The BLZ Submission to the NIST 2011 LRE: Data Collection, System Development and Performance

Luis Javier Rodríguez-Fuentes[1], Mikel Penagarikano[1], Amparo Varona[1], Mireia Diez[1], Germán Bordel[1], Alberto Abad[2], David Martínez[3], Jesus Villalba[3], Alfonso Ortega[3], Eduardo Lleida[3]

[1] *GTTS, Department of Electricity and Electronics, University of the Basque Country UPV/EHU, Spain*
[2] *L^2F - Spoken Language Systems Lab, INESC-ID Lisboa, Portugal*
[3] *Aragon Institute for Engineering Research (I3A), University of Zaragoza, Spain*

luisjavier.rodriguez@ehu.es

Abstract

This paper describes the most relevant features of a collaborative multi-site submission to the NIST 2011 Language Recognition Evaluation (LRE), consisting of one primary and three contrastive systems, each fusing different combinations of 13 state-of-the-art (acoustic and phonotactic) language recognition subsystems. The collaboration focused on collecting and sharing training data for those target languages for which few development data were provided by NIST, and on defining a common development dataset to train backend and fusion parameters and select the best fusions. Official and post-key results are presented and compared, revealing that the greedy approach applied to select the best fusions provided suboptimal but very competitive performance. Several factors contributed to the high performance attained by BLZ systems, including the availability of training data for low resource target languages, the reliability of the development dataset (consisting only of data audited by NIST), the diversity of modeling approaches, features and datasets in the systems considered for fusion, and the effectiveness of the search for optimal fusions.

Index Terms: Spoken Language Recognition, NIST 2011 LRE, Multiclass Discriminative Fusion, Greedy Search

1. Introduction

BLZ (Bilbao-Lisboa-Zaragoza) was a three-site team that made a joint submission to the NIST 2011 Language Recognition Evaluation (LRE) [1]. The three research groups are GTTS from the University of the Basque Country (EHU), the Spoken Language Systems Laboratory (L^2F) from INESC-ID Lisboa and the Aragon Institute for Engineering Research (I3A) from the University of Zaragoza.

The NIST 2011 LRE featured 24 target languages. Nine of them had not been used as target languages in previous evaluations. The main novelty of the NIST 2011 LRE was the focus on the discrimination between pairs of languages. A new performance metric was defined taking into account only the 24 language pairs for which system performance (assuming a perfect calibration) was worst. Thus, all the target languages should be suitably modeled and the availability of training and development data for all of them was one of the keys to obtain good results under the new metric.

The collaboration for a joint submission was motivated by a previous work [2], which successfully exploited the complementarity of systems based on different approaches or features, or trained on different datasets. This time the efforts focused on collecting and sharing data for the newly added target languages (for which few development data were delivered by NIST), and on defining a common development set to allow the estimation of backend and fusion parameters on independent data (i.e. not used to train models) and the selection of the best fusions. The FoCal[1] and Bosaris[2] toolkits were used to estimate and apply backend and fusion models, and to evaluate language recognition performance, respectively.

The BLZ submission to the NIST 2011 LRE consisted of one primary and three contrastive systems, built upon 13 subsystems, some of them implementing cutting edge approaches. A greedy search for the best overall fusion of subsystems was applied to define the primary system. Contrastive systems were developed to explore variants to the selection algorithm and to the standard backend configuration.

The paper is organized as follows. Section 2 describes the datasets used for system development. The main features of the subsystems developed at each site, along with the backend and fusion approaches on which the submission relies and the search procedure defined to find optimal fusions are described in Section 3. Finally, official and post-key results are presented and briefly discussed in Section 4.

2. Training and development data

2.1. Data collection for the newly added target languages

NIST provided a development dataset specifically collected for the 2011 LRE, including 100 30-second segments for each one of the nine newly added target languages (except for Lao, for which only 93 segments were provided), containing either conversational telephone speech (CTS) or narrow-band broadcast news speech (NB-BN). The dataset was augmented with 10- and 3-second segments, automatically extracted from the original 30-second segments. The resulting dataset, hereafter called *lre11*, was randomly split into two disjoint subsets, each having approximately half the segments for each language: (1) *lre11-train*, used to train specific models for the newly added languages; and (2) *lre11-dev*, used to estimate backend and fusion parameters for the joint submission, and to evaluate the performance of single subsystems and fusions during development (see Section 3 for details).

The lre11-train subset contained around 25 minutes of speech per target language, which did not allow to train robust models. So, additional training data were collected for the newly added languages. Voice-Of-America (VOA) data provided for the 2009 NIST LRE were explored in first place, starting from the labels provided by NIST. Music and fragments in English were automatically detected and filtered out, retaining only telephone-channel speech fragments. Around two hours of Lao were extracted this way. Databases distributed

[1] FoCal toolkit: http://sites.google.com/site/nikobrummer/focal
[2] Bosaris toolkit: http://sites.google.com/site/bosaristoolkit

by the LDC were explored in second place. Some of them contained conversational telephone speech (LDC2006S45 for Arabic Iraqi and LDC2006S29 for Arabic Levantine), whereas others contained broadcast news with fragments of telephone speech (LDC2000S89 and LDC2009S02 for Czech). In both cases, segments containing telephone speech were extracted with no further processing.

The remaining materials were extracted from wideband broadcast news recordings, dowsampled to 8 kHz and applied the *Filtering and Noise Adding Tool*[3] (FANT) to get a frequency characteristic as defined by ITU for telephone equipment. The COST-278 Broadcast News database [3] 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 [4]. Finally, new broadcasts were *captured* from video archives in TV websites to get speech segments in Arabic Maghrebi (Arrabia TV, http://www.arrabia.ma) and Polish (Telewizja Polska, TVP INFO, http://tvp.info). TV broadcasts were fully audited, so that only those segments that were subjectively judged as containing clean speech were selected for training. We were not able to collect by any means additional training materials for Punjabi. Hereafter, the dataset collected for the newly added target languages will be called *BLZ-train*.

2.2. Training data

Besides the shared datasets, BLZ partners were free to use any other data for building their systems (an interesting way to get complementary systems for fusion), with a single constraint: not using development data for training. In all cases, training data comprised CTS from previous LRE and other sources (e.g. Switchboard), narrowband speech segments extracted from VOA broadcasts provided by NIST for the 2009 LRE, the lre11-train and the BLZ-train datasets. Each training subset featured a different language/dialect (including target and non-target languages) and/or source. Each site applied different criteria and filtering options, such as to keep the size of the datasets as small as possible, to limit the amount of data from repeated speakers, etc. Finally, EHU, L^2F and I3A defined 66, 43 and 61 training subsets, respectively (see [5] for details).

2.3. Development data

To make the process of tuning systems as robust and reliable as possible, development data comprised only segments audited by NIST. To cover all the target languages, the evaluation sets of the NIST 2007 and 2009 LREs (using only the segments corresponding to NIST 2011 LRE target languages), together with the lre11-dev subset, as defined in Section 2.1, were used. Three development subsets were defined: *dev30*, *dev10* and *dev03*, corresponding to nominal durations of 30, 10 and 3 seconds, containing 8539, 8343 and 8290 segments, respectively. Target languages showed large differences in the number of segments among each other. In particular, the newly added target languages were the less populated, with around 50 segments each, and thereby, they were the most likely to suffer from overtraining and/or robustness issues.

3. Systems

3.1. EHU subsystems

The EHU team developed two acoustic and three phonotactic language recognition subsystems. The two acoustic subsystems used a 56-dimensional feature vector, consisting of 7 static MFCC and 49 Shifted Delta Cepstrum (SDC) coefficients, under a 7-2-3-7 configuration. A gender independent 1024-mixture GMM was estimated on the training dataset and used as Universal Background Model (UBM). For each input utterance, UBM-MAP adaptation was applied and the zero-order and centered and normalized first-order Baum-Welch statistics were computed. The first acoustic subsystem applied the Linearized Eigenchannel GMM (LE-GMM) approach, also known as *Dot-Scoring* [6], including channel compensation [7]. The second acoustic subsystem applied the total variability iVector approach, as described in [8], but starting from the channel-compensated sufficient statistics computed for the Dot-Scoring subsystem. The EHU phonotactic subsystems followed a phone-lattice-SVM approach, using the TRAPs/NN phone decoders developed by the Brno University of Technology (BUT) for Czech (CZ), Hungarian (HU) and Russian (RU) [9]. Phone lattices were used to compute expected counts of phone n-grams, up to trigrams. Counts were stacked in a single vector and an L2-regularized L1-loss Support Vector Machine (SVM) classifier was estimated and applied, by means of LIBLINEAR [10].

3.2. L^2F subsystems

L^2F developed two acoustic and four phonotactic language recognition subsystems. One of the acoustic subsystems followed the GSV approach [11], using PLP-RASTA and SDC features, under a 7-1-3-7 configuration. A 1024-mixture GMM-UBM was trained on approximately 150 randomly selected speech segments per training subset. SVM language models were trained on MAP-adapted GMM supervectors by means of LIBLINEAR [10], using a linear kernel based on the Kullback-Leibler (KL) divergence. An iVector subsystem was also developed, following the approach described in [8], using the same PLP-RASTA SDC features and the same 1024-mixture GMM-UBM developed for the GSV subsystem. The total variability matrix was estimated on zero and first-order sufficient statistics computed on the training dataset, according to [12], the dimension of the total variability subspace being 400. Four phone-lattice-SVM phonotactic subsystems were developed, using L^2F phone decoders for European Portuguese (PT), Brazilian Portuguese (BR), European Spanish (ES) and American English (EN), based on the hybrid ASR system AUDIMUS [13]. Reduced vectors including only the counts of the 10000 most frequent n-grams (up to trigrams) were used. For each target language and each phone decoder, an L2-regularized SVM classifier was trained on the corresponding set of training vectors, by means of LIBLINEAR [10].

3.3. I3A subsystems

The I3A team developed two acoustic subsystems. The first one followed the implementation of the iVector approach described in [8]. Acoustic vectors included 7 static MFCC and 49 SDC coefficients computed under a 7-1-3-7 configuration. Vocal Tract Length Normalization and Cepstral Mean and Variance Normalization were applied in MFCC computation. A 2048-mixture GMM-UBM was used. Both the GMM-UBM and the total variability matrix were trained on the whole training dataset. The distributions of iVectors for individual languages were modeled by Gaussian distributions with a single within-class full covariance matrix shared by all the languages. Only target languages were modeled in this step, using (when possible) 500 speech segments per language. The second I3A subsystem followed the Joint Factor Analysis (JFA) approach

[3] http://dnt.kr.hs-niederrhein.de/download.html

described in [14], using the same 56-dimensional acoustic features and the same 2048-mixture GMM-UBM developed for the iVector subsystem. Two factors were defined, one for the language and one for the channel. Thus, a channel compensated model for each language was obtained. The whole training dataset was used to estimate model parameters. Finally, each utterance was scored via linear scoring, as proposed in [15].

3.4. The BLZ submission

The BLZ submission is summarized in Table 1. Backend and fusion parameters were estimated and applied separately for each nominal duration, under two different configurations: (1) Gaussian backend, training datasets: dev10 + dev30 for 10- and 30-second segments, dev03 + dev10 + dev30 for 3-second segments; and (2) *zt-norm* + Gaussian backend, training datasets: dev30 for 30-second segments, dev10 for 10-second segments and dev03 for 3-second segments. The second configuration was only applied to the third contrastive system, because it yielded a slight improvement on the development set.

Table 1: BLZ primary and contrastive systems: configuration (see details in the paper) and fused subsystems.

		Subsystems		
System	**Config**	**EHU**	L^2F	**I3A**
Pri	(1)	Phone-CZ Phone-HU Phone-RU DotScoring	Phone-PT Phone-BR	iVector JFA
Con1	(1)	Phone-CZ Phone-HU Phone-RU DotScoring iVector	Phone-PT Phone-BR Phone-EN Phone-ES GSV iVector	iVector JFA
Con2	(1)	Phone-RU	Phone-ES	JFA
Con3	(2)	Phone-CZ Phone-HU Phone-RU DotScoring	Phone-PT Phone-BR	iVector JFA

The EHU, L^2F and I3A subsystems produced 66, 43 and 24 scores, respectively (one score per trained model). These scores were taken as input by the backend, which output 24 log-likelihoods, one per target language. A Gaussian backend (preceded by a *zt-norm* for the third contrastive system) was applied in all cases. Finally, the resulting $N \times 24$ log-likelihood values (N: number of subsystems) were fused to get 24 calibrated scores for which a minimum expected cost Bayes decision was made, according to application-dependent language priors and costs. Calibration/fusion models were estimated by applying linear logistic regression under a multiclass paradigm [16], by means of the FoCal toolkit.

3.5. Selection of subsystems

To select the best combinations of subsystems, the development set was split in two halves, the first one being used to estimate backend and fusion parameters and the second to generate a set of trials, on which the performance measure, as defined in the Evaluation Plan, was computed, using the Bosaris toolkit. In fact, to have a more robust measure of system performance, 10 random partitions (always the same) were defined and the average performance was computed on them. This strategy pursued (via random subset selection) the same goal than a 2-fold cross-validation strategy, but providing a better balance between the size of the evaluation subset (large enough for the results to be reliable) and the number of partitions considered in the average (for statistical significance).

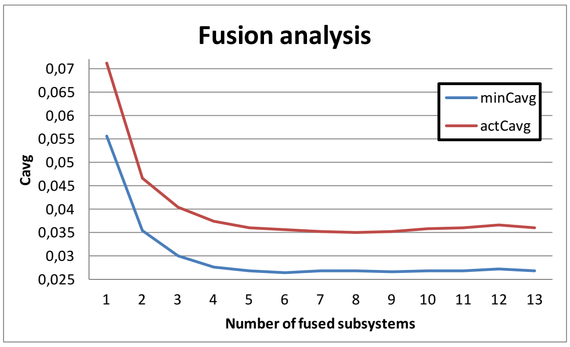

Figure 1: Actual and minimum average cost (2011 LRE definition) on the development set (30-second segments) for the optimal fusions of k subsystems according to a greedy algorithm.

Decisions were made based on system performance for the subset of 30-second segments, applying a Gaussian backend and discriminative multiclass fusion. An exhaustive search to find the best combination of k subsystems out of 13 becomes unfeasible because of the huge computational cost it involves for values of k greater than 4. Thus, a faster greedy strategy was applied: the best combination of k subsystems was determined by extending the best combination of $k - 1$ subsystems with each one of the available subsystems, and the combination that yielded the best performance was selected. Though this should generally lead to suboptimal solutions, we found that the best greedy k-combinations for k = 1, 2, 3 and 4 matched the optimal ones (those previously found with an exhaustive search). The actual and minimum average cost (C_{avg}) for the optimal combinations under the greedy approach are graphically shown in Figure 1. For the primary system, the best overall combination was selected according to the evolution of the actual cost. The combination involving eight subsystems was chosen because the actual cost, which had monotonically decreased to that point, began to increase for combinations of higher order. The first contrastive system fused all the subsystems, aiming to check whether it outperformed the selection approach described above; the second consisted of the best fusion of 3 subsystems containing one subsystem per site, as a kind of minimal (computationally less expensive) approach; finally, the third contrastive system fused the same set of subsystems selected for the primary system, but under a different backend configuration.

4. Results

The official results obtained by BLZ systems in all the evaluation tracks, in terms of the *traditional* C_{avg} and the new C_{avg} defined for the NIST 2011 LRE, are shown in Table 2. Note that both measures are highly correlated, suggesting that systems performing best for the 24 most confusable language pairs are also the best for all the language pairs. It seems that the new measure used in the NIST 2011 LRE does not provide additional information to that provided by the traditional pooled measure. Instead, it introduces some uncertainty when comparing systems, since the set of 24 pairs yielding the worst min-C_{avg} is generally different for each one. Note also that the average cost attained on the development set (see Figure 1) was remarkably lower than that found on the evaluation set. A mismatch between the development and evaluation datasets, that could be critical for some languages with few or less reliable data, may explain this result and may also explain the poor calibration achieved by most systems and the success of the zt-norm in the 30-second track (BLZ Contrastive System 3).

Table 2: Official NIST 2011 LRE results for the BLZ systems.

	All pairs		**24 worst pairs**	
30s	min-C_{avg}	act-C_{avg}	min-C_{avg}	act-C_{avg}
Pri	0.0081	0.0156	0.0573	0.0884
Con1	0.0079	0.0159	0.0568	0.0919
Con2	0.0099	0.0165	0.0658	0.0914
Con3	0.0071	**0.0139**	0.0509	**0.0764**
10s	min-C_{avg}	act-C_{avg}	min-C_{avg}	act-C_{avg}
Pri	0.0259	0.0346	0.1103	0.1343
Con1	0.0243	**0.0328**	0.1089	**0.1310**
Con2	0.0346	0.0423	0.1371	0.1543
Con3	0.0262	0.0347	0.1133	0.1322
3s	min-C_{avg}	act-C_{avg}	min-C_{avg}	act-C_{avg}
Pri	0.0982	0.1185	0.2382	0.2709
Con1	0.0909	**0.1073**	0.2250	**0.2511**
Con2	0.1116	0.1207	0.2521	0.2638
Con3	0.1128	0.1426	0.2705	0.3160

It must be noted that BLZ systems attained very competitive performance in the 30-second track, for which the selection of subsystems was optimized. In fact, BLZ systems were among the best systems submitted to the NIST 2011 LRE. Besides using cutting edge approaches in some of the fused subsystems, high performance can be partly explained by the effort devoted to collecting and designing suitable datasets for training and development, partly by the use of diverse heterogenous subsystems for fusion and partly by the effectiveness of the greedy algorithm applied to select the best combination of subsystems for fusion. Regarding this, note that the subset of subsystems found optimal on the development set (BLZ primary) outperformed the fusion of all the subsystems (BLZ contrastive 1) in the 30-second track. The surprisingly high performance attained by the BLZ third contrastive system tell us just about the importance of a smart use of the development data, but it may be also related to an error in the computation of I3A iVector scores for 10- and 3-second segments (remind that under configuration (1), the backend and fusion parameters for the 30-second track were estimated on dev10 + dev30), which also explains the degraded performance of BLZ systems in those tracks when compared to other submissions.

Post-key results were obtained by applying the greedy search for the best fusion on the subset of 30-second segments of the evaluation dataset, under the backend configuration (1). The best fusion obtained this way yielded min-C_{avg} = 0.0576 and act-C_{avg} = 0.0776 and included just 4 subsystems: EHU-Phone-CZ, EHU-Phone-RU, EHU-iVector and I3A-JFA. This is the maximum performance that can be attained with the official BLZ scores just by selecting the best possible combination of subsystems under configuration (1). The best fusion found on the development set was much more populated (i.e. more costly) and provided worse performance, which tell us again about a mismatch between the development and evaluation datasets. Finally, to further explore the potential performance attainable with BLZ systems, we ran the greedy search again using the amended scores of the I3A iVector subsystem and the backend configuration (2). This approach led to improved performance in the 30-second track: min-C_{avg} = 0.0522 and act-C_{avg} = 0.0709, and included 6 subsystems: the same of the BLZ primary system except for EHU-Dot-Scoring and L^2F-Phone-BR.

5. Acknowledgements

The work of EHU was funded by the Spanish MICINN under project TIN2009-07446. The work of Alberto Abad was supported by FCT (INESC-ID multi-annual funding) through the PIDDAC Program funds and the EU-funded project *EUTV Adaptive Media Channels*. The work of I3A was funded by the Spanish MICINN under project TIN2011-28169-C05-02.

6. References

[1] *The NIST 2011 LRE Plan*, http://www.nist.gov/itl/iad/mig/upload/LRE11_EvalPlan_releasev4.pdf.

[2] 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. Fränti, “Multi-site Heterogeneous System Fusions for the Albayzin 2010 Language Recognition Evaluation,” in *Proceedings of IEEE ASRU*, 2011.

[3] A. Vandecatseye, J.-P. Martens, J. Neto, H. Meinedo, C. Garcia-Mateo, J. Dieguez, 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 2004*, Lisbon, Portugal, 2004, pp. 873–876.

[4] L. J. Rodriguez-Fuentes, M. Penagarikano, A. Varona, M. Diez, and G. Bordel, “The Albayzin 2010 Language Recognition Evaluation,” in *Proceedings of Interspeech*, 2011, pp. 1529–1532.

[5] L. J. Rodríguez-Fuentes, M. Penagarikano, A. Varona, M. Diez, G. Bordel, A. Abad, D. Martínez, J. Villalba, A. Ortega, and E. Lleida, “The BLZ Systems for the 2011 NIST Language Recognition Evaluation,” in *NIST 2011 LRE Workshop Booklet*, Atlanta, USA, December 6-7 2011.

[6] A. Strasheim and N. Brümmer, “SUNSDV system description: NIST SRE 2008,” in *NIST 2008 SRE Workshop Booklet*, 2008.

[7] N. Brümmer, A. Strasheim, V. Hubeika, P. Matejka, L. Burget, and O. Glembek, “Discriminative Acoustic Language Recognition via Channel-Compensated GMM Statistics,” in *Proceedings of Interspeech*, Brighton, UK, 2009, pp. 2187–2190.

[8] 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.

[9] P. Schwarz, “Phoneme recognition based on long temporal context,” Ph.D. dissertation, Faculty of Information Technology, Brno University of Technology, 2008.

[10] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, “LIBLINEAR: A library for large linear classification,” *Journal of Machine Learning Research*, vol. 9, pp. 1871–1874, 2008, software available at http://www.csie.ntu.edu.tw/∼cjlin/liblinear.

[11] W. M. Campbell, D. E. Sturim, and D. A. Reynolds, “Support vector machines using GMM supervectors for speaker verification,” *IEEE Signal Processing Letters*, vol. 13, no. 5, pp. 308–311, 2006.

[12] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel, “A Study of Inter-Speaker Variability in Speaker Verification,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 16, no. 5, pp. 980–988, July 2008.

[13] H. Meinedo, A. Abad, T. Pellegrini, I. Trancoso, and J. Neto, “The L2F Broadcast News Speech Recognition System,” in *Proceedings of FALA 2010: VI Jornadas en Tecnologia del Habla and II Iberian SLTech Workshop*, Vigo (Spain), 10-12 November 2010.

[14] P. Kenny, “Joint Factor Analysis of Speaker and Session Variability: Theory and Algorithms,” CRIM, Tech. Rep. CRIM-06/08-13, 2005, available at http://www.crim.ca/perso/patrick.kenny/.

[15] O. Glembek, L. Burget, N. Dehak, N. Brümmer, and P. Kenny, “Comparison of Scoring Methods Used in Speaker Recognition with Joint Factor Analysis,” in *Proceedings of IEEE ICASSP*, Taipei, April 2009, pp. 4057–4060.

[16] N. Brümmer and D. van Leeuwen, “On calibration of language recognition scores,” in *Proceedings of Odyssey - The Speaker and Language Recognition Workshop*, 2006, pp. 1–8.