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

L. F. D'Haro[1], R. Cordoba[1], C. Salamea[1,2], J. D. Echeverry[1]

[1] Speech Technology Group – Dpto. de Ing. Electrónica – E.T.S.I. Telecomunicación - Universidad Politécnica de Madrid. Ciudad Universitaria s/n 28040, Madrid, Spain

[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 C_{avg} (we also present results for the official metric during the evaluation, F_{act}). 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 C_{avg}.

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 sub-space 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 sub-systems 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

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

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.

	Train	Dev	Eval
No. Files	5115	458	941
No. of clean files	3231	252	409
No. of noisy files	1884	206	532

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 (CMNV) 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_{act} 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:

$$p(m|f) = \sum_{\forall s} p(m|s,f)\, p(m|f) \qquad (1)$$

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.

$$LLR_f^m = log \frac{p(x_f|m)}{\frac{1}{M-1}\sum_{\forall n \neq m} p(x_f|n)} \quad m = 1, \dots, M \qquad (2)$$

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_{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 posteriors 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.

	Hungarian		Czech		Russian	
	C_{avg} (Im)	**F_{act} (Im)**	**C_{avg} (Im)**	**F_{act} (Im)**	**C_{avg} (Im)**	**F_{act} (Im)**
Baseline	8.97	17.64	9.62	18.19	10.02	18.60
Phone-PCA 25	7.98 (11.0)	15.89 (9.9)	8.40 (12.7)	16.56 (9.0)	8.40 (16.2)	16.32 (12.3)
Phone HLDA 25	8.38 (6.6)	16.36 (7.3)	8.41 (12.6)	16.88 (7.2)	8.02 (20.0)	16.50 (11.3)
States PCA 60	**6.95** (22.5)	**14.39** (18.4)	**7.59** (21.1)	**15.17** (16.6)	**7.20** (28.1)	**14.84** (20.2)

Table 2. PLLR results using different dimensionality reduction techniques and comparing with the use of state-phones.

In Table 2 we can see the results. We show the C_{avg} and the F_{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_{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.

	Hungarian		Czech		Russian	
	C_{avg} (Im)	**F_{act} (Im)**	**C_{avg} (Im)**	**F_{act} (Im)**	**C_{avg} (Im)**	**F_{act} (Im)**
Phone + PCA 30 + SDPC 1_5_5	**6.74** (15.6)	**14.15** (11.0)	7.86 (6.4)	15.16 (8.5)	7.11 (15.4)	14.90 (8.7)
States + PCA 35 + SDPC 1_5_3	**6.57** (5.5)	**13.97** (2.9)	7.33 (2.8)	14.70 (0.6)	7.00 (3.4)	14.75 (3.1)

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_{avg} for the three recognizers is 26.8%-23.8%-30.1% respectively.

3.4 Classifier 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 Multi-Class 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 10.9% compared to system 1. System 3 shows that using just one PLLR system instead of the phonotactic gives a clear improvement (10.5% from 2 to 3, 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 are additional improvements of 24.5% and 27.4% over Baseline 2. Also, comparing 2 and 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 phone 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 CMNV 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), MA2VICMR (Comunidad Autónoma de Madrid, S2009/TIC-1542) projects, and the European project Sim-ple4All (under grant agreement No. 287678).

Subsystem Type	Configuration	Dev		Eval	
		C_{avg}(%)	F_{act}	C_{avg} (%)	F_{act}
MFCC-SDC	400iv, 512 Gauss.	6.50	12.24	6.95	14.68
RPLP-SDC	400iv, 512 Gauss	6.54	12.34	7.36	14.73
Phonotactic	400iv	6.94	13.37	9.85	18.14
PLLR_Hung	States + PCA-35 + SDPC 1_5_3	**4.85**	**10.88**	**6.57**	**13.97**
PLLR_Czec	States + PCA-35 + SDPC 1_5_3	5.23	11.18	7.33	14.70
PLLR_Russ	States + PCA-35 + SDPC 1_5_3	5.67	11.52	7.00	14.75

Table 4. Best results for each subsystem on the dev and evaluation sets

Fusion	MFCC	RPLP	Phono Hung	PLLR Hung	PLLR Russ	PLLR Czec	Dev		Eval	
							C_{avg}	F_{act}	C_{avg}	F_{act}
Baseline 1	X	X					4.83	6.09	5.39	7.77
Baseline 2	X	X	X				2.88	3.89	4.85	6.48
System 1	X	X					3.87	5.21	5.03	7.04
System 2	X	X	X				2.78	3.54	4.48	5.88
System 3	X	X		X			2.78	3.17	4.01	5.43
System 4				X	X	X	2.88	3.50	4.44	6.41
System 5			X	X	X	X	2.36	3.23	4.13	6.07
System 6	X	X		X	X	X	**2.15**	2.83	3.66	5.29
System 7	X	X	X	X	X	X	2.25	**2.64**	**3.52**	**5.18**

Table 5. Fusion results for the different subsystems on the dev and evaluation sets

REFERENCES

[1]. L. F. D'Haro, R. Córdoba. 2012. *"The GTH-LID System for the Albayzin LRE12 Evaluation"*. In Proc. Iberspeech 2012, pp. 528-539.

[2]. N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, 2011. *"Front-End Factor Analysis For Speaker Verification"*, IEEE Transactions on Audio, Speech and Language Processing, Vol. 19 (4), May 2011.

[3]. L. F. D'Haro, R. Cordoba, M. A. Caraballo, J. M. Pardo. 2013. *"Low-Resource Language Recognition using a Fusion of Phoneme Posteriorgram Counts, Acoustic and Glottal-based I-Vectors"*. ICASSP 2013, Vancouver, Canada. May 26-31, 2013.

[4]. L. F. D'Haro, O. Glembek, O. Plchot, P. Matejka, M. Soufifar, R. Cordoba, J. Cernocký. *"Phonotactic Language Recognition using i-vectors and Phoneme Posteriorgram Counts",* in Proc. of Interspeech 2012.

[5]. H. Li; B. Ma; K A. Lee, *"Spoken Language Recognition: From Fundamentals to Practice,"* Proceedings of the IEEE , vol.101, no.5, pp.1136,1159, May 2013

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

[7]. M. Diez, Luis J. Rodriguez-Fuentes, M. Penagarikano, A. Varona, G. Bordel, 2013. *"Using Phone Log-Likelihood Ratios as Features for Speaker Recognition"*, Interspeech 2013; Lyon, France.

[8]. P.A. Torres-Carrasquillo, E. Singer, M.A. Kohler, R.J. Greene, D.A. Reynolds, and J.R. Deller Jr., 2002. *"Approaches to language identification using Gaussian mixture models and shifted delta cepstral feature,"*, in Proc. International Conferences on Spoken Language Processing (ICSLP), Sept. 2002, pp. 89–92.

[9]. L. J. Rodriguez-Fuentes, N. Brümmer, M. Penagarikano, A. Varona, G. Bordel, M. Diez, 2013. *"The Albayzin 2012 Language Recognition Evaluation"*, Interspeech 2013; Lyon, France.

[10]. J. Pelecanos and S. Sridharan, 2001. *"Feature warping for robust speaker verification",* in Proc. ISCA Odyssey, Crete, Greece, Jun. 2001.

[11]. S. O. Sadjadi, M. Slaney, and L. Heck. 2013. *"MSR Identity Toolbox vs 1.0"*, Microsoft Research. [Online] Available at http://research.microsoft.com/

[12]. J. Rajnoha, and P. Pollák. 2011. *"ASR systems in Noisy Environment: Analysis and Solutions for Increasing Noise Robustness"*. Radio engineering, Vol. 20, No. 1, April 2011, pp. 74-84.

[13]. F. Hönig, G. Stemmer, C. Hacker, and F. Brugnara. 2005. *"Revising Perceptual Linear Prediction (PLP)"*. In Eurospeech 2005, p. 2997-3000.

[14]. M. Diez, Luis J. Rodriguez-Fuentes, M. Penagarikano, A. Varona, G. Bordel, 2013. *"Dimensionality Reduction of Phone Log-Likelihood Ratio Features for Spoken Language Recognition"*, Interspeech 2013; Lyon, France, 25-29 aug., 2013;

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

[16]. P. Schwarz, 2009. *"Phoneme Recognition based on Long Temporal Context, PhD Thesis"*, Brno University of Technology, 2009. [Online] Available at http://speech.fit.vutbr.cz/software/phoneme-recognizer-based-long-temporal-context

[17]. H. Wang, C.-C. Leung, T. Lee, B. Ma, and H. Li, 2013. *"Shifted-Delta MLP Features for Spoken Language Recognition"*, IEEE Signal Processing Letters, Vol. 20, No. 1, January 2013.

[18]. L. Burget. 2004. *"Combination of Speech Features Using Smoothed Heteroscedastic Linear Discriminant Analysis"*, Proc. 8th International Conference on Spoken Language Processing, pp. 2549-2552.

[19]. K. Han, and J.; Pelecanos, J., 2012. *"Frame-based phonotactic Language Identification"*, IEEE Spoken Language Technology Workshop (SLT), pp. 303-306.

[20]. A. Hatch and A. Stolcke, 2006. *"Generalized linear kernels for one-versus-all classification: application to speaker recognition"*, in Proc. ICASSP Vol 5, page V.