Procesamiento del Lenguaje Natural, Revista nº 51, septiembre de 2013, pp 153-160 recibido 30-04-2013 revisado 16-06-2013 aceptado 21-06-2013

Language Recognition on Albayzin 2010 LRE using PLLR features

Reconocimiento de la Lengua en Albayzin 2010 LRE utilizando características PLLR

M. Diez, A. Varona, M. Penagarikano,
L.J. Rodriguez-Fuentes, G. Bordel
University of the Basque Country, UPV/EHU
GTTS, Department of Electricity and Electronics
amparo.varona@ehu.es

Resumen: Los así denominados Phone Log-Likelihood Ratios (PLLR), han sido introducidos como características alternativas a los MFCC-SDC para sistemas de Reconocimiento de la Lengua (RL) mediante iVectors. En este artículo, tras una breve descripción de estas características, se proporcionan nuevas evidencias de su utilidad para tareas de RL, con un nuevo conjunto de experimentos sobre la base de datos Albayzin 2010 LRE, que contiene habla multi-locutor de banda ancha en seis lenguas diferentes: euskera, catalán, gallego, español, portugués e inglés. Los sistemas de iVectors entrenados con PLLRs obtienen mejoras relativas significativas respecto a los sistemas fonotácticos y sistemas de iVectors entrenados con características MFCC-SDC, tanto en condiciones de habla limpia como con habla ruidosa. Las fusiones de los sistemas PLLR con los sistemas fonotácticos y/o sistemas basados en MFCC-SDC proporcionan mejoras adicionales en el rendimiento, lo que revela que las características PLLR aportan información complementaria en ambos casos.

Palabras clave: Reconocimiento de la Lengua, Phone Log-Likelihood Ratios, iVectors

Abstract: Phone Log-Likelihood Ratios (PLLR) have been recently proposed as alternative features to MFCC-SDC for iVector Spoken Language Recognition (SLR). In this paper, PLLR features are first described, and then further evidence of their usefulness for SLR tasks is provided, with a new set of experiments on the Albayzin 2010 LRE dataset, which features wide-band multi speaker TV broadcast speech on six languages: Basque, Catalan, Galician, Spanish, Portuguese and English. iVector systems built using PLLR features, computed by means of three open-source phone decoders, achieved significant relative improvements with regard to the phonotactic and MFCC-SDC iVector systems in both clean and noisy speech conditions. Fusions of PLLR systems with the phonotactic and/or the MFCC-SDC iVector systems led to improved performance, revealing that PLLR features provide complementary information in both cases.

Keywords: Spoken Language Recognition, Phone Log-Likelihood Ratios, iVectors

1. Introduction

In the last years, two complementary types of Spoken Language Recognition (SLR) systems prevail: (1) those using *low-level* (typically, short-term spectral) features; and (2) those using *high-level* (typically, phonotactic) features. Among the first type of systems, the so called Total Variability Factor Analysis approach (also known as *iVector* approach) has been recently introduced, using Mel-Frequency Cepstral Coefficients and Shifted Delta Cepstra (MFCC-SDC) features (Dehak et al., 2011b). The iVector approach maps high-dimensional input data, typically a Gaussian Mixture Model (GMM) supervector, to a low-dimensional feature vector (an iVector), hypothetically retaining most of the relevant information.

Due to its high performance and low complexity, the iVector approach has become a state-of-the-art technique. Besides MFCC-SDC, other alternative features have been already tested under this approach, such as prosodic features (pitch, energy and dura-

ISSN 1135-5948

tion) (Martínez et al., 2012) or speaker vectors from subspace GMM (Plchot et al., 2012). It has been reported that these systems alone do not yield outstanding results, but performance improves significantly when fusing them with a system based on spectral features.

Among *high-level* approaches, best results are reported for the so called *Phone-Lattice-SVM* approach (Campbell, Richardson, and Reynolds, 2007), which uses expected counts of phone n-grams (computed on phone lattices provided by phone decoders) as features to feed a Support Vector Machine (SVM) classifier.

There have been some efforts to use phonotactic features under the iVector approach. In (Soufifar et al., 2012), expected counts of phone n-grams are used as features, reaching the same performance as state-of-the-art phonotactic systems. In (DHaro et al., 2012), phone posteriorgrams (instead of phone lattices) are used to estimate n-gram counts, and the iVector approach is then applied to reduce the high-dimensionality of the resulting feature vectors. Both approaches yield reasonable good results, and the latter is reported to fuse well with a SLR system based on short-term spectral features.

Best results are usually obtained by fusing several acoustic and phonotactic systems. Increasingly sophisticated fusion and calibration techniques have been applied, including generative Gaussian backends (Singer et al., 2003; BenZeghiba, Gauvain, and Lamel, September 2009) and discriminative logistic regression (Brümmer and van Leeuwen, 2006; Brümmer and de Villiers, 2011; Penagarikano et al., 2012).

The development of SLR technology has been largely supported by NIST Language Recognition Evaluations (LRE) (NIST LRE, 2011), held in 1996 and every two years since 2003. As a result, the datasets produced and distributed for such evaluations have become standard benchmarks to test the usefulness of new approaches. NIST LRE datasets consist mostly of narrow-band (8 kHz) conversational telephone speech.

Aiming to fill the gap of SLR technology assessment for wide-band broadcast speech, the Albayzin LREs have been organized (Rodriguez-Fuentes et al., 2010; Rodriguez-Fuentes et al., 2011), with the support of the Spanish Thematic Network on Speech Technologies (RTTH, 2006) and the ISCA Special Interest Group on Iberian Languages (SIG-IL). For the Albayzin 2008 LRE, the four official languages spoken in Spain: Basque, Catalan, Galician and Spanish, were used as target languages. In (Varona et al., 2010) an in depth study was carried out, the main verification system being obtained from the fusion of an acoustic system and 6 phonotactic subsystems.

The set of Iberian languages was completed in the Albayzin 2010 LRE by adding Portuguese as target language. Due to its international relevance and its pervasiveness in broadcast news, English was also added as target language in the Albayzin 2010 LRE. A new condition was introduced, depending on the presence of background noise, music and/or conversations (overlapped speech), leading to two additional tracks which involved clean speech and a mix of clean and noisy speech, respectively.

In a previous work (Diez et al., 2012), we proposed and evaluated the use of log-likelihood ratios of phone posterior probabilities, hereafter called Phone Log-Likelihood Ratios (PLLR), as alternative features to MFCC-SDC under the iVector approach. We found very promising results in language recognition experiments on the NIST 2007 and 2009 LRE datasets.

In this paper, a more detailed study of the PLLR features is undertaken. A new set of experiments has been carried out on the Albayzin 2010 LRE dataset (Rodriguez-Fuentes et al., 2012) to prove their effectiveness. Three iVector systems have been built using three open-source phone decoders to compute the PLLR features. These systems are compared to (and fused with) various state-of-the-art baseline systems, namely: (1) an acoustic iVector system using MFCC-SDC as features; and (2) three Phone-Lattice-SVM systems built on the same decoders used to compute the PLLR features.

The rest of the paper is organized as follows. Section 2 provides some background and describes the computation of the phone log-likelihood ratios used as features in this work. Section 3 describes the experimental setup. Section 4 presents results and compares the performance of the proposed approach to that of state-of-the-art approaches. Finally, conclusions are given in Section 5.

2. Phone Log-Likelihood Ratio (PLLR) features

In (Biadsy, Hirschberg, and Ellis, 2011), a new dialect recognition approach mixing acoustic and phonetic information was presented, based on the assumption that certain phones are realized in different ways across dialects. Acoustic models were trained for different phonetic categories, based on the phonetic segmentation provided by a phone decoder. Scores were computed based on differences between acoustic models corresponding to the same phonetic category in different dialects.

That work encouraged us to search for similar but more sophisticated approaches. After exploring the possibility of using phone posteriors at the frame level to smooth the phonetic segmentation, we came to the idea of using phone posteriors alone as features. The non-Gaussian distribution of phone posteriors was addressed by transforming phone posteriors into phone log-likelihood ratios, which carry the same information but show approximately Gaussian distributions, as illustrated in Figure 1. Under this configuration, phone models perform as a sort of reference system and phone log-likelihood ratios at a given frame can be interpreted as the *location* of the speech segment being analyzed in the space defined by those models.

To compute the PLLRs, let us consider a phone decoder including N phone units, each of them represented typically by means of a model of S states. Given an input sequence of acoustic observations X, we assume that the acoustic posterior probability of each state s $(1 \leq s \leq S)$ of each phone model i $(1 \leq i \leq N)$ at each frame t, $p_{i,s}(t)$, is output as side information by the phone decoder. Then, the acoustic posterior probability of a phone unit i at each frame t can be computed by adding the posteriors of its states:

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

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

$$LLR_i(t) = \log \frac{p_i(t)}{\frac{1}{(N-1)} \sum_{\forall j \neq i} p_j(t)} \qquad i = 1, ..., N \qquad (2)$$

The resulting N log-likelihood ratios per frame are the PLLR features considered in our approach.

3. Experimental setup

3.1. PLLR iVector system

As a first step to get the PLLR features, we applied 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) (Schwarz, 2008), which include 42, 58 and 49 phonetic units, respectively, plus 3 non-phonetic units. Note that BUT decoders represent each phonetic unit by a three-state model and output the transformed posterior probabilities $p_{i,s}(t)$ (Diez et al., 2012) as side information, for each state s of each phone model i at each frame t.

Before computing PLLR features, the three non-phonetic units —*int* (intermittent noise), *pau* (short pause) and *spk* (non-speech speaker noise)— were integrated into a single 9-state non-phonetic unit model. Then, a single posterior probability was computed for each phone i $(1 \leq i \leq N)$, according to Equation 1. Finally, the log-likelihood ratio for each phone i was computed according to Equation 2. In this way, we get 43, 59 and 50 PLLR features per frame using the BUT decoders for Czech, Hungarian and Russian, respectively.

As shown in (Diez et al., 2012), adding first order dynamic coefficients improved significantly the performance of the PLLR-based iVector system. Therefore, PLLR+Δ were used as features also in this work. Voice activity detection was performed by removing the feature vectors whose highest PLLR value corresponded to the integrated non-phonetic unit. A gender independent 1024-mixture GMM (Universal Background Model, UBM) was estimated by Maximum Likelihood using the NIST 2011 LRE training set. The total variability matrix (on which the iVector approach relies) was estimated as in (Dehak et al., 2011a), using only target languages in the NIST 2011 LRE training set. A generative modeling approach was applied in the iVector feature space (as in (Martínez et al., 2011)), the set of iVectors of each language being modeled by a single Gaussian distribution. Thus, the iVector scores were compu-

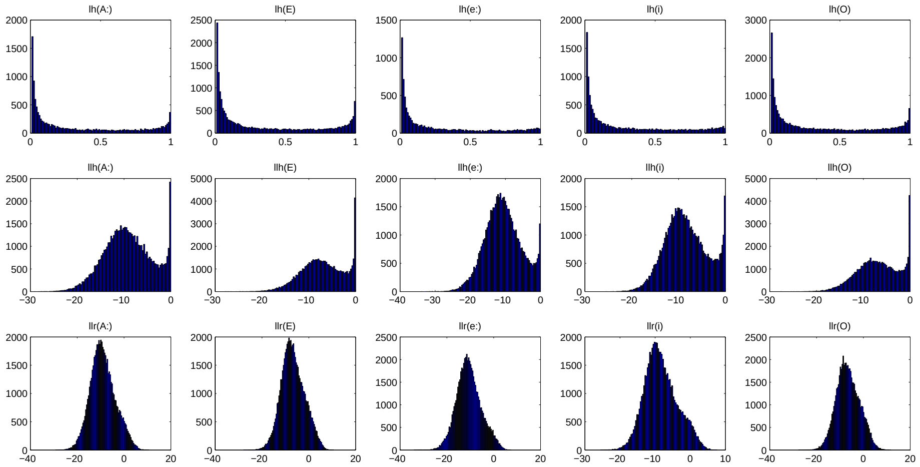

Figure 1: Distributions of frame-level likelihoods (*lh*, first row), log-likelihoods (*llh*, second row) and log-likelihood ratios (*llr*, third row) for five Hungarian phones (A:, E, e:, i and O).

ted as follows:

$$score(f, l) = N(w_f; \mu_l, \Sigma) \quad (3)$$

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

3.2. MFCC-SDC iVector system

In this case, the concatenation of MFCC and SDC coefficients under a 7-2-3-7 configuration was used as acoustic representation. Voice activity detection, GMM estimation and total variability matrix training and scoring were performed as in the PLLR iVector approach.

3.3. Phonotactic systems

The three phonotactic systems used in this work have been developed under the phone-lattice-SVM approach (Campbell, Richardson, and Reynolds, 2007) (Penagarikano et al., 2011). Given an input signal, an energy-based voice activity detector was applied in first place, which split and removed long-duration non-speech segments. Then, the BUT TRAPs/NN phone decoders for Czech, Hungarian and Russian (Schwarz, 2008) were applied. Regarding channel compensation, noise reduction, etc. the three systems relied on the acoustic front-end provided by BUT decoders.

Phone posteriors output by BUT decoders were converted to phone lattices by means of HTK (Young et al., 2006) along with the BUT recipe (Schwarz, 2008). Then, expected counts of phone n-grams were computed using the *lattice-tool* of SRILM (Stolcke, 2002). Finally, a SVM classifier was applied, SVM vectors consisting of expected frequencies of phone n-grams (up to $n = 3$). A sparse representation was used, which involved only the most frequent features according to a greedy feature selection algorithm (Penagarikano et al., 2011). L2-regularized L1-loss support vector regression was applied, by means of LIBLINEAR (Fan et al., 2008).

3.4. Dataset

The Albayzin 2010 LRE dataset (KALAKA-2) contains wide-band 16 kHz TV broadcast speech signals for six target languages. The Albayzin 2010 LRE (Rodriguez-Fuentes et al., 2011) 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 defined 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

	Clean Speech			Noisy Speech		
	Hours	# 30s segments		Hours	# 30s segments	
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

Table 1: Albayzin 2010 LRE: Distribution of training data (hours) and development and evaluation data (# 30s segments).

in Table 1. 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, specifically 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 1. For further details, see (Rodriguez-Fuentes et al., 2012).

3.5. Fusion

The *FoCal* multiclass toolkit was applied to perform the calibration and fusion of SLR systems (Brümmer and du Preez, 2006).

3.6. Evaluation 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 (C_{LLR}) (Brümmer and du Preez, 2006).

4. Results

Table 2 shows the performance of the baseline systems (the acoustic MFCC-SDC and phonotactic systems) and the proposed approach (using the BUT Czech (CZ), Russian (RU) and Hungarian (HU) decoders) on the Albayzin 2010 LRE closed-set clean-speech and noisy-speech 30-second task. Regarding clean-speech, most of the systems performed similarly, except for the proposed PLLR iVector system when trained on the HU decoder PLLR features, which clearly stands out as the best single system, yielding 1.41 $C_{avg} \times 100$, which means a 33 % relative improvement with regard to the MFCC-SDC-based iVector approach and a 40 % relative improvement with regard to the respective HU phonotactic approach. Performance differences across decoders were found on both PLLR and phonotactic approaches (e.g. the performance of the phonotactic RU system degraded with regard to that of other phonotactic systems).

When focusing on the noisy speech condition, differences in performance were more noticeable. MFCC-SDC-based iVector system attained great performance (3.95 $C_{avg} \times 100$), but was once again outperformed by the HU PLLR iVector system (3.17 $C_{avg} \times 100$). All PLLR iVector systems outperformed their respective phonotactic counterparts (yielding between 5 % and 56 % relative improvements).

Since the HU PLLR iVector system showed the best performance among individual systems, we selected the HU decoder-based systems to analyze system fusions. Table 3 shows the performance of different fusions involving the baseline MFCC-SDC iVector system, the HU phonotactic system and the HU PLLR iVector system (for a better comparison, single system results are also included in Table 3). All pairwise fusions yielded high performance. The fusion of the MFCC-SDC iVector and phonotactic systems, led to great improvements with regard to single system performance (1.10 $C_{avg} \times 100$). A similar figure was achieved by the fusion of the PLLR iVector and phonotactic system (1.09 $C_{avg} \times 100$), closely followed by the fusion of the MFCC-SDC and PLLR iVector systems (1.20 $C_{avg} \times 100$). The fusion of the three systems yielded great performance: 0.97 $C_{avg} \times 100$, meaning a 31 % relative impro-

		Clean		Noisy	
System		$C_{avg} \times 100$	C_{LLR}	$C_{avg} \times 100$	C_{LLR}
MFCC-SDC iVector		2.12	0.176	3.95	0.325
CZ	Phonotactic	2.15	0.215	7.00	0.664
	PLLR iVector	2.33	0.223	6.66	0.546
HU	Phonotactic	2.35	0.218	7.28	0.621
	PLLR iVector	**1.41**	**0.127**	**3.17**	**0.308**
RU	Phonotactic	2.85	0.244	6.54	0.571
	PLLR iVector	2.34	0.225	4.38	0.352

Table 2: $C_{avg} \times 100$ and C_{LLR} performance for the baseline systems, the PLLR iVector system and different fusions on the Albayzin 2010 LRE primary task on clean and noisy speech.

		Clean		Noisy	
System		$C_{avg} \times 100$	C_{LLR}	$C_{avg} \times 100$	C_{LLR}
MFCC-SDC iVector (a)		2.12	0.176	3.95	0.325
HU	Phonotactic (b)	2.35	0.218	7.28	0.621
	PLLR iVector (c)	**1.41**	**0.127**	**3.17**	**0.308**
Fusion	(a)+(b)	1.10	0.106	2.43	0.211
	(a)+(c)	1.20	0.109	2.65	0.227
	(b)+(c)	1.09	0.092	2.65	0.228
	(a)+(b)+(c)	**0.97**	**0.086**	**1.86**	**0.168**
Fusion	ALL (7 systems, Table 2)	0.82	0.075	1.74	0.169

Table 3: $C_{avg} \times 100$ and C_{LLR} performance for the baseline systems, the PLLR iVector system and different fusions on the Albayzin 2010 LRE primary task on clean and noisy speech.

vement with regard to the best individual system (PLLR HU iVector). Finally, the fusion of all the systems led to the best result: 0.82 $C_{avg} \times 100$, that is, a 41 % relative improvement with regard to the PLLR HU system.Note, however, that this improvement was achieved by fusing 7 systems, more than two times the number of systems used to obtain the second best result.

Results for the noisy-speech condition are consistent with the ones attained on clean-speech. 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 MFCC-SDC 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 $C_{avg} \times 100$, meaning a 41 % relative improvement with regard to the best individual system. Once again, the PLLR iVector system seems to provide complementary information to baseline systems under all configurations. The fusion of the 7 subsystems shown in Table 3 yielded again the best result on the noisy speech condition: 1.74 $C_{avg} \times 100$, that is, a 45 % relative improvement with regard to the best individual system.

Table 4 shows the confusion matrix for the fusion of PLLR HU iVector, HU phonotactic system and MFCC-SD iVector system on the clean condition of the Albayzin 2010 LRE. As expected, the most confused languages were Spanish and Galician, followed by Spanish and Catalan, and Galician and Catalan. On the other hand, significantly low miss and false alarm probabilities were reached for the Basque, Portuguese and English languages.

5. *Conclusions and future work*

In this paper, further evidence of the suitability of Phone Log-Likelihood Ratio (PLLR) features for improving SLR performance under the iVector approach has been presented. The performance of a PLLR-based iVector system has been compared to that of two baseline acoustic (MFCC-SDC-based iVector) and phonotactic (Phone-Lattice-SVM) systems, using the Albayzin 2010 LRE dataset

		Target Language					
		Basque	Catalan	English	Galician	Portuguese	Spanish
Test audio	Basque	0.00	0.00	0.00	0.00	0.00	0.00
	Catalan	0.00	1.34	0.00	0.00	1.34	1.34
	English	0.00	0.00	0.00	0.00	0.00	0.00
	Galician	2.48	3.31	0.00	3.31	0.00	14.05
	Portuguese	0.00	0.00	0.00	0.00	0.00	0.00
	Spanish	0.00	0.00	0.00	7.20	0.00	0.80

Table 4: Confusion matrix for the fusion of the PLLR HU iVector, Phonotactic HU and MFCC-SDC iVector systems on the clean condition of the Albayzin 2010 LRE. *Miss probabilities (%)* are shown in the diagonal and *false alarm probabilities (%)* out of the diagonal.

as benchmark. The PLLR-based iVector system not only outperformed the baseline systems, but also proved to contribute complementary information in pairwise fusions with both of them. Finally, the fusion of the three approaches led to very competitive performance. The high performance achieved on noisy speech conditions opens a new track for PLLR features, which will be explored in future work on other databases, such as the Albayzin 2012 LRE dataset, featuring speech on more noisy and challenging conditions.

6. Acknowledgments

This work has been supported by the University of the Basque Country under grant GIU10/18 and project US11/06 and by the Government of the Basque Country under program SAIOTEK (project S-PE12UN55). M. Diez is supported by a research fellowship from the Department of Education, Universities and Research of the Basque Country Government.

References

BenZeghiba, M. F., J. L. Gauvain, and L. Lamel. September 2009. Language Score Calibration using Adapted Gaussian Back-end. In *Proceedings of Interspeech 2009*, pages 2191–2194, Brighton, UK.

Biadsy, Fadi, Julia Hirschberg, and Daniel P. W. Ellis. 2011. Dialect and accent recognition using phonetic-segmentation supervectors. In *Interspeech*, pages 745–748.

Brümmer, N. and J. du Preez. 2006. Application-Independent Evaluation of Speaker Detection. *Computer, Speech and Language*, 20(2-3):230–275.

Brümmer, N. and D.A. van Leeuwen. 2006. On calibration of language recognition scores. In *Proceedings of Odyssey - The Speaker and Language Recognition Workshop*, pages 1–8.

Brümmer, Niko and Edward de Villiers. 2011. The BOSARIS Toolkit: Theory, Algorithms and Code for Surviving the New DCF. In *Proceedings of the NIST 2011 Speaker Recognition Workshop*, Atlanta (GA), USA, December.

Campbell, W. M., F. Richardson, and D. A. Reynolds. 2007. Language Recognition with Word Lattices and Support Vector Machines. In *Proc. IEEE ICASSP*, pages 15–20.

Dehak, N., P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet. 2011a. Front-end factor analysis for speaker verification. *IEEE Transactions on ASLP*, 19(4):788–798, May.

Dehak, N., P. A. Torres-Carrasquillo, D. A. Reynolds, and R. Dehak. 2011b. Language Recognition via i-vectors and Dimensionality Reduction. In *Interspeech*, pages 857–860.

DHaro, L.F., O. Glembek, O. Plocht, P. Matejka, M. Soufifar, R. Cordoba, and J. Cernocky. 2012. Phonotactic Language Recognition using i-vectors and Phoneme Posteriogram Counts. In *Proceedings of the Interspeech 2012*, Portland, USA.

Diez, M., A. Varona, M. Penagarikano, L.J. Rodríguez Fuentes, and G.Bordel. 2012. On the Use of Phone Log-Likelihood Ratios as Features in Spoken Language Recognition. In *Proc. IEEE Workshop on SLT*, Miami, Florida, USA.

Fan, R.E., K.W. Chang, C.J. Hsieh, X.R. Wang, and C.J. Lin. 2008. LIBLINEAR: A Library for Large Linear Classification. *J. Machine Learning Research*, 9:1871–1874.

Martínez, D., L. Burget, L. Ferrer, and N.S. Scheffer. 2012. iVector-based Prosodic System for Language Identification. In *Proceedings of ICASSP*, pages 4861–4864, Japan.

Martínez, D., O. Plchot, L. Burget, O. Glembek, and P. Matejka. 2011. Language Recognition in iVectors Space. In *Proceedings of Interspeech*, pages 861–864, Firenze, Italy.

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

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

Penagarikano, M., A. Varona, L.J. Rodriguez-Fuentes, and G. Bordel. 2011. Dimensionality Reduction for Using High-Order n-grams in SVM-Based Phonotactic Language Recognition. In *Interspeech*, pages 853–856.

Plchot, O., M. Karafiát, N. Brümmer, O. Glembek, P. Matejka, and E. de Villiers J. Cernocký. 2012. Speaker vectors from Subspace Gaussian Mixture Model as complementary features for Language Identification. In *Odyssey: The Speaker and Language Recognition Workshop*, pages 330–333.

Rodriguez-Fuentes, L. J., M. Penagarikano, G. Bordel, and A. Varona. 2010. The Albayzin 2008 Language Recognition Evaluation. In *Proceedings of Odyssey 2010: The Speaker and Language Recognition Workshop*, pages 172–179, Brno, Czech Republic.

Rodriguez-Fuentes, L. J., M. Penagarikano, A. Varona, M. Diez, and G. Bordel. 2011. The Albayzin 2010 Language Recognition Evaluation. In *Proceedings of Interspeech*, pages 1529–1532, Firenze, Italia.

Rodriguez-Fuentes, L. J., M. Penagarikano, A. Varona, M. Diez, and G. Bordel. 2012. 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.

RTTH, 2006. *Spanish Network on Speech Technology*. Web (in Spanish): http://lorien.die.upm.es/∼lapiz/rtth/.

Schwarz, P. 2008. *Phoneme recognition based on long temporal context*. Ph.D. thesis, Faculty of Information Technology, Brno University of Technology, Brno, Czech Republic.

Singer, E., P. A. Torres-Carrasquillo, T. P. Gleason, W. M. Campbell, and D. A. Reynolds. 2003. Acoustic, Phonetic and Discriminative Approaches to Automatic Language Identification. In *Proceedings of Eurospeech (Interspeech)*, pages 1345–1348, Geneva, Switzerland.

Soufifar, M., S. Cumani, L. Burget, and J. Cernocky. 2012. Discriminative Classifiers for Phonotactic Language Recognition with iVectors. In *Proc. IEEE ICASSP*, pages 4853–4856.

Stolcke, A. 2002. SRILM - An extensible language modeling toolkit. In *Interspeech*, pages 257–286.

Varona, Amparo, Mikel Penagarikano, Luis Javier Rodriguez Fuentes, Mireia Diez, and Germán Bordel. 2010. Verification of the four spanish official languages on tv show recordings. In *XXV Congreso de la Sociedad Espaï¿$\frac{1}{2}$ola para el Procesamiento de Lenguaje Natural (SEPLN)*, Valencia, Spain, 8-10 September.

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