OPTIMIZING PLLR FEATURES FOR SPOKEN LANGUAGE RECOGNITION

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

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

ABSTRACT

Phone Log-Likelihood Ratios (PLLR) have been recently introduced as features for spoken language and speaker recognition systems. This representation has proven to be an effective way of retrieving acoustic-phonotactic information into frame-level vectors, which can be easily plugged into state-of-the-art systems. In a previous work, we began the search of reduced representations of PLLRs, as a mean of reducing computational costs. In this paper, we extend this search, by looking for the optimal compromise between feature vector size and system performance. Results achieved by Principal Component Analysis projection on the PLLR space are extensively analyzed. Also, to evaluate the effect of using larger temporal contexts, a Shifted Delta transformation is applied (and its optimal configuration explored) on highly reduced sets of PCA-projected PLLR features, leading to further performance improvements over the best PCA-projected PLLR set.

Index Terms— Spoken Language Recognition, Phone Log-Likelihood Ratios, Total Variability Factor Analysis.

I. INTRODUCTION

Spoken Language Recognition (SLR) is Pattern Recognition task which consists of recognizing the language spoken in an utterance by computational means. In the last years, there is an increasing need of SLR technology, as new applications are emerging and/or becoming popular, such as multilingual conversational systems [1], spoken language translation [2], multilingual speech recognition [3], spoken document retrieval [4], etc.

The general structure of a SLR system involves four stages (see Figure 1):

- Feature/token extraction, which aims to concentrate in few and, as far as possible, independent (that is, uncorrelated) parameters the information relevant to the classification task.
- Applying a classifier which scores feature/token sequences with regard to models of target languages.
- Applying a backend to normalize/calibrate the resulting scores, which allows us to use a single threshold for all the target languages and makes the system work at the desired application point.
- Making a hard decision (which depends on the task).

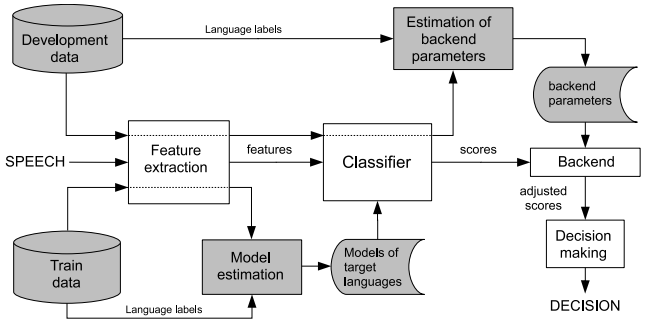

Fig. 1. Structure of a Spoken Language Recognition (SLR) system.

Two main complementary types of systems are commonly applied in SLR tasks [5]: acoustic systems and phonotactic systems. In acoustic systems, the target language is modeled with information taken from the spectral characteristics of the audio signal. The acoustic approach known as Gaussian Mixture Model - Universal Background Model (GMM-UBM) [6], which makes use of Mel-Frequency Cepstral Coefficients (MFCC) and Shifted Delta Cepstrum (SDC) features [7], is the baseline for many other developments. Currently, most state-of-the-art SLR systems are based on an approach derived from Joint Factor Analysis, known as *Total Variability Factor Analysis* or, more briefly, *i-vector* approach[8] [9].

Phone Log-Likelihood Ratios (PLLR) have been recently introduced as features for language [10] and speaker recognition systems [11]. These features provide acoustic-phonetic information in a sequence of frame-level vectors, which makes them suitable to plug into traditional approaches based on the widely used MFCC or Perceptual Linear Prediction (PLP) features.

In a previous work, given the high dimensionality of the PLLR feature vectors (in contrast with MFCC or PLP representations), we started the search of a reduced the Phone Log-Likelihood Ratio features representation [12]. Several supervised and unsupervised dimensionality reduction techniques were applied on the phone posterior probability space

by merging or selecting phones. Besides, for comparison purposes, Principal Component Analysis (PCA) was also applied on the PLLR space. It was found that the use of PCA not only reduced the computational costs (by reducing the size of the feature vector), but also enhanced the performance of the system.

In this paper, we extend that work by searching for an optimal compromise between the feature vector size and the performance of the system. A complete set of experiments is carried out on the NIST 2007 Language Recognition Evaluation (LRE) dataset (which features 14 target languages), by applying PCA projection into different dimensionalities. Moreover, we also test the effect of the Shifted-Delta transformation over the reduced PCA-projected PLLR features, with the purpose of considering larger temporal contexts. As in the SDC feature based systems [7], the SD-PLLR features are specified by four parameters N-d-P-k: N is the number of coefficients from which derivatives are computed at each frame, d determines the size of analysis windows (consisting of $2 \cdot d + 1$ frames) to compute de derivatives, P is the shift (number of frames) between two consecutive analysis windows and k is the number of analysis windows whose delta coefficients are concatenated to form the final feature vector. Optimal configuration of SD-PLLRs is found on the NIST 2007 LRE dataset and then applied to the NIST 2011 LRE dataset (which features 24 target languages).

It must be noted that NIST LREs, held in 1996 and every two years since 2003, have largely supported the development of SLR technology [13]. As a result, the datasets produced and distributed for such evaluations have become standard benchmarks to prove the usefulness of new approaches. These datasets consist of two types of signals: (1) narrow band, 8 kHz Conversational Telephone Speech (CTS) (NIST LRE 1996-2007); and (2) Narrow-Band Broadcast Speech (telephone calls included in broadcasts) coming from Voice of America (VOA) radio broadcasts (NIST 2009 LRE and 2011 LRE).

The rest of the paper is organized as follows. Section II gives details about the Phone Log-Likelihood Ratio extraction procedure. Section III describes the systems developed for this work: feature extraction, modeling and scoring. Section IV provides details about the experimental setup, the datasets and the evaluation measures. Section V presents the results achieved by using PCA-projected PLLR features and studies the gains achieved by optimizing the SD transformation. Finally, conclusions are given in Section VI.

II. PLLR 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:

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

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

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

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 [14].

III. I-VECTOR SYSTEM

Under the i-vector modeling assumption, an utterance GMM supervector (stacking the means of a GMM which is estimated by MAP adaptation of the UBM to the input utterance) is defined as:

$$M = m + Tw \tag{3}$$

where M is the utterance dependent GMM mean supervector, m is the utterance independent mean supervector, T is the so called total variability matrix (a low-rank rectangular matrix) and w is the so called i-vector (a normally distributed low-dimensional latent vector). That is, M is assumed to be normally distributed with mean m and covariance TT^t. The latent vector w can be estimated from its posterior distribution conditioned to the Baum-Welch statistics extracted from the utterance and using the UBM. The i-vector approach maps high-dimensional input data (the GMM supervector) to a low-dimensional feature vector, hypothetically maintaining most of the relevant information [8] [9].

IV. EXPERIMENTAL SETUP

IV-A. System Configuration

The open software Temporal Patterns Neural Network (TRAPs/NN) phone decoder , developed by the Brno University of Technology (BUT) for Hungarian (HU) was used as a first step to compute the PLLR features. This decoder is trained on the Hungarian SpeechDat(E) Database (8kHz), containing 10 hours of speech from 1000 Hungarian speakers (511 males, 489 females), recorded over the Hungarian fixed telephone network.

The BUT decoder takes into account three non-phonetic units: *int* (intermittent noise), *pau* (short pause) and *spk* (non-speech speaker noise) along with 58 phonetic units. For each unit, a three-state model is used, so three posterior probabilities per frame are calculated.

The BUT decoders produce a sequence of numbers representing the posterior probabilities $p^t_{i,s}$ for each one of

the three states s of each phone i at each frame t, encoded in the following way:

$$x_{i,s}(t) = \sqrt{-2 \log p_{i,s}(t)} \quad (4)$$

Thus, the posterior probability $p^t_{i,s}$ can be obtained as follows:

$$p_{i,s}(t) = e^{-\frac{(x_{i,s}(t))^2}{2}} \quad (5)$$

Before computing log-likelihood ratios, we integrate the non-phonetic units *int, pau* and *spk* into a single 9-state model. Then, a single posterior probability is computed for each unit i $(1 \leq i \leq N)$, by adding the posterior probabilities of all the states in the corresponding model (Equation 1). Finally the log-likelihood ratio for each unit is computed according to Equation 2. In this manner we get 59 log-likelihood ratios at each frame t. This feature vectors are augmented with first order dynamic coefficients [10].

Voice activity detection is performed by removing the feature vectors whose highest PLLR value correspond to the non-phonetic unit feature.

For the i-vector system, a gender independent 1024-mixture UBM is estimated by the Maximum Likelihood criterion on the training dataset, using binary mixture splitting, orphan mixture discarding and variance flooring. The total variability matrix T is estimated as in [15], but using only data from target languages, according to [9].

A generative modeling approach is applied in the i-vector feature space [9], the distribution of i-vectors of each language being modeled by a single Gaussian distribution. Thus, the i-vector scores are computed as follows:

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

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

IV-B. Datasets

NIST 2007 LRE

The NIST 2007 LRE [16] defined a spoken language recognition task for conversational speech across telephone channels, involving 14 target languages. In this work, NIST 2007 LRE dataset is used for development purposes (for tuning system parameters) because, although being smaller than the ones used in later evaluations, it is diverse and representative enough to make consistent decisions.

Training and development data used in this work were limited to those distributed by NIST to all 2007 LRE participants: (1) the Call-Friend Corpus; (2) the OHSU Corpus provided by NIST for the 2005 LRE; and (3) the development corpus provided by NIST for the 2007 LRE. A set of 23 languages/dialects was defined for training, including target and non-target[1] languages. For development purposes, 10 conversations per language were randomly selected, and the remaining conversations (amounting to around 968 hours) were used for training. Development conversations were further divided into 30-second speech segments. The total number of 30-second segments was 3073. Results reported in this paper have been computed on the subset of 30-second speech segments of the test set for the closed-set condition (2158 segments), which was the primary task in the NIST 2007 LRE.

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

NIST 2011 LRE

The NIST 2011 LRE [17] 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 different sources. A set of 66 languages/dialects was defined for training [18]. Each of them was mapped either to a target language or to non-target languages[2]. 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 [10] were used. The whole development dataset consists of 13663 segments.

The NIST 2011 LRE is the largest and most challenging evaluation of SLR systems carried out up to date, making it a good choice for benchmarking SLR technology.

IV-C. 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} [19]; and (3) the primary measure C^{24}_{avg} used to evaluate system performance in the NIST 2011 LRE [17], 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.

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

[2] 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.

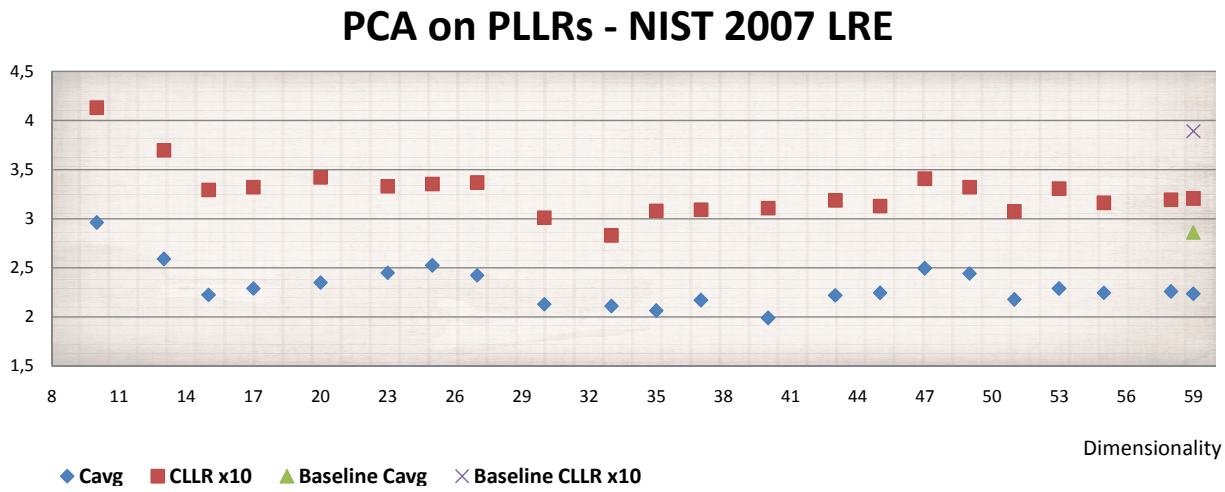

Fig. 2. C_{avg} and $10 \times C_{\text{LLR}}$ performance for the PLLR-based baseline system (with feature dimensionality=59) and systems trained on the set of PLLR features obtained after PCA projection into different dimensionalities, on the NIST 2007 LRE primary task

V. RESULTS

V-A. Complete study on the NIST 2007 LRE dataset

Principal Component Analysis

Results in [12] showed that PCA applied on the PLLR feature space could enhance the performance of an i-vector system. Therefore, the first set of experiments carried o int his work aims to find the optimal dimensionality to which the PLLR set could be reduced. Besides, this set of experiments also aims to search for the point where degradation would start harming system performance. Figure 2 shows the C_{LLR} and C_{avg} values for several systems trained on PLLR features after being projected by means of PCA to different dimensionalities.

The baseline system (without PCA) achieves 2.86 C_{avg} and 0.390 C_{LLR}. Note that performance is significantly enhanced even by simply PCA projecting the features into the same dimensionality, that is, without performing dimensionality reduction. This could be due to the whitening of the data obtained by means of PCA, that decorrelates the feature space, adapting it to the diagonal covariance GMMs that are used to model the data.

It can be seen that performance does not degrade significantly for a wide range of PCA projection dimensionalities. Four main ranges could be identified. The first one would be between 59 and 47. The second, in which best results can be found, ranges between 47 and 27, finding that 33 is the optimal dimensionality to which this set of features could be reduced, achieving 2.11 C_{avg} and 0.283 C_{LLR}. After that, between 27 and 15, a small degradation seems to affect the system. Finally, for values below 15 (at which 2.23 C_{avg} and 0.330 C_{LLR} is achieved) performance degrades markedly, revealing a significant loss in the information retrieved by the system.

Shifted Delta PLLR features

The second set of experiments focuses on checking whether the application of a Shifted-Delta transformation could provide information when applied over the PLLR features. To avoid unmanageable high dimensionality SD-PLLR feature sets, we started from PCA-projected PLLRs of the smallest dimension that didn't harm seriously system performance.

Table I. SLR performance of an i-vector system based on SD-PLLR features using different N values on the NIST 2007 LRE 30s test set.

PCA Dim		C_{avg}	C_{LLR}
13	PLLR	2.59	0.370
	SD- PLLR 13-2-3-7	1.71	0.260
15	PLLR	2.23	0.330
	SD- PLLR 15-2-3-7	1.94	0.264
17	PLLR	2.29	0.332
	SD- PLLR 17-2-3-7	1.73	0.241

As explained in Section I, the SD-PLLR features are specified by four parameters $N - d - P - k$. First, we optimized the number of coefficients from which derivatives are computed at each frame, N the other three parameters being set to common values in traditional MFCC acoustic systems ($d = 2$, $P = 3$ and $k = 7$). Given that 15 was the lowest dimension found without significant degradation, the SD-PLLR representation was tested for values around that dimensionality (13, 15 and 17). Table I shows a significant improvement over the PCA-projected PLLR set used as baseline. Best results on this dataset are achieved using the 13-2-3-7 configuration.

Given the improvement achieved by means of SD-PLLRs, SD configurations were further explored, by trying different values of the shift P. Results in Table II show a significant sensitivity to this parameter. again, the 13-2-3-7

Table II. SLR performance of an i-vector system based on SD-PLLR features, using different P values, on the NIST 2007 LRE 30s test set.

SD-PLLR configuration	C_{avg}	C_{LLR}
13-2-1-7	2.39	0.346
13-2-2-7	1.91	0.279
13-2-3-7	1.71	0.260
13-2-4-7	2.02	0.297
13-2-5-7	2.46	0.347

configuration was found to be optimal.

Table III. SLR performance of an i-vector system based on SD-PLLR features, using different d values, on the NIST 2007 LRE 30s test set.

SD-PLLR configuration	C_{avg}	C_{LLR}
13-1-3-7	2.04	0.286
13-2-3-7	1.71	0.260
13-3-3-7	2.03	0.277

Finally, the parameter d was also optimized, by considering values around the baseline $d = 2$. As shown in Table III, neither $d = 1$ nor $d = 3$ outperform $d = 2$, so the configuration 13-2-3-7 was considered optimal on the NIST 2007 LRE dataset.

V-B. Performance on the NIST 2011 LRE dataset

The optimal configuration of SD-PLLRs found on NIST 2007 LRE (13-2-3-7) was then applied on the NIST 2011 LRE dataset, with the purpose of checking the suitability of this choice on an independent test set. Table IV presents results for two PLLR-based systems on the NIST 2011LRE benchmark. The relative improvement achieved when using the PLLR features reduced by means of PCA and expanded to larger temporal contexts by means of SD computation was of 21% in terms of C_{avg} with regard to the baseline (without PCA) system.

Table IV. SLR performance of i-vector systems based on PLLR baseline and SD-PLLR features, on the NIST 2011 LRE 30s test set.

System	C_{avg}	C_{LLR}	$\%C_{avg}^{24}$
Baseline	5.18	0.982	12.12
SD-PLLR 13-2-3-7	4.10	0.826	10.48

VI. CONCLUSIONS

In this work, we have explored the improvements in SLR performance that could be achieved by PCA projection of the PLLR features into different dimensionalities. The study, developed on the NIST 2007 and 2011 LRE benchmarks, reveals that the features can be projected into a wide range of dimensionalities, enhancing the performance of the system, due in part to the decorrelation of the feature space achieved by applying Principal Component Analysis. For the Hungarian BUT decoder used in this work, best results are achieved by projecting PLLRs into 33 dimensions, where the system attained a 26% relative improvement in terms of C_{avg} with regard to the baseline system.

On the other hand, the application of Shifted-Delta over the PCA-projected set of PLLR features, proved to be effective, reaching 1.73 C_{avg} and 4.10 C_{avg}, for the NIST 2007 and 2011 LRE datasets, which means 40% and 21% relative improvements with regard to using the PLLR features, respectively.

VII. REFERENCES

[1] V. W. Zue and J. R. Glass, "Conversational Interfaces: Advances and Challenges," *Proceedings of the IEEE, Special Issue on Spoken Language Processing*, vol. 88, no. 8, pp. 1166–1180, August 2000.

[2] A. Waibel, P. Geutner, L. M. Tomoyiko, T. Schultz, and M. Woszczyna, "Multilinguality in speech and spoken language systems," *Proceedings of the IEEE, Special Issue on Spoken Language Processing*, vol. 88, no. 8, pp. 1181–1190, August 2000.

[3] Bin Ma, Cuntai Guan, Haizhou Li, and Chin-Hui Lee, "Multilingual Speech Recognition with Language Identification," in *Proceedings of ICSLP (Interspeech)*, 2002, pp. 505–508.

[4] Nicola Bertoldi and Marcello Federico, "Cross-Language Spoken Document Retrieval on the TREC SDR Collection," in *Advances in Cross-Language Information Retrieval, Lecture Notes in Computer Science, Volume 2785/2003*. 2003, pp. 476–481, Springer.

[5] P.A. Torres-Carrasquillo, E. Singer, T. Gleason, A. McCree, D.A. Reynolds, F. Richardson, and D.E. Sturim, "The MITLL NIST LRE 2009 language recognition system," in *Proceedings of ICASSP 2010*, 2010, pp. 4994–4997.

[6] Eddie Wong and Sridha Sridharan, "Methods to Improve Gaussian Mixture Model Based Language Identification System," in *Proceedings of ICSLP (Interspeech)*, 2002, pp. 93–96.

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

[8] N. Dehak, P. A. Torres-Carrasquillo, D. A. Reynolds, and R. Dehak, "Language Recognition via i-vectors and Dimensionality Reduction," in *Interspeech*, 2011, pp. 857–860.

[9] 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.

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

[11] Mireia Diez, Mikel Penagarikano, Amparo Varona, Lius Javier Rodríguez-Fuentes, and Germán Bordel, "University of the Basque Country Systems for the NIST 2012 Speaker Recognition Evaluation," in *Proceedings of the NIST-SRE 2012*, Orlando, USA, December 11-12 2012.

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

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

[14] *Free PLLR computation software, https://sites.google.com/site/gttspllrfeatures/home*.

[15] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on ASLP*, vol. 19, no. 4, pp. 788–798, May 2011.

[16] A. F. Martin and A. N. Le, "NIST 2007 Language Recognition Evaluation," in *Proceedings of Odyssey 2008 - The Speaker and Language Recognition Workshop, paper 016*, 2008.

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

[18] 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*, Portland, Oregon, USA, 9-13 September 2012.

[19] N. Brümmer and J. du Preez, "Application-Independent Evaluation of Speaker Detection," *Computer, Speech and Language*, vol. 20, no. 2-3, pp. 230–275, 2006.