INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Exploiting phone log-likelihood ratio features for the detection of the native language of non-native English speakers

Alberto Abad[1,2], *Eugénio Ribeiro*[1,2], *Fábio Kepler*[1], *Ramon Astudillo*[1,3], *Isabel Trancoso*[1,2]

[1]L[2]F - Spoken Language Systems Lab, INESC-ID Lisboa
[2]IST - Instituto Superior Técnico, University of Lisbon
[3]Unbabel Inc.
alberto.abad@l2f.inesc-id.pt

Abstract

Detecting the native language (L1) of non-native English speakers may be of great relevance in some applications, such as computer assisted language learning or IVR services. In fact, the L1 detection problem closely resembles the problem of spoken language and dialect recognition. In particular, log-likelihood ratios of phone posterior probabilities, known as Phone LogLikelihood Ratios (PLLR), have been recently introduced as features for spoken language recognition systems. This representation has proven to be an effective way of retrieving acoustic-phonotactic information at frame-level, which allows for its use in state-of-the-art systems, that is, in i-vector systems. In this paper, we explore the use of PLLR-based i-vector systems for L1 native language detection. We also investigate several linear and non-linear L1 classification schemes on top of the PLLR i-vector front-ends. Moreover, we compare PLLR based systems with both conventional phonotactic systems based on n-gram modelling of phoneme sequences and acoustic-based i-vector systems. Finally, the potential complementarity of the different approaches is investigated based on a set of system fusion experiments.

Index Terms: computational paralinguistics, native language recognition, PLLR, i-vectors.

1. Introduction

This paper presents INESC-ID's system for the Native Language (N) Sub-Challenge of the Computational Paralinguistics Challenge (ComParE) 2016 [1]. The task consists of identifying the native language (L1) of non-native English speakers. Detecting the L1 of the speakers is relevant for spoken language applications, since it provides information about the users that can be used to improve the interaction and the application performance. For instance, L1-specific ASR models can be used to improve recognition accuracy. Also, cultural information deduced from the identified L1 can lead to a more personal and context-aware dialog. Finally, accurate L1 detection can also play a role in software tools aimed at Computer Assisted Language Learning (CALL).

The data for the challenge is the ETS Corpus of Non-Native Spoken English, which contains 45-second answers from speakers with eleven different L1 backgrounds – Arabic, Chinese, French, German, Hindi, Italian, Japanese, Korean, Spanish, Telugu and Turkish – in the context of the TOEFL iBT® assessment. 3300 instances (41.3 hours) were selected for training, 965 (12.1 hours) for the development set, and 867 (10.8 hours) for testing.

In this paper, we explore the performance of i-vector systems based on Phone Log-Likelihood Ratios (PLLR) [2] on this task. We opted for this approach since it has been recently introduced and proved very effective in similar tasks. We also explore acoustic-based i-vector systems, and phonotactic systems based on Phone Recognition followed by Language Modelling (PRLM) for a matter of comparison. Furthermore, we investigate linear and non-linear models for language characterization based on i-vectors as an alternative to the simple, but effective, single Gaussian mixture model proposed in [3]. Finally, we also explore the combination of multiple subsystems through fusion based on logistic regression.

The following section presents some previous work related to the task. After that, Sections 3, 4, and 5 thoroughly describe the features, the classifiers, and the calibration and fusion approaches explored in this work for L1 identification. Results are presented and discussed in Section 6. Finally, the conclusions of the work are stated in Section 7.

2. Related Work

Automatic native language identification is a relatively recent task. For textual data, the most common approaches explore features related to spelling errors and the quality of writing, such as character, word, and POS n-grams, function words, and dependency relations [4, 5, 6, 7].

On the other hand, for spoken interactions, the literature is still very scarce. Different aspects of L2 speech may be explored, both at the segmental and supra-segmental level. The first level concerns the mispronunciations that are mainly due to the fact that some L2 phonemes are missing from the L1 inventory of speech sounds, which causes the non-native speakers to often replace an L2 phoneme by an L1 phoneme with a similar place or manner of articulation. Teaching L2 pronunciation traditionally focuses on segmentals. Supra-segmental features, related to intonation, are particularly hard to acquire for adults in L2. Hence, such features are also often used by humans to identify the native language. In [8], for instance, the authors use both cepstral and prosodic features, to identify 3 South Indian languages as L1 in non-native English speech.

The native language identification task is similar to the language, accent, and dialect identification tasks in Spoken Language Recognition (SLR). Thus, it makes sense to describe some of the most successful approaches used on those tasks. SLR approaches can be generally classified according to the source of information that they rely on. The most successful

Work partially supported by FCT project UID/CEC/50021/2013 and EU project H2020-EU.3.7. 653587

http://dx.doi.org/10.21437/Interspeech.2016-1491

systems are based on the exploitation of the acoustic [3, 9] and phonotactic [10, 11] characteristics of each language. While the first govern how a given language sounds, the latter are the rules that govern the possible phone combinations in a language. Usually, the combination of different sources of knowledge and systems of different characteristics tends to provide increased language recognition performances [12]. Recently, a new set of features known as Phone Log-Likelihood Ratios (PLLR) have been introduced for SLR [2]. These features convey frame-by-frame acoustic-phonetic information, which can be used in conventional acoustic systems like those based on the well-known Total Variability Factor Analysis (i-vector) approach [13]. The use of PLLR in SLR has been proven to lead to one of the best individual system results reported on relevant benchmarks [14].

3. Features for L1 recognition

In this work we decided to use acoustic-phonetic, acoustic, and phonotactic features for L1 recognition. The specific features of each category are described below.

3.1. Acoustic-Phonetic Features

In terms of acoustic-phonetic features, we use PLLR features as described in [14]. This section presents a brief summary of the PLLR extraction process, including post-processing stages, together with implementation details of the phonetic decoders.

3.1.1. PLLR definition

Considering a phone decoder that provides frame-by-frame phone posteriors p_i for each phone unit ($1 \leq i \leq N$), so that $\sum_{i=1}^{N} p_i = 1$ and $p_i \in [0, 1]$, the PLLR features are computed from these phone posteriors as follows [15]:

$$r_i = \text{logit}(p_i) = \log \frac{p_i}{(1-p_i)} \quad i = 1, \ldots, N. \qquad (1)$$

One of the main advantages of the transformation of phone posteriors into PLLRs is that it allows for the gaussianization of the resulting features, which makes them more suitable for typical GMM modelling. However, as pointed out in [16], the PLLR feature space as defined by Equation 1 is bounded, which limits the distribution of the features. In order to avoid the bounding effect, PLLRs are projected as described in [14]. Then, Principal Component Analysis (PCA) is applied to decorrelate the parameters and to reduce the feature dimensionality. After that, shifted delta cepstra (SDC) coefficients [17] are obtained. Following [18], the SDC configuration for PLLR features is 13-2-3-7, resulting in a feature vector of 104 components.

3.1.2. Phonetic classifiers

The phonetic classifiers used in this work are part of our hybrid Automatic Speech Recognition (ASR) system, AUDIMUS [19]. The phonetic models are neural networks of the MultiLayer Perceptron (MLP) type trained to estimate the posterior probabilities of the different phonemes of a specific language for a given input speech frame (and its context). In this case, we have used four language-dependent phonetic decoders: European Portuguese (*pt*), Brazilian Portuguese (*br*), European Spanish (*es*) and American English (*en*). For each phonetic decoder, an independent set of PLLR features is obtained based on the generated frame-by-frame posterior probabilities. In practice, frames that have silence as the most probable class are removed.

Each of the recognizers combines four MLP outputs trained with Perceptual Linear Prediction features (PLP, 13 static + first derivative), PLP with log-RelAtive SpecTrAl speech processing features (PLP-RASTA, 13 static + first derivative), Modulation SpectroGram features (MSG, 28 static) and Advanced Front-End from ETSI features (ETSI, 13 static + first and second derivatives). Each MLP network is characterized by the size of its input layer that depends on the particular parameterization and the frame context size (13 for PLP, PLP-RASTA and ETSI; 15 for MSG), the number of units of the two hidden layers (500), and the size of the output layer. In this case, only monophone units are modelled, resulting in MLP networks of 41 (39 phonemes +1 silence + 1 breathing) soft-max outputs in the case of *en*, 39 for *pt* (38 phonemes + 1 silence), 40 for *br* (39 phonemes + 1 silence) and 30 for *es* (29 phonemes + 1 silence). The output size corresponds to the length of the PLLR feature vectors before dimensionality reduction.

The language-dependent MLP networks were trained with different amounts of annotated data. For the *pt* acoustic models, 57 hours of Broadcast News (BN) down-sampled data and 58 hours of mixed fixed-telephone and mobile-telephone data were used. The *br* models were trained with around 13 hours of BN down-sampled data. The *es* networks used 36 hours of BN down-sampled data and 21 hours of fixed-telephone data. The *en* system was trained with the HUB-4 96 and HUB-4 97 down-sampled data sets, that contain around 142 hours of TV and Radio Broadcast data.

3.2. Acoustic Features

Acoustic features typically used for SLR have also been adopted in this work. In particular, we used shifted delta cepstra (SDC) of Mel-frequency Cepstrum Coefficients (MFCC) [17]. First, 7 MFCC static features are obtained and SDC features with a 7-1-3-7 configuration are computed, resulting in a feature vector of 56 components. Then, frames of each segment that are simultaneously labeled as silence by the four previously described language-dependent phonetic decoders are removed. Finally, cepstral mean normalization is applied in a per segment basis.

3.3. Phonotactic Features

In order to model the phonotactics of each target native language, a phonetic tokenization is obtained using the previously described classifiers. Likewise, we use the same 4 language-dependent ASR systems: *pt*, *br*, *es* and *en*. In this case, a decoding process is performed for each speech sequence (in contrast to simple frame-by-frame posterior probability computation). The AUDIMUS decoder is based on a weighted finite-state transducer (WFST) approach [20]. A phone-loop grammar with minimum phoneme duration of three frames is used to obtain the phonetic sequences.

4. Front-end models for L1 characterization

4.1. The i-vector front-end

Total-variability modelling [13] has emerged as one of the most powerful approaches to the problems of speaker and language verification. In this approach, the variability present in the high-dimensional GMM supervector is jointly modelled as a single low-rank total-variability space. The low-dimensionality total variability factors extracted from a given speech segment form a vector, named i-vector, which represents the speech segment

in a very compact and efficient way. Thus, the total-variability modelling is used as a factor analysis based front-end extractor. The success of i-vector based speaker recognition has motivated the investigation of its application to other related fields, including language recognition [3, 9], where it has become the current de facto standard for acoustic SLR. In this work, we have developed i-vector based LR sub-systems very similar to the one in [3], where the distribution of i-vectors for each language is modelled with a single Gaussian.

It is worth mentioning that, as an alternative to the i-vector approach, classifiers based on feed-forward networks were also trained for acoustic and acoustic-phonetic features. These approaches employed various techniques to account for the variable length of the feature matrices including Convolutional and Recurrent neural networks (CNN, RNN). In general these approaches led to poor results compared to i-vector modelling.

4.1.1. Total variability and i-vector extraction

The first step of i-vector system development consists of training a GMM-UBM. In this case, a GMM-UBM of 1024 mixtures is trained using all the training data available for the challenge. Then, the total variability factor matrix (**T**) is estimated according to [21]. The dimension of the total variability sub-space is fixed to 400. Next, zero and first-order sufficient statistics of the training set are used for training **T**. In order to do so, 10 Expectation-Maximization (EM) iterations of consecutive Maximum Likelihood (ML) and minimum divergence estimation updates are applied. The covariance matrix is not updated in any of the EM iterations. The estimated **T** matrix is used for extraction of the total variability factors of the processing speech segments as described in [21]. Additionally, we apply i-vector centering and whitening [22] that is known to contribute to a reduction of the channel variability. Finally, the resulting factor vectors are normalized to be of unit length, which we will henceforth refer to as i-vectors.

4.1.2. Language modelling and scoring

Like in [3], all the extracted i-vectors of each target L1 language are used to train a single mixture Gaussian distribution with full covariance matrix shared across different target languages. As an alternative to this approach, Log-linear and non-linear classifiers based on feed-forward networks were also investigated. In fact, it could be observed that the i-vector front-end already provided a very good separation of the classes which led to similar results for the different modelling techniques. For this reason, experimental results on alternative classifiers on the top of i-vectors are not reported. Finally, for a given test i-vector, each Gaussian model is evaluated and log- likelihood scores are obtained. The 11 likelihoods of the 11 L1 target languages form a vector of scores that is used for later calibration and fusion.

4.2. PRLM-LR sub-systems

The Phone Recognition followed by Language Modelling (PRLM) systems used in this work exploit the phonotactic information extracted by the four individual tokenizers described previously. For each target L1 language and for each tokenizer a different phonotactic *n-gram* language model is trained using the phonetic sequences of the challenge training data set. For that purpose, the SRILM toolkit has been used[1] and 3-gram back-off models smoothened using Witten-Bell discounting are obtained. During test, the phonetic sequence of a given speech signal is extracted with the phonetic decoders and the likelihood of each target language model is evaluated. Like the i-vector sub-systems, the likelihoods of the 11 L1 target languages form a vector of scores that is later used for calibration and fusion.

Similarly to [23], we tried likelihood scores obtained with phonotactic models of an arbitrary set of languages trained on external data as possibly discriminant features for L1 recognition. This approach, however, did not reveal useful for this task.

5. Calibration and Fusion Back-End

Calibration and fusion was carried out using a combination of linear Gaussian Back-Ends (GBE) followed by a Linear Logistic Regression (LLR). GBE were applied after every single subsystem to transform the score-vector $\mathbf{x}_i$ into a 11-element log-likelihood vector $\mathbf{s}_i$, corresponding to each of the target languages, using the following equation:

$$\mathbf{s}_i = \mathbf{A}_i \mathbf{x}_i + \mathbf{o}_i, \tag{2}$$

where $\mathbf{A}_i$ is the transformation matrix for system i and $\mathbf{o}_i$ is the offset vector. Notice that in this work the dimension of $\mathbf{x}_i$ for all the considered sub-systems is 11. Nevertheless, we kept the GBEs given that they contributed for improved language identification in the development experiments.

Then, LLR was used to fuse the log-likelihood outputs generated by the linear GBEs of the selected sub-systems to produce fused log-likelihoods **l** as follows:

$$\mathbf{l} = \sum_i \alpha_i \mathbf{s}_i + \mathbf{b}, \tag{3}$$

where α_i is the weight for sub-system i and **b** is the language-dependent shift. For this challenge, the language with the highest fused log-likelihood is the hypothesized L1 language.

During the development of our systems, the GBEs and the LLR fusion parameters were trained and evaluated on the development set using a kind of 2-fold cross-validation [24]: development data was randomly split in two halves, one for parameter estimation and the other for assessment. This process was repeated using 10 different random partitions so that the mean and variance of the systems' performance could be computed. This method allowed for a comparison and ranking of the different sub-systems under study. Then, for the trial submissions, no partition was made and all the development data was used to simultaneously calibrate the GBEs and the LLR fusion. Calibration was carried out using the FoCal Multi-class Toolkit [2].

6. Experimental Results

6.1. Baseline System

Similarly to previous years, the baseline system proposed for the Nativeness challenge employs the ComParE features set. This comprises 6373 features resulting from the computation of various functionals over low-level descriptor (LLD) contours. The features are computed with openSMILE [25] employing the configuration file `IS13_ComParE.conf`. All features were normalized to the mean and standard deviation of the training set. The classifier used is a Support Vector Machine (SVM) with epsilon insensitive loss, and a fixed ϵ of 1.0. Sequential Minimal Optimization (SMO) is used as the training algorithm.

[1] `http://www-speech.sri.com/projects/srilm/`

[2] `https://sites.google.com/site/nikobrummer/focalmulticlass`

The optimal complexity was set to 10^{-2}, based on the development set. The SVM implementation in WEKA 3 [26] was used for this purpose. See [1] for a complete description.

Table 1 shows the baseline performance in terms of Accuracy, Unweighted Average Recall (UAR) and Recall for each of the languages. Table 2 also provides the confusion matrix across languages for comparison purposes.

Table 1: Accuracy, UAR, and per language Recall of the baseline and our best submitted system over the development set.

Metric	Baseline	L^2F submission
Accuracy	45%	84%
UAR	45%	84%
Recall (ARA)	33%	86%
Recall (CHI)	45%	94%
Recall (FRE)	36%	83%
Recall (GER)	64%	91%
Recall (HIN)	56%	77%
Recall (ITA)	48%	87%
Recall (JPN)	42%	82%
Recall (KOR)	35%	81%
Recall (SPA)	32%	75%
Recall (TEL)	51%	75%
Recall (TUR)	48%	89%

Table 2: Baseline Confusion Matrix over the development set (rows: reference; columns: hypothesis).

	ARA	CHI	FRE	GER	HIN	ITA	JPN	KOR	SPA	TEL	TUR
ARA	**29**	3	5	7	5	5	6	6	7	6	7
CHI	4	**38**	5	4	5	2	5	10	6	4	1
FRE	11	7	**29**	8	0	4	3	1	11	0	6
GER	5	3	5	**55**	1	7	1	2	5	1	0
HIN	4	1	1	0	**47**	2	2	2	2	21	1
ITA	6	2	9	6	6	**46**	0	4	10	1	4
JPN	4	13	4	2	2	1	**36**	11	10	1	1
KOR	4	19	1	2	2	3	14	**32**	5	3	5
SPA	6	11	15	6	2	4	9	9	**32**	1	5
TEL	2	0	2	2	24	2	2	2	2	**43**	2
TUR	6	5	5	5	2	6	7	8	5	0	**46**

6.2. Proposed System

Table 3 presents the results obtained by our phonotactic and i-vector approaches on the challenge's development set. Notice that these results are on the complete development set, that is, the back-end cross-validation strategy described previously was not applied to obtain these results. The first thing to notice is that both approaches were able to surpass the baseline. However, the i-vector approach did so by a much larger margin. In this sense, the fusion of the 4 phonotactic sub-systems obtained 63.3% UAR, which represents an improvement of 18.2 percentage points over the baseline. On the other hand, even the worse individual i-vector approach was able to improve the baseline by a large margin. It is worth noticing that all the individual i-vector sub-systems based on PLLR features outperformed the one based on acoustic features. Moreover, the combination of the 4 PLLR based sub-systems provides a remarkable improvement with respect to the conventional i-vector acoustic subsystem. Nevertheless, the fusion of the 5 i-vector sub-systems resulted in additional performance gains. This combination of PLLR and MFCC i-vector approaches corresponds to the L^2F primary submission to the challenge.

Table 3: UAR [%] and Accuracy [%] results obtained by the phonotactic and i-vector approaches on the development set.

	UAR [%]	Acc [%]
Baseline	45.1	44.9
Phonotactic (BR)	46.4	46.2
Phonotactic (EN)	51.4	51.4
Phonotactic (ES)	50.0	49.8
Phonotactic (PT)	53.1	53.1
Phonotactic (All) **(I)**	63.3	63.2
i-vectors (MFCC) **(II)**	76.2	76.3
i-vectors (BR-PLLR)	76.9	76.9
i-vectors (EN-PLLR)	79.2	79.2
i-vectors (ES-PLLR)	77.6	77.4
i-vectors (PT-PLLR)	80.6	80.5
i-vectors (ALL-PLLR) **(III)**	83.0	82.9
(I) + (II)	78.6	78.7
(II) + (III)	84.6	84.6

Table 4 shows the confusion matrix of the L^2F system submitted to the challenge over the development set. The most confused classes are the same as in the baseline system (Table 2), namely, between Telugu and Hindi, although with a lower frequency. On the other hand, the large confusion shown by the baseline system over Chinese, Japanese, and Korean is not a problem for the proposed system.

Finally, the proposed L^2F system achieves 81.3% UAR in the challenge test set, in contrast to the 47.5% of the baseline system, which is again surpassed by a large margin. The performance drop with respect to the development set results can be partially due to a slight over-fitting of the back-end estimation. Nevertheless, we consider these results very promising.

Table 4: Confusion Matrix of the i-vector system over the development set (rows: reference; columns: hypothesis).

	ARA	CHI	FRE	GER	HIN	ITA	JPN	KOR	SPA	TEL	TUR
ARA	**77**	0	3	1	0	1	1	0	1	0	2
CHI	0	**78**	0	1	0	1	2	0	1	1	0
FRE	3	0	**64**	2	0	2	2	0	5	0	2
GER	2	1	2	**78**	0	0	0	1	0	0	1
HIN	0	0	0	0	**67**	0	0	0	0	16	0
ITA	1	0	5	2	0	**79**	1	1	3	0	2
JPN	1	1	1	0	0	0	**70**	8	4	0	0
KOR	2	4	1	1	0	0	5	**77**	1	0	0
SPA	2	1	2	1	0	5	4	5	**77**	1	2
TEL	0	0	0	0	18	0	0	0	0	**65**	0
TUR	0	1	1	3	1	2	0	2	1	0	**84**

7. Conclusions

This paper explored the use of PLLR-based i-vector systems for native language detection, building on the good results this method achieves for the closely related task of spoken language and dialect recognition. Results on the Native Language (N) Sub-Challenge of the Computational Paralinguistics Challenge (ComParE) 2016 confirm the potential of the approach outperforming the baseline by a large margin. A possible cause for the observed performance differences is the fact that the i-vector approach is able to better leverage the information present in large data-sets. As future work direction, it would be worth investigating approaches to identify the segments in each sentence that provide a better L1 discrimination.

8. References

[1] B. Schuller, S. Steidl, A. Batliner, J. Hirschberg, F. K. Burgoon, A. Baird, A. Elkins, Y. Zhang, E. Coutinho, and K. Evanini, "The interspeech 2016 computational paralinguistics challenge: Deception, sincerity & native language," in *Proceedings of Interspeech*, 2016.

[2] M. Diez, A. Varona, M. Penagarikano, L. J. Rodriguez-Fuentes, and G. Bordel, "On the use of phone log-likelihood ratios as features in spoken language recognition," in *Spoken Language Technology Workshop (SLT), 2012 IEEE*. IEEE, 2012, pp. 274–279.

[3] D. Martınez, O. Plchot, L. Burget, O. Glembek, and P. Matejka, "Language recognition in ivectors space," *Proceedings of Interspeech, Firenze, Italy*, pp. 861–864, 2011.

[4] S.-M. J. Wong and M. Dras, "Exploiting parse structures for native language identification," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2011, pp. 1600–1610.

[5] J. Brooke and G. Hirst, "Measuring interlanguage: Native language identification with l1-influence metrics," in *LREC*, 2012, pp. 779–784.

[6] J. R. Tetreault, D. Blanchard, A. Cahill, and M. Chodorow, "Native tongues, lost and found: Resources and empirical evaluations in native language identification," in *COLING*, 2012, pp. 2585–2602.

[7] S. Malmasi and M. Dras, "Language transfer hypotheses with linear svm weights," in *EMNLP*, 2014, pp. 1385–1390.

[8] R. K. Guntur and R. Krishnan, "Influence of mother tongue on english accent," in *ICON-2014*, 2014.

[9] 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.

[10] M. A. Zissman *et al.*, "Comparison of four approaches to automatic language identification of telephone speech," *IEEE Transactions on Speech and Audio Processing*, vol. 4, no. 1, p. 31, 1996.

[11] H. Li, B. Ma, and C.-H. Lee, "A vector space modeling approach to spoken language identification," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 1, pp. 271–284, 2007.

[12] L. J. Rodriguez-Fuentes, M. Penagarikano, A. Varona, M. Diez, G. Bordel, D. Martinez, J. Villalba, A. Miguel, A. Ortega, E. Lleida, A. Abad, O. Koller, I. Trancoso, P. Lopez-Otero, L. Docio-Fernandez, C. Garcia-Mateo, R. Saeidi, M. Soufifar, T. Kinnunen, T. Svendsen, and P. Franti, "Multi-site heterogeneous system fusions for the albayzin 2010 language recognition evaluation," in *IEEE Automatic Speech Recognition and Understanding Workshop*, Waikoloa, HI, USA, Dec. 2011, pp. 377–382.

[13] 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, 2011.

[14] M. Diez, A. Varona, M. Penagarikano, L. J. Rodriguez-Fuentes, and G. Bordel, "New insight into the use of phone log-likelihood ratios as features for language recognition." in *INTERSPEECH*, 2014, pp. 1841–1845.

[15] ——, "On the complementarity of phone posterior probabilities for improved speaker recognition," *Signal Processing Letters, IEEE*, vol. 21, no. 6, pp. 649–652, 2014.

[16] ——, "On the projection of pllrs for unbounded feature distributions in spoken language recognition," *Signal Processing Letters, IEEE*, vol. 21, no. 9, pp. 1073–1077, 2014.

[17] P. A. Torres-Carrasquillo, E. Singer, M. A. Kohler, R. J. Greene, D. A. Reynolds, and J. R. Deller Jr, "Approaches to language identification using gaussian mixture models and shifted delta cepstral features." in *INTERSPEECH*, 2002.

[18] M. Diez, A. Varona, M. Penagarikano, L. J. Rodriguez-Fuentes, and G. Bordel, "Optimizing pllr features for spoken language recognition," in *Pattern Recognition (ICPR), 2014 22nd International Conference on*. IEEE, 2014, pp. 779–784.

[19] H. Meinedo, A. Abad, T. Pellegrini, I. Trancoso, and J. Neto, "The l2f broadcast news speech recognition system," *Proc. Fala*, pp. 93–96, 2010.

[20] D. Caseiro and I. Trancoso, "A Specialized On-The-Fly Algorithm for Lexicon and Language Model Composition," *IEEE Transactions on Audio, Speech and Lang. Proc.*, vol. 14, no. 4, Jul. 2005.

[21] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel, "A study of interspeaker variability in speaker verification," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 16, no. 5, pp. 980–988, 2008.

[22] D. Garcia-Romero and C. Y. Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems," in *Interspeech*, 2011.

[23] E. Ribeiro, J. Ferreira, J. Olcoz, A. Abad, H. Moniz, F. Batista, and I. Trancoso, "Combining multiple approaches to predict the degree of nativeness," in *Interspeech*, 2015.

[24] L. J. Rodrıguez-Fuentes, M. Penagarikano, A. Varona, M. Dıez, G. Bordel, A. Abad, D. Martınez, J. Villalba, A. Ortega, and E. Lleida, "The blz systems for the 2011 nist language recognition evaluation," 2012.

[25] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*. ACM, 2010, pp. 1459–1462.

[26] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: an update," *ACM SIGKDD explorations newsletter*, vol. 11, no. 1, pp. 10–18, 2009.