Dimensionality Reduction of Phone Log-Likelihood Ratio Features for Spoken Language Recognition

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

GTTS (http://gtts.ehu.es), Department of Electricity and Electronics, ZTF/FCT
University of the Basque Country UPV/EHU, Barrio Sarriena, 48940 Leioa, Spain
mireia.diez@ehu.es

Abstract

In a previous work, we introduced the use of log-likelihood ratios of phone posterior probabilities, called Phone Log-Likelihood Ratios (PLLR) as features for language recognition under an iVector-based approach, yielding high performance and promising results. However, the high dimensionality of the PLLR feature vectors (with regard to MFCC/SDC features) results in comparatively higher computational costs. In this work, several supervised and unsupervised dimensionality reduction techniques are studied, based on either fusions or selection of phone posteriors, finding that PLLR feature vectors can be reduced to almost a third of their original size attaining similar performance. Finally, Principal Component Analysis (PCA) is also applied to the original PLLR vector as a feature projection method for comparison purposes. Results show that PCA stands out among all the techniques studied, revealing that it does not only reduce computational costs, but also improves system performance significantly.

Index Terms: Spoken Language Recognition, iVectors, Phone Log-Likelihood Ratios, Phonetic Broad Classes, Principal Component Analysis

1. Introduction

Log-likelihood ratios of phone posterior probabilities, hereafter called Phone Log-Likelihood Ratios (PLLR), have been recently introduced as alternative features to the traditional Mel Filter Cepstral Coefficients / Shifted Delta Cepstrum (MFCC/SDC) for Spoken Language Recognition (SLR) tasks under an iVector approach [1], achieving competitive performance as a stand-alone system and providing significant relative improvements when fused with state-of-the-art acoustic and/or phonotactic approaches, which reveals the complementarity of the features with regard to both approaches.

PLLR features can be plugged into traditional acoustic systems, by simply replacing the MFCC/SDC features, since PLLRs provide acoustic-phonetic information in a sequence of frame-level feature vectors. The availability of open source decoders, like the open-software Temporal Patterns Neural Network (TRAPs/NN) phone decoders, developed by the Brno University of Technology (BUT) [2] to get phoneme posteriors, and the free software to compute PLLR features [3], provides a handy framework to facilitate the use of the features under different system configurations.

Nevertheless, the high dimensionality of PLLR features with regard to acoustic representations, can pose a computational problem when dealing with certain approaches. Acoustic feature vectors usually range from 7 to 19 MFCC (which are then augmented with Delta coefficients, or used to compute Shifted Delta Cepstrum, depending on the approach) [4], [5], whereas the number of PLLR features depends on the number of phonetic units of the decoder, which in the case of BUT decoders amounts to 43 units for Czech, 59 units for Hungarian or 50 for Russian (which are also augmented with Delta coefficients to optimize performance) [1]. This work deals with the dimensionality issue, by studying different reduction techniques that can be applied to PLLR features.

There is a handful of works in different fields of speech recognition literature, aiming to reduce the number of phone models into smaller broad classes, either by clustering or by selection techniques. On the one hand, some works apply supervised clustering, which requires knowledge of the language/phonemes to define phone families, as in [6], where several phone sets are defined for a PRLM approach used in a SLR task, or in [7], where a reduced phone set is also used to reduce n-gram counts on a phonotactic SLR iVector approach. On the other hand, unsupervised clustering based on different distance metrics like the confusion among phonemes [8] or mutual information based merging and selection [9] have also been applied to improve speech or language recognition. In this work, different supervised and unsupervised methods have been tested on phone posterior probabilities, before computing the PLLR features. Finally, the widely known Principal Component Analysis (PCA) has been also tested in the space of PLLR features.

The rest of the paper is organized as follows. Section 2 describes the baseline system, including the computation of the phone log-likelihood ratios used as features and the iVector approach. In Section 3, the dimensionality reduction techniques applied in this work are briefly described. Section 4 describes the experimental setup. Section 5 presents language recognition results on the NIST 2007 and 2011 datasets and compares the performance of the proposed approaches. Finally, conclusions are given in Section 6.

2. System Description

2.1. Phone Log-Likelihood Ratio features

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:

$$p(i|t) = \sum_{\forall s} p(i|s,t) \quad (1)$$

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

probabilities as follows:

$$LLR(i|t) = \log \frac{p(i|t)}{\frac{1}{(N-1)}(1-p(i|t))} \qquad i = 1, ..., N \quad (2)$$

The resulting N log-likelihood ratios per frame are the PLLR features considered in our approach. Free software to compute them can be found in [3].

2.2. 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, Hungarian and Russian [2], 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 $p_{i,s}(t)$ [1] 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 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, using the BUT decoders for Czech, Hungarian and Russian, we get 43, 59 and 50 PLLR features per frame, respectively.

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

$$\mathbf{M} = \mathbf{m} + \mathbf{T}\mathbf{w} \quad (3)$$

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

A generative modeling approach can be applied in the iVector feature space (as in [11]), the set of iVectors of each language being modeled by a single Gaussian distribution. Thus, the iVector scores are computed as follows:

$$score(f, l) = N(w_f; \mu_l, \Sigma) \quad (4)$$

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. Dimensionality Reduction Techniques

3.1. Supervised Techniques

In phonotactic SLR approaches, it is a common practice to take advantage of the phonetic knowledge to reduce the set of phone units [6], [7]. Different clusterings can be performed in the phone posterior probability space (on which PLLRs are computed) based on expert knowledge. Four different phone sets were considered in this study:

- *Family-R*: The set of Reduced (R) phones used in [7] to reduce the number of n-gram counts in a phonotactic approach.
- *Family-SL*: A set of phonemes defined by merging all Short and Long (SL) phonemes. Vowels belonging to the same regions in the IPA charts were also merged.
- *Family-MP*: A set of phonemes defined according to phonetic categories following IPA charts. Phones produced with the same Manner and Place (MP) of articulation were merged.
- *Family-M*: A more generic phonetic classification, where consonants produced with the same Manner (M) of articulation were merged. Vowels belonging to the same regions in the IPA charts were also merged.

For each of the above families, phones included in the same phonetic class were used to define a single unit by adding the posteriors obtained in Equation 1, before computing the log-likelihood ratios.

3.2. Unsupervised Techniques

Supervised clustering poses some problems: knowledge of each language is needed to define suitable phone sets, and the dimensionality is constrained to a certain range according to the nature of the language (that is, we are not free to choose an arbitrary dimensionality). Unsupervised clustering techniques, instead, are more flexible and can be easily tuned to define set of phones of arbitrary dimensions [8], [9]. In this work, the following criteria have been applied:

- *Correlation*: An iterative clustering algorithm is used. In each step, the algorithm merges the closest phone pair (or phone group pair) according to the correlation among the phone posterior probabilities.
- *Frequency*: The N phones with the highest posterior probabilities overall in the training set are selected as most relevant, and therefore used as (reduced) phone set.

Finally, PCA was also tested. Since PCA is an orthogonal transformation that is assumed to deal with normally distributed data ranging in $[-\infty, \infty]$, it was not a suitable transformation to be applied on the phone posterior probability space, which ranges in [0,1]. Instead, PCA can be directly applied on the normally distributed PLLR space, which ranges in $[-\infty, \infty]$.

4. Experimental Setup

4.1. Datasets

4.1.1. NIST 2007 LRE

The NIST 2007 LRE [12] 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[1]; (2) the OHSU Corpus provided by NIST for the 2005 LRE[2]; and (3) the development corpus provided by NIST for the 2007 LRE[3]. A set of 23 languages/dialects was defined for training, including target and non-target[4] languages. For development purposes, 10 conversations per language were randomly selected, and the remaining

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

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

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

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

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.

4.1.2. NIST 2011 LRE

The NIST 2011 LRE [13] involved a pairwise language detection task with 24 target languages 9 of which had been never used as target languages in previous NIST evaluations. Development data specifically collected for these 9 languages, including 100 30-second segments per language, were randomly split into approximately two half disjoint subsets: the first half was used to train specific models for the new languages, and the second half was used to estimate backend and fusion parameters.

To train more robust models for the target languages, we added data from databases distributed by the Linguistic Data Consortium (LDC) (LDC2006S45 for Arabic Iraqi, LDC2006S29 for Arabic Levantine and LDC2000S89 + LDC2009S02 for Czech). The remaining materials were extracted from wide-band broadcast news recordings, dowsampling them to 8 kHz: COST278 Broadcast News database [14] was used to get speech segments for Czech and Slovak; Arabic MSA was extracted from Al Jazeera broadcasts included in the KALAKA-2 database created for the Albayzin 2010 LRE [15]; broadcasts were also *captured* from video archives in TV websites to get speech segments in Arabic Maghrebi (Arrabia TV) and Polish (Telewizja Polska, TVP INFO). We were not able to collect additional training materials for Panjabi by any means.

A set of 66 languages/dialects was defined for training [16]. Each of them was mapped either to a target language or to non-target languages[5]. The training dataset also includes 2007 CTS and 2009 VOA signals [1]. The whole training dataset for the NIST 2011 LRE benchmark amounts to 1953 hours.

For development purposes, the second half of the audited segments provided for new target languages, along with the NIST 2007 and 2009 evaluation datasets, and 30-second signals used for development in 2007 and 2009 [1] were used. The whole development dataset consists of 13663 segments.

Performances reported in this paper have been computed on the 30-second closed-set condition of the test set (primary evaluation task).

4.2. System Configuration

As shown in [1], 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 training sets. The total variability matrix (on which the iVector approach relies) was estimated as in [10], using only target languages in the training sets. For the first bunch of experiments, carried out to compare different dimensionality reduction techniques (see Section 5.1), 5 iterations were used to compute the total variability matrix. For the final results (see Section 5.2), 10 iterations of the algorithm were applied for matrix estimation to optimize results.

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

4.3. Backend and Fusion

The backend setup was separately optimized for each dataset. A ZT-norm followed by a discriminative Gaussian backend was applied in experiments on the NIST 2007, whereas a generative Gaussian backend was applied in experiments on the NIST 2011 LRE dataset. Discriminative multiclass calibration/fusion models were estimated on the development set and applied to scores after the backend. The FoCal toolkit was used to estimate and apply the backend and calibration/fusion models [17].

4.4. 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, (2) the Log-Likelihood Ratio Cost C_{LLR} [17]; and (3) the primary measure C_{avg}^{24} used to evaluate system performance in the NIST 2011 LRE [13], which first computes pairwise minimum and actual costs for all pairs of target languages, and then averages the actual cost for the 24 pairs with the highest minimum cost.

5. Results

5.1. Comparison of Dimensionality Reduction Techniques

Table 1 shows results for the baseline system trained on PLLR features obtained with the Hungarian (HU) decoder and without any dimensionality reduction, along with results for different dimensionality reduction techniques.

Table 1: $\%C_{\text{avg}}$ and C_{LLR} performance for the PLLR iVector system when each dimensionality reduction techniques, on the NIST 2007 LRE primary task.

HU PLLR System				Dim	$\%C_{\text{avg}}$	C_{LLR}
Baseline				59	**2.86**	**0.389**
Supervised	Merge Phones	Family	R	33	3.07	0.422
			SL	31	3.46	0.467
			MP	23	**2.98**	**0.426**
			M	14	4.22	0.580
Unsupervised	Merge Phones	Correlation		23	3.76	0.523
	Select Phones	Frequency		23	3.56	0.480
	PLLR Projection	PCA		23	**2.45**	**0.333**

Focusing on the supervised clustering results, we see that the 33-dimensional Family-R feature set gets litle degradation with regard to the baseline system (3.07% vs 2.86% in terms of C_{avg}). The system based on the Family-MP phone set defined for this work, reaches even better performance (2.98% C_{avg}) in spite of comprising the information to a smaller 23-dimensional feature vector. On the contrary, the results obtained with the systems trained on Family-SL (31-dimensional) and Family-M (14-dimensional) phone-sets, show a clear performance degradation (3.46% and 4.22% C_{avg}, respectively). Therefore, Family-MP clustering provides a phone set that reduces the feature vector to almost a third of the original PLLR vector size, with almost no performance degradation.

For the unsupervised methods, and given the previous analysis, results are shown for the same dimensionality of the Family-MP approach (23-dimensional systems) in order to allow meaningful comparisons. The systems trained on sets defined by phone merging according to correlation achieve worse

performance (3.76% C_{avg}) than the systems based on supervised dimensionality reduction techniques. Systems trained on phones selected according to frequency, achieve better results than those based on correlation, but still worse that those attained by supervised approaches (3.56% C_{avg}).

Finally, results are also shown for the system that applies PCA directly on the PLLR features, revealing that it even outperforms the baseline system, attaining 2.45% C_{avg} and 0.333 in terms of C_{LLR}.

5.2. Results using Multiple Decoders

In this section, results are first presented for the NIST 2007 LRE database and then for the more challenging NIST 2011 LRE database. Results are presented for the baseline phone set and the two best reduction approaches: Family-MP (supervised clustering) and PCA (unsupervised projection). The former achieves similar performance as the baseline system, while keeping the phonetic dependence of each unit of the feature vector, which makes the merging approach suitable also for phonotactic approaches [6], [7]. The latter provides a significant gain in performance, which enhances the competitiveness of the system. Note that, for a fair comparison of both approaches, PCA dimensionality was not optimized, but chosen according to the Family-MP phone set.

Unlike acoustic systems, PLLR-based systems can take advantage of the use of different decoders, and perform fusion for an optimal performance, like other phonotactic approaches do [18], [4], [5]. In this Section, results for individual and fused PLLR systems based on Czech (CZ), Hungarian (HU) and Russian (RU) phone decoders are presented.

5.2.1. Results on NIST 2007 LRE

As in Table 1, individual system performances reported in Table 2 show little degradation when using the Family-MP phone-set and a remarkable improvement when PCA is applied. When the three PLLR systems (each based on a different decoder) are fused, performance improves significantly. The fusion of baseline systems attains 2.09% C_{avg}, whereas the fusion of Family-MP system reaches 2.24% C_{avg} and the fusion of PCA projected systems achieves a remarkable 1.79% C_{avg}.

Table 2: $\%C_{\text{avg}}$ and C_{LLR} performance for PLLR iVector baseline system, and systems using PLLR features reduced to the Family-MP set and projected with PCA, for each of the BUT decoders, and their fusion, on the NIST 2007 LRE primary task.

PLLR System		$\%C_{\text{avg}}$	C_{LLR}
Baseline	CZ (43+Δ)	4.18	0.550
	HU (59+Δ)	2.66	0.382
	RU (50+Δ)	4.08	0.549
	CZ+HU+RU	**2.09**	**0.299**
Family-MP	CZ (25+Δ)	4.55	0.619
	HU (23+Δ)	3.08	0.424
	RU (21+Δ)	4.30	0.598
	CZ+HU+RU	**2.24**	**0.313**
PCA	CZ (25+Δ)	3.12	0.432
	HU (23+Δ)	2.17	0.320
	RU (21+Δ)	3.29	0.451
	CZ+HU+RU	**1.79**	**0.240**

5.2.2. Results on NIST 2011 LRE

In Table 3, the performance of the baseline and the reduced dimensionality PLLR systems on the NIST 2011 LRE database is presented. Results are consistent with those reported on the NIST LRE 2007 database. Performance of individual systems slightly degrades with regard to the baseline system when using the Family-MP phone set. On the contrary, performance improves when PLLRs are projected using PCA.

Table 3: $\%C_{\text{avg}}$, C_{LLR} and $C_{avg}^{24} \times 100$ performance for PLLR iVector baseline system, and systems using PLLR features reduced to Family-MP set and projected with PCA, for each of the BUT decoders, and their fusion, on the NIST 2011 LRE primary task.

PLLR System		$\%C_{\text{avg}}$	C_{LLR}	$\%C_{avg}^{24}$
Baseline	CZ (43+Δ)	5.31	0.978	12.46
	HU (59+Δ)	5.18	0.982	12.12
	RU (50+Δ)	4.70	0.898	11.27
	CZ+HU+RU	**3.79**	**0.720**	**9.10**
Family-MP	CZ (25+Δ)	5.53	1.054	13.62
	HU (23+Δ)	5.40	1.015	12.64
	RU (21+Δ)	5.13	0.961	11.57
	CZ+HU+RU	**3.82**	**0.693**	**9.79**
PCA	CZ (25+Δ)	4.46	0.855	11.20
	HU (23+Δ)	4.48	0.877	10.88
	RU (21+Δ)	4.20	0.803	11.01
	CZ+HU+RU	**3.21**	**0.634**	**8.45**

Focusing on the results when the three decoder-specific systems are fused, note that no performance loss is observed when fusing the Family-MP systems with regard to fusing the baseline systems (3.82% C_{avg} vs 3.79% C_{avg}). Performance attained with the fusion of PCA based systems, stands out as the best result once again, achieving 3.21% C_{avg}.

6. Conclusions

Several dimensionality reduction techniques have been studied for a PLLR-iVector system. Results show that, using a supervised phone merging criteria based on phonetic knowledge, PLLR feature vectors can be reduced up to almost a third of the original size (reaching a dimensionality comparable to the one of the MFCC-SDC features), attaining similar performance. Since merging is performed in the phone posterior probability space, the resulting components keep the phonetic dependence/meaning, which makes the merging criteria useful also for phonotactic approaches.

Regarding unsupervised methods, results have consistently shown that applying PCA in the PLLR feature space not only reduces the computational cost, but also improves system performance significantly.

Finally, the study has also revealed that PLLR systems can take advantage of the specific acoustic-phonetic information provided by different decoders. PLLR systems built on different phone decoders can be fused to get significant gains in performance.

7. Acknowledgments

We thank Mehdi Soufifar for giving us details of the phone merging criterion used in [7].

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-PE12UN055); Mireia Diez is supported by a 4-year research fellowship from the Department of Education, University and Research of the Basque Country.

8. References

[1] 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.

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

[3] *Free PLLR computation software*. [Online]. Available: https://sites.google.com/site/gttspllrfeatures/home

[4] 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.

[5] 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.

[6] T. Kempton and R. K. Moore, "Language Identification: Insights from the Classification of Hand Annotated Phone Transcripts," in *Proc. Odyssey 2008 - The Speaker and Language Recognition Workshop*, Stellenbosch, South Africa, January 21-24 2008.

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

[8] A. Zgank, B. Horvat, and Z. Kacic, "Data-driven generation of phonetic broad classes, based on phoneme confusion matrix similarity," *Speech Comunication*, vol. 47, no. 3, pp. 379–393, September 2005.

[9] C. S. Kumar, H. Li, R. Tong, P. Matejka, L. Burget, and J. Cernocký, "Tuning Phone Decoders For Language Identification," in *Proceedings of the workshop on Human Language Technology*, Stroudsburg, PA, USA, 2010, pp. 4861–4864.

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

[11] 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.

[12] 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.

[13] C. Greenberg, A. Martin, and M. Przybocki, "The 2011 NIST Language Recognition Evaluation," in *Proceedings of Interspeech*, Portland, Oregon, 2012.

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

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

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

[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.