for the baseline systems fused with the HU PLLR iVector system (light gray columns). Note that P_{miss} and P_{fa} are computed by dividing the number of miss and false alarm errors by the number of target and non-target trials, respectively (in this regard, the NIST 2007 LRE primary task involves 2158 target trials and 28054 non-target trials). In all cases, the PLLR iVector system made errors to decrease in more that 20%, up to almost 50%.

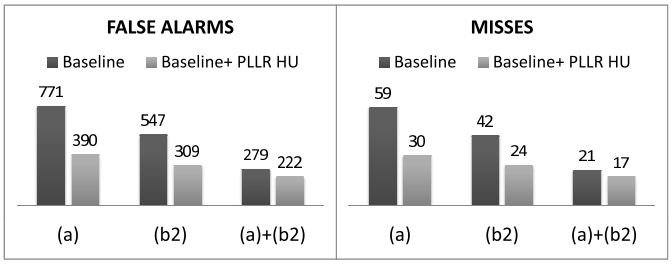

Fig. 3. False alarm and miss errors on the NIST LRE 2007 primary task for baseline systems: (a) acoustic MFCC-SDC iVector system, (b2) Phone-Lattice-SVM system and (a)+(b2) the fusion of the two latter, taken alone (dark gray) and fused with (c2) the HU PLLR iVector system (light gray).

To measure the statistical significanc of performance improvements (in terms of C_{avg}, which is the primary performance measure in this work), a series of two-tailed paired T-tests was carried out [47], which gives an idea of the variability of performance improvements (and thus, the robustness of such improvements) across randomly define sets of data. To that end, the NIST LRE 2007 evaluation dataset was split into 20 language-balanced disjoint random subsets. Then, C_{avg} values were computed on each subset for baseline systems (a), (b2) and (a)+(b2) and for the same systems fused with the PLLR HU system: (a)+(c2), (b2)+(c2) and (a)+(b2)+(c2). Figure 4 shows the mean and the confidenc interval at 95% confidenc level of the relative C_{avg} improvements, revealing that they are statistically significan in all cases.

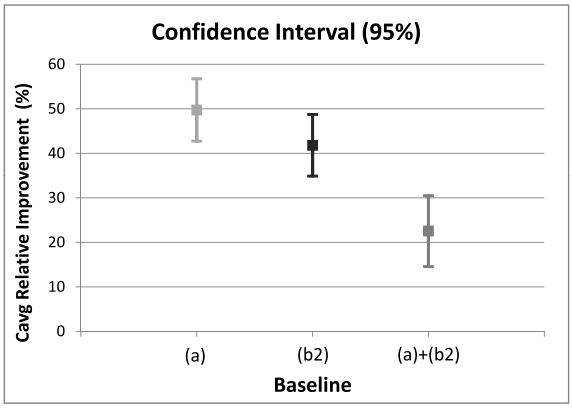

Fig. 4. Means of C_{avg} relative improvements and their corresponding intervals at 95% confidenc level, on the NIST LRE 2007 primary task, when fusing the HU PLLR iVector system (c2) with baseline systems: (a) acoustic MFCC-SDC iVector system, (b2) Phone-Lattice-SVM system and (a)+(b2) the fusion of the two latter.

VIII. Results

To save the reader an unmanageable floo of results which would not contribute new insights, in this Section results will be only reported for a single decoder. Based on the performance attained on the NIST 2007 LRE dataset, we will only report results for the BUT decoder for Hungarian, using PLLR features and their firs order dynamic coefficients Note, however, that better performance could be attained by optimizing the setup for each dataset, as previous results using phonotactic systems suggest [26] [27]. In particular, the best choice of phone decoder would depend on how well the acoustic-phonetic space of target languages is covered by the inventory of phonetic units.

A. NIST 2009 LRE

Table VIII shows the performance of the baseline systems and the proposed approach on the NIST 2009 LRE primary evaluation task. Note that the three systems performed more or less the same, the PLLR-iVector system being slightly better than the baseline systems. On the other hand, fusion performance was consistent with the results obtained for the NIST 2007 LRE dataset. The fusion of the phonotactic and MFCC-SDC iVector systems provided the best pairwise performance, followed closely by the fusion of the PLLR iVector and phonotactic systems. The fusion of the PLLR iVector and MFCC-SDC iVector systems was also competitive. Once again, the best performance was achieved when fusing the three systems: 1.48 $C_{avg} \times 100$.

TABLE VIII
EER, $C_{avg} \times 100$ AND C_{LLR} PERFORMANCE FOR THE BASELINE PHONOTACTIC AND IVECTOR SYSTEMS, THE PLLR IVECTOR SYSTEM AND THE FUSION OF THEM, ON THE LRE09 PRIMARY EVALUATION TASK.

System			EER	$C_{avg} \times 100$	C_{LLR}
MFCC-SDC iVector		(a)	2.63	2.70	0.535
HU	Phonotactic	(b)	2.44	2.49	0.502
	PLLR iVector	(c)	**2.43**	**2.42**	**0.505**
Fusion	(a)+(b)		1.63	1.67	0.346
	(a)+(c)		1.79	1.79	0.392
	(b)+(c)		1.66	1.69	0.357
	(a)+(b)+(c)		**1.47**	**1.48**	**0.321**

B. NIST 2011 LRE

Table IX shows the EER, C_{avg} and C_{LLR} performance of the baseline systems and the proposed approach on the NIST 2011 LRE primary evaluation task. Table X shows the performance in terms of the actual and minimum C_{avg}^{24} (primary metric used in the NIST 2011 LRE).

TABLE IX
EER, $C_{avg} \times 100$ AND C_{LLR} PERFORMANCE FOR THE PHONOTACTIC AND IVECTOR BASELINE SYSTEMS, THE PLLR IVECTOR SYSTEM AND THE FUSION OF THEM, ON THE LRE11 PRIMARY EVALUATION TASK.

System			EER	$C_{avg} \times 100$	C_{LLR}
MFCC-SDC iVector		(a)	5.95	5.96	1.088
HU	Phonotactic	(b)	7.10	7.15	1.280
	PLLR iVector	(c)	**5.21**	**5.18**	**0.982**
Fusion	(a)+(b)		4.35	4.34	0.823
	(a)+(c)		3.99	4.00	0.789
	(b)+(c)		4.40	4.39	0.829
	(a)+(b)+(c)		**3.70**	**3.63**	**0.714**

TABLE X
MIN $C_{avg}^{24} \times 100$ AND ACTUAL $C_{avg}^{24} \times 100$ PERFORMANCE FOR THE PHONOTACTIC AND IVECTOR BASELINE SYSTEMS, THE PLLR IVECTOR SYSTEM AND THE FUSION OF THEM, ON THE LRE11 PRIMARY EVALUATION TASK.

System			min $C_{avg}^{24} \times 100$	$C_{avg}^{24} \times 100$
MFCC-SDC iVector		(a)	11.63	13.56
HU	Phonotactic	(b)	12.49	14.28
	PLLR iVector	(c)	**9.83**	**12.12**
Fusion	(a)+(b)		7.98	10.43
	(a)+(c)		7.78	10.19
	(b)+(c)		7.75	10.31
	(a)+(b)+(c)		**6.68**	**9.14**

For this benchmark, the PLLR iVector approach stands out among single systems, attaining 5.18 $C_{avg} \times 100$, which means 13% relative improvement with regard to the MFCC-SDC iVector approach and 28% relative improvement with regard to the phonotactic approach. Fusion performance was consistent with these results: the best pairwise fusion involved the two iVector systems. The next best pairwise fusion depended on the metric, but both combinations (MFCC-SDC iVector system + phonotactic system and PLLR iVector system + phonotactic system) attained similar performance. In any case, as for the two previous benchmarks, the PLLR iVector system seemed to provide complementary information to both baseline systems. Finally, the fusion of the three systems yielded the best performance: 3.63 $C_{avg} \times 100$ and 9.14 $C_{avg}^{24} \times 100$.

C. Albayzin 2010 LRE

Table XI shows the performance of the baseline systems and the proposed approach on the Albayzin 2010 LRE closed-set clean-speech 30-second task. In this case, the best single system was the proposed PLLR iVector system, yielding 1.41 $C_{avg} \times 100$, which means a 33% relative improvement with regard to the acoustic iVector approach and a 40% relative improvement with regard to the phonotactic approach.

The fusion of the acoustic and PLLR iVector systems yielded high performance also on this benchmark. Fusing the phonotactic and acoustic iVector systems yielded almost the same performance than fusing the phonotactic and PLLR iVector systems. Finally, as in previous experiments, the fusion

TABLE XI
EER, $C_{avg} \times 100$ AND C_{LLR} PERFORMANCE FOR THE BASELINE SYSTEMS, THE PLLR IVECTOR SYSTEM AND DIFFERENT FUSIONS ON THE ALBAYZIN 2010 LRE PRIMARY TASK ON CLEAN SPEECH.

System			EER	$C_{avg} \times 100$	C_{LLR}
MFCC-SDC iVector		(a)	2.25	2.12	0.176
HU	Phonotactic	(b)	2.37	2.35	0.218
	PLLR iVector	**(c)**	**1.31**	**1.41**	**0.127**
Fusion	(a)+(b)		1.06	1.10	0.106
	(a)+(c)		1.18	1.20	0.109
	(b)+(c)		0.91	1.09	0.092
	(a)+(b)+(c)		**1.00**	**0.97**	**0.086**

of the three systems yielded the best performance: 0.97 $C_{avg} \times 100$.

Table XII shows the performance of the baseline systems and the proposed approach on the Albayzin 2010 LRE closed-set noisy-speech 30-second task. As on the clean-speech condition, the best single system was the proposed PLLR iVector system, yielding 3.17 $C_{avg} \times 100$, a 20% relative improvement with regard to the acoustic iVector approach and a 56% relative improvement with regard to the phonotactic approach.

TABLE XII
EER, $C_{avg} \times 100$ AND C_{LLR} PERFORMANCE FOR THE BASELINE SYSTEMS, THE PLLR IVECTOR SYSTEM AND DIFFERENT FUSIONS ON THE ALBAYZIN 2010 LRE PRIMARY TASK ON NOISY SPEECH.

System			EER	$C_{avg} \times 100$	C_{LLR}
MFCC-SDC iVector		(a)	3.85	3.95	0.325
HU	Phonotactic	(b)	7.48	7.28	0.621
	PLLR iVector	**(c)**	**3.38**	**3.17**	**0.308**
Fusion	(a)+(b)		2.28	2.43	0.211
	(a)+(c)		2.80	2.65	0.227
	(b)+(c)		2.58	2.65	0.228
	(a)+(b)+(c)		**1.75**	**1.86**	**0.168**

The fusion of the acoustic and PLLR iVector systems yielded the best pairwise performance. As on the clean-speech condition, the fusion of the phonotactic and acoustic iVector systems yielded the same performance than the fusion of the phonotactic and PLLR iVector systems, and the best fusion involved the three systems, with 1.86 $C_{avg} \times 100$. Once again, the PLLR iVector system seems to provide complementary information to baseline systems under all configurations

D. Overall Performance Comparison

To have a wider perspective, in this Section we present results reported by other authors on the same databases. To allow easy comparisons, the most relevant results obtained with PLLR features are also included in Tables XIII–XVI (in boldface).

1) NIST 2007 LRE: Many results have been published in the literature for the primary task of the NIST 2007 LRE (30-second speech segments, closed-set condition). Table XIII shows some of them, reported by the Massachusetts Institute of Technology (MIT) [21] [18], the Brno University of Technology (BUT) [48], the Institute for Infocomm Research (IIR) [49] and the Laboratoire d'Informatique pour la Mecanique et les Sciences de l'Ingnieur (LIMSI) [50].

Results in Table XIII suggest that quite similar performance can be attained by applying either acoustic or phonotactic approaches. Best performance has been reported when fusing several acoustic and phonotactic subsystems [21]. When compared to other approaches, the PLLR iVector system stands out for providing high-performance fusions, thanks to its complementarity with both acoustic and phonotactic approaches.

TABLE XIII
REPORTED RESULTS ON THE PRIMARY TASK OF THE NIST 2007 LRE

Approach	Model	EER	$C_{avg} \times 100$
Acoustic	GMM-MMI [21]	–	2.10
	GSV-SVM [21]	–	1.92
	Discriminative GMM-MAP [48]	–	1.74
Phonotactic	HU, Phone-SVM, lattices [49]	1.84	–
	HU, Phone-SVM, lattices [18]	2.40	–
	EN, Phone-SVM, lattices [18]	1.80	–
PLLR	**HU, iVector, generative**	2.69	2.66
Fusions	2 acoustic subsystems [21]	–	1.55
	2 phonotactic subsystems [21]	–	1.55
	3 phonotactic subsystems [50]	–	0.90
	4 subsystems [21]	0.93	0.97
	acoustic+phonotactic+PLLR	0.80	0.82

2) NIST 2009 LRE: Table XIV shows results on the primary task of the NIST 2009 LRE (30-second speech segments, closed-set condition), reported by MIT [1] [11] [51], BUT [28] [52] [13] [14], Politecnico di Torino [22] and LIMSI [50]. Again, best performance was attained when fusing several acoustic and phonotactic subsystems [1] [14] [50] [52] [22].

TABLE XIV
REPORTED RESULTS ON THE PRIMARY TASK OF THE NIST 2009 LRE

Approach	Model	EER	$C_{avg} \times 100$
Acoustic	GMM-MMI [11]	2.30	–
	SVM-GSV [1]	–	2.30
	JFA [28]	–	2.02
	iVector (LDA+WCCN) [11]	2.40	–
	iVector, generative [13]	–	3.09
	iVector (Logistic reg.) [14]	–	2.35
Phonotactic	EN, Phone-SVM [1]	–	2.34
	HU, Phone-SVM [52]	–	3.85
	RU, Phone-SVM [52]	–	3.03
PLLR	**HU, iVector, generative**	2.43	2.42
Fusions	2 acoustic subsystems [1]	–	2.00
	2 acoustic subsystems [14]	–	1.78
	3 phonotactic subsystems [52]	–	2.39
	3 phonotactic subsystems [50]	–	1.99
	3 subsystems [1]	–	1.64
	36 subsystems [22]	–	1.16
	acoustic+phonotactic+PLLR	1.47	1.48

For this database, the performance of the PLLR iVector system is among the best of single systems, and the fusion of the two baseline systems and the PLLR iVector system yields the second best reported result on this task (the best one corresponding to the fusion of 36 subsystems).

3) NIST 2011 LRE: Recently, some results have been published for the primary task of the NIST 2011 LRE (30-second speech segments, closed-set condition). Table XV shows those reported by MIT [23], BUT [29], the Institute for Infocomm Research (I^2R) [53] and the Bilbao-Lisboa-Zaragoza (BLZ) team [54]. For this dataset, performance is reported in terms of the new metric define by NIST (C_{avg}^{24}). Note that the PLLR iVector system yields state-of-the-art performance. However, performance figure reported on NIST 2011 LRE tasks strongly depend on the training and development materials available (due to the scarcity of data for some target languages) [29]. Therefore, no clear conclusions

regarding features and methodologies can be drawn from performance differences found on this dataset.

TABLE XV
REPORTED RESULTS ON THE PRIMARY TASK OF THE NIST 2011 LRE

Approach	Model	$C_{avg} \times 100$	$C_{avg}^{24} \times 100$	
			min	actual
Acoustic	iVector (LDA) [23]	4.15	–	8.90
	iVector (HLDA) [29]	–	–	10.35
	GMM-SVM [53]	–	10.41	–
Phonotactic	RU, PCA [29]	–	–	14.32
	HU, n-gram iVector [29]	–	–	15.42
PLLR	**HU, iVector, generative**	5.18	9.83	12.12
Fusions	5 subsystems [23]	3.30	–	7.00
	8 subsystems [54]	3.35	5.09	7.64
	3 subsystems [29]	–	–	8.47
	3 subsystems [53]	–	9.02	–
	acoustic+phonotactic+PLLR	3.63	6.68	9.14

4) Albayzin 2010 LRE: Table XVI shows results reported for the primary task of the Albayzin 2010 LRE (30-second speech segments, closed-set condition). In this case, the PLLR iVector system outperforms all the state-of-the-art systems reported so far, even fusions of several subsystems.

TABLE XVI
REPORTED RESULTS ON THE PRIMARY TASK OF THE ALBAYZIN 2010 LRE

Condition	Approach	Model	$C_{avg} \times 100$
Clean	Acoustic	JFA [55]	1.86
		GMM-MMI [55]	4.33
	Phonotactic	HU, Phone-ML [55]	4.41
		EN, Phone-ML [3]	3.26
	PLLR	**HU, iVector, generative**	1.41
	Fusions	8 subsystems [55]	1.84
		5 subsystems [56]	1.77
		2 subsystems [57]	1.81
		acoustic+phonotactic+PLLR	0.97
Noisy	Fusions	5 subsystems [56]	3.90
		2 subsystems [57]	2.53
		acoustic+phonotactic+PLLR	1.86

This result is very interesting, since the Albayzin 2010 LRE dataset consists of wide-band (16kHz) speech signals, which means that systems reported in the literature may have taken advantage of the additional information present in the 4-8kHz band, whereas all the systems reported in this paper (including the acoustic MFCC-SDC iVector system) have been built and evaluated on downsampled (8kHz) speech signals.

IX. DISCUSSION

In the following paragraphs, we suggest reasons that may explain the high performance attained by PLLR features in all benchmarks and discuss practical issues that make them suitable to improve state-of-the-art spoken language recognition systems.

PLLR features are extracted at every frame, telling us about the probability that a speech segment around the considered frame matches the corresponding phone model. It is expected that segments actually containing a given phone will yield the highest posterior for that phone, though posteriors for similar phones or for phones appearing in the surrounding contexts may be also high. So, the set of PLLRs at any given frame could provide an indirect way of characterizing the spectral content of a speech segment through the use of acoustic-phonetic models.

On the other hand, the MFCC-SDC representation commonly used in the iVector SLR systems essentially describes a spectral trajectory spanning about 200 milliseconds. Phone decoders are also based on spectral features, either directly modeling sequences of them or featuring a left-right topology of states, which makes them also capable of matching sequences of spectral features. In this regard, PLLR features may be providing the same information than MFCC-SDCs. However, though indirect, PLLR features may provide a more robust characterization of the spectral content than MFCC-SDCs, since many different phone models are *simultaneously* contributing to such characterization. In the experiments reported above, PLLRs led to better performance than MFCC-SDCs under the iVector approach, and attending to fusion performance, they also provided complementary information.

In the PLLR-iVector approach presented in this paper, phonotactic information conveyed by PLLR sequences has been left aside. PLLRs have been used just in the same way as acoustic features, i.e. by modeling frame-level distributions, firs through a GMM and then by means of an iVector approach which maps GMM statistics into a reduced dimensionality total variability space. However, except for the NIST 2007 LRE benchmark, PLLR-iVector systems yielded the same or better performance than phonotactic systems starting from the same information source (phone posterior probabilities). Differences in performance were even more significan on the Albayzin 2010 LRE datasets, in both clean-speech and noisy-speech conditions. This could be due to different reasons: on the one hand, test segments in the Albayzin 2010 LRE contain continuous multi-speaker speech (not sparse single-speaker speech fragments, as for NIST datasets); on the other hand, test segments were extracted from wide-band broadcast recordings and downsampled to 8 kHz, whereas BUT phone decoders applied in this work were trained on telephone speech (which makes them more suitable for NIST LRE datasets). Probably, under mismatch conditions such as those of the Albayzin 2010 LRE dataset (specially in the presence of background noise), n-gram counts are less reliable than PLLR distributions. This suggests that acoustic-phonetic information provided by PLLRs may serve the SLR task in a more effective way than n-gram counts estimated on phone lattices extracted from the same phone posteriors. In any case, these two ways of using phone posteriors seem to be complementary, since fusions of PLLR-iVector and phonotactic systems consistently led to significan performance improvements.

From a practical point of view, PLLR features can simply replace the MFCC-SDC features (or whatever other features) used in acoustic SLR systems. The remaining modules could be kept unchanged. In fact, this is what we did with the acoustic iVector system presented in this work. A SLR system based on PLLR features could thus be easily and rapidly deployed. The public availability and high performance (also in terms of computation time) of BUT phone decoders for Czech, Hungarian and Russian makes it very simple to compute PLLR features and search for the choice of decoder that best fit any

target application[10]. Finally, attending to the results presented in this paper, a PLLR-based iVector system could outperform other state-of-the-art acoustic and phonotactic systems, but even in the case it wouldn't, or if the highest performance was required, including it in a multi-system fusion would very likely improve performance, since it has proven to be complementary to both acoustic and phonotactic systems.

X. Conclusions

In this work, the so called Phone Log-Likelihood Ratios (PLLR) have been proposed as features for spoken language recognition. The proposed features have been analysed and interpreted, highlighting their potential application to represent the sounds of spoken language at the acoustic-phonetic level and their complementarity with regard to other features. From a practical point of view, PLLRs are naturally obtained from existing phone decoders and can be easily integrated in state-of-the-art language recognition systems.

An iVector system trained on PLLR features has consistently yielded competitive performance on four different benchmarks, outperforming an iVector system trained on MFCC-SDC features in all cases and yielding the same or better performance than a Phone-Lattice-SVM system on 3 out of 4 benchmarks. The fusion of the system with acoustic and phonotactic systems yielded between 27% and 66% relative improvements with regard to baseline systems, and the fusion of the three approaches yielded the best results in all cases, outperforming the fusion of phonotactic and MFCC-SDC iVector systems and revealing that the proposed PLLR features actually provide complementary information to both acoustic and phonotactic features.

References

[1] P. Torres-Carrasquillo, E. Singer, T. Gleason, A. McCree, D. Reynolds, F. Richardson, and D. Sturim, "The MITLL NIST LRE 2009 Language Recognition System," in *ICASSP*, 2010, pp. 4994–4997.

[2] N. Brümmer and D. van Leeuwen, "On calibration of language recognition scores," in *Proceedings of Odyssey - The Speaker and Language Recognition Workshop*, 2006, pp. 1–8.

[3] L. J. Rodriguez-Fuentes, M. Penagarikano, A. Varona, M. Diez, G. Bordel, D. Martínez, J. Villalba, A. Miguel, A. Ortega, E. Lleida, A. Abad, O. Koller, I. Trancoso, P. Lopez-Otero, L. Docio-Fernandez, C. Garcia-Mateo, R. Saeidi, M. Soufi ar, T. Kinnunen, T. Svendsen, and P. Franti, "Multi-site Heterogeneous System Fusions for the Albayzin 2010 Language Recognition Evaluation," in *Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop (ASRU) 2011*, Hawaii, USA, December 2011.

[4] P. A. Torres-Carrasquillo, E. Singer, M. A. Kohler, R. J. Greene, D. A. Reynolds, and J. R. Deller, "Approaches to Language Identificatio using Gaussian Mixture Models and Shifted Delta Cepstral Features," in *Proceedings of ICSLP*, 2002, pp. 89–92.

[5] F. Castaldo, D. Colibro, E. Dalmasso, P. Laface, and C. Vair, "Acoustic Language Identificatio Using Fast Discriminative Training," in *Proceedings of Interspeech*, Antwerp, Belgium, August 2007, pp. 346–349.

[6] L. Burget, P. Matejka, and J. Cernocky, "Discriminative Training Techniques for Acoustic Language Identification" in *Proceedings of IEEE ICASSP*, vol. I, Toulouse, France, May 2006, pp. 209–212.

[7] W. Shen and D. Reynolds, "Improved GMM-Based Language Recognition using Constrained MLLR Transforms," in *Proceedings of IEEE ICASSP*, 2008, pp. 4149–4152.

[10]To help implementing PLLR-based approaches, we provide a sample script for PLLR feature extraction based on BUT decoders: https://sites.google.com/site/gttspllrfeatures/home

[8] P. Kenny, "Joint factor analysis of speaker and session variability: Theory and algorithms," CRIM, Tech. Rep. Technical Report CRIM-06/08-13, 2005, [Online]. Available: http://www.crim.ca/perso/patrick.kenny/.

[9] F. Castaldo, S. Cumani, P. Laface, and D. Colibro, "Language Recognition Using Language Factors," in *Proceedings of Interspeech*, Brighton, UK, September 2009, pp. 176–179.

[10] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification" *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 4, pp. 788 –798, may 2011.

[11] N. Dehak, P. A. Torres-Carrasquillo, D. A. Reynolds, and R. Dehak, "Language Recognition via i-vectors and Dimensionality Reduction," in *Proceedings of the Interspeech 2011*, Florence, Italy, August 27-31 2011, pp. 857–860.

[12] D. Martínez, O. Plchot, L. Burget, O. Glembek, and P. Matejka, "Language Recognition in iVectors Space," in *Proceedings of Interspeech*, Firenze, Italy, 2011, pp. 861–864.

[13] D. Martínez, L. Burget, L. Ferrer, and N. Scheffer, "iVector-based Prosodic System for Language Identification" in *Proceedings of ICASSP*, Japan, 2012, pp. 4861–4864.

[14] O. Plchot, M. Karaf át, N. Brümmer, O. Glembek, P. Matejka, and E. de Villiers J. Cernocký, "Speaker vectors from Subspace Gaussian Mixture Model as complementary features for Language Identification" in *Odyssey 2012: The Speaker and Language Recognition Workshop*, Singapore, June 2012, pp. 330–333.

[15] W. Campbell, T. Gleason, J. Navratil, D. Reynolds, W. Shen, E. Singer, and P. Torres-Carrasquillo, "Advanced Language Recognition using Cepstra and Phonotactics: MITLL System Performance on the NIST 2005 Language Recognition Evaluation," in *Proceedings of Odyssey 2006 - The Speaker and Language Recognition Workshop*, San Juan, Puerto Rico, June 2006, pp. 1–8.

[16] R. Tong, B. Ma, H. Li, and E. S. Chng, "A Target-Oriented Phonotactic Front-End for Spoken Language Recognition," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 17, no. 7, pp. 1335–1347, September 2009.

[17] M. Penagarikano, A. Varona, L. J. Rodríguez-Fuentes, and G. Bordel, "Improved Modeling of Cross-Decoder Phone Co-occurrences in SVM-based Phonotactic Language Recognition," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 8, pp. 2348–2363, November 2011.

[18] F. Richardson and W. Campbell, "Language recognition with discriminative keyword selection," in *Proceedings of ICASSP*, 2008, pp. 4145–4148.

[19] M. Penagarikano, A. Varona, L. Rodriguez-Fuentes, and G. Bordel, "Dimensionality Reduction for Using High-Order n-grams in SVM-Based Phonotactic Language Recognition," in *Proceedings of Interspeech 2011*, Florence, Italy, August 28-31 2011, pp. 853–856.

[20] M. Soufi ar, S. Cumani, L. Burget, and J. H. Cernocky, "Discriminative Classifier for Phonotactic Language Recognition with iVectors," in *Proceedings of ICASSP*, Kyoto, Japan, 2012, pp. 4853–4856.

[21] P. Torres-Carrasquillo, E. Singer, W. Campbell, T. Gleason, A. McCree, D. Reynolds, F. Richardson, W. Shen, and D. Sturim, "The MITLL NIST LRE 2007 language recognition system," in *Proceedings of Interspeech*, 2008, pp. 719–722.

[22] F. Castaldo, D. Colibro, S. Cumani, E. Dalmasso, P. Laface, and C. Vair, "Loquendo-Politecnico di Torino system for the 2009 NIST Language Recognition Evaluation," in *ICASSP*, 2010, pp. 5002–5005.

[23] E. Singer, P. A. Torres-Carrasquillo, D. A. Reynolds, A. McCree, F. Richardson, N. Dejak, and D.Sturim, "The MITLL NIST LRE 2011 Language Recognition System," in *Odyssey 2012: The Speaker and Language Recognition Workshop*, Singapore, June 2012, pp. 209–215.

[24] N. F. Chen, W. Shen, and J. P. Campbell, "A Linguistically-Informative Approach to Dialect Recognition Using Dialect-Discriminating Context-Dependent Phonetic Models," in *Proceedings of IEEE ICASSP*, 2010, pp. 5014–5017.

[25] F. Biadsy, J. Hirschberg, and D. P. W. Ellis, "Dialect and accent recognition using phonetic-segmentation supervectors," in *Proceedings of Interspeech*, Firenze, Italia, August 28-31 2011, pp. 745–748.

[26] M. Penagarikano, A. Varona, M. Diez, L. J. Rodriguez Fuentes, and G. Bordel, "Study of Different Backends in a State-Of-the-Art Language Recognition System," in *Interspeech 2012*, Portland, Oregon, USA, 9-13 September 2012.

[27] M. Penagarikano, A. Varona, L. J. Rodriguez Fuentes, M. Diez, and G. Bordel, "The EHU Systems for the NIST 2011 Language Recognition Evaluation," in *Interspeech 2012*, Portland, Oregon, USA, 9-13 September 2012.

[28] Z. Jancík, O. Plchot, N. Brummer, L. Burget, O. Glembek, V. Hubeika, M. Karaf át, P. Matejka, T. Mikolov, A. Strasheim, and J. Cernocký, "Data selection and calibration issues in automatic language recognition - investigation with BUT-AGNITIO NIST LRE 2009 system," in *Proc. Odyssey 2010 - The Speaker and Language Recognition Workshop*, 2010, pp. 215–221.

[29] N. Brümmer, S. Cumani, O.Glembek, M. Karaf át, P. Matejka, J. Pesán, O. Plchot, M. Soufi ar, E. de Villiers, and J. Cernocký, "Description and analysis of the Brno 276 system for LRE2011," in *Odyssey 2012: The Speaker and Language Recognition Workshop*, Singapore, June 2012, pp. 216–223.

[30] P. Schwarz, "Phoneme recognition based on long temporal context," Ph.D. dissertation, Faculty of Information Technology, Brno University of Technology, http://www.fit.vutb .cz/, Brno, Czech Republic, 2008.

[31] W. M. Campbell, F. Richardson, and D. A. Reynolds, "Language Recognition with Word Lattices and Support Vector Machines," in *Proceedings of IEEE ICASSP*, Honolulu, Hawaii, USA, 2007, pp. 15–20.

[32] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Lui, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK Book (for HTK Version 3.4)*. Cambridge, UK: Entropic, Ltd., 2006.

[33] A. Stolcke, "SRILM - An extensible language modeling toolkit," in *Proceedings of Interspeech*, November 2002, pp. 257–286.

[34] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "LIBLINEAR: A Library for Large Linear Classification" *Journal of Machine Learning Research*, vol. 9, pp. 1871–1874, 2008, software available at http://www.csie.ntu.edu.tw/∼cjlin/liblinear.

[35] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas, "Score normalization for text-independent speaker verificatio systems," *Digital Signal Processing*, vol. 10, no. 1-3, pp. 42–54, January 2000.

[36] N. Brümmer and J. du Preez, "Application-Independent Evaluation of Speaker Detection," *Computer, Speech and Language*, vol. 20, no. 2-3, pp. 230–275, April-July 2006.

[37] N. Brümmer, L. Burget, J. Cernocky, O. Glembek, F. Grezl, M. Karafiat D. van Leeuwen, P. Matejka, P. Schwarz, and A. Strasheim, "Fusion of heterogeneous speaker recognition systems in the STBU submission for the NIST speaker recognition evaluation 2006," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 7, pp. 2072–2084, 2007.

[38] A. F. Martin and A. N. Le, "NIST 2007 Language Recognition Evaluation," in *Proceedings of Odyssey 2008 - The Speaker and Language Recognition Workshop, paper 016*, 2008.

[39] A. Martin and C. Greenberg, "The 2009 NIST Language Recognition Evaluation," in *Odyssey 2010 - The Speaker and Language Recognition Workshop, paper 030*, Brno, Czech Republic, 2010, pp. 165–171.

[40] A. Vandecatseye, J.-P. Martens, J. P. Neto, H. Meinedo, C. Garca-Mateo, J. Dieguez-Tirado, F. Mihelic, J. Zibert, J. Nouza, P. David, M. Pleva, A. Cizmar, H. Papageorgiou, and C. Alexandris, "The COST278 pan-european broadcast news database," in *Proceedings of LREC*, Lisbon, Portugal, 2004.

[41] L. J. Rodriguez-Fuentes, M. Penagarikano, A. Varona, M. Diez, and G. Bordel, "The Albayzin 2010 Language Recognition Evaluation," in *Proceedings of Interspeech*, Firenze, Italia, August 28-31 2011, pp. 1529–1532.

[42] L. J. Rodriguez-Fuentes, M. Penagarikano, A. Varona, M. Diez, and G. Bordel, "KALAKA-2: a TV broadcast speech database for the recognition of Iberian languages in clean and noisy environments," in *Proceedings of the LREC*, Istanbul, Turkey, 23-25 May 2012.

[43] *The 2011 NIST Language Recognition Evaluation Plan (LRE11)*, http://www.nist.gov/itl/iad/mig/upload/LRE11_EvalPlan_releasev1.pdf.

[44] FoCal, *Toolkit for Evaluation, Fusion and Calibration of statistical pattern recognizers*, 2008, http://sites.google.com/site/nikobrummer/focal.

[45] F.-H. Liu, R. M. Stern, X. Huang, and A. Acero, "Efficien cepstral normalization for robust speech recognition," in *Proceedings of the workshop on Human Language Technology*, ser. HLT '93. Stroudsburg, PA, USA: Association for Computational Linguistics, 1993, pp. 69–74.

[46] J. Pelecanos and S. Sridharan, "Feature Warping for Robust Speaker Verification" in *2001: A Speaker Odyssey - The Speaker Recognition Workshop*, 2001, pp. 213–218.

[47] L. Gillick and S. Cox, "Some statistical issues in the comparison of speech recognition algorithms," in *Porc. ICASSP*, 1989.

[48] N. Brümmer, A. Strasheim, V. Hubeika, P. Matejka, L. Burget, and O. Glembek, "Discriminative Acoustic Language Recognition via Channel-Compensated GMM Statistics," in *Proceedings of Interspeech*, Brighton, UK, September 2009, pp. 2187–2190.

[49] R. Tong, B. Ma, H. Li, and E. S. Chng, "Selecting Phonotactic Features for Language Recognition," in *Proceedings of Interspeech*, September 2010, pp. 737–740.

[50] M. F. BenZeghiba, J. L. Gauvain, and L. Lamel, "Fusing Language Information from Diverse Data Sources for Phonotactic Language Recognition," in *Odyssey 2012: The Speaker and Language Recognition Workshop*, Singapore, June 2012, pp. 346–352.

[51] F. S. Richardson and W. M. Campbell, "NAP for high level language identification" in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2011)*, Prague, Czech Republic, May 22-27 2011, pp. 4392–4395.

[52] T. Mikolov, O. Plchot, O. Glembek, P. Matejka, L. Burget, and J. Cernocky, "PCA-based Feature Extraction for Phonotactic Language Recognition," in *Proceedings of Odyssey 2010 - The Speaker and Language Recognition Workshop*, 2010, pp. 251–255.

[53] C. H. You, H. Li, E. Ambikairajah, K. A. Lee, and B. Ma, "Bhattacharyya-based GMM-SVM System with Adaptive Relevance Factor for Pair Language Recognition," in *Odyssey 2012: The Speaker and Language Recognition Workshop*, Singapore, June 2012, pp. 338–345.

[54] L. J. Rodriguez Fuentes, M. Penagarikano, A. Varona, M. Diez, G. Bordel, A. Abad, D. Martinez, J. Villalba, A. Ortega, and E. Lleida, "The BLZ Submission to the NIST 2011 LRE: Data Collection, System Development and Performance," in *Interspeech 2012*, Portland, Oregon, USA, 9-13 September 2012.

[55] D. Martínez, J. Villalba, A. Miguel, A. Ortega, and E. Lleida, "I3A Language Recognition System for Albayzin 2010 LRE," in *Proceedings of Interspeech*, Firenze, Italy, 2011, pp. 849–852.

[56] L. J. Rodriguez Fuentes, A. Varona, M. Diez, M. Penagarikano, and G. Bordel, "Evaluation of Spoken Language Recognition Technology Using Broadcast Speech: Performance and Challenges," in *Odyssey 2012: The Speaker and Language Recognition Workshop*, Singapore, June 25-28, 2012.

[57] A. Abad, O. Koller, and I. Trancoso, "The L2F Language Verificatio Systems for Albayzin 2010 Evaluation," in *VI Jornadas en Tecnologías del Habla and II Iberian SLTech Workshop*, Vigo, Spain, 10-12 November 2010, pp. 383–388.

Mireia Diez was born in Barakaldo, Spain, in 1985. She received the M.Sc. in Electronic Engineering from the University of the Basque Country (UPV/EHU) in 2009.
From 2010 she is pursuing the Ph.D. degree at the same university under a research grant. Her research interests include language recognition and speaker recognition.

Amparo Varona was born in Barakaldo, Spain, in 1970. She received the M.Sc. and Ph.D. degrees in Physics from the University of the Basque Country (UPV/EHU) in 1993 and 2000, respectively.
From 1994 to 1996 she was at the Department of Electricity and Electronics, UPV/EHU, under a research grant. From 1996 to 2003 she was Assistant Professor and since 2003 she has been Associate Professor of Computer Science in the same department. Her past research activities include language modeling and efficien search for ASR. Her current research interests include spoken document retrieval, language recognition and speaker recognition.

Mikel Penagarikano was born in Zumarraga, Spain, in 1973. He received the M.Sc. degree in Physics from the University of the Basque Country (UPV/EHU), Leioa, Spain, in 1996. He is currently pursuing the Ph.D. degree at the same university.
From 1997 to 2000 he was at the Department of Electricity and Electronics, UPV/EHU, under a research grant. Since 2000, he has been Assistant Professor of Computer Science in the same department. His research interest focuses on developing efficien software architectures for speech processing applications, such as ASR, language recognition, speaker recognition, etc.

Luis Javier Rodriguez-Fuentes was born in Bilbao in 1968. He received the M.Sc. and Ph.D. degrees in Physics from the University of the Basque Country (UPV/EHU) in 1991 and 2004, respectively.
From 1993 to 1996 he was at the Department of Electricity and Electronics, UPV/EHU, under a research grant. Since 1996, he has been Assistant Professor of Computer Science in the same department. His past research activities include acoustic modeling, spontaneous speech modeling and speaker adaptation for ASR. His current research interests include spoken document retrieval, language recognition and speaker recognition.

German Bordel was born in Bilbao in 1961. He received the M.Sc. and Ph.D. degrees in Physics from the University of the Basque Country (UPV/EHU) in 1985 and 1996, respectively.
In 1988, after two years working on microprocessor systems for control, he joined the Department of Electricity and Electronics, UPV/EHU, as Assistant Professor of Computer Science, where since 2008 he is Associate Professor. His interests include software architectures, web-based applications, spoken document retrieval and speaker recognition.