# Post-evaluation analysis and improvements with the L$^2$F Language Verification System submitted to NIST LRE 2009

*Alberto Abad*[1] *and Isabel Trancoso*[1,2]

[1]L$^2$F - Spoken Language Systems Lab, INESC-ID / [2]IST, Lisboa, Portugal
{Alberto.Abad,Isabel.Trancoso}@l2f.inesc-id.pt

## Abstract

The INESC-ID's Spoken Language Systems Laboratory (L$^2$F) Language Verification system submitted to the 2009 NIST Language Recognition evaluation is introduced in this paper. Then, as a sequence of the evaluation workshop and post-analysis of the results, the set of modifications that lead to significant performance gains are reported. Main differences between the original *submitted* system and the *post-evaluation* system consist of: 1) the kind and amount of training and development data considered for language model training and calibration and fusion, 2) the improvement of the acoustic based sub-systems and the reduction of the number of sub-systems that compose the whole system, and 3) the application of a better calibration and fusion scheme. Contrastive results of the *submitted* and the *post-evaluation* language recognition system for the different conditions in the evaluation are provided.

## 1. Introduction

The National Institute of Standards and Technology (NIST) has organized in the last years a series of evaluations in some relevant speech processing topics devoted to encourage language research activities.

In the 2009 NIST Language Recognition Evaluation (LRE09) the objective is to detect whether a target language is in fact spoken in a given speech segment. The number of possible target languages is 23. Three distinct test conditions are proposed depending on the possible set of competitive/non-target languages: "closed-set" (the set of non-target languages is the set of LRE09 target languages, minus the target language), "open-set" (the same as "closed-set", plus other "unknown" languages) and "language-pair" (the non-target language is a single language). Detailed information on the LRE09 campaign can be found in the evaluation plan document [1].

Language recognition (LR) approaches can generally be classified according to the kind of source of information that they rely on. The most successful systems are based on the exploitation of the acoustic phonetics, that is the acoustic characteristics of each language, or the phonotactics which are the rules that govern the phone combinations in a language.

This paper summarizes the LR system developed by the INESC-ID's Spoken Language Systems Laboratory (L$^2$F) for the LRE09 campaign and the post-evaluation efforts devoted to improve the LR system. Next Section 2 presents a description of the data used in this work. Section 3 describes the *submitted* language recognition system, starting by the phonotactic modules (subsection 3.1) and the acoustic ones (subsection 3.2). Then, the series of modifications and improvements introduced to the submitted system are described in Section 4. Finally, language verification results are provided for the *submitted* and for the *post-evaluation* systems for the different conditions.

## 2. Training, calibration and testing data

Data from previous evaluations and new data from Voice of America (VOA) radio broadcast [1] was made available for LR training and development.

### 2.1. Data for acoustic and phonotactic modeling

Language recognition acoustic models and phonotactic models used for the evaluation have been trained using *only* data from the VOA3 corpus provided for this evaluation.

For all target languages, approximately 15 hours of data from VOA3 automatically labeled as telephone data were extracted. Segments were classified according to their length in sets of approximately 30, 10 and 3 seconds. The number of files of each duration is approximately the same in every language.

A telephone band detector processing was applied to automatically classify the data for which this type of classification was not available. First, speech-non-speech segmentation was applied to the training data [2]. Then two scores were obtained, by averaging frame-based scores over the speech segment. The scores are band-energy ratios around 3400 Hz upper-bound of telephone band (similar to [3]) and 400 Hz lower-bound. Finally, the scores obtained for each speech segment were compared to fixed thresholds.

Notice that VOA3 includes data for all the 23 possible target languages of LRE09, except for the case of American English and Indian English that are not distinguished. We could find around 4.5 hours of Indian English in data sets of previous evaluations, but it was considered insufficient compared to the 15 hours used for all the other languages. Additionally, we were not very sure of the impact of using a different source of data just for one of the target languages. This was the motivation for using a unique set of data for training English models, both American and Indian without distinction.

Finally, an additional data set of approximately 15 hours was also extracted for "other" languages present in VOA3. This "other" languages data set was used to train phonotactic and acoustic models for a general language class corresponding to the "unknown" languages that are not part of the set of 23 possible target languages of this evaluation. Table 1 summarizes the data used for training the L$^2$F language recognition system.

### 2.2. Calibration and fusion data

Data from three different sources has been used for calibration and fusion of the LR system: VOA2 and VOA3 segments audited by LDC, VOA3 non-audited segments (like the ones of the training set, but different segments) and segments from previous

| Lang | 30 | 10 | 3 | Tot |
|---|---|---|---|---|
| amha | 688 | 685 | 685 | 2058 (14.7h) |
| bosn | 647 | 657 | 657 | 1961 (14.1h) |
| cant | 917 | 894 | 894 | 2705 (15.5h) |
| creo | 792 | 788 | 788 | 2368 (14.7h) |
| croa | 336 | 641 | 339 | 1316 (11.3) |
| dari | 902 | 907 | 907 | 2716 (15.1h) |
| engl(*) | 979 | 977 | 977 | 2933 (15.6h) |
| fars | 643 | 634 | 634 | 1911 (14.2h) |
| fren | 794 | 790 | 790 | 2374 (14.9h) |
| geor | 664 | 2000 | 664 | 3328 (14.1h) |
| haus | 764 | 759 | 759 | 2282 (14.7h) |
| hind | 653 | 654 | 654 | 1961 (14.3h) |
| kore | 994 | 998 | 998 | 2990 (15.8h) |
| mand | 1094 | 1102 | 1102 | 3298 (16.1h) |
| pash | 844 | 844 | 844 | 2532 (15.0h) |
| port | 762 | 749 | 749 | 2260 (14.9h) |
| russ | 636 | 649 | 649 | 1934 (14.4h) |
| span | 550 | 545 | 545 | 1640 (13.9h) |
| turk | 619 | 623 | 623 | 1865 (14.3h) |
| ukra | 1085 | 1088 | 1088 | 3261 (16h) |
| urdu | 696 | 704 | 704 | 2104 (14.5h) |
| viet | 985 | 982 | 982 | 2949 (15.6h) |
| other | 679 | 681 | 681 | 2041 (14.3h) |
| total | 17723 | 19351 | 17713 | 54787 (338h) |

Table 1: *Number of training speech segments extracted from the VOA3 corpus of each target language and total duration. (*) American English and Indian English are not distinguished.*

LRE evaluation sets. For every target language, approximately 4 hours of data have been selected and also split in 30 seconds, 10 seconds and 3 seconds segment duration.

Distinguished sets were used for American English and Indian English. Additionally, a set of approximately 6.9 hours of "other" languages (including the non-target languages of the training set and some additional ones) has been collected.

The total calibration and fusion corpus is composed of 19346 segments: 7815 of 30 seconds, 5911 of 10 seconds and 5620 of 3 seconds. A summary of this development data set is shown in Table 2.

### 2.3. Testing data

The LRE09 test set is used for LR assessment. The corpus is composed of 41793 speech segments: 14166 of 30 seconds, 13847 of 10 seconds and 13780 of 3 seconds.

## 3. The L$^2$F LRE *submitted* system

The complete L$^2$F language recognition system is the result of the fusion of eight language verification scores provided by 8 individual sub-systems: 4 phonotactic and 4 acoustic-based. In this section the 8 sub-systems and the calibration and fusion steps are described.

### 3.1. The PRLM-LR systems

The PRLM (Phone Recognition followed by Language Modeling) systems used for LRE09 exploit the phonotactic information extracted by four parallel tokenizers: European Portuguese, Brazilian Portuguese, European Spanish (Castilian) and American English. The tokenizers are MultiLayer Perceptrons (MLP) trained to estimate the posterior probabilities of the different

| Lang | LDC | VOA3 | LREold | Tot |
|---|---|---|---|---|
| amha | 1.4h | 2.6h | — | 4h |
| bosn | 1.6h | 2.5h | — | 4.1h |
| cant | — | 2.7h | 1h | 3.7h |
| creo | 1.6h | 2.6h | — | 4.2h |
| croa | 1.5h | 2h | — | 3.5h |
| dari | 1.6h | 2.6h | — | 4.2h |
| engl.a | — | 2h | 2h | 4h |
| engl.i | — | — | 4h | 4h |
| fars | — | 2.5h | 2h | 4.5h |
| fren | 1.6h | 2.6h | — | 4.2h |
| geor | 1.1h | 2.5h | — | 3.6h |
| haus | 1.6h | 2.6h | — | 4.2h |
| hind | — | 2.5h | 2h | 4.5h |
| kore | — | 2.9h | 2h | 4.9h |
| mand | — | 2.8h | 3h | 5.8h |
| pash | 1.6h | 2.6h | — | 4.2h |
| port | 1.4h | 2.6h | — | 4h |
| russ | — | 2.5h | 3h | 5.5h |
| span | — | 2.5h | 4h | 6.5h |
| turk | 1.6h | 2.5h | — | 4.1h |
| ukra | 1.6h | 2.8h | — | 4.4h |
| urdu | — | 2.5h | 1h | 3.5h |
| viet | — | 2.8h | 3h | 5.8h |
| other | — | 4.9h | 2h | 6.9h |
| total | 18.2h | 61.1h | 29h | 108.3h |

Table 2: *Development data set composed of different data sources: audited VOA2 and VOA3 data (LDC), non-audited voa3 data (VOA3) and previous LRE data sets (LREold).*

phonemes for a given input speech frame (and its context).

#### *3.1.1. Feature extraction*

The system combines four MLP outputs trained with Perceptual Linear Prediction features (PLP, 13 static + first derivative), PLP with log-RelAtive SpecTrAl speech processing features (RASTA, 13 static + first derivative), Modulation SpectroGram features (MSG, 28 static) and the Advanced Font-End from ETSI (ETSI, 13 static + first and second derivatives).

#### *3.1.2. Phonetic tokenizers/classifiers*

For this evaluation, it was necessary to re-train our phonetic classifiers with Broadcast News (BN) data downsampled at 8kHz, since our original classifiers were developed for BN data at 16 kHz.

The European Portuguese classifier was trained with 57 hours of BN data, and 58 hours of mixed fixed-telephone and mobile-telephone data. The Brazilian Portuguese classifier was trained with around 13 hours of BN data. The Spanish classifier used 14 hours of BN data. Finally, the English system was trained with the HUB-4 96 and HUB-4 97 data sets, that contain around 142 hours of TV and Radio Broadcast data.

The size of the neural networks of each tokenizer differs due to the different amounts of training data. In the case of the output layer, its size corresponds to the number of phonetic units of each language, plus silence (no additional sub-phonetic or context-dependent units have been considered [4]).

#### *3.1.3. Phonotactics modeling*

For every phonetic tokenizer, the phonotactics of each target language is modeled with a 3-gram model. For that purpose the

SRILM toolkit has been used [5].

In both training and test, the raw phonotactic sequence obtained by each tokenizer was filtered, in order to avoid spurious phone recognitions. Concretely, phones that appeared only once in the middle of long sequences of identical phones were deleted.

### 3.2. The GSV-LR systems

Acoustic methods for LR are usually preferred to phonotactic approaches since they are not limited by the need of well-trained phonetic tokenizers. Recently, a method generally known as GMM supervectors (GSV) [6] has been shown to be a successful approach for both speaker verification and language verification tasks.

GSV-based approaches consist of a mapping of each speech utterance to a high-dimensional vector and the use of these high-dimensional vectors for training and classification with a support vector machine (SVM). The mapping to the high-dimensional space is the result of stacking in a single supervector the parameters (usually the means) of an adapted GMM to the characteristics of a given speech segment. In language recognition, a binary SVM classifier is trained for each target language with supervectors of the target language as positive examples and supervectors of other non-target languages as negative examples. During test, the supervector of the testing speech utterance must be also obtained and a score for each target language is obtained with the binary classifiers.

The four GSV-LR sub-systems that compose the complete L$^2$F language recognition system are slight variations of the GSV approach. Concretely, two of the GSV systems differ in the linear kernel considered (different normalization of the Gaussian mixture means in their projection to the high dimensional space). The last two systems are derivations of the previous GSV, where the SVM models parameters are pushed back to the GMM domain as proposed in [7].

#### *3.2.1. Feature extraction*

The extracted features are Perceptual Linear Prediction static features with log-RelAtive SpecTrAl speech processing (RASTA), and a stacked vector of shifted delta cepstra (SDC) of the same RASTA features. Concretely, 7 RASTA static features and a 7-1-3-7 SDC parameter configuration are computed, resulting in a final feature vector of 56 components.

#### *3.2.2. GMM UBM and SVM modeling*

A GMM universal background model of 256 mixtures was trained with approximately 20 hours of speech randomly selected from the 30 seconds training speech segments.

Five iterations of Maximum a posteriori (MAP) adaptation are performed for each speech segment to obtain the high-dimensional vector of size 56x256. Then, previously to SVM training (or classification) the high-dimensional vectors are normalized in two different ways resulting in two different GSV-SVM sub-systems.

Linear SVM classifiers are trained for each target language (and for the two different mean normalizations) with the lib-SVM toolkit [8]. For each target language, all the training segments/supervectors are used as positive examples. The negative examples were randomly selected among the training data from the other languages, in order to achieve approximately 1.2 times the number of positive examples.

### 3.3. Calibration and fusion

Linear logistic regression (LLR) fusion and calibration of the 8 sub-systems has been done with the FoCal Multiclass Toolkit [9]. For each evaluation condition ("closed-set", "open-set" and different "language-pairs"), a separate calibration and fusion has been trained for the 30, 10 and 3 seconds length segments.

In both the "closed-set" and "open-set" condition, the same score is used for both American and Indian English. However, notice that in the data used for calibration and fusion these varieties are distinguished. Thus, we expected that some discriminative information can be extracted from the relations with the other languages.

In addition to the models trained in the 8 sub-systems for the 23 different target-languages, an additional model for every system was trained with the "other" languages set. The score obtained by these models is used for representing the "unknown" language score in the "open-set" condition.

The scores obtained for the two languages of interest in the "language-pair" test condition were used to train fusion and calibration also with the FoCal Multiclass Toolkit.

## 4. The L$^2$F LRE *post-evaluation* system

After the evaluation Workshop and the analysis of the results, we focused on the improvement of the submitted system. In order to do that, we decided to apply simple modifications that did not essentially affect the architecture and the characteristics of the original language recognition system. Thus, the modifications were mainly aimed to correct some erroneous decisions (training data selection), to improve and reduce the number of GSV-LR sub-systems and to modify the calibration stage.

### 4.1. Training corpora selection

Data management –selection and filtering of the data for training and calibration– was even more important this year than in previous LRE editions due to the characteristics of the VOA corpus. Thus, selecting segments with a large variety of speakers was critical. It was shown during the evaluation that speaker clustering methods for rejecting frequent speakers was very convenient to assure speaker variability. Another issue related with the data was the relatively frequent presence of English in non-audited VOA3 data of any language.

In addition to these common problems, we noticed two errors in our submitted system related with the data. The first and more critical one is that the English data we selected from VOA3 was in fact not American as we expected, but it was English spoken by African speakers. Second, the decision of not training specific language models for Indian English resulted in an error since very poor performance was achieved detecting Indian with the submitted system.

Hence, in order to improve the quality of our training data set after the evaluation we decided to use a modified sub-set of the development data set of Table 2 for training. The reason is that the original development set was expected to be better since a larger amount of audited data was included (reducing miss-labeling errors), the speakers diversity was augmented, data from different sources than VOA3 was included and specific data for Indian English was available for language training. Additionally, the original "American English" segments of VOA3 were replaced with real "American English" speech (instead of English of African speakers). Finally, the *post-evaluation* training set was a random selection of 120 segments of 30 seconds duration per each language. Notice the difference on the amount

of training data (1 hour per language) compared to the original set of Table 1.

### 4.2. GSV-LR systems improvements

The performance of the *submitted* LR system was mainly ruled by the phonotactic systems. It is for that reason that we focused on improving the acoustic based sub-systems.

#### *4.2.1. Silence rejection and normalization*

Low-energy frame rejection and mean and variance feature normalization was incorporated to the GSV-LR front-end described in 3.2.1. Silence segmentation is obtained with the alignment generated by a simple bi-Gaussian model of the log energy distribution computed for each speech segment.

#### *4.2.2. Number of mixtures and sub-systems reduction*

First, the number of Gaussian mixture components was increased from 256 to 1024. Second, it was verified that the use of the two different kernels (supervector normalizations) did not provide any noticeable performance improvement. It was for that reason that we decided to keep only the GSV systems based on the Kullback-Leibler (KL) divergence [6] in the *post-evaluation* system. Thus, the total number of sub-systems of the *post-evaluation* system is reduced to six.

#### *4.2.3. NAP for channel compensation*

The Nuisance Attribute Projection (NAP) approach [10] is a compensation method aimed at removing nuisance attribute-related dimensions in high-dimensional spaces via projections. In the *post-evaluation* system we apply NAP to the conventional GSV approach (not the pushing-back scoring method). NAP projections were trained with a sub-set of the *post-evaluation* training set (50 segments per language). We used a nuisance space of dimension 32.

### 4.3. Gaussian back-end and development data

In the *post-evaluation* LR system the scores of each individual sub-system are processed by a Gaussian back-end prior to the LLR calibration and fusion. Separate back-ends are also trained with the FoCal toolkit for every evaluation condition and segment length.

Since it was used part of the original development data for language model training, we decided to use the evaluation test set for back-end and LLR training through a random 5-fold cross-validation process.

## 5. Language recognition results

The L$^2$F *submitted* system to the LRE09 competition is compared to the *post-evaluation* system described in previous Section 4 . Additionally, for better comparison purposes, a new calibration for the *submitted* system has been trained. Like in the *post-evaluation* system a random 5-fold cross-validation strategy using the test data is applied for training the LLR calibration and fusion of the eight sub-systems. Results for this new calibrated LR system are reported as *submitted*$^*$. Average cost LR performances (as defined in [1]) are shown in Table 3.

First it should be noticed that a considerable improvement – ranging from 12% to more than 20% – is achieved due to the *optimistic* calibration process involving the testing data. However, great performance gains are still achieved as a result of the improvements introduced in the *post-evaluation* system in all conditions and categories. These are particularly noticeable for longer segment durations. For instance, relative cost reductions of 46.7% and 36.9% are obtained for 30 seconds duration in the closed and open-set conditions respectively. The use of only 30 seconds segments in the training set of the *post-evaluation* system might partially explain these results. Table 3 also shows a generalized higher relative improvement of the *post-evaluation* system for the closed-set than for the open-set condition.

| C | T | submit | submit$^*$ | post-eval |
|---|---|---|---|---|
| closed | 30 | 0.0407 | 0.0346 (15.0%) | 0.0217 (46.7%) |
| | 10 | 0.0781 | 0.0618 (20.9%) | 0.0517 (33.8%) |
| | 3 | 0.1692 | 0.1430 (15.5%) | 0.1377 (18.6%) |
| open | 30 | 0.0582 | 0.0507 (12.9%) | 0.0367 (36.9%) |
| | 10 | 0.0935 | 0.0792 (15.3%) | 0.0673 (28.0%) |
| | 3 | 0.1865 | 0.1590 (14.7%) | 0.1513 (18.9%) |

Table 3: *Average cost performance for each of the three segment duration categories (T), and for the closed-set and open-set conditions (C). Relative performance improvements with respect to the submitted system are shown in brackets.*

## 6. Summary and conclusions

Improvements to the L$^2$F Language Recognition system submitted to the NIST LRE 2009 campaign have permitted remarkable recognition gains for all evaluation categories and conditions, achieving comparable performances to the best present LR systems. Particularly noticeable improvements have been obtained in the closed-set 30 seconds segment duration condition with an average cost performance of 0.0217. It is worth to mention that the *post-evaluation* system makes use of only 24 hours of data for language models training ($\sim$1 hour per target language) in contrast to the 338 hours of the *submitted* system.

## 7. References


[1] "The 2009 NIST Language Recognition Evaluation Plan (LRE09)", URL: http://www.itl.nist.gov/iad/mig/tests/lre/2009/.

[2] Meinedo, H. and Neto, J., "Audio Segmentation, Classification and Clustering in a Broadcast News Task", in Proc. ICASSP 2003, Hong Kong, Apr 2003.

[3] Plchot, O. et al, "Acquisition of Telephone Data from Radio Broadcasts with Applications to Language Recognition: Technical Report", URL: http://www.nist.gov/speech/tests/lre/2009/radio_broadcasts.pdf.

[4] Abad, A. and Neto, J., "Incorporating Acoustical Modelling of Phone Transitions in an Hybrid ANN/HMM Speech Recognizer", in Proc. INTERSPEECH-08, Brisbane, Australia, Sep 2008.

[5] Stolcke, A., "SRILM - An Extensible Language Modeling Toolkit", in Proc. ICSLP 2002, Denver, Colorado, Sep 2002.

[6] Campbell, W. M., Sturim, D. E. and Reynolds, D. A., "Support vector machines using GMM supervectors for speaker verification" IEEE Signal Process. Letters, vol. 13(5), pp. 308-311, 2006.

[7] Campbell, W. M. , "A covariance kernel for SVM language recognition", in Proc. ICASSP 2008, Las Vegas, USA.

[8] Chang, C.-C. and Lin, C-J, "LIBSVM - A Library for Support Vector Machines", URL: http://www.csie.ntu.edu.tw/ cjlin/libsvm/index.html.

[9] Brummer, N., "FoCal Multiclass Toolkit", URL: http://niko.brummer.googlepages.com/focalmulticlass.

[10] Campbell, W.M. et al, "SVM based Speaker Verification using GMM Supervector Kernel and NAP Variability Compensation", in Proc. ICASSP 2006, Toluse, France, May 2006.