New Insight into the use of Phone Log-Likelihood Ratios as Features for Language Recognition

Mireia Diez, Amparo Varona, Mikel Penagarikano, Luis Javier Rodriguez-Fuentes and German Bordel

GTTS, Department of Electricity and Electronics,
University of the Basque Country, UPV/EHU, 48940, Leioa, Spain
mireia.diez@ehu.es

Abstract

Phone Log-Likelihood Ratio (PLLR) features have been recently introduced as an effective way of making use of frame-level phone posteriors in language and speaker recognition systems. In this paper, a deep insight into PLLR features is made and further evidence of the usefulness of these features in spoken language recognition tasks is provided, with a new set of experiments carried out on the NIST 2007 LRE dataset, combining the latest progresses made in optimizing the features. PLLR features are projected into a subspace that enhances the information retrieved by the system. Then, dimensionality reduction is performed on the projected subspace by means of Principal Component Analysis, and shifted deltas are computed on the reduced features to optimize performance. Figures attained are among the best reported so far on the NIST 2007 LRE dataset.

Index Terms: Spoken Language Recognition, Phone Log-Likelihood Ratios, Feature Projection, i-vectors, Shifted Delta

1. Introduction

Nowadays, most Language Recognition (LR) technologies benefit from the combination of acoustic and phonotactic approaches [1, 2, 3]. Both approaches take advantage of different types of information present in acoustic signals. On the one hand, acoustic features like Mel-Frequency Cepstral-Coefficients (MFCC) or Perceptual Linear Prediction (PLP) features, model the spectral characteristics of the audio signal at the frame level. On the other hand, phonotactic approaches are usually trained on top of the output of phone decoders, which produce sequences or lattices of tokens that carry phonetic, prosodic or even word-level information, that can be effectively used to characterize the spoken language.

Recent works have introduced a different way of extracting acoustic-phonetic information by using frame level phone log-posteriors [4], or Phone Log-Likelihood Ratios (PLLR) [5], to obtain frame-by-frame feature vectors. The approach allows for testing and assembling the features in conventional acoustic systems like those based on the well-known *Total Variability Factor Analysis* (i-vector) approach [6], which has been successfully applied to the task [7] and has become state-of-the-art in LR technology.

This work gives an insight into the use of the PLLRs, showing the latest improvements made on the extraction and processing of these features. With the aim of eliminating bounds present in the original feature space, PLLRs are projected into a hyper-plane which enhances the information retrieved by the system [8]. Then, the feature set dimensionality is reduced by means of Principal Component Analysis (PCA) so that Shifted Deltas (SD) can be computed on top of them. As with MFCC features [9], the use of SDs over the original PLLRs has proven to be very effective [10, 11], leading to one of the best results reported so far on the selected benchmark.

The study is carried out on the well known National Institute of Standard Technology (NIST) 2007 Language Recognition Evaluation (LRE) database. NIST LREs started in 1996 and have been held every two years since 2003 [12], setting an excellent benchmark for research in the field, and the datasets provided have been widely used as baseline by the community.

The rest of the paper is organized as follows: Section 2 provides details about the PLLR feature computation and post-processing. Section 3 describes the experimental setup, including the dataset, the language modeling, the backend applied and the evaluation measures used in this work. In Section 4, results are presented and compared to the ones reported by other authors on the same benchmark. Finally, Section 5 outlines the conclusions.

2. Insight into PLLR Features

2.1. Definition of PLLR Features

Let us consider a phone decoder that provides frame-by-frame phone posteriors p_i for each phone unit ($1 \leq i \leq N$), so that $\sum_{i=1}^{N} p_i = 1$ and $p_i \in [0, 1]$. The nature of the posteriors makes them lie in the subset of R^N known as the $(N-1)$-dimensional standard simplex:

$$\Delta^{(N-1)} = \{\mathbf{p} \in \mathbb{R}^N \mid \mathcal{F}(\mathbf{p}) = \sum_{i=1}^{N} p_i - 1 = 0 \ \wedge \ p_i \geq 0 \, \forall i\} \quad (1)$$

The PLLR features are computed from these phone posteriors as follows [13]:

$$r_i = \text{logit}(p_i) = \log \frac{p_i}{(1-p_i)} \qquad i = 1, ..., N \quad (2)$$

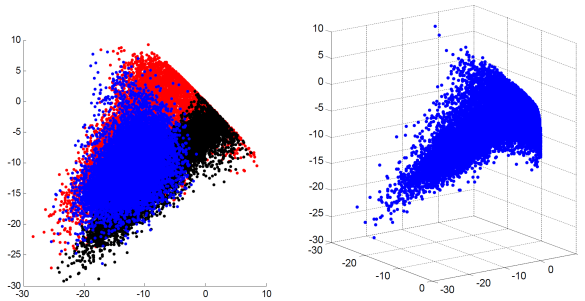
Thus, the hyper-surface $\mathcal{S}$ where the PLLRs lie can be derived from Equations 1 and 2 as:

$$\mathcal{S}^{(N-1)} = \left\{ \mathbf{r} \in \mathbb{R}^N \;\middle|\; \mathcal{G}(\mathbf{r}) = \sum_{i=1}^{N} \frac{1}{1+e^{-r_i}} - 1 = 0 \right\} \quad (3)$$

where $\mathcal{G}(\mathbf{r})$ is the implicit hyper-surface function. These PLLR features have been succesfully used for language and speaker recognition [5, 13].

2.2. Projected PLLR Features

In a previous work, we showed how the transformation of phone posteriors into PLLRs deals with the gaussanization of the features for each individual phone model [8]. However, the (N-1) standard simplex defined in Eq. 1, in which phone posteriors lie, is determined by a set of bounds that are still transferred into the PLLR feature space, making the hyper-surface $\mathcal{S}$ asymptotically perpendicular to the basis of PLLRs [8]. This restricts the areas where PLLRs are confined and therefore limits the distributions of the features. The bounds are clearly displayed when analyzing multi-dimensional distributions. Figure 1 illustrates two and three-dimensional distributions for some sets of PLLRs computed from a subset of signals from NIST 2007 LRE dataset, where limits are clearly displayed.



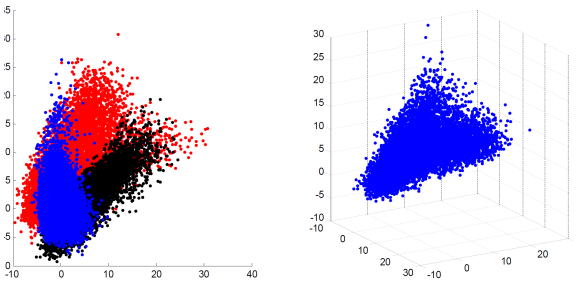
(a) (b)

Figure 1: Distribution of PLLRs for (a) three pairs of phones, *a:* vs *E* (red), *i* vs *i:* (black) and *dz* vs *h* (blue) and (b) the set of phones (*a:*, *E*, *O*), computed with the BUT HU phone decoder.

With the aim of obtaining a smoother representation of the features and to avoid this bounding effect, PLLRs are projected into the hyper-plane tangential to the top of the convex hyper-surface $\mathcal{S}$ (the point where all the posteriors take the same value $p_i = \frac{1}{N}$), using the projection matrix:

$$P = \mathbb{I} - \hat{\mathbf{1}}' * \hat{\mathbf{1}} \tag{4}$$

where $\hat{\mathbf{1}} = \frac{1}{\sqrt{N}}[1_1, 1_2, ..., 1_N]$ and $\mathbb{I}$ stands for the Identity matrix. Figure 2 displays the distributions of the projected features for the same set of PLLRs used in Figure 1, where bounds are no longer present.



(a) (b)

Figure 2: Distribution of the PLLRs shown in Figures 1(a) and 1(b) after projecting them into the defined hyper-plane.

After projecting the features, Principal Component Analysis (PCA) is applied to decorrelate the parameters, making them more suitable for the diagonal covariance Gaussian Mixture Model (GMM) used as Universal Background Model (UBM) in the i-vector approach (see subsection 3.2). Note that, as the projection method makes the features lie in a $(N-1)$-dimensional hyper-plane, the dimensionality of the feature vectors after PCA is reduced by one (the dimension corresponding to the null eigenvalue is removed).

Finally, the feature vector is augmented with first order dynamic coefficients.

2.3. Dimensionality Reduction and Shifted Deltas

First order Δ coefficients reflect short-term speech spectral dynamics which do not capture long term variations. Instead, Shifted Delta (SD) coefficients are longer-term temporal features, which better reflect the dynamics of the features. The use of SDs has proven to be a very effective way of improving LR performance [9]. Recent works have also shown the benefits of using SDs over the original PLLR features [11, 14].

Before computing SDs on the features, the dimensionality of the PLLR vector (without first order dynamic coefficients) should be reduced, in order to obtain a reasonably small feature set (to maintain an affordable computational cost of the system, and not to face the lack of data to get reliable estimates of the parameters). For that purpose, this work makes use of PCA, which besides decorrelating the feature space, allows for reducing the feature set dimensionality, while minimizing the degradation of the system. The PCA dimensionalities used for experimentation (23 and 13) were selected according to previous works [10, 14].

The SD-PLLR features are specified by four parameters N-d-P-k: N is the number of coefficients from which derivatives are computed at each frame, d determines the size of analysis windows (consisting of $2 \cdot d + 1$ frames) to compute the derivatives, P is the shift (number of frames) between two consecutive analysis windows and k is the number of analysis windows whose delta coefficients are concatenated to form the final feature vector. Following previous studies [14], the SD configuration applied in this work is 13-2-3-7.

3. Experimental Setup

3.1. Phone Posterior Extraction

The open software Temporal Patterns Neural Network (TRAPs/NN) phone decoder for Hungarian, developed by the Brno University of Technology (BUT) [8], was used to obtain phone posteriors. The BUT decoder for Hungarian includes 58 phonetic and 3 non-phonetic units. The non-phonetic units were combined and treated as a single unit model. Voice activity detection was performed by removing the feature vectors whose highest PLLR value correspond to the integrated non-phonetic unit.

3.2. i-vector System

For each PLLR-based system, a gender independent 1024-mixture GMM was used as UBM, and estimated by Maximum Likelihood using the training set from the dataset. The *Total Variability Factor Analysis approach* maps high-dimensional GMM supervectors into low-dimensional vectors, or *i-vectors*. For each utterance, the GMM supervector is modeled as:

$$M = m + Tw \tag{5}$$

where m is the utterance independent mean supervector, T is the total variability matrix and w is the normally distributed low-dimensional latent vector or i-vector. The total variability matrix T was estimated as in [6], using only target languages from the training set. A generative modeling approach was applied in the i-vector feature space (as in [7]), the set of i-vectors of each language being modeled by a single Gaussian distribution. Thus, the i-vector scores were computed as follows:

$$score(f, l) = N(w_f; \mu_l, \Sigma) \quad (6)$$

where w_f is the i-vector for target signal f, μ_l is the mean i-vector for language l and Σ is a common (shared by all languages) within-class covariance matrix.

3.3. Backend and Fusion

A ZT-norm followed by a discriminative Gaussian backend was applied. The *FoCal* toolkit was used to estimate and apply the backend and calibration/fusion models [15].

3.4. Dataset

The NIST 2007 LRE [16] defined 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) the OHSU Corpus provided by NIST for the 2005 LRE; and (3) the development corpus provided by NIST for the 2007 LRE. A set of 23 languages/dialects was defined for training, including target and non-target languages (french was the only non-target language used for NIST 2007 LRE). For development purposes, 10 conversations per language were randomly selected. 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. 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.

3.5. Performance Measures

In this work, systems are compared in terms of: (1) the average cost performance C_{avg} as defined in NIST evaluations up to 2009 and (2) the Log-Likelihood Ratio Cost function C_{LLR} [17]. Equal Error Rate is also used for comparative purposes.

4. Results

4.1. Projected PLLR Features

Table 1: $C_{\text{avg}} \times 100$ and C_{LLR} performance for the Baseline , Projected PLLR and Projected PLLR + PCA approaches.

PLLR System	$\%C_{\text{avg}}$	C_{LLR}
Baseline	2.66	0.389
Projection	2.31	0.320
Projection + PCA 58	2.10	0.310

Table 2 shows a C_{avg} and C_{LLR} performance comparison between three systems: the baseline using the original PLLR features, the one based on the projected features and a third one trained on the projected features after applying PCA to decorrelate the parameters. Performance is significantly enhanced by the projection of the features, going from 2.66 to 2.31 in terms of $C_{\text{avg}} \times 100$, whereas PCA still provides a further gain in performance, reaching 2.10 $C_{\text{avg}} \times 100$, which means a 21% relative improvement overall.

4.2. Shifted Delta PLLRs

Results attained with several sets of PLLR features, reduced to different dimensionalities by means of PCA, are shown in Table 2. As expected, system performance suffers some degradation when the set of features is reduced to smaller dimensions. But the use of shifted deltas on the smallest set under a 13-2-3-7 configuration led to the best performance: 1.52 $C_{\text{avg}} \times 100$.

Table 2: $C_{\text{avg}} \times 100$ and C_{LLR} performance for the Projected PLLR + PCA, Projected PLLR + PCA reduced and Projected PLLR + PCA reduced + SD approaches.

PLLR System	$\%C_{\text{avg}}$	C_{LLR}
Projection + PCA 58	2.10	0.310
Projection + PCA 23	2.16	0.316
Projection + PCA 13	2.43	0.330
Projection + PCA 13 + SD	1.52	0.225

Figure 3 shows graphically the overall performance gain attained with the final Projected PLLR + PCA reduced + SD system: a 43% improvement with regard to the baseline approach.

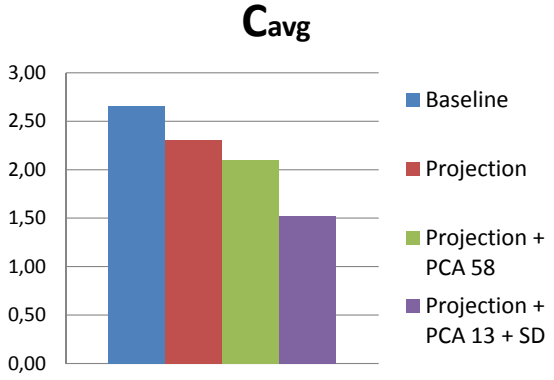


Figure 3: $C_{\text{avg}} \times 100$ performance for the Baseline (blue), Projected PLLR (red), Projected PLLR + PCA 58 (green) and Projected PLLR + PCA 13 + SD (purple) approaches.

4.3. Performance comparison

For comparative purposes, Table 3 shows some of the results published on the literature for the primary task (30 second, closed-set) of the NIST 2007 LRE. Figures show how the approach presented in this paper outperforms the single systems reported on the Table, as well as some of the pairwise fusions. In particular, the Projected HU + SD approach outperforms the acoustic i-vector (a), the phone-SVM lattice phonotactic (b) and the PLLR HU+SD (non-projected) approaches, attaining 47%, 27% and 13% relative improvements in terms of C_{avg}, respectively. The fusion of the Projected + PCA reduced + SD PLLR, acoustic (a) and phonotactic (b) systems leads (as far as we know) to the best result reported to date on this benchmark.

Table 3: Performance figures reported on the primary task of the NIST 2007 LRE.

Model			EER	$C_{avg} \times 100$
GMM-MMI [18]			–	2.10
GSV-SVM [18]			–	1.92
Discriminative GMM-MAP [19]			–	1.74
P-gram i-vectors [20]			–	3.15
Acoustic i-vectors [20]			–	2.40
Acoustic i-vectors [7]			–	1.91
Acoustic i-vectors [21]			2.59	2.61
HU, Phone-SVM, lattices [22]			1.84	–
HU, Phone-SVM, lattices [23]			2.40	–
EN, Phone-SVM, lattices [23]			1.80	–
Acoustic, i-vector [5]		(a)	2.75	2.85
HU, Phone-SVM, lattices [5]		(b)	1.95	2.08
PLLR HU+SD, i-vector [14]			1.70	1.75
Projected PLLR HU+SD, i-vector		**(c)**	**1.44**	**1.52**
Fusions	2 subsystems [7]		–	1.66
	2 acoustic subsystems [18]		–	1.55
	2 phonotactic subsystems [18]		–	1.55
	2 subsystems [20]		–	1.25
	4 subsystems [18]		0.93	0.97
	3 phonotactic subsystems [24]		–	0.90
	(a+b+c)		**0.60**	**0.75**

5. Conclusions

In this work, we have presented an in-depth study of PLLR features, aiming to show the latest progress made on PLLR post-processing. In order to avoid bounded feature distributions, the original PLLRs have been projected into a hyper-plane, enhancing the information retrieved by the system. Then, dimensionality reduction has been performed by means of PCA, and shifted deltas have been computed on the reduced features to optimize performance.

Results presented on the 2007 LRE database show that the projection method and the application of SDs on the reduced set of projected PLLR features has provided an overall 43% improvement with regard to the system trained on the original PLLR feature set. A performance comparison has also been performed with regard to results published by other authors on the same database, revealing that the developed system yields among the best performance figures reported so far on the considered benchmark.

6. Acknowledgements

This work has been supported by the University of the Basque Country under grants GIU13/28 and PES13/54.

7. References

[1] 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.

[2] N. Brümmer, S. Cumani, O.Glembek, M. Karafiát, P. Matejka, J. Pesán, O. Plchot, M. Soufifar, 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.

[3] 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.

[4] H. Wang, C.-C. Leung, T. Lee, B. Ma, and H. Li, "Shifted-Delta MLP Features for Spoken Language Recognition," *IEEE Signal Process. Lett.*, vol. 20, no. 1, pp. 15–18, 2013.

[5] M. Diez, A. Varona, M. Penagarikano, L. J. Rodriguez-Fuentes, and G. Bordel, "On the Use of Log-Likelihood Ratios as Features in Spoken Language Recognition," in *IEEE Workshop on Spoken Language Technology (SLT 2012)*, Miami, Florida, USA, December 2012.

[6] 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.

[7] 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.

[8] M. Diez, A. Varona, M. Penagarikano, L. Rodriguez-Fuentes, and G. Bordel, "On the projection of PLLRs for Unbounded Feature Distributions in Spoken Language Recognition," *IEEE Signal Processing Letters*, vol. 21, no. 9, pp. 1073–1077, Sept 2014.

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

[10] M. Diez, A. Varona, M. Penagarikano, L. J. Rodríguez-Fuentes, and G. Bordel, "Dimensionality Reduction of Phone Log-Likelihood Ratio Features for Spoken Language Recognition," in *Proceedings of Interspeech 2013*, Lyon, France, August 2013.

[11] L. D'Haro, R. Cordoba, C. Salamea, and J. Echeverry, "Extended Phone Log-Likelihood Ratio Features and Acoustic-Based i-vectors for Language Recognition," in *Proceedings of IEEE ICASSP*, Florence, Italy, May 2014.

[12] *NIST LRE*, http://www.itl.nist.gov/iad/mig/tests/lre/.

[13] M. Diez, A. Varona, M. Penagarikano, L. Rodriguez-Fuentes, and G. Bordel, "On the Complementarity of Phone Posterior Probabilities for Improved Speaker Recognition," *IEEE Signal Processing Letters*, vol. 21, no. 6, pp. 649–652, 2014.

[14] M. Diez, A. Varona, M. Penagarikano, L. J. Rodríguez-Fuentes, and G. Bordel, "Optimizing PLLR Features for Spoken Language Recognition," in *Proceedings of ICPR 2014*, Stockholm, Sweden, August 2014.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] L. D'Haro, O.Glembek, O. Plchot, P. Matejka, M. Soufifar, R. Cordoba, and J. Cernocký, "Phonotactic Language Recognition using i-vectors and Phoneme Posteriogram Counts," in *Proceedings of the Interspeech 2012*, Portland, Oregon, September 9-13 2012.

[21] M. Li and N. S., "Simplified supervised i-vector modeling with application to robust and efficient language identification and speaker verification," *Computer Speech and Language*, 2014.

[22] 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.

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

[24] 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.