The L^{2}F Language Verification System for NIST LRE 2009

Alberto Abad[1] *and Isabel Trancoso*[12]

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

Abstract

This paper presents a description of the INESC-ID's Spoken Language Systems Laboratory (L^{2}F) Language Verification system submitted to the 2009 NIST Language Recognition evaluation. The L^{2}F system is composed by the fusion of eight individual sub-systems: four phonotactic systems and four acoustic based methods. Language recognition results have been submitted for the "closed-set", "open-set" and all the "language-pair" conditions except for the English American versus English Indian pair. The data used for training both the phonotactic and the acoustic models, the calibration and fusion development data and measurements of the processing time of the complete system during test are also described in this document.

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.

Acoustic systems model each language short-term acoustics by means of stochastic models/classifiers such as Gaussian mixtures models (GMM), Neural Networks (NN) or Support Vector Machines (SVM). Phonotactic systems typically use language dependent stochastic grammars to model phonemes or broad categories of phonemes extracted by a tokenizer.

This paper presents the LR system developed by the INESC-ID's Spoken Language Systems Laboratory (L^{2}F) for the LRE09 campaign. The submitted system is composed by the fusion of eight individual LR sub-systems: four phonotactic and four acoustic-based. The four phonotactic systems are in fact a Parallel Phone Recognition and Language Modeling (PPRLM) system that exploits the phonotactic information extracted by four parallel tokenizers: European Portuguese, Brazilian Portuguese, European Spanish (Castilian) and American English. The four acoustic methods are based on what is generally known as GMM supervectors (GSV).

The next Section 2 presents a brief description of the data used for training, calibration and fusion. Section 3 describes the submitted language recognition system, starting by the phonotactic modules (subsections 3.1) and the GSV ones (subsection 3.2). Calibration and fusion of the sub-systems is described in 3.3. Measurements of the computational deployment in the processing of the evaluation data set are also provided in section 3.4. Finally, Section 4 presents our main conclusions.

2. Training, calibration and fusion data

Details on the data made available for this evaluation can be found in [1].

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. Training 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.

It is worth noting that telephone classification was initially provided only for some languages of the VOA3 data set. This was the motivation for using a telephone band detector 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. In order to adjust the thresholds, the VOA3 data provided with telephone segmentation was used for informal validation. Since a relatively large amount of data was available in VOA3 for every language, it was decided to adjust the thresholds to assure a reduced number of false alarms in exchange for a possible increase of false rejections.

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, a data set equivalent to the target languages one was extracted for “other” languages present in VOA3 with telephone segmentation already provided. Concretely, Albanian, Azerbaijani, Bangla, Greek, Kurdish, Macedonian, Serbian and Uzbek data were merged in a unique “other” data set of approximately 15 hours also. This “other” languages data set was later 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.

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 used for each target language and total duration extracted from the VOA3 corpus. For model training, American English and Indian English are not distinguished.*

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 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.

Notice that 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 described previously, plus Arabic, German, Italian and Japanese included in previous evaluation data sets) has been collected.

The total calibration and fusion corpus is composed of 19,346 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.

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: *Description of the development data set used for fusion and calibration of the LR system with different data sources: audited VOA2 and VOA3 data (LDC), non-auditted voa3 data (VOA3) and previous LRE data sets (LREold).*

3. The L^{2}F Language Recognition 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. Additionally, measurements of the processing time are provided.

3.1. The PRLM-LR systems

The PRLM systems used for LRE09 exploit the phonotactic information extracted by four parallel tokenizers. The key aspect of this type of systems is the need for robust phonetic classifiers that generally need to be trained with word-level or phonetic level transcriptions. At INESC-ID, we have been working for several years on Large Vocabulary Continuous Speech Recognition (LVCSR) using hybrid recognizers, combining Artificial Neural Networks and Hidden Markov models (ANN/HMM), the so-called connectionist paradigm. Although our first LVCSR system was initially developed for European Portuguese, it was recently ported to the Brazilian variety, and now also covers European Spanish (Castilian) and American English.

The tokenization of the input speech data in both training and testing sets is done with the neural networks that are part of our hybrid recognition (AUDIMUS). This type of recognizer is generally composed by one or more phoneme classification networks, particularly MultiLayer Perceptrons (MLP), that estimate the posterior probabilities of the different phonemes for a given input speech frame (and its context).

3.1.1. Feature extraction

In this evaluation, 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). The first three types of feature were part of the AUDIMUS version trained for Broadcast News (BN) data. The last feature representation was introduced for this evaluation, since it was considered adequate for the kind of data present at LRE09.

3.1.2. Phonetic tokenizers/classifiers

For this evaluation, it was necessary to re-train our phonetic classifiers with 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 [5]. 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. The context windows of the MLP networks trained with PLP, RASTA and ETSI features is fixed to 13, while a context of 15 frames was considered more appropriate for MSG features. The European Portuguese networks have two hidden layers of 1500 weights and an output layer of 39 weights. The Brazilian Portuguese networks have also two hidden layers of 600 units and an output layer of 40 units. The size of the Castilian network is 700 weights for the two hidden layers and 32 weights for the output layer. The English networks use two hidden layers of 1500 weights and an output of 39 units. Notice, that the size of the output layer corresponds to the number of phonetic units of each language, plus silence (no additional sub-phonetic or context-dependent units have been considered [6]).

3.1.3. Phonotactics modeling

For every phonetic tokenizer, the phonotactics of each target language is modelled with a 3-gram model. For that purpose the SRILM toolkit has been used [7].

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.

During test, for each target language and for each tokenizer, the corresponding phonotactic models are used to obtain a target language verification score for every speech segment.

3.2. The GSV-LR systems

Acoustic methods for LR are usually prefered to phonotactic approaches since they are not limited by the need of well-trained phonetic tokenizers. Recently, a method generally known as GMM supervectors [8, 9, 10] 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 done by means of a Bayesian adaptation of a universal background model (GMM UBM) to the characteristics of a given speech segment. In general, the mixture means of the resulting adapted GMM are stacked in a single supervector and used as the high-dimensional vector that represents the 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^2F language recognition system are slight variations of the GSV approach.

Concretely, two of the GSV systems differ in the normalization applied to 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 [10]. As a result, a positive and a negative "discriminative" GMM is obtained for each target language. These GMMs can be used directly to obtain GMM log-likelihood ratios. This approach is considered to obtain significant improvements relative to the conventional GSV method when testing with short speech segments that might result in a poor MAP adaptation. Additionally, the scores are obtained by simple log-likelihood ratio computation, which avoids the need for generating a supervector for every test segment.

3.2.1. Feature extraction

The extracted features are Perpetual 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.

This type of features may be justified by the fact that RASTA features are known to be a robust representation for speech processing applications [11]. On the other hand, it has also been shown that the use of SDC features (created by stacking delta cepstra computed across several frames) allows improved performances in LR tasks [12]. The selected front-end showed remarkable improvements compared to other evaluated feature representations during the development of the systems.

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. Experiments were made using a higher dimensionality which confirmed a better performance. However, the considerable increase in both processing time and storage space and the tight schedule motivated a more conservative decision.

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. Normalization details are explained in next sub-section.

Linear SVM classifiers are trained for each target language (and for the two different mean normalizations) with the libSVM toolkit [13]. 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.

During test, the target language dependent SVMs are used to obtain a score for each testing speech segment (converted to supervector form) and target language.

3.2.3. Mixture mean normalizations

In the recent literature, various mixture mean normalizations previous to SVM training/classification can be found. For this campaign, two slightly different mixture mean normalizations were tested, resulting in two independent sub-systems:

$$\hat{\mu}_{i,1} = \frac{\sqrt{w_i}}{\sigma_i}(\mu_i - \mu_{ubm_i}) \qquad (1)$$

$$\hat{\mu}_{i,2} = \frac{\mu_i - \mu_{ubm_i}}{\sigma_i} + \mu_{ubm_i} \qquad (2)$$

where $\hat{\mu}_{i,1}$ and $\hat{\mu}_{i,2}$ are the normalized means corresponding to the i-th gaussian mixture with mean μ_i, weight mixture w_i and standard deviation σ_i, and μ_{ubm_i} is the i-th mixture mean of the GMM UBM.

Our experiments did not reveal significant differences between these two normalizations. However, no performance drop was observed in the fusion step due to the combination of both of them and it was decided to kept them in the final system.

Notice that in the case of the two systems based on GMMs derived from the the SVM parameters as described in [10], an adequate "de-normalization" step was applied to build the GMMs.

3.3. Calibration and fusion

Linear logistic regression fusion and calibration of the 8 sub-systems has been done with the FoCal Multiclass Toolkit [14].

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. Although it is very likely that it is not a good solution, we still expect 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. Since there is a considerable reduced amount of data of each individual target language, it is very likely that these language pair detectors are considerably worse calibrated.

3.4. Processing time

The evaluation tests were run in a cluster of computers under the Condor framework for parellelization of tasks.

In order to approximately estimate the computational time, we have separated a reduced set of data from the evaluation test set (100 files amounting 2114 seconds in total) and we have run the language recognition tests in an Intel Quadcore CPU Q6600 @ 2.40GHz machine with 8 GBytes of DDR2-667 MHz. The various processing steps of the submitted LR system have been run in pipeline mode.

The first step is feature extraction for the 4 streams of the PRLM sub-systems and also the SDC of the GSV methods. The feature extraction step is approximately 0.11xRT. It is worth noting at this point that feature extraction also includes conversion from sphere format to raw data and amplitude scaling.

Obtaining the phonetic sequences of the four tokenizers takes 1.05xRT. Notice that none of the tokenizations was run in parallel and that neural network models are loaded for every trial and accounted for in the total time. 80 % of the tokenization time is due to the English and Portuguese tokenizers, since they have considerably larger neural networks. Obtaining LR scores for each trial for each target language and for each phonetic tokenizer takes an additional 0.3xRT.

The extraction of the evaluation supervectors for both normalizations takes 0.19xRT. SVM classification for every target language and both normalizations takes approximately 1xRT. The need to evaluate the scores of 23 (plus the "unknown") languages for each segment considerable increases this step. The GSV approach with pushed back GMMs is quite computational expensive due to the need for evaluating two GMMs for each target language. The cost of the GSV method with GMMs (using both normalizations) is 2.5xRT.

Finally, the calibration and fusion step consists of a simple linear combination which is almost inexpensive.

Table 3 summarizes the processing time spent in every step in the evaluation for the reduced evaluation set and the total time spent for this set. The total time of the complete L^{2}F language recognition system is 5.16xRT. However it is worth noticing that this estimation is extremely pessimistic since all the processes were run in a pipeline mode (some of them could have been easily parallelized in a single machine). Additionally, in the processing steps which imply the use of models (NN, SVMs, phonotactic...), the time for model loading has been taken into account in every trial, for every system and for every target language.

	FE	PE	SE	PS	SS	GS	TOT
Time	242	2218	400	631	2093	5326	10910
RT	0.11	1.05	0.19	0.3	1.0	2.51	5.16

Table 3: *Summary of the processing time for a reduced set of 2214 seconds. In the first row, time in seconds. In the last row, number of real time. The several steps are: feature extraction (FE), phonetic extraction (PE), supervector extraction (SE), phonotactic scoring (PS), SVM scoring (SS), GMM scoring (GS) and total time (TOT).*

4. Summary and conclusions

In this NIST language recognition evaluation campaign, we have presented a system based on the fusion of 8 individual language recognition systems.

Our main objective in participating in this evaluation, our first in NIST, was to introduce ourselves to the language recognition community, to explore the recently proposed methods and to learn as much as possible. In this sense, independently of the final results, our participation was already quite successful.

The large amount of data accumulated in previous LRE campaigns raised hard challenges for new participants like us.

Our previous involvement in an LR evaluation campaign, ALBAYZIN'08 [15], was devoted to the four official languages of Spain. In both this previous campaign and the current one, we have never been able to obtain with the GSV approach comparable results to our PPRLM systems. These results contrast with what is claimed in the literature, and strengthen the conviction that our systems can be significantly improved.

Different approaches may be taken to address this problem, namely by adopting more robust parametrization techniques (including speaker normalization), by applying channel normalization methods and by using higher dimension models.

The range of applications of language recognition is very wide. In the framework of the Vidi-Video European project, whose goal is to build a video search engine with a 1000-element thesaurus, LR is applied as a pre-processing stage, before speech recognition. This allows restricting the extraction of linguistic cues to video segments that are spoken in a language for which the recognizer is trained.

5. Acknowledgments

The authors would like to thank to David Matos and Tiago Luís for their support with the processing machines and data management issues. We also would like to thank to the organizers of the 2009 NIST language recognition evaluation for their availability for solving our doubts and problems. This work was partially funded by the European project Vidi-Video.

6. 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] Meinedo, H., Caseiro, D., Neto, J. and Trancoso, I., "Audimus.media: a broadcast news speech recognition system for the European Portuguese language", in Proc. PROPOR 2003, Faro, Portugal, 2003.

[5] Abad, A., Meinedo, H. and Neto, J., "Automatic classification and transcription of telephone speech in radio broadcast data", in PROPOR'2008 - International Conference on Computational Processing of the Portuguese Language, Springer, Aveiro, Portugal, Sep 2008.

[6] 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.

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

[8] 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.

[9] Castaldo, F. et al, "Acoustic Language Identification using Fast Discriminative Training", in Proc. INTERSPEECH 2007, Antwerp, Belgium, Sep 2007.

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

[11] Hermansky, H. and Morgan, N., "RASTA processing of speech", IEEE Transactions on Speech and Audio Processing, Vol. 2(4), pp 578-589, Oct 1994.

[12] Torres-Carrasquillo, P. A. et alt., "Approaches to Language Identification using Gaussian Mixture Models and Shifted Delta Cepstral Features", in Proceedings of ICSLP 2002, pp. 89-92, Denver, Colorado, Sep 2002.

[13] 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.

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

[15] "Plan de Evaluación de Sistemas ALBAYZIN-08 Verificación de la Lengua (ALBAYZIN-08 VL)", URL: http://jth2008.ehu.es/Plan_Albayzin-08_VL_final.pdf.