EXTENDED PHONE LOG-LIKELYHOOD RATIO FEATURES AND ACOUSTIC-BASED I-VECTORS FOR LANGUAGE RECOGNITION

L. F. D’Haro¹, R. Cordoba¹, C. Salamea¹,², J. D. Echeverry¹

² Universidad Politécnica Salesiana del Ecuador Calle Vieja 12-30 y Elia Liut, Casilla 26, Cuenca, Ecuador
{lfdharo,cordoba, csalamea,jdec}@die.upm.es

ABSTRACT
This paper presents new techniques with relevant improvements added to the primary system presented by our group to the Albayzin 2012 LRE competition, where the use of any additional corpora for training or optimizing the models was forbidden. In this work, we present the incorporation of an additional phonotactic subsystem based on the use of phone log-likelihood ratio features (PLLR) extracted from different phonotactic recognizers that contributes to improve the accuracy of the system in a 21.4% in terms of Cavg (we also present results for the official metric during the evaluation, Fact). We will present how using these features at the phone state level provides significant improvements, when used together with dimensionality reduction techniques, especially PCA. We have also experimented with applying alternative SDC-like configurations on these PLLR features with additional improvements. Also, we will describe some modifications to the MFCC-based acoustic i-vector system which have also contributed to additional improvements. The final fused system outperformed the baseline in 27.4% in Cavg.

Index Terms—Phone Log-Likelihood Ratios, SDC, dimensionality reduction.

1 INTRODUCTION
In this paper, we describe several modifications we have done to the final system that we presented for the Albayzin 2012 LRE evaluation [1]. In this evaluation, our primary system outperformed the other systems thanks mainly to the fusion of different subsystems: 1) an acoustic system based on MFCC-SDC features, 2) a phonotactic system based on trigram posteriorgram counts, and 3) an acoustic system based on RPLP-SDC features. As most of current state-of-the-art LID systems, all these subsystems make use of subspace projections in the form of i-vectors [2] that were calibrated and fused using multiclass logistic regression. In [3] it was shown that one of the main advantages of our system was the use of RPLP (Revised PLP) features which allowed the incorporation of noise-robust features, and the use of a phonotactic i-vector based system that uses non-sparse n-gram counts estimated using the posterior probabilities output of a phoneme recognizer and trained using subspace multinomial models [4]. Finally, our best system is based on the fusion of the scores of four different subsystems allowing the integration of various levels of perceptual cues as it is recommended in [5].

In this paper, we will describe new enhancements done to the final system presented in the Albayzin evaluation. The main change has been the incorporation of a new kind of phonotactic subsystem that uses Phone Log-Likelihood Ratios features (PLLR) that have proved to improve both language [6] and speaker recognition systems [7]. Later, we will show how these features can be extended to provide a better performance thanks to the use of likelihood ratios at a phone state level, instead of a phone level, the addition of new coefficients based on the Shifted Delta Cepstra (SDC) philosophy [8], which we called Shifted Delta PLLR Coefficients (SDPC) and the use of PCA and HLDA dimensionality reduction techniques.

The paper is organized as follows: Section 2 describes the database used for the evaluation. Section 3 explains each subsystem while section 4 shows the fusion results. Finally, section 5 presents the conclusions and future work.

2 EVALUATION AND DATA DESCRIPTION
The Albayzin LRE 2012 was an international evaluation organized by the Software Technologies Working Group of the University of the Basque Country with the collaboration of Niko Brümmer from Agnitio Research, in the context of the IberSpeech 2012 conference [9]. In comparison with its previous editions, this evaluation was more difficult as it changed the application domain from TV broadcast speech to any kind of speech found on Internet, without providing training data for some of the target languages (a common situation for low-resource languages) in two of the four conditions, and forbidding the use of any additional database. The provided audio files were extracted from
music, noise and any kind of non-human sounds. All audio conditions, number of speakers, etc. The files might contain YouTube videos, with different length durations, channel conditions, number of speakers, etc. The files might contain music, noise and any kind of non-human sounds. All audio files used in our experiments were 16 KHz@16 bits in contrast with [3] where we used 8KHz@16 bits.

Table 1 shows the statistics of the database and the number of files used in our setup and experiments. We will show results only on the main condition of the evaluation, i.e. plenty-closed, where the target languages were: Spanish, Catalan, Basque, Galician, Portuguese, and English.

<table>
<thead>
<tr>
<th>No. Files</th>
<th>Train</th>
<th>Dev</th>
<th>Eval</th>
</tr>
</thead>
<tbody>
<tr>
<td>5115</td>
<td>458</td>
<td>941</td>
<td></td>
</tr>
<tr>
<td>3231</td>
<td>252</td>
<td>409</td>
<td></td>
</tr>
<tr>
<td>1884</td>
<td>206</td>
<td>532</td>
<td></td>
</tr>
</tbody>
</table>

Table 1. Dataset statistics

### 3 SYSTEM DESCRIPTION

In this section, we will describe each of the subsystems that we used and fused for creating the final systems.

#### 3.1 MFCC-SDC Acoustic System + i-vectors

For this subsystem, for each audio file we extract 12 MFCC coefficients (including C0) from 24 Mel filter banks plus the energy for each frame. As Voice Activity Detector (VAD) we used the output from the BUT Hungarian phone recognizer suppressing all segments marked as silence or noise in the output. Then, a RASTA filter was applied to reduce short-term noise variations in each frequency sub-band followed by a short-term Cepstral Mean and Variance Normalization (CMVN) normalization (instead of using global CMVN as [3]). After that, every 10 ms speech frame was mapped to a 56-dimensional feature vector generated from the concatenation of SDC features using the 7-1-3-7 configuration. Then, in comparison with [3], we included feature warping [10] after removing the non-speech frames using the toolkit available at [11]. Finally, i-vectors of 400 dimensions and using 512 Gaussians were trained following the same algorithm reported in [2], which is the optimum configuration. With respect to the same subsystem reported in [3], the use of the short-term CMVN and feature warping allowed us to improve the C$_{avg}$ and F$_{si}$ in 10.4% and 7.2% relative respectively.

#### 3.2 RPLP-SDC Acoustic System + i-vectors

Proposed in [12] and [13], the Revised PLP (RPLP) features can be seen as a hybrid approach between calculating MFCC and PLP features, combining the best of both and providing as result noise-robust features. In [3] we showed that these features highly contributed to improve the final system and performed better than the MFCC subsystem in spite of using the same configuration, i.e. number of Gaussians, i-vector dimension, SDC, etc.

In this work, the same modifications applied to the MFCC subsystem (i.e. the use of short-term CMNV and Feature Warping) did not provide any improvement. Therefore we kept the same subsystem reported in [3].

#### 3.3 Phone Log-Likelihood Ratio (PLLR) Features

In [6] and [7] it is shown that the PLLR features can be successfully used for language and speaker recognition tasks. Its success is probably due to the simplicity of its calculation and because they can be easily integrated with the i-vector framework where the PLLR can be seen as an alternative to the acoustic MFCC-SDC features. On the other hand, as proved in [7], other alternative features as frame-level posteriors or phone log-posteriors (which are usually provided by phone recognizers) are not suitable for tasks where the features are assumed to be Gaussian-distributed. In contrast, the transformation from log posteriors into log-likelihood ratios (LLR) provides final distributions that are nearly Gaussian. In order to calculate the PLLR features [14], the acoustic posterior probability of a phone unit $m$ at each frame $f$, is calculated by summing up the posteriors of its corresponding states:

$$p(m|f) = \sum_{\forall \varepsilon} p(m|s, f) p(s|f)$$

Then, the log-likelihood ratios at each frame $f$ can be computed from posterior probabilities using equation (2) where it is assumed a classification task with flat priors.

$$LLR^n_f = log \frac{1}{M-1} \sum_{m=1}^{M} \frac{p(x_f|m)}{p(x_f|m)}$$

Finally, the resulting M log-likelihood ratios per frame are stacked together to create the Phone Log-Likelihood Ratio (PLLR) features. For our system, these features were created using the open-source toolkit available in [15]. After that, an i-vector system similar to the one described in sections 3.1 and 3.2 was trained on the PLLR features.

#### 3.3.1 Phone Recognizers

As explained above, in order to calculate the PLLR features we used the Hungarian, Czech, and Russian phone decoders developed by the Brno University of Technology (BUT) [16]. These phone decoders use a three-state model per phone, which means that three posterior probabilities per unit are given at each frame. Since these posterior probabilities are encoded by default, we applied some simple mathematical formulas to decode them (see section 4.2 in [7] for further details).
3.3.2 Baseline PLLR features

Following the same approach mentioned in [14], before computing the PLLR features, the three non-phonetic units of the BUT phone recognizers, i.e., int, pau, and spk, are fused into a single non-phonetic unit. Then, a unified posterior probability is computed for each phone model by adding the posterior probabilities of all the states in the corresponding phone model (Eq. 1). Finally, the log-likelihood ratios were computed using eq. 2. In this way, for the Hungarian phone recognizer we have 59 PLLR features, 50 for Russian, and 43 for Czech. As in [14], the use of first order deltas provided us a relative improvement of 3.4% in \( C_{\text{avg}} \) and the use of different kind of phone mappings did not provide improvements for any of the phone recognizers. Therefore, our baselines are given using the complete phone set for all the recognizers, including the delta features, using i-vectors of 400 dimensions, and UBMs with 512 Gaussians.

3.3.3 Modification using States

The first modification we tried over the baseline PLLR features was to use the likelihood ratio of each individual state as a feature instead of summing up the posterior probabilities of the corresponding phone-states (Eq. 1). The motivation was to take advantage of the information encoded in the transitions between phones as well as between states which also provides discriminative information between languages. The caveat is the dimensionality problem: since each phone has three states per phone, the final PLLR vector for each frame is of dimension 177 for the Hungarian phone recognizer, of 129 for the Czech, and 150 for the Russian decoder. We dealt with this problem using dimensionality reduction techniques as we will see in the next section.

3.3.4 Dimensionality reduction techniques

Following the results reported in [14] and [17], where the accuracy of a LID system was improved thanks to the dimensionality reduction of the PLLR features using PCA, for our experiments we also tested different dimensionality reduction techniques such as HLDA [18]. In this case, the dimensionality reduction was applied for the baseline PLLR features as well as for the state-based PLLR features.

In Table 2 we can see the results. We show the \( C_{\text{avg}} \) and the \( F_{\text{act}} \) values with relative improvements over the baseline in parenthesis. The number in PCA/HLDA means the optimum dimension. The conclusion is that state-PLLR provides significant improvements over phone-PLLR in all cases (between 9.6% and 14.3% relative in \( C_{\text{avg}} \)). Also, PCA is better than HLDA except in one case.

3.3.5 Modification using SDPC parameters

In [19] it is shown that the use of stacked coefficients are useful in the context of having noisy files and using similar phone log likelihood ratios as the ones used in this work. Here, stacked frames created from borrowing concepts from the SDC coefficients helped to compensate the potential drawbacks resulting from using the short-term phone information from the PLLR features, since it is possible to capture longer-term statistics. We apply the windowing concepts from SDC to the PLLR features, obtaining what we call Shifted Delta PLLR Coefficients (SDPC) and then we apply a PCA projection as in [17] because in this case dimensionality reduction is a must with the high dimensionality vectors that we have to manage (for instance, 177 states in the Hungarian recognizer with a SDC 1_5_3 will result in a vector of dimension 708). We compared using first the PCA reduction and then stacking the SDPC or first stacking the SDPC and then applying PCA. The last option provided consistent worse results and was discarded.

We also experimented different configurations for the SDPC parameters (i.e. N-d-P-K) and PCA dimension (the \( N \) in our case) in order to capture long-term information. The best result so far was obtained for the configuration PCA-30 and 1-5-3 for the other SDPC parameters, which captures 200 ms. On the other hand, we obtained very similar results with other configurations to have a longer range. In this case, we think that the increase in dimensionality is the reason of this best result with \( K=3 \).

Table 3 shows our best results. The relative improvements in parenthesis are comparisons with the best results without using SDPC from Table 2.

<table>
<thead>
<tr>
<th>Phone + PCA</th>
<th>Hungarian</th>
<th>Czech</th>
<th>Russian</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phone</td>
<td>6.74 (15.6)</td>
<td>14.15 (11.0)</td>
<td>7.86 (6.4)</td>
</tr>
<tr>
<td>HDLA</td>
<td>7.11 (15.4)</td>
<td>14.90 (8.7)</td>
<td></td>
</tr>
</tbody>
</table>

Table 3. PLLR results using SDPC parameters

In summary, we obtain the best results using the state-based approach and SDPC provides improvements in all cases. In comparison with the baseline, the relative improvement in \( C_{\text{avg}} \) for the three recognizers is 26.8%-23.8%-30.1% respectively.
### 3.4 Classifiers and calibration back-end

As classifier for all our subsystems, we used a Multiclass logistic regression, and for calibration and fusion, a Gaussian Back-end followed by a Discriminative MultiClass Logistic Regression. Previously, the input i-vectors were conditioned by within-class covariance normalization (WCCN, [20]) and length normalized. In Table 4 we can see the best results for the individual subsystems. The Hungarian PLLR subsystem is the best one, even better than the best acoustic (5.5% relative).

## 4 FUSION RESULTS

In Table 5 we can see the results for fusing all our systems. All improvements from now on will be relative in $C_{avg}$. The first two lines are our baselines, i.e. the systems presented in [3] using the acoustic modules (but using the original 16 kHz audio files) and our previous phonotactic system (see [3] for more details). System 1 uses the acoustic modules, but with feature warping and local CMVN in the MFCC subsystem with an improvement of 6.7% over the baseline.

System 2 adds the phonotactic system with an improvement of 7.6% over the baseline and of 20.3% from 1 to 3. In System 4, where all three PLLR modules are included, we can observe that the result is even better than the two acoustic modules together, 11.7% improvement over System 1, which is quite relevant. In System 5, the combination of the phonotactic and the PLLR modules provides an additional improvement of 7.0% over System 4. In System 6 and System 7, there is an improvement of 21.4% due to the PLLR subsystems.

### 5 CONCLUSIONS AND FUTURE WORK

In this paper we have described different improvements to a language recognition system. The main change is the incorporation of a phonotactic system based on the use of state-based phon log likelihood ratios and PCA as dimensionality reduction technique. The relative improvement over the baseline with no PCA and phone-based PLLR is between 23.8% and 30.1% thanks to the use of the state-based approach and SDPC parameters. With PCA the state-based approach improves the phone-based between 9.6% and 14.3% relative. Besides, the inclusion of SDPC coefficients after the PCA projection provided additional improvements as we showed in Table 3, being the result that one PLLR subsystem is better than the best acoustic one.

On the other hand, feature normalizations (i.e. short-term CMVN and feature warping) to the MFCC system contributed to improve this acoustic subsystem in 6.7%.

The fusion of all the subsystems has shown many interesting conclusions described in Section 4, which can be summarized in that the PLLR modules have contributed to an improvement of 21.4% in $C_{avg}$ to the final system.

As future work, we will investigate new techniques to reduce the redundancy of information between adjacent frames when using the PLLR-based features. We also want to reduce the total number of states by merging the less frequent phones, especially for the Hungarian recognizer.

## 6 ACKNOWLEDGMENT

This work has been supported by TIMPANO (TIN2011-28169-C05-03), INAPRA (MICINN, DPI2010-21247-C02-02), MA2VIMCR (Comunidad Autónoma de Madrid, S2009/TIC-1542) projects, and the European project Simple4All (under grant agreement No. 287678).
REFERENCES


[15]. PLLLR computation software. [Online]. Freely available at: https://sites.google.com/site/gttspllrfeatures/home


