I3A Language Recognition System for Albayzin 2010 LRE

David Martínez, Jesús Villalba, Antonio Miguel, Alfonso Ortega and Eduardo Lleida

Aragon Institute for Engineering Research (I3A), University of Zaragoza (UZ), Spain

(david | villalba | amiguel | ortega | lleida)@unizar.es

Abstract

This paper describes the two systems submitted to the Albayzin 2010 Language Recognition Evaluation by I3A. This evaluation is similar to the one organized by NIST every 2 years, but the languages to be recognized are those spoken in the Iberian peninsula (Spanish, Catalan, Basque, Galician and Portuguese) plus English. Both submissions are a fusion of five phonotactic and three acoustic subsystems. The only difference between them is the normalization and fusion of the scores. State-of-the art methods for Language Recognition are adapted to and investigated in the KALAKA-2 database. Our primary system was ranked in the first position of the evaluation.

Index Terms: language recognition, phonotactic LRE, acoustic LRE, channel compensation, discriminative training

1. Introduction

Language Recognition (LR) has experimented a huge development in the last years. To compare the quality of the different LR systems around the world, NIST has coordinated several evaluations (1996, 2003, 2005, 2007 and 2009)[1]. In year 2008, the Spanish Thematic Network on Speech Technologies (RTTH) organized a similar one [3], and Albayzin LRE 2010 is the second edition [1]. This evaluation was presented within the conference FALA 2010[2], Vigo, in November 2010. The main difference with NIST is that the languages to be recognized are Spanish, Catalan, English, Basque, Galician and Portuguese, and they are extracted from multi-speaker TV broadcast recordings. The database used for this purpose is KALAKA-2 [2].

The two systems submitted by I3A are fusion of five phonotactic subsystems (PRLM) and three acoustic subsystems. These methods have been previously used for classification and in this paper we show one approach to adapt and fuse them in a LR task. Both systems are identical except for the normalization and fusion methods used at the back-end. In the first submission, we make a T-norm of scores and perform a discriminative fusion. In the second, we make a ZT-norm of scores, and a generative Gaussian backend is followed by a discriminative fusion. We have tested them in the closed- and open-set conditions for clean speech.

Among the five phonotactic subsystems we include four built with the Brno phoneme recognizer [8], based on Neural Networks and Hidden Markov Models (ANN/HMM), and one built with the I3A phoneme recognizer, based on Gaussian Mixture Models and HMMs (GMM/HMM). In the first case an N-best strategy has been slected to train the Language Models (LM), while in the second case we have used triphonemes, that is, phonemes with context, and only the 1-best hypothesis.

[1] http://www.itl.nist.gov/iad/mig/tests/lre

[2] http://fala2010.uvigo.es

Among the three acoustic subsystems we include a GMM classifier trained via Maximum Likelihood (ML) with the Expectation Maximization (EM) algorithm [9], a discriminatively trained Maximum Mutual Information (MMI) classifier [4], and a Joint Factor Analysis (JFA) system similar to the one used in [7].

The rest of the paper is organized as follows: Section 2 specifies the data and parametrization; Section 3 describes the acoustic, phonotactic and fusion methods; in Section 4, results obtained in the evaluation are analysed; and Section 5 gives the conclusions.

2. Data and Parametrization

The data used for training our systems come from the training part of KALAKA-2 database, with the exception of the training of the phone recognizers. We have used phoneme recognizers trained on Czech, with the Czech SpeechDat-E database [10], on Hungarian, with the Hungarian SpeechDat-E database [11], on Russian, with the Russian SpeechDat-E database [12], on English, with the TIMIT database [13], and on Spanish, with the Albayzin [14] and Speech Dat Car [15] databases.

Calibration of the results was performed with the development part of KALAKA-2 database.

The features used for the acoustic systems are Mel Frequency Cepstral Coefficients (MFCC) concatenated to their Shifted Delta Cepstra Coefficients (SDC) [16]. 6 MFCCs plus energy are extracted from the audio, cepstral mean normalization is applied, and the SDC with a 7-1-3-7 configuration are calculated. After that, these features are transformed with a Short Time Gaussianization (STG) [17].

3. System Description

Both submitted systems are a fusion of eight subsystems: three acoustics and five phonotactics. We review each of them in the following subsections.

3.1. Acoustic Systems

3.1.1. GMM ML subsystem

In the GMM ML subsystem one GMM model for each language is calculated. A splitting process from 1 Gaussian up to 2048 is performed, in which every Gaussian is split in 2 in every step, and then several iterations of the EM algorithm are run. This generative method tries to maximize the likelihood of the data for each class.

3.1.2. GMM MMI subsystem

Starting from the GMM ML model, we perform a discriminative re-training based on MMI to obtain the final models. 10 iterations are run. Unlike ML, this method tries to maximize

the posterior probability of recognizing all training utterances given the labelled data [4]. The objective function is

$$F_{\text{MMI}}(\lambda) = \sum_{r=1}^{R} \log \frac{p_\lambda(\chi_r|s_{\text{r}})^{K_{\text{r}}} P(s_{\text{r}})^{K_{\text{r}}}}{\sum_{\forall s} p_\lambda(\chi_r|s)^{K_r} P(s)^{K_{\text{r}}}} \quad (1)$$

where $p_\lambda(\chi_r|s_{\text{r}})$ is the likelihood of r-th training segment, χ_r, given the correct language identity of the segment, s_r, and model parameters λ. R is the number of training segments, and the denominator represents the overall probability density, $p_\lambda(\chi_r)$. We consider the prior probabilities of all classes equal and drop the prior terms $P(s_{\text{r}})$ and $P(s)$. Usually, segment likelihood $p_\lambda(\chi_r|s)$ is computed as simple multiplication of frame likelihoods incorrectly assuming statistical independence of feature vectors. Factor $0 < K_r < 1$, which increases the confusion between hypothesis represented by numerator and denominator, can be considered as a compensation for underestimating segment likelihoods caused by this incorrect assumption.

3.1.3. GMM JFA subsystem

This system is based on a JFA for the mean of the models, following the principles of [5]. Two factors have been defined, one for the language and one for the channel. Thus, a channel compensated model for each language is obtained. This is a two-level hierarchy model, since we assume a different GMM that generates every speech segment, and we also assume that for every speech segment, this GMM has been generated by a sub-model. Then, for the speech segment *s*, we have

$$M_s = t_{l(s)} + U x_s \quad (2)$$

where $t_{l(s)}$ are the *language location vectors*, x_s is a vector of C segment-dependent *channel factors*, and U is a 114688-by-C *factor loading matrix*, which translates the channel factors from their low dimensional space to the high dimensional space where the model M_s lies (56*2048=114688).

The $t_{l(s)}$ matrix is obtained by MAP adaptation from a UBM model with mean m_0 and covariance matrix Σ, in the following way

$$t_{l(s)k} = \frac{\sum_s f_{sk}}{\tau + \sum_s n_{sk}} \quad (3)$$

being n_{sk} and f_{sk} the zero and first order statistics respectively, for the *k*th Gaussian component.

The U matrix and the channel factors are calculated according to a ML criterion, using the EM algorithm iteratively, in a similar way to [5]. The scoring of each utterance is computed via linear scoring, as proposed in [6].

3.2. Phonotactic Systems

Five PRLM sub-systems [19] in different languages have been fused in a Parallel PRLM fashion (PPRLM). Four of them use the Brno University of Technology (BUT) phoneme recognizer, based on ANN/HMM and Temporal Patterns (TRAPS) with Split Temporal Context (STC) [8], and are trained on Czech, Hungary, Russian and English. The other one uses the phoneme recognizer of the I3A, which is based on GMM/HMM with conventional MFCC and is trained on Spanish. In the latter, phonemes are taken with right and left context, so we will call the recognition unit subphoneme instead. However, we will keep only the central unit for the posterior step, that is, the phoneme without context. The output phonemes are used to train a LM for each one of the target languages with the SRILM toolkit [18].

All LMs are built with an interpolated Witten-Bell discounting method. We use 4-grams for training and testing in all cases, except for testing with the Spanish LM, where we use 3-grams. The reason is that we saw a better performance with this configuration. In addition, for the four phoneme recognizers based on GMM/ANN, we make use of the N-best hypothesis to get more information out of the acoustic signal. Specifically, we create a 100-best list to train and a 5-best list to test. The idea is similar to [20], where lattices are used. However, with the N-best hypothesis we look for a reduction in complexity.

3.3. Fusion for the Primary Submission

In our primary submission, the scores coming from each subsystem are T-Normalized [22] and fused. For the closed-set condition, another T-Normalization is applied after the fusion. The fusion is also a calibration [21] and the fused log-likelihood vector is

$$\mathbf{l}'(x_t) = \sum_{k=1}^{K} \alpha_k \mathbf{l}_k(x_t) + \beta \quad (4)$$

where the coefficients α_k and β are calculated via a discriminative Linear Logistic Regression (LLR) training, using the FoCal Multi-class toolkit [21], and $\mathbf{l}_k(x_t)$ is the output of system k when input in time t is x_t.

3.4. Fusion for the Alternative Submission

In this submission, we investigate the ZT-Normalization [22] technique, combined with a Gaussian Back-End (GBE) followed by a discriminative LLR fusion [21]. In the closed-set condition, scores after the GBE and after the LLR fusion are again T-Normalized.

In a GBE, the likelihood scores are obtained from multivariate Gaussians, with target language specific means and one common full covariance matrix. As explained in [23], the GBE can be seen as an affine transformation. The linear part of the transform is the same as a Linear Discriminant Analysis (LDA) transform, which tries to maximize the ratio of between-class to within-class variance. The translation part of the affine transform is equivalent to the calibration task of setting language dependent thresholds. The decision made by the GBE corresponds to the following normal distribution function:

$$\delta_l(\mathbf{x}) = (\mathbf{x} - \mu_l)^t \Sigma^{-1} (\mathbf{x} - \mu_l) \quad (5)$$

where μ_l is the mean for language l, Σ is the common covariance matrix, $\mathbf{x}$ is the input score and $\delta_l(\mathbf{x})$ is the transformed output. The posterior discriminative LLR increased the calibration to the system.

4. Analysis of Results

The performance of our systems is measured in terms of C_{avg}, which is a cost parameter defined in [1] in a similar way to the NIST evaluations. This parameter is the one to be minimized by our system. The description of the results will be focused on the primary system, since the results of the alternative system follow the same trend but with higher error rates. Comparison with the rest of sites can be found in [1].

4.1. Primary System - Closed Clean (CC)

In Fig. 1 the results of the primary system are shown, for the clean speech, closed set condition, and 30, 10 and 3 s tasks.

C_{avg} is 0.0184, 0.0418 and 0.0943, respectively. The 30 s test of this condition is the one used to rank systems in the evaluation.

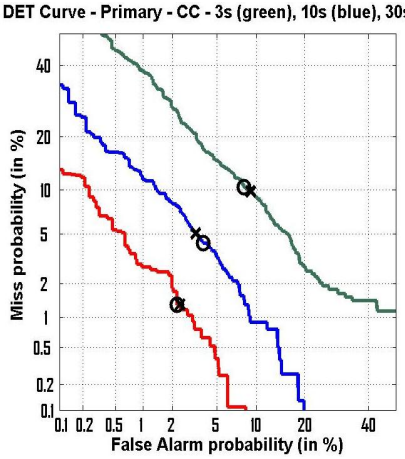

Figure 1: *DET Curves for Primary System CC condition. C_{avg} is marked with 'x' and C^*_{avg} with 'o'*

If we analyse Table 1, we see a very good performance recognizing all the languages for the 30 s CC condition, having a total P_{miss} of 0.0131. The highest error is for Galician with a P_{miss} of 0.050. However, if we look at the false-alarm probabilities, we check a good global performance, but a low one when discriminating between Spanish and Galician. The probability of saying that the transmitted language is Galician when it is really Spanish is 0.232 and of saying that it is Spanish when it is really Galician is 0.413. Several reasons could be considered for this behaviour, but we think that the most important is, after listening some of the recordings, the fact that many Galician speakers are Spanish-native speakers. Therefore, their Galician accents are very influenced by the Spanish one.

Table 1: *Error Rates for CC 30 s condition in the primary system. Target languages L_t in columns and segment languages L_n in rows. Labels mean SPA=Spanish, CAT=Catalan, ENG=English, BAS=Basque, GAL=Galician and POR=Portuguese*

$P_{fa}(L_t, L_n)$	Target Language L_t					
L_n	SPA	CAT	ENG	BAS	GAL	POR
SPA	–	0.016	0.000	0.000	0.232	0.000
CAT	0.020	–	0.000	0.000	0.007	0.007
ENG	0.000	0.000	–	0.000	0.000	0.000
BAS	0.000	0.000	0.000	–	0.000	0.000
GAL	0.413	0.016	0.000	0.000	–	0.000
POR	0.000	0.000	0.000	0.000	0.000	–
$P_{miss}(L_t)$	0.008	0.013	0.000	0.008	0.050	0.000
Avg. $P_{fa}(L_t)$	0.087	0.006	0.000	0.000	0.048	0.001
Avg. P_{miss} = 0.0131						
Avg. P_{fa} = 0.0237						

People speaking several languages are a general problem for language recognizers and it should be taken into account when training LR systems. One ideal solution would be to train models for native and non-native speakers independently, as if they were different languages. But recording a database that take this into consideration is difficult. So, we should think in other approaches to face this problem, like discriminative techniques that place more gaussians (in GMM systems) at the borders between these languages for a better characterization of these areas of the vector space. For more acoustically different languages, such as English and Basque, confusion rates are 0.

In the center column of Table 2 results for each individual subsystem presented in the 30 s condition of the evaluation are detailed. We can see that the subsystem that performs the best is the JFA. On the other hand, the PRLM_ES and PRLM_EN do not give good results by themselves. A post evaluation driven by the GTTS group of the Basque Country University (the organizers) revealed that we could have obtained better results using only a T-norm followed by a GBE in the backend. This shows the great importance of this stage in a LR system. These results are in the right column of Table 2.

Table 2: *Left, C_{avg} for the individual subsystems of the primary submission for the CC 30 s condition. Right, C_{avg} after post-evaluation*

Subsystem	C_{avg}	Post Evaluation C_{avg} GTTS
JFA	0.0357	0.0186
ML	0.0855	0.0637
MMI	0.0598	0.0433
PRLM_CZ	0.0569	0.0480
PRLM_HU	0.0501	0.0441
PRLM_RU	0.0547	0.0472
PRLM_EN	0.2618	0.2576
PRLM_ES	0.1474	0.1321

4.2. Primary System - Open Clean (OC)

Results in terms of C_{avg} are 0.0307, 0.0644 and 0.1202, for the 30, 10 and 3 s conditions, respectively. The corresponding DET curve can be seen in Fig. 2. We check that the performance of the system has dropped slightly, as expected after introducing Out-Of-Set (OOS) languages The average P_{miss} drops to 0.0193, but more dramatic is the decrease in the average P_{fa}, which drops to 0.0422, i.e. a relative decrease of 79% compared to the 30s CC condition. These results are in Table 3. The main confusion of OOS languages is with English.

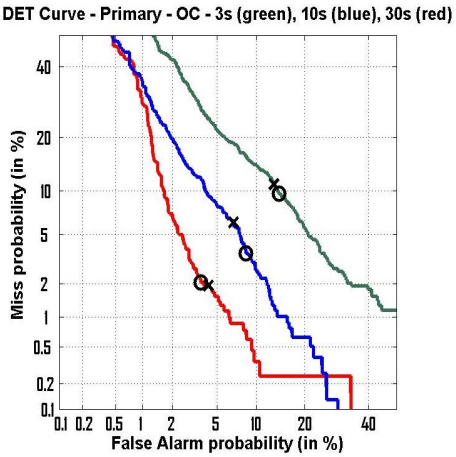

Figure 2: *DET Curves for Primary System OC condition*

4.3. Alternative System - CC and OC

For the alternative system, based on a different backend explained in section 3.4, a reduction on the accuracy of the system is experimented. For the CC condition, C_{avg} is 0.0238, 0.0498 and 0.1087, for the 30, 10 and 3 s tasks respectively. For the OC condition, C_{avg} is 0.0373, 0.0635 and 0.1309, respectively.

Table 3: *Error Rates for OC 30 s condition in the primary system. Target languages L_t in columns and segment languages L_n in rows. Labels of languages are SPA=Spanish, CAT=Catalan, ENG=English, BAS=Basque, GAL=Galician, POR=Portuguese and OOS=Out-Of-Set*

$P_{fa}(L_t, L_n)$	Target Language L_t					
L_n	SPA	CAT	ENG	BAS	GAL	POR
SPA	–	0.008	0.000	0.000	0.208	0.000
CAT	0.013	–	0.000	0.000	0.000	0.000
ENG	0.000	0.000	–	0.000	0.000	0.000
BAS	0.000	0.000	0.000	–	0.000	0.000
GAL	0.446	0.008	0.000	0.000	–	0.000
POR	0.000	0.000	0.000	0.000	0.000	–
OOS	0.026	0.064	0.188	0.052	0.003	0.094
$P_{miss}(L_t)$	0.024	0.013	0.007	0.008	0.050	0.013
Avg.$P_{fa}(L_t)$	0.092	0.003	0.000	0.000	0.042	0.000
Avg.$P_{fa}(L_t + L_o)$	0.066	0.028	0.075	0.021	0.026	0.038
Avg. P_{miss} = 0.0193						
Avg. P_{fa} = 0.0422						

5. Conclusions

In this edition of Albayzin LRE, I3A participated for the first time in a LR Evaluation. We built several state-of-the-art systems that were adapted to and tested in the KALAKA-2 database. For each submission, all the systems were finally fused into one, in which the characteristics of each were combined to improve the performance. We used acoustic and phonotactic systems, because the information they provide is different and, as we have seen, they contributed complementarily to the final decision. For the ranking of systems in the evaluation only the primary submission of the CC 30 s condition was considered and our system was the most competitive.

Two submissions that differ only in the backend were presented. The final performance depended dramatically on the back stage, and in the post-evaluation, we could check that our two implementations could have worked better. Therefore, more experimentation in this area is needed.

In global, our system presented very low P_{miss} values and also low P_{fa} values. However, we detected a big confusion between Spanish and Galician, mainly caused by the fact that many Galician speaker were non-native speakers and their accent was influenced by the Spanish language. Multilingual speakers are a degradation source for LR systems and can be an interesting point to be studied in future investigations.

6. Acknowledgements

We would like to thank GTTS for his big work organizing Albayzin 2010 LRE, and also the organization of Fala 2010 for supporting this evaluation.

This work was funded by the Spanish Ministry of Science and Innovation under project TIN2008-06856-C05-04.

7. References

[1] L.J. Rodríguez-Fuentes, M. Penagarikano, A. Varona, M. Díez, G. Bordel, "Overview of the Albayzin 2010 Language Recognition Evaluation: Database Design, Evaluation Plan and Preliminary Analysis of Results", in FALA 2010, VI Jornadas en Tecnologa del Habla and II Iberian SLTech Workshop, Vigo.

[2] KALAKA-2. Speech Database created for the Albayzin 2010 Language Recognition Evaluation, organized by the Spanish Network on Speech Technology. Produced by the Software Technologies Working Group (GTTS, http://gtts.ehu.es), University of the Basque Country.

[3] L.J. Rodríguez-Fuentes, M. Penagarikano, G. Bordel, and A. Varona, "The Albayzin 2008 Language Recognition Evaluation", in Proceedings of Odyssey 2010: The Speaker and Language Recognition Workshop, Brno, Czech Republic, 28 June - 1 July 2010.

[4] D. Povey, "Discriminative Training for Large Vocabulary Speech Recognition", Ph.D. thesis, Cambridge University, July 2004.

[5] P. Kenny, "Joint Factor Analysis of Speaker and Session Variability: Theory and Algorithms", Technical Report CRIM-06/08-13, CRIM, 2005, http://www.crim.ca/perso/patrick.kenny/FAtheory.pdf.

[6] O. Glembek, L. Burget, N. Dehak, N. Brümmer, and P. Kenny, "Comparison of Scoring Methods used in Speaker Recognition with Joint Factor Analysis", in Proc. ICASSP, Taipei, Apr. 2009.

[7] N. Brümmer, A. Strasheim, V. Hubeika, P. Matějka, P. Schwarz, J. Černocký, "Discriminative Acoustic Language Recognition via Channel-Compensated GMM Statistics", in Proc. Interspeech 2009, Brighton, GB.

[8] P. Schwarz, "Phoneme Recognition Based on Long Temporal Context", Ph.D. Thesis, Brno University of Technology, 2009. http://speech.fit.vutbr.cz/cs/software/phoneme-recognizer-based-long-temporal-context.

[9] A.P. Dempster, N.M Laird, D.B. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm", Journal of Royal Statistical Society, Series B (Methodological), Vol. 39, No.1, 1977, pp.1-38.

[10] http://www.fee.vutbr.cz/SPEECHDAT-E/sample/czech.html

[11] http://www.fee.vutbr.cz/SPEECHDAT-E/sample/hungarian.html

[12] http://www.fee.vutbr.cz/SPEECHDAT-E/sample/russian.html

[13] http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC93S1

[14] A. Moreno, D. Poch, A. Bonafonte, E. Lleida, J. Llisterri, J.- B. M. no, and C. Nadeu, "Albayzin Speech Database: Design of the Phonetic Corpus", in Proceedings of the 3rd European Conference on Speech Communication and Technology (Eurospeech-Interspeech), Berlin, Germany, September 1993.

[15] A. Moreno, B. Lindberg, C. Draxler, G. Richard, K. Choukri, S. Euler, and J. Allen, "Speech Dat Car. A Large Speech Database for Automotive Environments", in Proceedings of the II Language Resources European Conference, Athens, Greece, June 2000.

[16] P.A. Torres-Carrasquillo, E. Singer, M.A. Kohler, R.J. Greene, D.A. Reynolds, and J.R. Deller Jr., "Approaches to Language Identification using Gaussian Mixture Models and Shifted Delta Cepstal Features", in Proc. International Conferences on Spoken Language Processing, Sept. 2002.

[17] J. Pelecanos and S. Sridharan, "Feature Warping for Robust Speaker Verification", Proc. Speaker Odyssey 2001 conference, June 2001.

[18] A. Stolcke, "SRILM - An Extensible Language Modeling Toolkit", in Proc. ICSLP, pp. 901-904, 2002. http://www.speech.sri.com/projects/srilm.

[19] M.A. Zissman, "Comparison of Four Approaches to Automatic Language Identification of Telepone Speech", IEEE Trans. Acoust., Speech Signal Processing, vol. 4, no. 1, pp. 31-44, 1996.

[20] J.L. Gauvain, A. Messaoudi, and H. Schwenk, "Language Recognition using Phoneme Lattices", in Proc. International Conferences on Spoken Language Processing (ICSLP), Sept. 2004, pp. 1283-1286.

[21] N. Brümmer, "FoCal Multi-class: Toolkit for Evaluation, Fusion and Calibration of Multi-class Recognition Scores - Tutorial and User Manual"-. http://sites.google.com/site/nikobrummer/focalmulticlass.

[22] R. Auckenhalter, M. Carey, and H. Lloyd-Thomas, "Score Normalization for Text-Independent Speaker Verification Systems", Digital Signal Processing, 10(1), 42-54.

[23] D.A. van Leeuwen and N. Brümmer, "Channel-Dependent GMM and Multi-Class Logistic Regression Models for Language Recognition", 2006 IEEE Odyssey: The Speaker and Language Recognition Workshop.