LOW-RESOURCE LANGUAGE RECOGNITION USING A FUSION OF PHONEME POSTERIORGRAM COUNTS, ACOUSTIC AND GLOTTAL-BASED I-VECTORS

L. F. D'Haro, R. Cordoba, M. A. Caraballo, J. M. Pardo

Speech Technology Group – Dpto. de Ing. Electrónica – E.T.S.I. Telecomunicación - Universidad Politécnica de Madrid.
Ciudad Universitaria s/n 28040, Madrid, Spain
{lfdharo,cordoba, macaraballo,pardo}@die.upm.es

ABSTRACT

This paper presents a description of our system for the Albayzin 2012 LRE competition. One of the main characteristics of this evaluation was the reduced number of available files for training the system, especially for the empty condition where no training data set was provided but only a development set. In addition, the whole database was created from online videos and around one third of the training data was labeled as noisy files. Our primary system was the fusion of three different i-vector based systems: one acoustic system based on MFCCs, a phonotactic system using trigrams of phone-posteriorgram counts, and another acoustic system based on RPLPs that improved robustness against noise. A contrastive system that included new features based on the glottal source was also presented. Official and post-evaluation results for all the conditions using the proposed metrics for the evaluation and the C_{avg} metric are presented in the paper.

Index Terms—LID system, noise robustness, scarce data, posteriorgram counts, i-vectors

1. INTRODUCTION

In this paper, we describe the final system that we presented for the Albayzin 2012 LRE evaluation [1]. In this evaluation, our primary system obtained very good results thanks to the fusion of 4 different subsystems: 1) Acoustic system based on MFCC-SDC features + RASTA + i-vectors, 2) Phonotactic system based on trigram posteriorgram counts + i-vectors, 3) Acoustic system based on RPLP-SDC features + RASTA + i-vectors, and 4) Prosodic system based on glottal source features + i-vectors. As most of current state-of-the-art LID systems, all our subsystems make use of subspace projections in the form of i-vectors [2] that were calibrated and fused using multiclass logistic regression. The main advantage of our system was the use of RPLP features and the incorporation of the phonotactic system that use non-sparse features from the posteriorgram output of a phoneme recognizer.

The paper is organized as follows: Section 2 describes the database and sets used for the evaluation. Section 3 explains each subsystem and results. Section 4 covers the fusion and calibration results. Finally, section 5 presents the conclusions and future work.

2. EVALUATION AND DATA DESCRIPTION

The Albayzin LRE 2012 was an evaluation organized by the Software Technologies Working Group of the University of the Basque Country and Niko Brümmer from Agnitio Research, as part of the IberSpeech 2012 conference [3]. This evaluation was more difficult than the previous ones as it changed the application domain from TV broadcast speech to any kind of speech found on Internet, without providing training data for some of the target languages (a common situation for low-resource languages), and forbidding the use of any additional database. The provided audio files were extracted from YouTube videos, with different length durations, channel conditions, number of speakers, etc. The files might contain music, noise and any kind of non-human sounds. Each audio file contained speech in a single language, except for signals corresponding to out-of-languages, which might contain speech in two or more languages but none of them were target languages. All files were 16 KHz@16 bits but we down-sampled them to 8 KHz. Table 1 shows the statistics of the database as well as the number of files used in our setup and experiments.

Two different conditions were proposed: Plenty and Empty with the aim of evaluating to what extent the availability of training materials (and thus specific models) for target languages affected system performance. For the plenty condition, training and dev. data was available, but for the empty condition only dev. Data was available. For the plenty condition the target languages were: Spanish, Catalan, Basque, Galician, Portuguese, and English. For the empty condition the target languages were: French, German, Greek, and Italian. Finally, for each condition, closed and open conditions were also proposed to check if the systems were able to identify out-of-set languages (OOL).

2.1. Plenty conditions

We divided the original dev. set into two subsets with a similar language distribution. The first one is the "Dev." set used to calibrate the system, and the second one is the "Test" set, which we used to obtain our pre-evaluation results. For the final evaluation, we added the "Test" set to the Train set to have more training data, but calibrated the system only with the "Dev." set.

		Closed				Open			
		Train	**Dev**	**Test**	**Eval**	**Train**	**Dev**	**Test**	**Eval**
Plenty (6 lang.)	No. Files	4656	458	457	941	5265	725	725	1477
	No. of clean files	3060	-	-	-	3060	-	-	-
	No. of noisy files	1596	-	-	-	1596	-	-	-
Empty (4 lang.)	No. Files	-	304	305	631	-	571	571	1167
	For our system	7400 (*)	304	305		10141 (**)	571	571	

Table 1. Dataset statistics for all the conditions

2.2. Empty conditions

Table 1 also shows the number of files used in our setup for training, development, test and evaluation. As there was no Train set, we reused the Train set from the plenty conditions and merged it with the development data available (replicated three times to give it more relevance). We did not apply any adaptation technique to the models from the plenty conditions. In Table 1, we can see that for (*) we merged the data from the plenty closed (PC) training data with the PC dev data and 3 times the empty closed (EC) dev data. In the same way, for (**), we merged the data from plenty open (PO) training data with the PO dev. data and 3 times the empty open (EO) dev. data.

As in the plenty conditions, for the final evaluation we also added the "Test" set to the Train set to have more training data, calibrating the system with the "Dev." set. For training the logistic regression classifier for the EC condition, we have unified the "Dev." and "Test" sets but the calibration was done only on the "Dev." Set. For the EO condition we have used 10141 files for training the LR classifier and the calibration was done on the 571 files.

3. SYSTEM DESCRIPTION

In this section, we will describe each of the subsystems that we used and fused for creating the final systems.

3.1. MFCC-SDC acoustic system + i-vectors

For this subsystem, we use SPRO5 [4] for extracting, for each file, 12 MFCC coefficients (including C0) from 24 Mel filter banks plus the energy for each frame. Finally, Cepstral mean and variance normalization was applied. The Voice Activity Detector (VAD) used for this subsystem (and also for the RPLP subsystem) was the output from the BUT Hungarian phone recognizer. Then, we suppressed all segments marked as silence or noise in the output. After that, every 10 ms speech frame was mapped to a 56-dimensional feature vector generated from the concatenation of SDC features [5] using the 7-1-3-7 configuration. Finally, a RASTA filter was applied to reduce short-term noise variations in each frequency sub-band. Then, we trained the universal background model (UBM) through five iterations of the EM-algorithm and using all the feature vectors coming from all the languages that appeared in the training set. Then, we extracted the i-vectors following the same algorithm reported in [2]. Finally, these i-vectors were normalized by first subtracting the mean of all the training i-vectors and then dividing them by its corresponding norm.

For the plenty conditions, 400 i-vectors provided better results with little difference between 512 and 1024 Gaussians, so we used 512 Gaussians. For the empty conditions, we used 64 Gaussians and 400 i-vectors.

3.2. RPLP-SDC acoustic system + i-vectors

Our goal with this subsystem was to introduce a new set of features which could be more robust against noise. In this case, we decided to use the RPLP (Revised PLP) features used in [6] and proposed in [7]. These features can be seen as a hybrid approach between MFCC and PLP, combining the best of both. The main advantage is that it performs a double suppression of spectral dynamics before calculating the cepstral coefficients and with less effect on the accuracy when modifying the number of FB bands, shape, and non-linearity scaling. .In [6] good improvements were found for ASR recognition in comparison with the standard features.

Finally, we applied a RASTA filter to these coefficients and then we applied SDC with the same configuration as for MFCC. In our experiments, there was little difference between 512 and 1024 Gaussians, so we used 512. We can see in Table 2 that RPLP results are better than those obtained with MFCC for both conditions.

3.3. Glottal source based system + i-vector

The goal of this subsystem was to check the viability of using glottal source features for language recognition based on the good results reported by [8] on speaking style identification using only prosodic information (74% accuracy rate) and by [9] for classifying expressive speech: a 95% for styled speech and 82% for emotional speech.

GlottHMM [10] is a vocoding toolkit recently developed for parametric speech synthesis. It is based on decomposing speech into the glottal source and vocal tract through glottal inverse filtering. In our case, we have used GlottHMM to extract only the F0 and the Harmonics to Noise Ratio (HNR) of the glottal source and then calculating the SDCs coefficients on these features. HNR is evaluated based on the ratio between the upper and lower smoothed spectral

envelopes and averaged across five frequency bands according to the equivalent rectangular bandwidth (ERB) scale. In contrast to the previous subsystems, the F0 information was used as VAD. Finally, we extracted the i-vectors using the same approach as for the acoustic subsystems.

For the plenty conditions, best results were obtained using 128 Gaussians, for the empty conditions only 16. This subsystem was used only for the contrastive systems.

3.4. Phonotactic system + i-vectors

For this subsystem, we used a novel approach to phonotactic LID reported in [11], where instead of using soft-counts based on phoneme lattices, we use posteriorgrams to obtain n-gram counts. In this approach, the high-dimensional vectors of counts are reduced to low-dimensional units for which we adapted the commonly used technique i-vectors. The reduction is based on subspace multinomial modeling (SMM, [12][13]) and is designed to work in the total-variability space. In comparison with other techniques based on soft-counts, the new features do not present sparse counts, and avoid the use of pruning techniques when creating the lattices. Reported results on NIST 2009 LRE showed better results compared to use soft-counts, and with very good results when fused with an acoustic i-vector LID system. In this evaluation, we tested again its robustness and success, as we will see in section 4.

Figure 1 shows the process of creating the vector of posteriorgram-based n-gram counts. In the figure, we consider the bigram counts for simplicity, but in our system we used trigrams. The process can be divided into four steps:

The first step is to tokenize speech by the means of running a phone-recognizer that, for each frame, provides the posterior probabilities of the phone occurrences. In our experiments, we used the BUT Hungarian phone recognizer.

The second step is to sum up and average the posterior probabilities for the frames that are considered to be within the same phoneme unit (e.g., A in the Figure). The phone boundaries are obtained by running Viterbi decoding on the posteriorgram. The averaged posteriorgrams provide a good de-correlation and smoothness for the resulting matrix.

The third step is to create the joint-posteriorgram – a sequence of matrices of joint probabilities for the n consecutive frames. Here, we take the averaged posteriorgram of each frame and we do the outer product with the posteriorgram of the previous frame. The process is repeated for all the phone-grams considering the n-1 history.

The final step is to sum up all frames (matrices) of the joint-posteriorgram. This way, we create a matrix of n-gram counts that is converted into a 1xD vector (where D is the total number of possible n-grams, in the case of trigrams is 33^3=35937, using 33 phonemes) and then used as a feature vector for training the i-vectors using SMMs.

For training the i-vectors, we have followed the algorithm reported in [14] with slight modifications in the EM step (see [11]). Finally, in our experiments, we have considered a set of 1089 multinomial models when using trigrams (i.e. considering all the possible number of bigram histories, 33x33, using 33 phones for the Hungarian recognizer). For the empty condition, we obtained the T matrix by using the created training set described in section 2.2 together with the development set and applying two epochs and two iterations for the EM i-vectors extraction process. Then, the new T matrix was used to extract the final i-vectors for all sets.

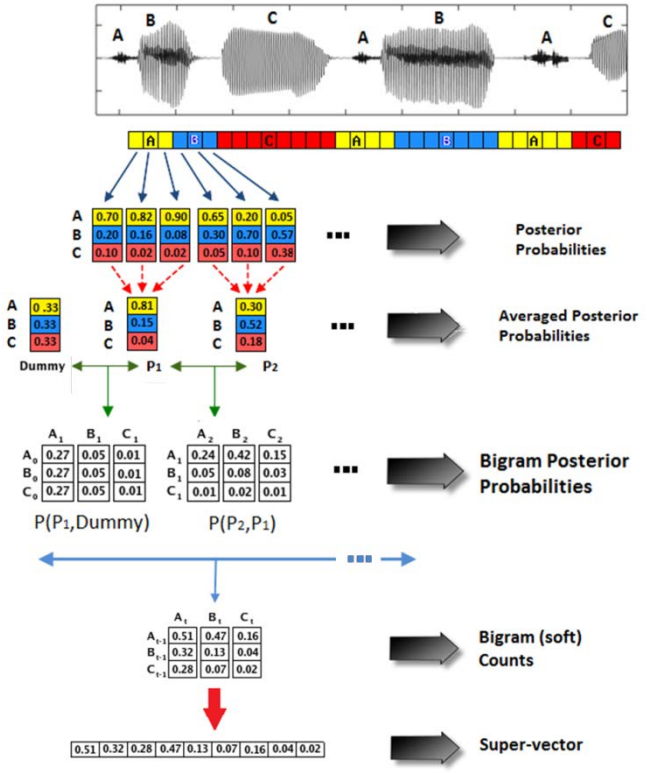

Figure 1. Procedure to generate posteriorgram counts

3.5. Classifier and calibration back-end

As classifier for all our subsystems, we used a Multiclass logistic regression that generates a different classifier for each language to recognize. Then, these classifiers were used to generate scores for the files in our development, test and evaluation sets. For calibration and fusion, a Gaussian Back-end followed by a Discriminative Multi-Class Logistic Regression was used to post-process the scores obtained before. Previously, the input vectors were conditioned by within-class covariance normalization (WCCN, [16]).

4. RESULTS, CONCLUSIONS & FUTURE WORK

Table 2 shows the results for each subsystem, for Test (pre-evaluation results) and Eval (final results with the evaluation data). The main metric for the evaluation was Fact as defined in the evaluation plan [15]. This metric measures the ratio by which the system has changed the prior confusion value (i.e. equally distributed probability among all languages), and allows to know how discriminative is a system

and how well calibrated it is. This way, an Fact value higher than one means that the system is performing worse than the naive system, and values close to zero are the goal, being 0 a perfect system. We also included the C_{avg} metric used in NIST evaluations for an easier comparison.

Table 3 shows the results for the different fusions we have tested during and after the evaluation. It also includes the Fusion weights estimated by the classifier back-end.

For the plenty conditions, considering Fact, we can see that the RPLP subsystem consistently outperforms the MFCC. The phonotactic system (Phon) performs better than both of them in Test but slightly worse in Eval. In any case, the fusion of RPLP+Phon outperforms the fusion of MFCC+RPLP, with Phon having more weight in the fusion. This proves that Phon provides complementary information. The fusion of MFCC+RPLP+Phon provides the best results in all cases, having Phon the highest weight. The Glot system did not provide additional improvements, probably because glottal information is more speaker-dependent than language-dependent.

Similar conclusions can be obtained for the empty conditions. As the data available is so small, results degrade but our proposed approach also obtained very good results in the competition. For example the Fact for the next best system in the EC condition was 0.262847 and for EO condition was 0.289229.

In summary, we have described the system that we presented for the Albayzin 2012 LRE evaluation. We have seen that the novel phonotactic system and the use of RPLP features have drastically improved the performance, obtaining the best results.

As future work, we will test new features to improve robustness against noise, as well as the incorporation of a discriminative selection of n-grams, that we have successfully developed in previous work [17], in order to reduce the feature vector size for the phonotactic subsystem.

5. ACKNOWLEDGMENT

This work has been supported by TIMPANO (TIN2011-28169-C05-03), INAPRA (MICINN, DPI2010-21247-C02-02), MA2VICMR (Comunidad Autonoma de Madrid, S2009/TIC-1542) projects, and the European project Simple4All (under grant agreement No. 287678).

Condition	Subsystem Type	Configuration	Closed				Open			
			Test		Eval		Test		Eval	
			Fact	C_{avg}(%)	Fact	C_{avg} (%)	Fact	C_{avg} (%)	Fact	C_{avg} (%)
Plenty	MFCC-SDC	400iv, 512 Gauss.	0.172519	9.25	0.174073	8.37	0.200035	10.50	0.175059	9.74
	RPLP-SDC	400iv, 512 Gauss	0.159279	7.62	0.162301	8.43	0.173536	10.20	0.160839	9.29
	Phonotactic	400iv, 1089 Gauss	0.138718	9.43	0.181393	9.85	0.163411	10.37	0.189176	11.30
	Glottal Source	400iv, 512 Gauss	0.668717	30.81	0.741573	32.90	0.718101	32.64	0.749976	34.82
Empty	MFCC-SDC	400iv, 64 Gauss.	0.075818	0.43	0.406734	14.71	0.092105	3.32	0.334499	16.63
	RPLP-SDC	400iv, 256 Gauss	0.038545	0.047	0.324332	11.33	0.050978	1.27	0.185699	10.24
	Phonotactic	400iv, 1089 Gauss	0.037714	0.17	0.498291	21.82	0.047180	2.40	0.296573	15.88
	Glottal Source	400iv, 16 Gauss.	0.082595	2.29	1.000507	47.18	0.162338	7.96	0.956041	43.61

Table 2. Best results for each subsystem on the test and evaluation sets

Condition	Type	System 1	System 2	System 3	System4	Closed			Open		
						Test	Eval		Test	Eval	
						Fact (Cavg %)	Fusion weights	Fact (Cavg %)	Fact (Cavg %)	Fusion weights	Fact (Cavg %)
Plenty	2 systems	MFCC-512G	RPLP-512G	-	-	0.096189 (5.39)	1.30;1.74	0.084610 (5.95)	0.111999 (7.71)	1.32;1.38	0.091928 (6.60)
	2 systems	Phon-1089G	RPLP-512G	-	-	0.072827 (4.43)	1.94;1.59	0.076544 (5.59)	0.079405 (5.25)	1.49;1.44	0.083729 (6.29)
	Primary	MFCC-512G	Phon-1089G	RPLP-512G	-	0.069258 (4.16)	0.84;1.73;1.14	**0.067717 (4.79)**	0.080184 (5.77)	0.99;1.36;0.79	**0.076513 (6.04)**
	Contrastive	MFCC-512G	Phon-1089G	Glot-128G	-	0.071393 (4.16)	1.58;1.91;0.26	0.076562 (5.50)	-	-	-
	Contrastive2	RPLP-512G	Phon-1089G	Glot-512G	-	-		-	0.079517 (5.37)	1.43;1.45;0.66	0.086465 (6.48)
	4 systems	MFCC-512G	Phon-1089G	RPLP-512G	Glot-128G	0.068014 (3.93)	0.84;1.70; 1.14;0.25	0.068009 (4.77)	0.078737 (5.63)	0.99;1.32; 0.81;0.66	0.079370 (6.04)
Empty	2 systems	Phon-1089G	RPLP-256G	-	-	0 (0)	1.42;1.36	0.180540 (9.55)	0 (0)	1.03;1.31	0.140641 (8.79)
	Primary	MFCC-64G	Phon-1089G	RPLP-256G	-	0 (0)	0.97;1.35;1.05	**0.141554 (7.53)**	0 (0)	0.83;0.99;1.10	**0.129796 (8.20)**
	Contrastive	RPLP-256G	Phon-1089G	Glot-16G	-	0 (0)	1.29;1.41;0.94	0.187794 (10.04)	0 (0)	1.30;1.03;0.67	0.141322 (8.70)
	4 systems	MFCC-64G	Phon-1089G	RPLP-256G	Glot-16G	0 (0)	0.96;1.33; 1.00;0.80	0.146676 (7.96)	0 (0)	0.83;0.99; 1.10;0.09	0.129708 (8.21)

Table 3. Summary of all systems presented to the evaluation and final results

6. BIBLIOGRAPHY

[1]. Luis Fernando D'Haro, Ricardo Córdoba. 2012. "The GTH-LID System for the Albayzin LRE12 Evaluation". In Proc. Iberspeech 2012, pp. 528-539.

[2]. Najim Dehak, Patrick Kenny, Réda Dehak, Pierre Dumouchel, and Pierre Ouellet, "Front-End Factor Analysis For Speaker Verification", IEEE Transactions on Audio, Speech and Language Processing, Vol. 19 (4), May 2011.

[3]. Iberspeech conference. Homepage 2012. http://iberspeech2012.ii.uam.es/

[4]. Guillaume Gravier (2012, November 30), Spro 5 available online at https://gforge.inria.fr/projects/spro/

[5]. 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 cepstral features," in Proc. International Conferences on Spoken Language Processing (ICSLP), Sept. 2002, pp. 89–92.

[6]. Rajnoha, J., and Pollák, P. 2011. "ASR systems in Noisy Environment: Analysis and Solutions for Increasing Noise Robustness". Radionegineering, Vol. 20, No. 1, April 2011, pp. 74-84.

[7]. Hönig, F., Stemmer, G., Hacker, C., Brugnara, F. "Revising Perceptual Linear Prediction (PLP)". In Eurospeech 2005, p. 2997-3000.

[8]. Nicolas Obin, "MeLos: Analysis and Modelling of Speech Prosody and Speaking Style", PhD, Ircam-UPMC, Paris, 2011.

[9]. Jaime Lorenzo-Trueba, Roberto Barra-Chicote, Tuomo Raitio, Nicolas Obin, Paavo Alku, Junichi Yamagishi, Juan M Montero. "Towards Glottal Source Controllability in Expressive Speech Synthesis", in Proc. of Interspeech 2012.

[10]. Raitio, T. and Suni, A. and Yamagishi, J. and Pulakka, H. and Nurminen, J. and Vainio, M. and Alku, P., "HMM-based speech synthesis utilizing glottal inverse filtering", Audio, Speech, and Language Processing, IEEE Transactions on, 9:153-165, IEEE, 2011.

[11]. Luis Fernando D'Haro, Ondrej Glembek, Oldrich Plchot, Pavel Matejka, Mehdi Soufifar, Ricardo Cordoba, Jan Cernocký. "Phonotactic Language Recognition using i-vectors and Phoneme Posteriorgram Counts", in Proc. of Interspeech 2012.

[12]. D. Povey, Lukas Burget et. al, 2011. "The Subspace Gaussian Mixture Model– a Structured Model for Speech Recognition", Computer Speech and Language, 25(2), pp. 404-439

[13]. D. Martínez, O. Plchot, L. Burget, O. Glembek, and P. Matejka, "Language Recognition in iVectors Space", in Proc. of Interspeech 2011.

[14]. Kockmann, et al, 2010. "Prosodic speaker verification using subspace multinomial models with intersession compensation," in Proc. of ICSPL, Makuhari, Chiba, Japan, 2010.

[15]. Luis Javier Rodriguez-Fuentes, Niko Brummer, et al.2012. Albayzin LRE 2012 evaluation plan available at http://iberspeech2012.ii.uam.es/images/PDFs/albayzin_lre12_evalplan_v1.3_springer.pdf

[16]. Hatch and A. Stolcke, 2006. "Generalized linear kernels for one-versus-all classification: application to speaker recognition," ICASSP Vol 5, page V.

[17]. R. Córdoba, L. F. D'Haro, F. Fernandez-Martinez, J. Macias-Guarasa, J. Ferreiros. 2007. "Language Identification based on n-gram Frequency Ranking". Proc. of Interspeech 2007, pp. 2137-2140. Antwerp, Belgium, 27-31 of August 2007.