University of the Basque Country (EHU) Systems for the 2011 NIST 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
University of the Basque Country, Spain

mikel.penagarikano@ehu.es

Abstract

This paper describes the systems developed by the Software Technologies Working Group (http://gtts.ehu.es) of the University of the Basque Country for the 2011 NIST Language Recognition Evaluation. Four different systems (one primary and three contrastive) were submitted, consisting of a fusion of five 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 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 LRE11. Systems were evaluated on 10 random partitions of the development set, using one half for estimating backend and fusion parameters and the other half for testing. The average cost as defined in the LRE11 evaluation plan was used as performance measure. The primary system yielded an actual average cost of 0.038 (± 0.002), being Hindi-Urdu, by far, the most challenging pair, with an actual average cost of 0.222.

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 2011 NIST Language Recognition Evaluation (LRE). Attending to preliminary evaluation on development data, this submission yields improved performance with regard to previous EHU submissions to NIST LRE in 2007 [1] and 2009 [2].

Currently, spoken language recognition systems can be classified under two main categories, depending on the features used to model target languages [3]: those using *low level* acoustic features and those using *high level* phonotactic features (recently, both approaches have been successfully mixed for a dialect recognition task [4]). 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 for the 2011 NIST LRE aims to take advantage from this complementarity, by combining both types of systems. Two acoustic and three phonotactic subsystems have been 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 2011 NIST LRE features 24 target languages, some of them already used in previous 2007 and/or 2009 LREs (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, Panjabi, Polish and Slovak) have been used as target languages for the first time in this evaluation.

The main novelty with regard to previous evaluations is the focus on the discrimination between pairs of languages (276 different pairs can be defined on a set of 24 target languages), which is emphasized with a new performance measure which takes into account *only the 24 most challenging language pairs*, i.e. those for which system performance is worst. This means that all the target languages should be suitably modeled and the discriminative power suitably balanced for all the pairs. In other words, if a single language was poorly modeled, a high number of confusable pairs (involving that language) could appear and cause performance to drop drastically. This is why the availability of training data to provide coverage for all the target languages (specially for those newly added in this evaluation) seemed critical to us. As for previous NIST evaluations, three test conditions are defined for three nominal durations of 30, 10 and 3 seconds. More detailed information about the 2011 NIST LRE can be found in [5].

The rest of the paper is organized as follows. Section 2 describes the datasets used for training and development, including details about the collection of training data for the target languages appearing for the first time in 2011. Section 3 describes the acoustic and phonotactic subsystems on which the EHU submission is based. Section 4 completes the picture by briefly describing the backend and fusion strategies and the subtle differences among the four systems submitted to 2011 NIST LRE. Finally, the average performance of individual subsystems and the fused systems on 10 random partitions of the development corpus are presented and briefly discussed in Section 5.

2. Train and development data

2.1. Data collection for the newly added target languages

NIST has provided a development dataset specifically collected for this evaluation, 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-

This work has been supported by the University of the Basque Country (EHU) under grant GIU10/18, by the Government of the Basque Country under program SAIOTEK (project S-PE10UN87) and by the Spanish MICINN under Plan Nacional de I+D+i (project TIN2009-07446, partially financed by FEDER funds).

second segments. Hereafter, we will refer to this dataset as *lre11*.

It is not strictly necessary to train models for all the target languages, since a backend can be trained and applied to map scores obtained for an arbitrary set of language models into log-likelihoods for the set of target languages. This means that lre11 may be used just to estimate the backend parameters for the newly added languages. By doing that, we assume that any target language can be parameterized and discriminated in terms of the available models.

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, and to evaluate system performance during development (see Section 5 for details).

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 2011 NIST 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 (that was the case of CTS databases and BN databases including telephone speech). When this was not possible, we used broadcast news speech, downsampled it to 8 kHz and applied the *Filtering and Noise Adding Tool* (FANT) [1] to filter speech data with a frequency characteristic as defined by ITU for telephone equipment [2].

The 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 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 COST278 Broadcast News database [6] 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 [7]. 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 reasonably clean speech segments were selected for training.

We were not able to collect by any means additional training materials for Panjabi, which means that a single model (trained on just 55 segments) was used for this language.

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

[2] Thanks to Alberto Abad from L^2F for doing all the filtering tasks on BN speech and VOA materials.

2.2. Train data

Train data have been obtained from several sources. Most of them were provided by NIST to LRE participants in past campaigns:

- Conversational telephone speech (CTS) from previous LRE: (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.
- Narrowband speech segments extracted from VOA broadcasts, which were provided by NIST for the 2009 LRE [8][9].
- The lre11-train corpus, as defined in Section 2.1, which amounts to half of the segments provided by NIST for the newly added target languages in the 2011 LRE.

As noted above, we considered the two following criteria: (1) there should be models for all the target languages, to prevent performance loss due to a lack of coverage; and (2) the amount of training data should be increased for the newly added languages (for which only the lre11-train corpus was available), to prevent robustness issues. Therefore, we collected additional training data for the newly added languages (we have already addressed this task in Section 2.1).

We ended up with 66 subsets, very heterogeneous in size and composition, corresponding to different languages/dialects, including target and non-target languages, and different sources (see Table 1). 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 2 shows the distribution of segments in the subset *dev30* with regard to the target languages and sources. Target languages show large differences in the number of segments amongst each other. The newly added target languages are the less populated (and thereby, the most likely to suffer from overtraining and/or robustness issues), with around 50 segments each.

3. The EHU Language Recognition Sub-systems

3.1. Acoustic Sub-systems

For the acoustic systems, 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 (Universal Background Model, UBM) was estimated by Maximum Likelihood on the training dataset, using binary mixture splitting, orphan mixture discarding and variance flooring. Finally, for each input utterance, UBM-MAP adaptation was applied and the centered zero-order and first-order Baum-Welch statistics were used as features.

Table 1: Training set: distribution of subsets (66), according to the language/dialect and source.

Source	Languages
LRE 2007 (CTS)	Bengali, English-American, English-Indian, Farsi, French, German, Hindi, Japanese, Korean, Mainland (Mandarin), Russian, Spanish-Caribbean, Spanish-Mexican, Spanish-NonCaribbean, Taiwan (Mandarin), Tamil, Thai, Urdu
LRE 2009 (VOA, CTS from BN)	Albanian, Amharic, Bangla, Creole, Dari, French, Georgian, Greek, Hausa, Hindi, Indonesian, Kinyarwanda/Kirundi, Korean, Lao, Mandarin, Ndebele, Oromo, Pashto, Persian/Farsi, Russian, Shona, Somali, Spanish, Swahili, Tibetan, Tigrigna, ttam (English), Turkish, Ukrainian, Urdu
LRE 2011 (lre11-train, CTS and/or BN)	Arabic-Iraqi, Arabic-Levantine, Arabic-Magrebi, Arabic-MSA, Czech, Lao, Panjabi, Polish, Slovak
LDC 2006S45 (CTS)	Arabic-Iraqi
LDC 2006S29 (CTS)	Arabic-Levantine
Arrabia TV (BN)	Arabic-Magrebi
Al Jazeera (BN)	Arabic-MSA
LDC 2000S89 (CTS from BN)	Czech
LDC 2009S02 (CTS from BN)	Czech
COST278 (BN)	Czech, Slovak
Telewizja Polska (BN)	Polish

Table 2: Development set (30-second segments): distribution with regard to the target language and source.

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
Panjabi	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

3.1.1. Dot Scoring Sub-system

The Linearized Eigenchannel GMM (LE-GMM) sub-system, that we briefly call *Dot-Scoring* sub-system, makes use of a linearized procedure to score test segments against target models [10]. The log-likelihood ratio between the target model and the UBM used for scoring can be approximated as follows:

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

where m_l denotes the vector of normalized means corresponding to language l and $\hat{x}_f$ is the vector of channel-compensated first-order statistics corresponding to the target signal f. Channel compensation was performed by using Niko Brümer's recipe [11]. The channel matrix was estimated using only data from target languages.

3.1.2. iVector Sub-system

The estimation of the total variability matrix T and the computation of iVectors started from the channel-compensated sufficient statistics obtained with the Dot-Scoring system. 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, computations were performed as in [12]. Once again, the total variability matrix was estimated using only data from target languages.

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

3.2. Phonotactic Sub-systems

Three phonotactic sub-systems 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) [13], were applied to perform phone tokenization. Non-phonetic units: *int* (intermittent noise), *pau* (short pause) and *spk* (non-speech speaker noise) were mapped to *sil* (silent pause). 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 lattices. Lattices, which encode multiple hypotheses with acoustic likelihoods, were then used to produce expected counts of phone n-grams, by means of HTK [14]. Finally, a Support Vector Machine 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 [15]. L2-regularized L1-loss support vector classification was applied, by means of LIBLINEAR [16], whose source code was slightly modified to get regression values.

4. The EHU submission

The EHU submission consists of one primary and three contrastive systems, fusing the 5 sub-systems described in Section 3 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 3). Note that backend and fusion were estimated and applied separately for each nominal duration. The four submitted systems have the same complexity and processing speed (see Section 4.1 for details).

Table 3: Main features of the EHU primary and contrastive systems. Backend and fusion were estimated and applied separately for each nominal duration.

System	***zt-norm***	**Backend & Fusion Train 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

Each sub-system produces 66 scores (one score per trained model). These scores are taken as input by the backend, which outputs 24 log-likelihoods, one per target language. A Gaussian backend, preceded by an optional *zt-norm* [17], has been applied in all cases. Though discriminative backends have been 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 are fused by applying linear logistic regression, under a multiclass paradigm, to get 24 calibrated scores for which a minimum expected cost Bayes decision is made, according to application-dependent language priors and costs. We have 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 has been used to estimate and apply the backend and calibration/fusion models [18, 19].

4.1. Processing times

Processing times were all measured on a computer with 2 Intel Xeon 5550 CPUs (x 4 cores x 2 turbo HT) running at 2.66GHz with 32GB of memory. Real-time factors for the five subsystems and the overall fused systems are shown in Table 4. For the iVector sub-system, the real-time factor only accounts for the iVector estimation from the total variability matrix and precomputed statistics, since it relies on the compensated statistics computed for the Dot-Scoring sub-system. Sub-processes with relatively small (negligible) run times, such as dot product, iVector scoring and SVM vector scoring, have not been taken into account. Processing times for the backend and fusion operations have been also omitted, since they are extremely fast. The overall fused systems run at 0.7295 times real time.

Table 4: Real-time factors of the five sub-systems and the corresponding sub-processes. The fused systems are obtained by sequentially running the five sub-systems, so the real-time factor is computed by adding the real-time factors of sub-systems.

Dot-Scoring		**0.0467**
	Acoustic Parameterization	0.0020
	Sufficient Statistics	0.0187
	Channel Compensation	0.0260
iVector		**0.0250**
Phone-SVM-CZ		**0.2114**
	Lattice Decoding	0.1267
	Expected Counts	0.0847
Phone-SVM-HU		**0.2300**
	Lattice Decoding	0.1517
	Expected Counts	0.0783
Phone-SVM-RU		**0.2164**
	Lattice Decoding	0.1327
	Expected Counts	0.0837
Fused Systems		**0.7295**

5. System performance: results on the development dataset

5.1. Evaluation methodology

To measure system performance, the development dataset can be split in two halves, the first 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, can be computed. Note, however, that if we consider a single partition, a positive or negative bias may be introduced. To have a more robust measure of system performance, we define 10 random partitions (always the same) and compute the average performance on them. This strategy pursues (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).

The above described strategy may introduce a positive bias if signals used for testing also appear in the subset used to estimate backend and fusion parameters. In this regard, note that for the 9 newly added target languages, 10-second segments in the development set were entirely extracted from 30-second segments, and 3-second segments were entirely extracted from 10-second segments. Moreover, we suspect that 10- and 3-second segments provided by NIST in the evaluation sets of the 2007 and 2009 LREs (which have been also included in the development set) were partly obtained using a similar procedure. This means that, for contrastive systems 1 and 3, whose development sets for nominal durations 10 and 3 consist of dev10 + dev30 and dev03 + dev10 + dev30, respectively, some signals may appear two or even three times. Due to these dependencies, performance results for contrastive systems 1 and 3 have been omitted [4].

5.2. Overall performance results

The actual and minimum average costs for the EHU primary system and the EHU contrastive system 2, along with the costs for the sub-systems involved in the corresponding fusions, in experiments on 10 fixed partitions of the development set, are shown in Table 5. The only difference between the primary and contrastive systems regards the introduction of a *zt-norm* before the backend in the contrastive system, which consistently leads to a slight (but not significant) improvement in performance.

[4] Regarding performance on 30-second segments, the contrastive system 1 is identical to the primary system, and the contrastive system 3 is identical to the contrastive system 2.

Table 5: Actual and minimum average costs for the EHU primary system, the EHU contrastive system 2 and the sub-systems involved in the respective fusions, in experiments on 10 fixed partitions of the development set.

	$C_{\text{avg}}^{\text{act}}$			$C_{\text{avg}}^{\text{min}}$		
	30s	10s	3s	30s	10s	3s
Primary	**0.038** ($\pm$0.002)	**0.084** ($\pm$0.005)	**0.209** ($\pm$0.009)	**0.029** ($\pm$0.002)	**0.067** ($\pm$0.003)	**0.179** ($\pm$0.007)
Dot-Scoring	0.071 ($\pm$0.005)	0.140 ($\pm$0.007)	0.276 ($\pm$0.010)	0.056 ($\pm$0.004)	0.116 ($\pm$0.005)	0.248 ($\pm$0.008)
iVector	0.086 ($\pm$0.006)	0.172 ($\pm$0.008)	0.304 ($\pm$0.010)	0.069 ($\pm$0.003)	0.144 ($\pm$0.005)	0.271 ($\pm$0.007)
Phone-SVM-CZ	0.078 ($\pm$0.005)	0.161 ($\pm$0.007)	0.321 ($\pm$0.011)	0.062 ($\pm$0.004)	0.139 ($\pm$0.007)	0.284 ($\pm$0.008)
Phone-SVM-HU	0.086 ($\pm$0.005)	0.160 ($\pm$0.006)	0.300 ($\pm$0.008)	0.068 ($\pm$0.004)	0.135 ($\pm$0.004)	0.264 ($\pm$0.003)
Phone-SVM-RU	0.073 ($\pm$0.005)	0.158 ($\pm$0.011)	0.300 ($\pm$0.009)	0.059 ($\pm$0.005)	0.133 ($\pm$0.008)	0.261 ($\pm$0.007)
Contrastive 2	**0.037** ($\pm$0.002)	**0.082** ($\pm$0.004)	**0.205** ($\pm$0.009)	**0.028** ($\pm$0.002)	**0.066** ($\pm$0.003)	**0.174** ($\pm$0.007)
Dot-Scoring	0.073 ($\pm$0.004)	0.135 ($\pm$0.008)	0.276 ($\pm$0.011)	0.056 ($\pm$0.003)	0.112 ($\pm$0.003)	0.243 ($\pm$0.008)
iVector	0.082 ($\pm$0.006)	0.169 ($\pm$0.007)	0.304 ($\pm$0.009)	0.067 ($\pm$0.005)	0.141 ($\pm$0.005)	0.271 ($\pm$0.008)
Phone-SVM-CZ	0.077 ($\pm$0.004)	0.160 ($\pm$0.008)	0.322 ($\pm$0.012)	0.061 ($\pm$0.004)	0.139 ($\pm$0.007)	0.282 ($\pm$ 0.007)
Phone-SVM-HU	0.082 ($\pm$0.004)	0.157 ($\pm$0.006)	0.297 ($\pm$0.008)	0.066 ($\pm$0.003)	0.132 ($\pm$0.004)	0.263 ($\pm$0.004)
Phone-SVM-RU	0.073 ($\pm$0.006)	0.154 ($\pm$0.009)	0.299 ($\pm$0.008)	0.058 ($\pm$0.005)	0.130 ($\pm$0.007)	0.262 ($\pm$0.007)

Systems are not perfectly calibrated, as the differences between $C_{\text{avg}}^{\text{act}}$ and $C_{\text{avg}}^{\text{min}}$ reveal. In fact, a perfect calibration may provide cost reductions ranging from 15% to 25%. Calibration issues probably arise due to a lack of samples in the development set for some of the target languages. This effect is more noticeable in these experiments than in the scores submitted to NIST 2011 LRE, because only one half of the development set (maybe unbalanced) has been used to estimate backend and fusion parameters. The dot-scoring sub-system consistently yields the best performance, followed by the Phone-SVM sub-systems for Russian, Czech and Hungarian. Finally, the iVector subsystem does not perform as well as expected, maybe because the precompensation issue commented in Section 3. In any case, the five sub-systems provide a reasonably good fusion. We found that the iVector sub-system contributed with complementary information, improving the fusion of the four other sub-systems, despite being trained on the same compensated statistics used for the dot-scoring sub-system.

5.3. Most challenging language pairs

Figure 1 shows the minimum and actual average costs for the 24 language pairs yielding the highest minimum average costs on dev30, when using the EHU primary system. The pair Hindi-Urdu yields, by far, the highest cost, with $C_{\text{avg}}^{\text{act}} = 0.222$, the next highest costs being one third that value: $C_{\text{avg}}^{\text{act}} \approx 0.073$ for the pairs English_American-English_Indian, Czech-Slovak and Arabic_Levantine-Arabic_Iraqui. There are only three more pairs with actual costs greater than 0.05: Dari-Farsi, Hindi-Panjabi and Panjabi-Urdu. In all cases, the involved pairs were expected to be highly confusable. However, the pair Hindi-Urdu accounts for a large fraction of the overall cost, so specific efforts should be devoted to analyze the reasons for this result and to study ways to improve the discrimination between both languages.

On the other hand, the EHU primary system seems to be reasonably well calibrated for some language pairs, but very poorly for other pairs. Table 6 shows the 5 worst calibrated pairs, in terms of the absolute and relative difference between $C_{\text{avg}}^{\text{act}}$ and $C_{\text{avg}}^{\text{min}}$. In all cases, one of the languages of the pair, or both, have few development utterances (less than 100, in most cases around 50, see Table 2), and *only half of them* (on average) are used to estimate backend and fusion parameters for each of the 10 partitions considered in these experiments. Therefore, as we suggested above, the lack of samples is the most plausible explanation for the calibration issues observed in Figure 1. We expect the scores to be better calibrated in the submission to NIST 2011 LRE, since backend and fusion parameters were estimated on *all* the development utterances.

Table 6: The 5 worst calibrated language pairs (on average) in absolute and relative terms, when using the EHU primary system on 10 fixed partitions of the development set (30-second segments).

Absolute $C_{\text{avg}}^{\text{act}} - C_{\text{avg}}^{\text{min}}$		**Relative** $C_{\text{avg}}^{\text{act}} / C_{\text{avg}}^{\text{min}} - 1$	
Arabic_Levantine, Arabic_MSA	0.032	Arabic_Levantine, Arabic_MSA	3.22
Lao, Thai	0.028	Polish, Slovak	2.88
Arabic_Iraqi, Arabic_Levantine	0.025	Bengali, Panjabi	2.24
Czech, Slovak	0.018	Bengali, English_Indian	2.24
Arabic_Iraqi, Arabic_MSA	0.018	Lao, Thai	1.75

6. Acknowledgements

We thank Alberto Abad from L^2F and the people from the University of Zaragoza, which collaborated with us for a joint submission to NIST 2011 LRE: their work, the enriching discussions and the fun time shared in a series of video-conferences have been key for the development of EHU systems. We also thank the NIST 2011 LRE organizers for their readiness to answer our questions.

7. References

[1] M. Penagarikano, G. Bordel, L. J. Rodriguez, and J. P. Uribe, "University of the Basque Country + Ikerlan System for NIST 2007 Language Recognition Evaluation," in *2007 NIST Language Recognition Evaluation (LRE) Workshop*, Orlando, Florida, USA, 2007.

[2] M. Penagarikano, A. Varona, M. Zamalloa, L. J. Rodriguez, G. Bordel, and J. P. Uribe, "University of the Basque Country + Ikerlan System for NIST 2009 Language Recognition Evaluation," in *2009 NIST Language Recognition Evaluation (LRE) Workshop*, Baltimore, MD, USA, 2009.

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

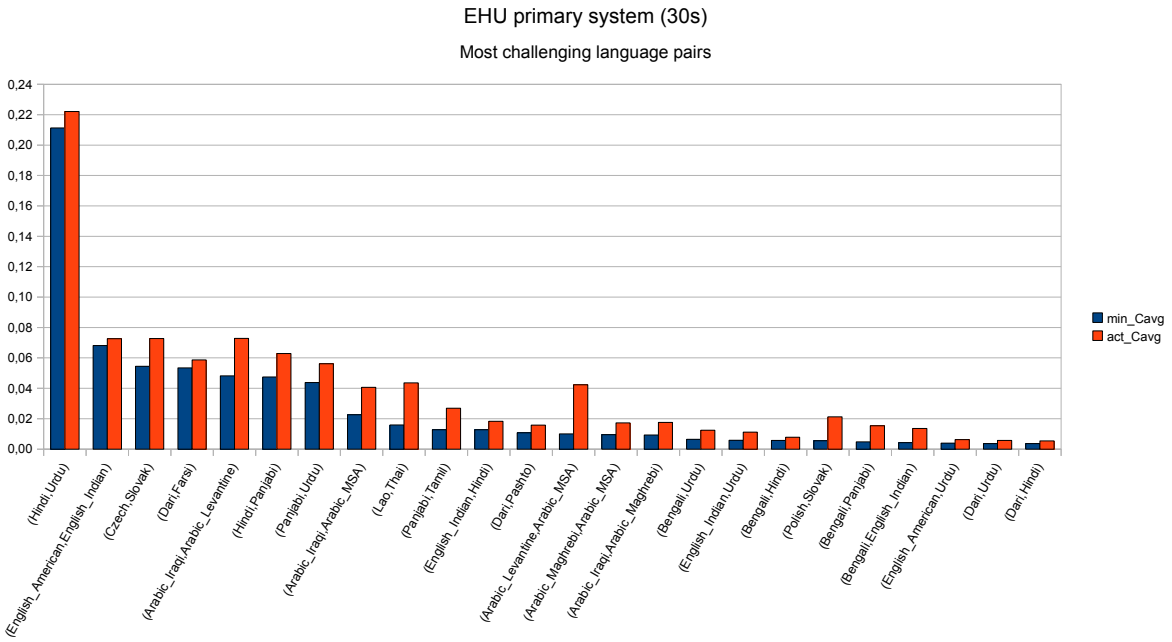

Figure 1: Minimum and actual average costs for the 24 language pairs yielding the highest minimum average costs when using the EHU primary system. Performance has been averaged on 10 fixed partitions of the development set (30-second segments).

[4] F. Biadsy, J. Hirschberg, and D. P. W. Ellis, "Dialect and accent recognition using phonetic-segmentation supervectors," in *Proceedings of Interspeech*, Firenze, Italia, August 28-31 2011, pp. 745–748.

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

[6] 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*, Lisbon, Portugal, 2004, pp. 873–876.

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

[8] *The 2009 NIST Language Recognition Evaluation Plan (LRE09)*, http://www.itl.nist.gov/iad/mig/tests/lre/2009/LRE09_EvalPlan_v6.pdf.

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

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

[11] 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.

[12] 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.

[13] P. Schwarz, "Phoneme recognition based on long temporal context," Ph.D. dissertation, Faculty of Information Technology, Brno University of Technology, http://www.fit.vutbr.cz/ schwarzp/publi/thesis.pdf, 2008.

[14] 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.

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

[16] 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.

[17] 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, January 2000.

[18] 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.

[19] N. Brümmer, L. Burget, J. Cernocky, O. Glembek, F. Grezl, M. Karafiat, D. van Leeuwen, P. Matejka, P. Schwarz, and A. Strasheim, "Fusion of heterogeneous speaker recognition systems in the STBU submission for the NIST speaker recognition evaluation 2006," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 7, pp. 2072–2084, 2007.