The L^{2}F Language Verification Systems for Albayzin-2010 Evaluation

Alberto Abad[1]*, Oscar Koller*[1]*, Isabel Trancoso*[1,2]

[1]L^{2}F - Spoken Language Systems Lab, INESC-ID Lisboa
[2]Instituto Superior Técnico, Lisboa, Portugal
alberto.abad@l2f.inesc-id.pt

Abstract

This paper presents a description of INESC-ID's Spoken Language Systems Laboratory (L^{2}F) Language Verification systems submitted to the ALBAYZIN-2010 evaluation. The primary submission consists of the fusion of six individual subsystems: one Gaussian supervector approach with support vector machines that relies on the acoustic characteristics extracted by a front-end of shifted deltas, and five individual Phone Recognition and Language Modeling detectors based on five different phone tokenizers. Additionally, two contrastive systems have been developed. Language detection results have been submitted for all the evaluation conditions for every system. The main particularity of the systems developed for this evaluation is that individual language models for clean and noisy conditions have been trained for each target language. Results for the different systems and evaluation conditions are reported.

1. Introduction

The "Red Temática en Tecnologías del Habla" (RTTH) has organized in the recent years a series of evaluations - so called ALBAYZIN evaluations - in some relevant speech processing topics devoted to encourage language research activities on the four official languages of Spain: Castilian, Catalan, Basque and Galician.

Similar to the well-known NIST Language Recognition Evaluation series, a Language Verification (LV) task was proposed in ALBAYZIN-08 with the objective of determining if each one of the four official languages of Spain was spoken (or not) in a given test file. In the new ALBAYZIN-2010 campaign the set of target languages is increased to cover also Portuguese and English.

This paper presents the LV systems developed by INESC-ID's Spoken Language Systems Laboratory (L^{2}F) for the ALBAYZIN-2010 campaign. A primary and two contrastive systems have been submitted, which differ in the number of employed sub-systems and in the followed back-end strategy for calibration. The *primary* system consists of the fusion of six different language detection sub-systems: an acoustic system based on Gaussian mixture model SuperVectors (GSV) [1] and five phonotactic Phone Recognition and Language Modeling (PRLM) [2] systems. Additionally, the *alt1* system explores a method for introducing segment duration score normalization to the calibration stage, while the *alt2* system is aimed at developing a simplified LV system. The next Section 2 presents a brief description of the task, the data provided for the evaluation and the evaluation metrics. Section 3 describes some commonalities of the systems developed (see Section 3.1) and details of each one of the six individual sub-systems: the GSV-LV system and the PRLM-LV detectors are described in Sections 3.2 and 3.3, respectively. Measurements of the computational deployment in the processing of the evaluation data set are also provided. The three submitted systems are described in Section 4. In Section 5 results obtained by the three systems in the different evaluation conditions with the development data set are presented. Finally, Section 7 presents our main conclusions.

2. ALBAYZIN-2010 LV: Task, Data and Metric Description

Detailed information on the ALBAYZIN-2010 LV campaign can be found in the evaluation plan document [3].

2.1. Task and Evaluation Conditions

The task consists of deciding whether a speech segment belongs to each one of the six target languages (Castilian, Catalan, Basque, Galician, Portuguese and English) or not. For each test signal six decision results (true or false) are produced together with a score, one for each of the target languages.

Four test evaluation conditions are proposed depending on the type of verification test (closed-set vs. open-set) and the type of speech (clean vs. noisy). In contrast to the closed mode, in the open mode speech segments from unknown languages different from the target ones may appear in the test data and are taken into account for the systems' assessment. The four evaluation conditions are referred to as closed-clean (CC), closed-noisy (CN), open-clean (OC) and open-noisy (ON).

2.2. Train, Development and Test Data

All the data provided for the ALBAYZIN-2010 evaluation are TV programs captured at 16 kHz. The training data set consists of more than 12 hours per target language, in several files of variable length separated in clean speech (more than 10 hours) and noisy speech (around 2 hours). The evaluation data set consists of 4992 files with speech of the six target languages and in other unknown languages of 3 different nominal durations: 3, 10 and 30 seconds. Additionally, a development data set consisting of 4950 files of similar characteristics to the evaluation set was provided with language identification, duration and type of speech labels.

2.3. Performance Metric

An average performance score based on the false positive and false alarm rates obtained by the evaluating systems is used. The performance score, hereinafter referred to as C_{avg}, is computed independently for each test length duration (3, 10 and 30 seconds). Further details about the metrics can be found in [3].

3. Language Verification Sub-system Description

3.1. Common Characteristics

3.1.1. Audio Data Pre-processing

The training data provided for each target language and for each speech type was pre-processed in order to segment long data files into a set of homogeneous reduced length speech segments. First, speech-non-speech (SNS) segmentation was applied [4]. The SNS module is a finite state machine that uses a binary Multi Layer Perceptron (MLP) trained with several hours of BN data to identify audio portions that do not contain speech, speech with too much noise or pure music. After this segmentation process, continuous speech segments (1 second of non-speech tolerance) of length above 8 seconds and below 40 seconds were selected. In the particular case of Castilian, these thresholds for segment filtering were fixed to 7 and 49 seconds. Notice, that this pre-processing segmentation was only applied to the training data and not to the development and evaluation data sets. Table 1 shows the amount of selected segments and the total duration in minutes for each target language and type of speech.

	clean		noisy	
	#segm	dur. [min]	#segm	dur. [min]
castilian	576	227.9	223	81.7
catalan	674	237.8	235	76.6
english	600	231.3	266	92.7
basque	722	260.7	268	80.8
galician	746	258.5	254	74.6
portuguese	583	222.3	233	83.3

Table 1: *Training data segmentation for each target language and speech type.*

3.1.2. Target Language Modeling

One of the main particularities shared among all the developed sub-systems is that a separate target language model was trained for clean and noisy speech. The two target models of each language are used to obtain two language-dependent scores for each speech test segment. Consequently, for every test segment a vector of 12 scores $\mathbf{x}_i$ is produced by every individual sub-system i.

3.1.3. Linear Gaussian Back-End

A linear Gaussian Back-End (GBE) follows every single sub-system to transform the 12 elements score-vector $\mathbf{x}_i$ to a 7 elements log-likelihood vector $\mathbf{s}_i$ (6 target languages plus 1 out-of-set language log-likelihoods):

$$\mathbf{s}_i = \mathbf{A}_i\mathbf{x}_i + \mathbf{o}_i \quad (1)$$

where $\mathbf{A}_i$ is the transformation matrix for system i and $\mathbf{o}_i$ is the offset vector.

A common characteristic of all the systems developed is that open-set and closed-set conditions have not been distinguished in back-end calibration (nor have they in the later fusion of the individual sub-systems). In other words, the same 7 log-likelihoods are produced independently of the type of verification test (closed-set or open-set) and they are used to obtain detection log-likelihood ratios and decisions using the adequate prior distributions over language classes in each verification test type.

3.2. GSV-LV sub-system

A method generally known as GSV [1] is known to be a successful approach for both speaker and language verification tasks. GSV-based approaches map each speech utterance to a high-dimensional vector space. Support Vector Machines (SVMs) are used for classification of test vectors within this space. The mapping to the high-dimensional space is achieved by stacking all parameters (usually the means) of an adapted GMM in a single supervector by means of a Bayesian adaptation of a universal background model (GMM-UBM) 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 is used by the binary classifier to generate a score for each target language.

3.2.1. Feature Extraction

The extracted features are shifted delta cepstra (SDC) [5] of Perceptual Linear Prediction features with log-RelAtive SpecTrAl speech processing (PLP-RASTA). First, 7 PLP-RASTA static features are obtained and mean and variance normalization is applied in a per segment basis. Then, SDC features (with a 7-1-3-7 configuration) are computed, resulting in a feature vector of 56 components. Finally, low-energy frames detected with the alignment generated by a simple bi-Gaussian model of the log energy distribution computed for each speech segment are removed.

3.2.2. Supervector Extraction and SVM Language Modeling

A GMM-UBM of 256 mixtures was trained with approximately 9 hours of speech randomly selected among the clean segments (around 1.5 hours per target language) of the training data set of Table 1.

One single iteration of Maximum a Posteriori (MAP) adaptation with relevance factor 16 is performed for each speech segment to obtain the high-dimensional vector of size 56x256.

The linear SVM kernel of [1] based on the Kullback-Leibler (KL) divergence is used to train the target language models with the LibLinear implementation of the libSVM tool [6]. For each target language and type of speech, all the training segments of that language are used as positive examples and all the segments from the other languages are used as negative background set.

3.2.3. Processing Time

Processing time measurements of the developed language recognition systems were carried out in a machine with two Quad Xeon 2.4GHz (E5530) processors with 48 GBytes of DDR3 RAM at 1333 MHz. Notice, however, that data is stored in a distributed file system with relatively slow transfer rates. Thus, disk access can become a bottleneck in some fast operations. The processing time was measured in a sub-set of 100 test files amounting to 1522.8 seconds. The feature extraction of the 100 files consumed 405 seconds, Bayesian adaptation and supervector extraction lasted 118 seconds, and scoring was performed in 6 seconds. These figures correspond to 0.35xRT approximately.

3.3. PRLM-LV Sub-systems

The Phone Recognition followed by Language Modeling (PRLM) systems used for ALBAYZIN-2010 exploit the phonotactic information extracted by five individual tokenizers: European Portuguese, Brazilian Portuguese, European Spanish (Castilian), American English and a mixed African/European Portuguese tokenizer using special mono-phonetic units [7]. The key aspect of this type of system is the need for robust phonetic classifiers that generally need to be trained with word-level or phonetic level transcriptions. In this case, the tokenizers are MultiLayer Perceptrons (MLP) trained to estimate the posterior probabilities of the different phonemes for a given input speech frame (and its context). For each target language and for each tokenizer a different phonotactic *n-gram* language model is trained. During test, the phonetic sequence of a given speech signal is extracted with the phonetic classifiers and the likelihood of each target language model is evaluated.

3.3.1. Phonetic Tokenizers

The tokenization of the speech data is done with the neural networks that are part of our hybrid Automatic Speech Recognition (ASR) system named AUDIMUS [8]. The tokenizers combine three MLP outputs trained with Perceptual Linear Prediction features (PLP, 13 static + first derivative), PLP with log-RelAtive SpecTrAl speech processing features (PLP-RASTA, 13 static + first derivative) and Modulation SpectroGram features (MSG, 28 static). A phone-loop grammar with phoneme minimum duration of two frames is used for phonetic decoding.

The networks were trained with different amounts of broadcast news (BN) annotated data. For the European Portuguese classifier, 57 hours of manually annotated data and more than 300 hours of automatically transcribed BN data were used. The Brazilian Portuguese classifier was trained with around 13 hours of BN data. The Spanish system used 14 hours of manually annotated data and 78 hours of automatically transcribed data. The English system was developed with the HUB-4 96 and HUB-4 97 data sets, that contain around 142 hours of TV and Radio Broadcast data. Finally, the mixed African/European system was trained with two times 6 hours of manually annotated broadcast news data containing both Portuguese varieties equally balanced. Through a particular training process (refer to [7]), this system is tuned to differentiate between the close Portuguese varieties.

The size of the input and hidden layers of the neural networks varies among the different parameterizations and languages, but in general all the MLPs are composed by two hidden-layers with a relatively small number of hidden units in order to accelerate the tokenization process. 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 [9]).

3.3.2. Phonotactics Modeling

For every phonetic tokenizer, the phonotactics of each target language for every type of speech condition (clean and noisy) is modeled with a 3-gram back-off model, that is smoothened using Witten-Bell discounting. For that purpose the SRILM toolkit has been used [10].

3.3.3. Processing Time

Using the previously described machine, the total time deployed in processing the 100 files sub-set when running the 5 PRLM systems in parallel is 245 seconds, which corresponds to 0.16xRT. When the PRLM systems are run one after the other, the total amount of processing time increases up to 936 seconds. The phonetic tokenization operations account for 60% to 80% of the processing time (depending on the network) and the rest of the time is consumed in the scores generation.

4. The L^{2}F Submissions

The L^{2}F submitted systems consist of the fusion of some of the sub-systems described in previous Section 3. Linear logistic regression (LLR) has been used to fuse the log-likelihood outputs generated by the linear GBEs of the individual sub-systems to produce fused likelihoods $\mathbf{l}$:

$$\mathbf{l} = \sum_i \alpha_i \mathbf{s}_i + \mathbf{b} \tag{2}$$

where α_i is the weight for sub-system i and $\mathbf{b}$ is the language-dependent shift.

The GBEs and the LLR fusion have been trained and tested with the development data set using a jack-knifing strategy: data is partitioned in 5 random sets and each one of the sets is once held out for testing and the other 4 sets are used to estimate the calibration and fusion parameters. The initial randomization of the data is iterated 5 times and a jack-knife scheme is repeated resulting in total in 25 sets of estimated back-end and fusion parameters that are averaged to obtain the final back-end. Calibration was carried out using the FoCal Multiclass Toolkit [11].

4.1. Primary System (primary)

The *primary* system consists of the fusion of the GSV subsystem and the five PRLM sub-systems. Segment length duration dependent and type of speech dependent back-ends were trained. That is, for each combination of type of speech (clean and noisy) and segment duration (30, 10 and 3 seconds) different GBE and LLR parameters are estimated using only the development data of the corresponding type and duration. In test –the evaluation data set was split in 30, 10 and 3 seconds– the back-ends trained with only clean speech are used for the CC and OC evaluation conditions. On the other hand, CN and ON language detection is performed using the back-ends trained with noisy speech. It is worth remembering that, as explained in section 3.1.3, the same back-end is used for closed-set and open-set conditions applying different language priors for log-likelihood ratio and decision generation.

4.2. First Contrastive System (alt1)

The objective of the *alt1* system is to investigate an alternative back-end method that incorporates segment duration normalization. Like in the primary system, the *alt1* system consists of the fusion of the six sub-systems described in Section 3 and uses segment length duration dependent and type of speech dependent back-ends. However, in contrast to the *primary* system, a segment length normalization strategy similar to the one described in [12] was considered. In [12] a duration-independent back-end that uses duration-information in the fusion as side-information is proposed.

$$\mathbf{l} = \sum_i \alpha_i \mathbf{s}_i + \mathbf{b} + \mathbf{C}\mathbf{d} \tag{3}$$

where $\mathbf{d}$ is the vector of durations (that may be different for each sub-system). Additionally, the scores of each individual system are augmented with multiplied versions of the duration

d^p ($\mathbf{x}_i d^p$), where d is the segment duration and p can take different values.

$$\mathbf{s}_{i,p} = \mathbf{A}_{i,p}(\mathbf{x}_i d^p) + \mathbf{o}_{i,p} \quad (4)$$

In the *alt1* system, the segment duration information is ignored as a side-information in the fusion process ($\mathbf{C} = \mathbf{0}$), but it is used to produce duration normalized scores with p values of $0, -1$ and $-1/2$, that correspond, to the original scores, the scores normalized by the duration and by the square root of the duration of each individual sub-system. The duration d is measured as the number of speech frames in the GSV sub-system or the number of decoded phones in the case of PRLM subsystems. Therefore, instead of 6 GBEs, 18 GBEs need to be estimated and the fused likelihood vector $\mathbf{l}$ is the result of the fusion of the "18 sub-systems".

$$\mathbf{l} = \sum_{i.p} \alpha_{i,p}\mathbf{s}_{i.p} + \mathbf{b} \quad (5)$$

Notice, that in contrast to [12], duration-dependent backends are estimated and, thus, we are applying a sort of within-class duration normalization. Consequently, like in the *primary* system, different GBE and LLR parameters for each type of speech and duration condition are estimated.

4.3. Second Contrastive System (alt2)

The aim of the *alt2* system is to assess the performance of a heavily simplified LV system. First, the number of sub-systems considered is reduced and only the GSV sub-system and the PRLM sub-system based on the European Spanish tokenizer – which was consistently the best performing one– are fused. Second, a single back-end is trained for all the conditions using all the development data ignoring the segment duration class and type of speech. A back-end that incorporates segment duration normalization scores like the one described above is used. Notice, that the *alt2* system generates the same scores and decisions for the OC and ON conditions and for the CC an CN conditions. Like in the previous submitted systems, the difference between the closed-set and open-set conditions relies on the target language priors applied.

5. Results on the Development Set

Table 2 presents the results obtained in the development set, for the primary and both contrastive systems. The three top-rows correspond to clean speech development data results and the last three rows to noisy speech. For the sake of clarity, 100xC_{avg} performance scores are reported.

With respect to the different submitted systems, similar detection performances of the *primary* and *alt1* systems can be observed. In fact, the *alt1* system consistently outperforms the *primary* system (except in the 10 seconds CC condition), showing the benefits of the within-class duration normalization method. However, we observed some calibration instabilities during the training of the calibration and fusion parameters (probably due to lack of data) that prevented us from presenting the *alt1* system as our primary submission. On the other hand, the *alt2* system shows a considerable lower performance, as we expected. However, in spite of its simplicity and the use of a general back-end for all conditions, it is still able to provide a quite significant language detection performance, particularly in closed-set and long segment duration conditions.

Regarding the evaluation conditions and segment duration, as it is well-known in LV tasks, the open mode is significantly more challenging than the closed one, and the use of longer segments contributes to smaller detection errors. The performance of all the submitted systems is also considerably affected by the speech quality. Significant performance degradations are obtained in noisy speech conditions, particularly, larger relative cost increase is observed in segments with longer durations.

System	30 sec		10 sec		3 sec	
	cl	op	cl	op	cl	op
primary	0.28	0.45	1.28	2.21	5.35	7.03
alt1	0.23	0.26	1.38	1.99	4.94	6.86
alt2	0.97	1.94	2.18	3.50	7.55	9.75
primary	1.32	1.88	2.02	3.61	6.73	7.90
alt1	0.92	1.16	1.82	2.33	5.08	7.07
alt2	1.90	3.09	4.71	6.87	12.64	15.50

Table 2: *100xC_{avg} performance on the ALBAYZIN-10 LV development set on closed-set and open-set mode and for clean (top three rows) and noisy speech (last three rows).*

6. Results on the Evaluation Set

Table 3 presents the results obtained in the evaluation set. Like in previous development results, 100xC_{avg} performance scores are reported. It must be noticed that an error in the PRLM American English sub-system that affected both the *primary* and *alt1* submissions was detected after submission: detection scores of the evaluation data were erroneously generated using the n-gram language models of the African Portuguese PRLM. In order to draw correct conclusions about the performance of the submitted systems, only the corrected results are provided here.

The most remarkable observation is the quite different language detection performance achieved by the *primary* and *alt1* systems with respect to the *alt2* and the results obtained in the development data set. While *primary* and *alt1* are still quite similar between them, a huge performance loss with respect to the development set is obtained, which is still more noticeable when they are compared to *alt2*. In longer segment duration conditions and particularly in noisy type of speech (30 seconds clean and 30 and 10 seconds noisy), *alt2* clearly outperforms the other submissions.

A post-evaluation analysis is still being conducted, however some preliminary explanations for these unexpected results can be provided. The *alt2* system mainly differs from the other submissions in two aspects: the number of sub-systems and the way the back-end parameters are estimated. With respect to the number of sub-systems, although an increased number of sub-systems does not necessarily imply improved detection, it is quite unlikely that it is the cause for large performance loss, specially when the individual sub-systems have been verified to provide significant language detection ability individually. The most likely reason for the different performance observed in development and evaluation data sets is the poor estimation of the back-end parameters due to the insufficient amount of data available for each evaluation condition, and the large number of back-end parameters. On the one hand, there are 1164 clean and 486 noisy development segments for every segment length duration. On the other hand, the back-end of the *primary* system is composed of around 550 parameters and the *alt1* has about three times this number of parameters. Given the fact that back-end parameters are individually estimated for each

segment duration and type of speech, we believe that an over-estimation problem to development data occurred. The fact that the most important performance degradations are observed in noisy conditions seems to verify this hypothesis. According to our current post-evaluation calibration experiments, language recognition improvements can be obtained by simply applying the back-end parameter estimation strategy of the *alt2* system to calibrate and fuse the six sub-systems.

System	30 sec		10 sec		3 sec	
	cl	op	cl	op	cl	op
primary	2.23	2.96	3.59	4.68	8.53	10.73
alt1	2.19	3.09	3.63	4.45	8.44	10.29
alt2	1.81	3.41	4.59	6.11	10.55	12.89
primary	4.16	7.00	8.10	9.81	12.73	15.51
alt1	4.03	8.39	7.54	9.48	12.17	16.09
alt2	2.53	4.75	6.36	9.36	13.42	16.54

Table 3: *$100xC_{avg}$ performance on the ALBAYZIN-10 LV evaluation set on closed-set and open-set mode and for clean (top three rows) and noisy speech (last three rows).*

7. Conclusions

In the ALBAYZIN 2010 language recognition evaluation campaign, the L^2F has presented a primary system based on the fusion of 6 individual language recognition systems (one acoustic and five phonotactics) and two additional contrastive systems. The estimated processing time of the primary system is approximately 0.51xRT. The performance achieved by the submitted systems in the different evaluation conditions in the development data set was outstanding. However, a considerable performance loss was verified in the evaluation data and in contrast to our expectations, the simplest submitted system resulted in the most robust language detector for most of the conditions. The main reason for the differences observed in development and evaluation is most likely due to weak back-end parameters estimation of the systems that applied duration and type of speech dependent back-end calibration and fusion. With increased development data, or with a different back-end scheme, we believe that the primary system would be able to provide significant language recognition improvements, closer to the development data results. Anyway, the reported systems were still able to provide promising language detection performances in the different evaluation conditions.

8. References

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

[2] Zissman, M., "Comparison of four approaches to automatic language identication of telephone speech", IEEE Transactions on Speech and Audio Processing, vol. 4(1), pp. 31-44, 1996.

[3] "The Albayzin 2010 Language Recognition Evaluation Plan (Albayzin 2010 LRE)", URL: http://jth2010.gts.tsc.uvigo.es/images/stories/pdfs/albayzin_lre10_evalplan_v2.pdf.

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

[5] 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.

[6] Lin, C-J, "LIBLINEAR - A Library for Large Linear Classification", URL: http://www.csie.ntu.edu.tw/ cjlin/liblinear/.

[7] Koller, O., Abad, A. and Trancoso, I. "Exploiting variety-dependent Phones in Portuguese Variety Identification", IEEE Odyssey 2010: The Speaker and Language Recognition Workshop, 2010.

[8] Meinedo, H., Alberto, A., Pellegrini, T., Neto, J. and Trancoso, I., "The L^2F Broadcast News Speech Recognition System", in Proc. FALA-2010, Vigo, Spain, Nov 2010.

[9] 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.

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

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

[12] van Leeuwen, D. and Gonzalez-Dominguez, J., "The TNO system for LRE-2009", The 2009 NIST Language Recognition Evaluation (LRE09) Workshop, Baltimore, US, Jun 2009.