The EHU Systems for the NIST 2011 Language Recognition Evaluation

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

mikel.penagarikano@ehu.es

Abstract

This paper describes the systems developed by the Software Technologies Working Group of the University of the Basque Country (EHU) for the NIST 2011 Language Recognition Evaluation (LRE). One primary and three contrastive systems were submitted, all of them fusing five component subsystems: a Linearized Eigenchannel GMM (LE-GMM) subsystem, an iVector subsystem and three phone-lattice-SVM subsystems based on the publicly available BUT decoders for Czech, Hungarian an Russian. The four submitted systems were identical except for the backend approach and the development dataset used to estimate the backend and fusion parameters. Multiclass discriminative fusion was performed separately for each nominal duration. A development set was defined, including the evaluation sets of LRE07 and LRE09 and the development data provided by NIST for 9 additional languages in the 2011 campaign. The official results, which were among the best submitted to the evaluation, are presented and briefly discussed. Post-key analyses are also addressed in the paper, including the performance attained by component subsystems and a study of their contribution to fusion performance by means of a greedy selection procedure.

Index Terms: Spoken Language Recognition, NIST 2011 LRE, Gaussian Backend, Multiclass Discriminative Fusion

1. Introduction

This paper describes the systems developed by the Software Technologies Working Group (GTTS, http://gtts.ehu.es) of the University of the Basque Country (EHU) for the NIST 2011 Language Recognition Evaluation (LRE). Currently, most approaches to spoken language recognition can be classified either as acoustic or phonotactic, depending on the features used to model target languages. Acoustic systems are based on short-time spectral characteristics of the audio signal, whereas phonotactic systems use sequences or lattices of tokens produced by phone recognizers. Both approaches provide complementary information and their fusion usually leads to the best results. The EHU submission to the NIST 2011 LRE aimed to take advantage from this complementarity, by combining both types of systems. Two acoustic and three phonotactic subsystems were fused: a Linearized Eigenchannel GMM (LE-GMM) subsystem, an iVector subsystem and three Phone-SVM subsystems based on the Brno University of Technology (BUT) phone decoders for Czech, Hungarian and Russian.

The NIST 2011 LRE featured 24 target languages, some of them already used in previous evaluations (Bengali, Dari, English American, English Indian, Farsi/Persian, Hindi, Mandarin, Pashto, Russian, Spanish, Tamil, Thai, Turkish, Ukranian and Urdu), whereas the remaining ones (Arabic Iraqi, Arabic Levantine, Arabic Maghrebi, Arabic MSA, Czech, Lao, Punjabi, Polish and Slovak) had been never used before as target languages. As for previous NIST evaluations, test segments of three nominal durations (30, 10 and 3 seconds) were evaluated separately. More detailed information about the NIST 2011 LRE can be found in [1].

The main novelty of the NIST 2011 LRE with regard to previous evaluations was the focus on the discrimination between pairs of languages (note that 276 different pairs can be defined on a set of 24 target languages). This was emphasized by defining a new performance metric which considered only the 24 language pairs for which system performance (assuming a perfect calibration) was worst. This meant that if a single language was poorly modeled, a high number of confusable pairs (involving that language) could appear and cause performance to drop drastically. Thus, the availability of training and development data to provide a suitable coverage of all the target languages (in particular, of those newly added in this evaluation) was critical to obtain good performance under the new metric.

The rest of the paper is organized as follows. The datasets used for training and development are described in Section 2. Section 3 describes the most relevant features of the acoustic and phonotactic subsystems on which the EHU submission to the NIST 2011 LRE was based, along with the backend and fusion approaches. Finally, the official results obtained by EHU systems in the NIST 2011 LRE and post-key experiments aiming to study fusion performance in detail are presented and briefly discussed in Section 4.

2. Train 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 of the newly added target languages, except for Lao, for which only 93 segments were provided. We augmented the dataset with 10- and 3-second segments extracted from the original 30-second segments. Hereafter, we will refer to this dataset as *lre11*.

For a better coverage of target languages, we randomly split lre11 into two disjoint subsets (each having approximately half the segments for each language): *lre11-train* was used to train specific models for the newly added languages, and *lre11-dev* was used to estimate backend and fusion parameters for the EHU submission. However, splitting lre11 in two halves may lead to data sparsity and robustness issues. Note that each subset amounted to around 25 minutes of speech per target language, which may be enough to estimate backend parameters, but probably not enough to train robust models. In the context of a joint submission to the NIST 2011 LRE, the INESC-ID Spoken Language Systems Laboratory (L^2F), the University of Zaragoza and the University of the Basque Country collaborated in order to share, acquire and, whenever necessary, filter speech data for the newly added languages. In some cases we collected telephone speech directly from the source. When this was not pos-

sible, we used broadcast news speech, downsampled it to 8 kHz and applied the *Filtering and Noise Adding Tool*[1] (FANT) to get a frequency characteristic as defined by ITU for telephone equipment[2].

The Voice-of-America (VOA) corpus used for the 2009 NIST LRE was explored in first place, starting from the labels provided by NIST. Music and fragments in English were automatically detected and filtered out, and telephone-channel speech fragments were extracted. Around two hours of Lao were extracted this way. Then we used databases distributed by the LDC, some of them containing conversational telephone speech (LDC2006S45 for Arabic Iraqi and LDC2006S29 for Arabic Levantine) and others containing 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, dowsampling them to 8 kHz and applying FANT to simulate a telephone channel. The COST-278 Broadcast News database [2] 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 [3]. Finally, broadcasts were also *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 subjectively judged to contain clean speech were selected for training. We were not able to collect by any means additional training materials for Punjabi, so that a single model (trained on just 55 30-second segments) was used for this language.

2.2. Training data

Training data included Conversational Telephone Speech (CTS) from previous LRE (Call-Friend, OHSU, NIST 2007 LRE development corpus) and narrow-band speech segments extracted from VOA broadcasts provided by NIST for the 2009 LRE [4]. For the newly added target languages, the lre11-train corpus and additional training data collected from several sources (see Section 2.1) were used. We ended up with 66 subsets, corresponding to different languages/dialects (including target and non-target languages) and different sources. We trained a different model on each subset, which means that models account not only for the spoken language but also for the channel and other factors related to the source from which the speech data were drawn.

2.3. Development data

The criterion applied to define the development set was making the process of tuning systems as robust and reliable as possible, so we decided to use only segments audited by NIST. To cover all the target languages, the evaluation sets of the NIST 2007 and 2009 LREs (only the segments corresponding to NIST 2011 LRE target languages), together with the lre11-dev subset, as defined in Section 2.1, were used. We defined three development subsets: *dev30*, *dev10* and *dev03*, corresponding to nominal durations of 30, 10 and 3 seconds, containing 8539, 8343 and 8290 segments, respectively. Table 1 shows the distribution of segments in *dev30* with regard to target languages and sources. Note that few development data (around 50 segments) were available for the newly added target languages, thereby being the most likely to suffer from overtraining and/or robustness issues.

Table 1: Development set (30-second segments): distribution with regard to target languages and sources.

Language	**LRE 2007 (eval)**	**LRE 2009 (eval)**	**LRE 2011 (lre11-dev)**	**Total**
Arabic Iraqi	-	-	48	48
Arabic Levantine	-	-	49	49
Arabic Maghrebi	-	-	54	54
Arabic MSA	-	-	51	51
Bengali	80	43	-	123
Czech	-	-	56	56
Dari	-	389	-	389
English American	80	896	-	976
English Indian	160	574	-	734
Farsi/Persian	80	390	-	470
Hindi	160	667	-	827
Lao	-	-	41	41
Mandarin	158	1015	-	1173
Punjabi	32	9	45	86
Pashto	-	395	-	395
Polish	-	-	46	46
Russian	160	511	-	671
Slovak	-	-	56	56
Spanish	240	385	-	625
Tamil	160	-	-	160
Thai	80	188	-	268
Turkish	-	394	-	394
Ukrainian	-	388	-	388
Urdu	80	379	-	459
Total	1470	6623	446	8539

3. The EHU Language Recognition Systems

3.1. Acoustic Subsystems

For the acoustic subsystems, the concatenation of 7 Mel-Frequency Cepstral Coefficients (MFCC) and the Shifted Delta Cepstrum (SDC) coefficients under a 7-2-3-7 configuration, were used as acoustic features. A gender independent 1024-mixture GMM was used as Universal Background Model (UBM). For each input utterance, UBM-MAP adaptation was applied. Finally, zero-order and centered and normalized first-order Baum-Welch statistics were computed.

3.1.1. Dot-Scoring Subsystem

The Linearized Eigenchannel GMM (LE-GMM) subsystem, that we briefly call *Dot-Scoring* subsystem, makes use of a linearized procedure to score test segments against target models [5]. The log-likelihood ratio between the target model and the UBM used for scoring can be approximated as follows:

$$score\,(f,l) = \log \frac{P\,(f|\lambda_l)}{P\,(f|\lambda_{ubm})} \approx \hat{m}_l^t \cdot \hat{x}_f \qquad (1)$$

where $\hat{m}_l$ denotes the centered and normalized channel-compensated MAP-means corresponding to language l, computed as follows:

$$\hat{m}_l = (\tau I + diag(n_l))^{-1}\,\hat{x}_l \qquad (2)$$

where $\tau = 16$ is the relevance factor, n_l are the zero-order statistics for language l and $\hat{x}_l$ and $\hat{x}_f$ are the channel-

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

[2] Thanks to Alberto Abad from L^2F for doing all the filtering tasks.

compensated first-order statistics corresponding to language l and target signal f, respectively. Channel compensation was performed by using Niko Brümmer's recipe [6]. The channel matrix was estimated using only data from target languages.

3.1.2. iVector Subsystem

The estimation of the total variability matrix T and the computation of iVectors started from the channel-compensated sufficient statistics computed for the Dot-Scoring subsystem. This is not the common procedure, since compensation is usually performed in the iVector space, but we had a hardware issue[3] and no time to reestimate Baum-Welch statistics for training the T matrix. We had the iVector software prepared, so we decided to go ahead with this alternative computation method. Except for the compensation of statistics, all the computations were performed as in [7]. The total variability matrix was estimated using only data from target languages. The iVector scores were computed as follows:

$$score\,(f,l) = \mathcal{N}\,(w_f; \mu_l, \Sigma) \tag{3}$$

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

3.2. Phonotactic Subsystems

Three phonotactic subsystems were developed under a phone-lattice-SVM approach. Given an input signal, an energy-based voice activity detector was applied in first place, which split and removed long-duration non-speech segments. Then, the Temporal Patterns Neural Network (TRAPs/NN) phone decoders developed by the Brno University of Technology (BUT) for Czech (CZ), Hungarian (HU) and Russian (RU) [8], were applied to perform phone tokenization. Regarding channel compensation, noise reduction, etc. the three subsystems relied on the acoustic front-end provided by BUT decoders.

BUT decoders were configured to produce phone posteriors that were converted to phone lattices by means of HTK [9] along with the BUT recipe, on which expected counts of phone n-grams were computed using the *lattice-tool* of SRILM [10]. Finally, a Support Vector Machine (SVM) classifier was applied, SVM vectors consisting of counts of features representing the phonotactics of an input utterance. In this work, phone n-grams up to $n = 3$ were used, weighted as in [11]. L2-regularized L1-loss support vector classification was applied, by means of LIBLINEAR [12], whose source code was slightly modified to get regression values.

3.3. The EHU submission

The EHU submission consisted of one primary and three contrastive systems, fusing the 5 subsystems described above, under four different configurations, depending on the type of backend and on the datasets used to estimate backend and fusion parameters for nominal durations 10 and 3 (see Table 2). Note that the backend and fusion models were estimated and applied separately for each nominal duration.

Each subsystem produced 66 scores (one score per trained model), that were taken as input by the backend, which output 24 log-likelihoods (one per target language). A Gaussian backend, preceded by an optional *zt-norm* [13], was applied in all cases. Though discriminative backends were also tried, the (generative) Gaussian backend outperformed them in most cases, probably due to a lack of samples which led to overtraining on the development set used in the experiments. Finally, the resulting 5×24 log-likelihood values were fused by applying linear logistic regression, under a multiclass paradigm, to get 24 calibrated scores for which a minimum expected cost Bayes decision was made, according to application-dependent language priors and costs. We also tried pairwise backends and fusions but they did not provide significant improvements with regard to the basic multiclass approach (much easier to implement). The *FoCal* toolkit was used to estimate and apply the backend and calibration/fusion models [14].

[3] We lost the LRE11 data (speech signals, statistics, etc.), due to a mechanical failure of a disk, two weeks before the submission deadline.

Table 2: Backend and fusion configuration for the EHU systems submitted to the NIST 2011 LRE.

System	***zt-norm***	**Backend & Fusion Training Dataset**		
		30s	10s	3s
Pri	No	dev30	dev10	dev03
Con1	No	dev30	dev10+dev30	dev03+dev10+dev30
Con2	Yes	dev30	dev10	dev03
Con3	Yes	dev30	dev10+dev30	dev03+dev10+dev30

4. Results

4.1. Overall Results

The actual and minimum average costs for the EHU systems are shown in Table 3, in terms of: (1) the C_{avg} computed for all the language pairs (full C_{avg}), which approximately matches the traditional C_{avg} used in previous LRE; and (2) the new C_{avg} computed on the 24 language pairs with the highest min-C_{avg}. These figures are among the best attained in the NIST 2011 LRE, especially in the 30-second track. No significant differences in performance can be found among the four developed systems, except for 3-second segments, where the first and third contrastive systems clearly outperformed the primary and second contrastive systems, probably because the dataset used to estimate backend and fusion parameters was more reliable. It is worth noting the poor calibration achieved in all cases (differences between minimum and actual C_{avg} are remarkable), a common issue for most of the systems submitted to the evaluation, which may reveal either a mismatch between the development and evaluation datasets or, more probably, just the fact that the development dataset was not large enough for some languages. Finally, applying a zt-norm before the backend did not significantly affect performance.

4.2. Post-key experiments

Results have been studied in detail by measuring the performance of component subsystems and fusions involving the three phonotactic subsystems and the two acoustic subsystems under the configuration applied for the primary system (see Table 4). The phonotactic subsystem based on the BUT decoder for Russian (Phone-RU) yielded the best performance among component subsystems: act-$C_{avg} \times 100$ = 13.76. The fusion of phonotactic subsystems yielded a quite interesting act-$C_{avg} \times 100$ = 10.13, outperforming the fusion of acoustic subsystems by more than 3 absolute points. However, acoustic subsystems did actually provide complementary information, as the complete fusion reveals, with a 12% additional improvement over the fusion of phonotactic subsystems.

The information provided by each subsystem to the complete fusion was further studied by means of a greedy selec-

Table 3: Official NIST 2011 LRE results for the EHU systems.

	All pairs (Full C_{avg})		**24 worst pairs** (New C_{avg})	
30s	min-C_{avg}	act-C_{avg}	min-C_{avg}	act-C_{avg}
Pri	0.0093	**0.0169**	0.0615	**0.0895**
Con1	0.0094	0.0169	0.0615	0.0895
Con2	0.0091	0.0175	0.0608	0.0909
Con3	0.0091	0.0175	0.0607	0.0907
10s	min-C_{avg}	act-C_{avg}	min-C_{avg}	act-C_{avg}
Pri	0.0337	0.0408	0.1299	0.1477
Con1	0.0323	0.0403	0.1244	0.1455
Con2	0.0331	0.0412	0.1272	0.1468
Con3	0.0314	**0.0395**	0.1236	**0.1436**
3s	min-C_{avg}	act-C_{avg}	min-C_{avg}	act-C_{avg}
Pri	0.1160	0.1288	0.2554	0.2725
Con1	0.1107	**0.1205**	0.2397	0.2534
Con2	0.1162	0.1286	0.2552	0.2705
Con3	0.1087	0.1206	0.2331	**0.2528**

Table 4: Performance (in terms of $C_{avg} \times 100$) of the EHU acoustic and phonotactic subsystems and partial and complete fusions on the subset of 30-second segments of LRE11.

	New $C_{avg} \times 100$		Full $C_{avg} \times 100$	
	min	act	min	act
Phone-CZ	12.15	14.02	2.97	3.76
Phone-HU	11.96	14.28	2.71	3.62
Phone-RU	11.38	13.76	2.57	3.46
Phonotactic	7.73	10.13	1.47	2.28
Dot-Scoring	11.62	14.18	2.19	3.17
iVector	11.58	14.15	2.60	3.50
Acoustic	11.18	13.30	2.00	2.85
All	6.15	8.95	0.93	1.69

tion strategy: 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 on the evaluation dataset was selected. This way, the minimum and actual C_{avg} evolved as shown in Figure 1, where EHUCZ, EHUHU and EHURU refer to the phonotactic subsystems based on the BUT decoders for Czech, Hungarian and Russian, respectively, EHUDOT refers to the Dot-Scoring subsystem and EHUIVGEN to the (generative) iVector subsystem. The highest relative improvement (above 25%) was found when fusing the Phone-RU and iVector subsystems (significantly, involving one phonotactic and one acoustic subsystems, which are known to complement well each other). Then, adding Phone-CZ provided a still remarkable 10% improvement, but further fusions only introduced small reductions in act-C_{avg}. According to Figure 1, the fusion Phone-RU + iVector + Phone-CZ seems to provide the best balance between the attained performance and the computational cost.

5. Acknowledgements

This work has been supported by the University of the Basque Country under grant GIU10/18 and project US11/06, by the Government of the Basque Country under project S-PE11UN065 and by the Spanish MICINN under project TIN2009-07446. Mireia Diez is supported by a 4-year research fellowship from the Department of Education, Universities and Research of the Government of the Basque Country.

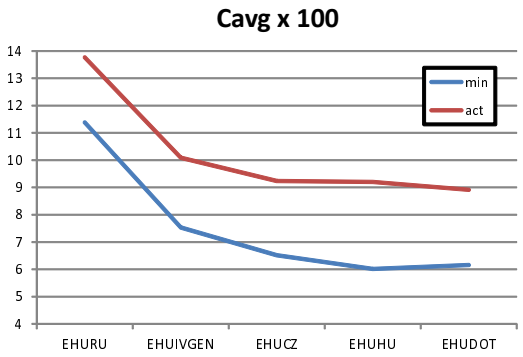

Figure 1: Actual and minimum C_{avg} on the development set (30-second segments) for the optimal fusions of k subsystems according to a greedy selection algorithm.

6. References

[1] *The 2011 NIST Language Recognition Evaluation Plan (LRE11)*, available at http://www.nist.gov/itl/iad/mig/upload/LRE11_EvalPlan_releasev4.pdf.

[2] 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 the LREC 2004*, 2004, pp. 873–876.

[3] 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.

[4] A. Martin and C. Greenberg, "The 2009 NIST Language Recognition Evaluation," in *Proceedings of Odyssey 2010: The Speaker and Language Recognition Workshop*, 2010, pp. 165–171.

[5] A. Strasheim and N. Brümmer, "SUNSDV system description: NIST SRE 2008," in *NIST Speaker Recognition Evaluation Workshop Booklet*, 2008.

[6] 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*, 2009, pp. 2187–2190.

[7] D. Martínez, O. Plchot, L. Burget, O. Glembek, and P. Matejka, "Language Recognition in iVectors Space," in *Proceedings of Interspeech*, 2011, pp. 861–864.

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

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

[10] A. Stolcke, "SRILM - an extensible language modeling toolkit," in *Proceedings of Interspeech*, 2002, pp. 257–286.

[11] F. Richardson and W. Campbell, "Language recognition with discriminative keyword selection," in *Proceedings of ICASSP*, 2008, pp. 4145–4148.

[12] 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.

[13] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas, "Score normalization for text-independent speaker verification systems," *Digital Signal Processing*, vol. 10, no. 1-3, pp. 42–54, 2000.

[14] 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.