The Albayzin 2008 Language Recognition Evaluation

Luis Javier Rodríguez-Fuentes, Mikel Penagarikano, Germán Bordel, Amparo Varona

Software Technologies Working Group (http://gtts.ehu.es)
Faculty of Science and Technology, University of the Basque Country
Barrio Sarriena s/n, 48940 Leioa, Spain
luisjavier.rodriguez@ehu.es

Abstract

The Albayzin 2008 Language Recognition Evaluation was held from May to October 2008, and their results presented and discussed among the participating teams at the 5th Biennial Workshop on Speech Technology [1], organized by the Spanish Network on Speech Technologies [2] in November 2008. In this paper, we present (for the first time) a full description of the Albayzin 2008 LRE and analyze and discuss recognition results. The evaluation was designed according to the test procedures, protocols and performance measures used in the NIST 2007 LRE. The KALAKA database [3], consisting of 16 kHz audio signals recorded from TV broadcasts, was created *ad-hoc* and used for the evaluation. The four official languages spoken in Spain (Basque, Catalan, Galician and Spanish) were taken as target languages, other (*unknown*) languages being also recorded to allow open-set verification tests. The best system, employing state-of-the-art technology, yielded $C_{avg} = 0,0552$ (around 5% EER) in closed-set verification tests on a set of 30-second segments. This reveals the difficulty of the task, despite using 16 kHz speech signals and having only four target languages. We plan to include also Portuguese and English as target languages for the next Albayzin 2010 LRE.

1. Introduction

The Albayzin 2008 Language Recognition Evaluation was organized and coordinated by the Software Technologies Working Group of the University of the Basque Country, as part of the activities of the Spanish Network on Speech Technology (SNST) [2]. Every two years, the SNST organizes a workshop which includes the so called *Albayzin* system evaluations. The 5th Biennial Workshop on Speech Technology included evaluations on three topics: speech translation, speech synthesis and language recognition (see [1] for details).

The main goal of the Albayzin 2008 LRE was to promote collaboration between research teams from Spain and Portugal that focus their research on language recognition. With this purpose in mind, a language verification task was designed, based on NIST 2007 LRE, but with only 4 target languages (Spanish, Catalan, Basque and Galician) and using 16 kHz audio signals (with medium to high SNR). Four different test conditions were defined, depending on development conditions (free vs. restricted) and the operation mode (open-set vs. closed-set). The closed-set restricted-development track was taken as reference to choose the best system.

The organizing team provided data for training, development and evaluation purposes, consisting on 16 kHz TV broadcast recordings. Training data amount to more than 8 hours of speech per target language. Development data consist of three subsets of speech segments, corresponding to three nominal durations: 30, 10 and 3 seconds. Each segment contains speech either in a target language, or in other *unknown* language from a set of languages not specified to participants. A key file with information about each segment was provided too. Finally, the evaluation set is structured exactly the same way as the development set, except for the segments in *unknown* languages, whose distribution is different.

The task was defined in the same terms as for NIST 2007 LRE [4]: *for each segment and each target language, systems must determine, via a hard decision and a score, whether or not the test segment contains speech of the target language*. As noted above, systems could be built under four possible test conditions. On the other hand, trials corresponding to the three subsets of segments, with nominal durations of 30, 10 and 3 seconds, were scored separately, thus resulting in 12 possible scorings (4 test conditions × 3 durations).

System performance was primarily measured by the cost function used in NIST evaluations, based on miss and false alarm rates between pairs of languages, which are then averaged to get a pooled cost. We also offered an alternative cost measure for those systems whose scores may be interpreted as log-likelihood ratios, the so called C_{LLR}, as defined in [5, 6]. Finally, Detection Error Tradeoff (DET) curves were used to graphically compare the global performance of systems.

This work has been supported by the Government of the Basque Country, under program SAIOTEK (project S-PE09UN47), and the Spanish MICINN, under Plan Nacional de I+D+i (project TIN2009-07446, partially financed by FEDER funds).

The rest of the paper is organized as follows. The language detection task is briefly defined in Section 2. The database, produced specifically and distributed to train, tune and evaluate language recognition systems in the Albayzin 2008 LRE, is described in Section 3. Section 4 defines the measures used to evaluate system performance, which are basically the same used in NIST 2007 LRE and are included here only for the sake of completeness. Section 5 addresses issues related to the organization of Albayzin 2008 LRE. Results are presented and discussed in Section 6, with special attention to the closed-set restricted-development condition (which was mandatory), but devoting some space to more detailed analyses. Finally, conclusions and future work are outlined in Section 7.

2. The language detection task

The language detection task was stated as for NIST evaluations [4]: *given a segment of speech and a language of interest (target language), determine whether or not that language is spoken in the segment, based on an automated analysis of the data contained in the segment*. Performance was computed by presenting the system a set of trials. Each trial comprises the following elements: (1) a segment of audio containing speech in a single language; (2) the target language; and (3) the non-target languages, that is, those languages that may be spoken in the segment. For each trial, the system must output: (1) a hard decision (yes/no) about whether or not the target language is spoken in the segment; and (2) a score indicating how likely is for the system that the target language is spoken in the segment, the higher the score the greater the confidence that the segment contains the target language. Participants may optionally state that their scores can be interpreted as log-likelihood ratios, in order to compute the alternative performance measure C_{LLR} defined in Section 4.2.

Regarding system development, two different conditions were defined: (1) *free development*, which means that any available materials can be used for system development; and (2) *restricted development*, which forces to use only the train and development materials provided in Albayzin 2008 LRE. It must be noted that restricted development implied that external materials could be used neither directly nor indirectly. For instance, acoustic models trained on an external acoustic database were not allowed. These conditions were defined with the aim to compensate for the advantage of research groups having lots of resources, with regard to those having almost nothing. The restricted-development condition tried to put all the groups at the starting line of having no previous data.

With regard to verification tests, open-set and closed-set operation modes were defined. In closed-set verification, the set of trials is limited to segments containing speech in one of the target languages, and scores are computed based on those trials. In open-set verification, scores are computed based on the whole set of trials for a given test, including those corresponding to segments containing speech in an *unknown* language. This way, systems could be designed specifically or optimized for a particular operation mode, and research teams could submit separate results for each operation mode. As we explain in Section 3, whereas the training set did not provide data for *unknown* languages, both the development and evaluation sets included segments in *unknown* languages (with different distributions). The set of *unknown* languages was not disclosed to participants.

With the aim to measure performance as a function of the available amount of speech, the development and evaluation sets were each divided into three subsets, containing segments of three nominal durations: 30, 10 and 3 seconds, respectively. Segment durations were not explicitly disclosed to participants, although they could be guessed from file sizes. Note that each segment is a fragment of an original TV broadcast recording, containing speech in a single language (from one or more speakers) mixed with fragments of silence or background noise, so the actual amount of speech is smaller than the nominal duration.

A test condition was determined by the operation mode (open-set vs. closed-set) and the development condition (free vs. restricted), so four test conditions were defined: (1) open-set / free-development (briefly, OF); (2) open-set / restricted-development (briefly, OR); (3) closed-set / free-development (briefly, CF); and closed-set / restricted-development (briefly, CR). Participants were allowed to submit multiple systems for each test condition, but were required to specify a single system as *primary*, the remaining being *contrastive*. Given a test condition, trials corresponding to each duration were scored separately, resulting in 12 possible scorings (4 test conditions × 3 durations).

3. Data

As noted above, in the free-development condition, participants could use any available materials to train and tune their systems. However, the evaluation focused on systems and results obtained by using just the materials provided by the organization. This was an attempt to compensate for the lack of data that may strongly limit the performance attainable by some participants.

A speech database, named KALAKA [3], was specifically designed, collected and built to support the Albayzin 2008 LRE. KALAKA allows us to build language recognition systems with four target languages: Basque, Catalan, Galician and Spanish. These are all official languages in Spain, though only Spanish is spoken in the whole territory, whereas the other three are spoken (with different usage levels) in specific regions. Due to the interaction between these languages, the task of distinguishing them could be more challenging than expected. In fact, one of the goals of the evaluation was to measure the accuracy that state-of-the-art language recognition systems could attain for this task.

KALAKA was designed by keeping in mind NIST evaluations [7] (in particular, NIST 2007 Language Recognition Evaluation [4]). There is, however, a significant difference: NIST LRE materials consisted of spontaneous conversations recorded at 8 kHz through telephone channels involving two speakers[1], whereas those of KALAKA were extracted from wide-band (44.1 kHz, stereo) TV show recordings, later converted to single-channel 16 kHz audio signals, including both planned and spontaneous speech in diverse environment conditions involving a varying number of speakers. Various types of TV shows were recorded, with prevalence of broadcast news, talk shows and debates.

After recording, fragments containing noisy speech, music, speech overlaps, etc. were discarded. Only speech fragments with a low level of background noise were validated for KALAKA. This task was performed by listening to and looking at audio signals. As a result, speech segments of indefinite length (each spoken in a single language by one or more speakers) were extracted from recorded materials and stored in (single channel, 16 kHz, 16 bits/sample, uncompressed PCM) WAV files. No further processing was applied to speech segments posted to the train dataset (see Table 1).

Table 1: Distribution of training segments per target language: number of segments (# seg), total duration (T) and average segment duration ($\bar{T}_{seg}$).

	Spanish	**Catalan**	**Basque**	**Galician**
# seg	282	278	342	401
T **(min)**	529	538	531	532
$\bar{T}_{seg}$ **(sec)**	112,55	116,12	93,16	79,60

Speech fragments posted to development and evaluation were taken as source to extract segments of fixed duration (30, 10 and 3 seconds), according to the following criteria:

1. Speech segments must be enclosed by a certain amount of silence (i.e. low-energy frames), which is included as part of the segments. This way, it is expected to catch natural segments and to avoid cutting words.

2. A 30-second segment is validated if and only if it contains a valid 10-second segment. Similarly, a 10-second segment is validated if and only if it contains a 3-second segment.

3. Segments can be slightly longer (but not shorter) than their nominal duration: 3-second segments are allowed to last up to 5 seconds; 10-second segments are allowed to last up to 12 seconds; and 30-second segments are allowed to last up to 33 seconds.

Development and evaluation data include utterances in target and *unknown* languages, so that open-set evaluations can be carried out. The development dataset consists of 1800 speech segments, distributed in three subsets, each containing 600 segments with nominal durations of 30, 10 and 3 seconds, respectively. Each subset consists of 120 segments per target language and 120 additional segments from *unknown* languages. The evaluation dataset has the same structure, except for the distribution of non-target languages (see Table 2).

Table 2: Distribution of segments (the same for each duration) for the *unknown* languages in the development and evaluation datasets.

	French	**Portuguese**	**English**	**German**
Devel	70	10	40	0
Eval	10	70	0	40

Summarizing, the training set contains around 9 hours of speech per target language, which amounts to around 36 hours of training data. The development and evaluation sets contains around 7.7 hours of speech each, distributed the same way: more than 90 minutes of speech per target language and more than 90 minutes of speech for *unknown* languages all together. The whole database amounts to around 50 hours of speech and is distributed (after direct request to authors) in three DVD (see [3] for details).

4. Performance measures

The language recognition task defined in this evaluation considers two types of errors: (1) *misses*, those for which the correct answer is *yes* but the system says *no*; and (2) *false alarms*, those for which the correct answer is *no* but the system says *yes*. Therefore, for any test condition the corresponding error rates can be computed as the fraction of target trials that are rejected (*miss rate*, P_{miss}) and the fraction of impostor trials that are accepted (*false alarm rate*, P_{fa}), and suitable cost functions can be defined as combinations of these basic error rates.

4.1. Average cost across target languages

Let assume that there are L target languages. Let $P_{miss}(i)$ be the miss rate computed on trials corresponding to target language i ($i \in [1, L]$), and $P_{fa}(i, j)$ the false alarm rate computed on trials corresponding to other language j (the index 0 representing *unknown* lan-

[1] In the latest NIST LRE (2009), most data were collected from *Voice of America* radio broadcasts, but only speech segments coming from telephone channels were used for testing purposes.

guages), that is, the fraction of trials corresponding to language j that are erroneously accepted as containing language i. The *pairwise cost* $C(i,j)$ is defined as follows:

$$\begin{aligned} C(i,j) &= C_{miss} \cdot P_{target} \cdot P_{miss}(i) + \\ & \quad C_{fa} \cdot (1 - P_{target}) \cdot P_{fa}(i,j) \end{aligned} \quad (1)$$

Note that the pairwise cost model depends on three application parameters: C_{miss}, C_{fa} and P_{target}. For this evaluation, the same values used in NIST 2007 LRE are applied:

$$\begin{aligned} C_{miss} &= C_{fa} = 1 \\ P_{target} &= 0.5 \end{aligned}$$

Pairwise costs are computed separately for each of the four test conditions and for each of the three segment duration categories. Finally, an average cost is defined by adding the contributions for all the combinations of target and non-target languages:

$$\begin{aligned} C_{avg} &= \frac{1}{L} \sum_{i=1}^{L} \{ C_{miss} \cdot P_{target} \cdot P_{miss}(i) \\ &+ \sum_{\substack{j=1 \\ j \neq i}}^{L} C_{fa} \cdot P_{non-target} \cdot P_{fa}(i,j) \\ &+ C_{fa} \cdot P_{OOS} \cdot P_{fa}(i,0) \} \end{aligned} \quad (2)$$

where $P_{non-target}$ is the prior probability of each non-target language (assuming a uniform distribution) and P_{OOS} the prior probability of *unknown* (Out-Of-Set) languages. In this evaluation, the following values are applied:

$$\begin{aligned} P_{OOS} &= \begin{cases} 0.0 & \text{closed-set condition} \\ 0.2 & \text{open-set condition} \end{cases} \\ P_{non-target} &= \frac{1 - P_{target} - P_{OOS}}{L - 1} \end{aligned}$$

The average cost C_{avg} is computed separately for each of the four test conditions and for each of the three segment duration categories, and serves as the main system performance measure in this evaluation. The scoring script of NIST LRE [8] (with some minor changes) was used to compute C_{avg}.

4.2. Log-Likelihood Ratio (LLR) average cost

As noted above, sites may optionally specify that their scores represent (or can be interpreted) as log-likelihood ratios. In such cases, it was planned to evaluate systems also in terms of the so called C_{LLR} [5], which is used as an alternative performance measure in NIST evaluations. C_{LLR} shows two important features: (1) it allows us to evaluate system performance globally by means of a single numerical value, which is somehow related to the area below the DET curve, provided that scores can be interpreted as log-likelihood ratios; and (2) C_{LLR} does not depend on application costs; instead, it depends on the calibration of scores, an important feature of detection systems. To compute C_{LLR}, the *FoCal* toolkit can be used [9].

Let $LR(X,i)$ be the *likelihood ratio* corresponding to segment X and target language i. The likelihood ratio can be expressed in terms of the conditional probabilities of X with regard to the alternative target and non-target hypotheses, as follows:

$$LR(X,i) = \frac{prob(X|i)}{prob(X|\neg i)} \quad (3)$$

Let consider an evaluation set E, consisting of the union of $L+1$ disjoint subsets: E_j ($j \in [1,L]$) containing segments in the target language j, and E_0 containing segments in *unknown* languages. Pairwise costs $C_{LLR}(i,j)$, for $i \in [1,L]$ and $j \in [0,L]$, are defined as follows:

$$C_{LLR}(i,j) = \begin{cases} \frac{1}{|E_i|} \sum\limits_{X \in E_i} \log_2(1 + LR(X,i)^{-1}) & j = i \\ \frac{1}{|E_j|} \sum\limits_{X \in E_j} \log_2(1 + LR(X,i)) & j \neq i \end{cases} \quad (4)$$

Finally, the average cost C_{LLR} is computed by adding the pairwise costs for all the combinations of target and non-target (including Out-Of-Set) languages, as follows:

$$\begin{aligned} C_{LLR} &= \frac{1}{L} \sum_{i=1}^{L} \{ P_{target} \cdot C_{LLR}(i,i) \\ &+ \sum_{\substack{j=1 \\ j \neq i}}^{L} P_{non-target} \cdot C_{LLR}(i,j) \\ &+ P_{OOS} \cdot C_{LLR}(i,0) \} \end{aligned} \quad (5)$$

The cost function C_{LLR} returns an unbounded non-negative value which can be interpreted as information bits, with lower values representing better performance, the value 0 corresponding to a perfect system and the value $\log_2(L)$ corresponding to a system which just relies on (uniform) priors, thus providing no information to decide a trial. Further details about the reasons for using and the interpretation of C_{LLR} can be found in [5, 6].

4.3. Graphical evaluation: DET curves

Detection Error Tradeoff (DET) curves [10] provide a straightforward way of comparing the global average performance of different systems for a given test condition. A DET curve is generated by computing P_{miss} and P_{fa} for a wide range of operation points (thresholds), based on the scores yielded by the analyzed system for a given test set. Besides C_{avg} and C_{LLR}, DET curves are used in NIST evaluations to support system performance comparisons. In this evaluation, DET curves are generated by means of NIST software [11].

5. Organizational issues

After training and development materials were sent to participants, there were 3 months for system development. Then, after evaluation data were sent to participants, there were 3 weeks for processing data and sending results, in a format similar to that used in NIST LRE, that is, a text file with a trial per line, each trial consisting of 6 fields: development condition, target language, operation mode, test file, decision and score. Since multiple systems could be submitted, a naming protocol was established, consisting of a site identifier, a test condition identifier (OF, OR, CF or CR) and a system identifier (primary, contrastive1, contrastive2, etc.). For each file of results, participants had to specify whether or not the scores may be interpreted as log-likelihood ratios. Finally, each participant was committed to send a complete description of their systems, with the aim to give readers a clear sense of what each system was about (methods, references, training data, processing speed, etc.). System ranking in each test condition and each duration subcategory was done taking into account C_{avg} values. DET curves and C_{LLR} values were computed only to allow more detailed analyses. The best system award was given to the best system in the CR condition on the subset of 30-second segments.

Table 3: Performance (C_{avg}) of primary and contrastive systems submitted to Albayzin 2008 LRE in the four test conditions (OF, OR, CF and CR) for the subset of 30-second segments.

	C_{avg}					
	OF-30	**OR-30**		**CF-30**	**CR-30**	
	pri	pri	con	pri	pri	con
T1	0,0946	0,1313	0,1110	0,0552	0,0778	0,0656
T2	0,1204	0,2787		0,0556	0,2420	
T3					0,2597	0,5389
T4					0,5035	

6. Results

There were 4 participants in the evaluation, three of them from Spain and one from Portugal. Since only one of them sent scores that may be interpreted as log-likelihood ratios, the alternative measure C_{LLR} was not computed and will not be considered in the analyses hereafter. As shown in Table 3, 13 systems were submitted, 6 from Team 1, 4 from Team 2, 2 from Team 3 and 1 from Team 4. Performance as a function of the available amount of speech will be presented in Section 6.1. The remaining analyses in this Section will deal with results on the subset of 30-second segments, for which the best figures are obtained. In particular, Table 3 shows the performance of the submitted systems on the subset of 30-second segments. Team 1 presented the best systems in all conditions, featuring state-of-the-art technology. Systems presented by Team 2, also featuring state-of-the-art technology, did only yield competitive performance in some conditions. Finally, systems presented by Teams 3 and 4 were not specifically designed for this task and they probably needed more tuning.

Focusing on results obtained by Teams 1 and 2, note that the best performance ($C_{avg} = 0,0552$, around 5% Equal Error Rate, in the CL condition) is not so good as those obtained in NIST LRE, which deal with much more data and target languages. Obviously, we are not comparing the *same systems* on two different tasks, but different systems on different tasks, and we cannot extract conclusions. Anyway, these results may indicate that the proposed task is, in fact, more difficult than expected, taking into account that we are dealing with 16 kHz (good quality) speech signals and just 4 target languages. This difficulty may be due to the presence of various sources of variability (speakers, environment, channel, etc.) but more probably to the phonetic and lexical similarity among the target languages, which evolved jointly in different regions of the Iberian Peninsula, being Castilian Spanish the shared and most influential language. On the other hand, system performance is remarkably worse (with almost two times the C_{avg}) in open-set than in closed-set verification tests. This reveals that a high number of false alarms are being detected for impostor trials.

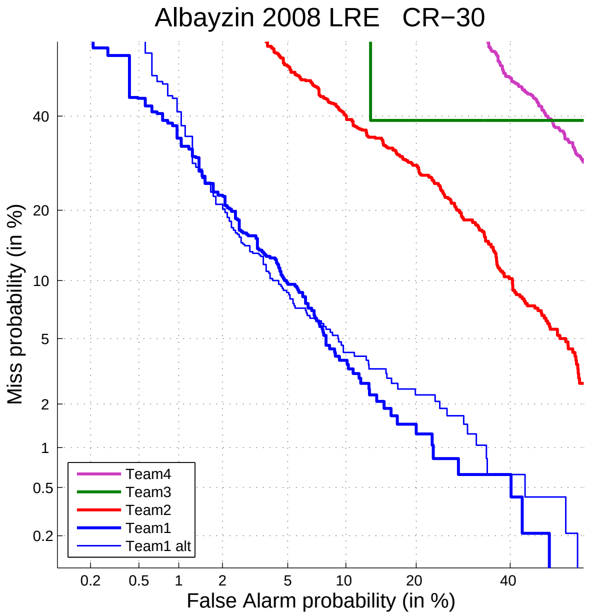

Figure 1: Pooled DET curves of systems submitted to Albayzin 2008 LRE in the CR condition for the subset containing 30-second segments.

Figure 1 shows DET curves of systems submitted to the CR condition for the subset of 30-second segments. The primary system submitted by Team 1 clearly yielded the best performance among all primary systems ($C_{avg} = 0,0778$), so the best system award was given to Team 1. Note, however, that the best performance in this condition was obtained by the contrastive system submitted by Team 1 ($C_{avg} = 0,0656$).

Regarding processing times, only three systems were reported to take more than 1×Real-Time: the primary systems of Team 1 for the OF and CF conditions (1,5×Real-Time), and the primary system of Team 3 for the CR condition (1,4×Real-Time). The other systems were reported to take under 0,4×Real-Time. The contrastive system of Team 3 for the CR condition was the fastest, with 0,004×Real-Time.

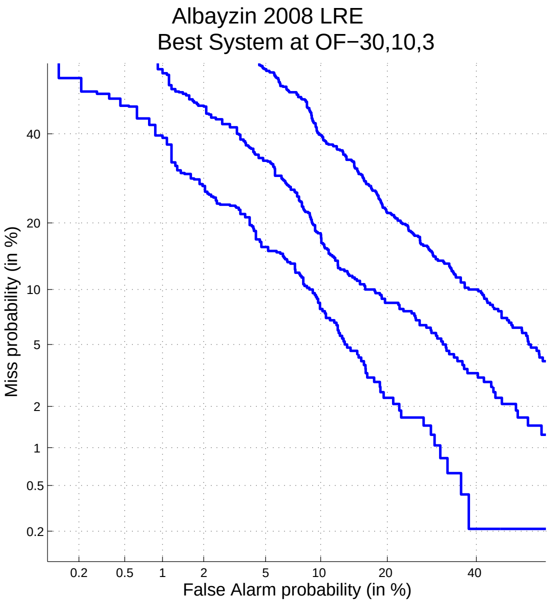

Figure 2: Pooled DET curves of the best system in the OF test condition for the subsets containing 30, 10 and 3-second segments.

6.1. Results per duration of test segments

Results in all conditions showed that, as expected, performance was consistently worse for 10-second than for 30-second test segments, and even worse for 3-second test segments. The DET curves for the best system in the OF condition are shown in Figure 2. Note that the EER for 3-second segments (around 20%) is two times the EER for 30-second segments (around 10%). Similar results were obtained for other systems and conditions.

6.2. Results per development condition

Systems applying development restrictions (i.e. using only those materials provided for Albayzin 2008 LRE, allowing neither external data nor subsystems trained on external data) performed consistently worse than systems freely developed on any available data. In the case of Team 1 (blue curves in Figure 3) the restricted system yielded quite remarkable results: $C_{avg} = 0,0778$ in closed-set verification tests on the set of 30-second segments, which means only 40% increased cost with regard to the free system. In the case of Team 2 (red curves in Figure 3), differences were much higher: the restricted system yielded more than 4 times the cost of the free system. Finally, note that the free systems of Teams 1 and 2 yielded almost the same C_{avg} performance, their DET curves being very close to each other.

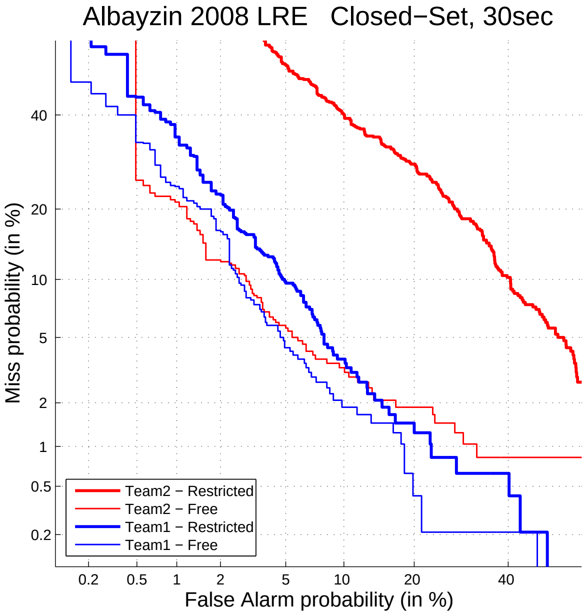

Figure 3: Pooled DET curves of systems corresponding to Team 1 and Team 2, in the CF and CR test conditions for the subset of 30-second segments.

6.3. Results per target language

Analyzing in detail the behaviour of language recognition systems for each target language would take a lot of time and effort. Here we simply inspect disaggregated results (per target language) for the best system in the CF and OF conditions on the set of 30-second segments (see DET curves in Figures 4 and 5). Clearly, system performance was not homogeneous when disaggregated for target languages. In the CF condition, the best recognition performance was obtained for Basque, whereas performance was quite similar for the three other target languages. In the OF condition, the presence of impostor trials with *unknown* languages had almost no effect in the performance for Basque (which yielded again the best performance) and Spanish, whereas the effect was quite remarkable for Catalan. The confusion among languages in both conditions has been depicted in Tables 4 and 5, where the error rates ($P_{miss}(i)$ in the diagonal, $P_{fa}(i, j)$

outside the diagonal) are expressed as grey levels (white for 0 and black for 1). Note the high P_{fa} associated to Catalan for impostor trials with *unknown* languages in the OF condition.

Table 4: Error rates ($P_{miss}(i)$ in the diagonal, $P_{fa}(i,j)$ outside the diagonal) for the best system in the CF condition on the subset of 30-second segments. The darker the cell means the higher the error rate.

		Target			
		Spanish	Catalan	Basque	Galician
Segment	Spanish	0.0750	0.0167	0.1250	0.0833
	Catalan	0.0083	0.1167	0.0083	0.0000
	Basque	0.0083	0.0000	0.0083	0.0000
	Galician	0.1167	0.0500	0.0083	0.1000

Table 5: Error rates ($P_{miss}(i)$ in the diagonal, $P_{fa}(i,j)$ outside the diagonal) for the best system in the OF condition on the subset of 30-second segments. The darker the cell means the higher the error rate.

		Target			
		Spanish	Catalan	Basque	Galician
Segment	Spanish	0.0833	0.0083	0.0667	0.0083
	Catalan	0.0083	0.1750	0.0000	0.0000
	Basque	0.0083	0.0000	0.0250	0.0000
	Galician	0.1083	0.0417	0.0000	0.1083
	Unknown	0.0667	0.4333	0.1083	0.1417

The high performance for Basque may be due to the different origins of Basque with regard to the other target languages (which are Romance languages). Basque has been influenced by Romance languages (specially by Spanish and French), but has completely different roots, and its lexicon is quite different from those of the other languages appearing in KALAKA. On the other hand, the high confusion of Catalan (and at a lower degree, also of Galician) with the *unknown* languages may be due to its similarity to French or Portuguese (note that all of them are Romance languages).

7. Conclusions and Future Work

In this work, the main features of the Albayzin 2008 Language Recognition Evaluation have been described, and results obtained by the submitted systems have been presented and discussed. The evaluation involved the four official languages spoken in Spain (Basque, Catalan, Galician and Spanish) as target languages. A database, consisting of 16 kHz audio signals taken from TV broadcasts, was created and used specifically for this evaluation. The best system, employing state-of-the-art technology and without development restrictions, yielded $C_{avg} = 0,0552$ in closed-set verification tests on 30-second speech segments. This reveals the difficulty of the task, maybe due to the presence of various sources of variability (speakers, environment, channel, etc.) but more probably to the phonetic and lexical similarity among the target languages, which evolved jointly in different regions of the Iberian Peninsula, being Castilian Spanish the shared and most influential language.

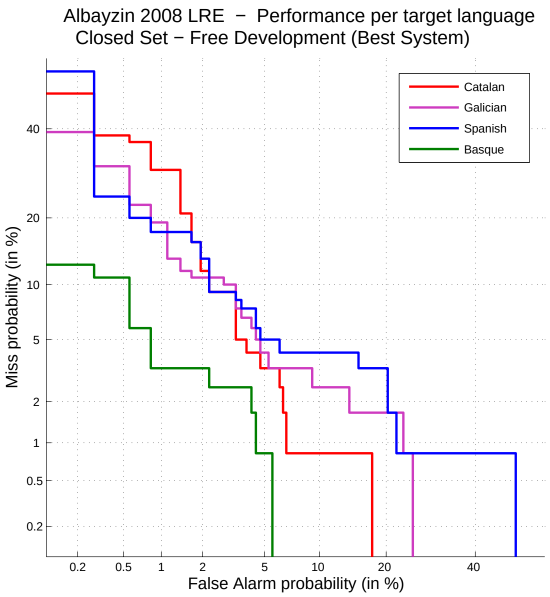

Figure 4: DET curves for target languages, using the best system in the CF test condition.

Results were analyzed in detail from different angles, taking the performance of the best system as reference. Performance was remarkably worse (with almost two times the C_{avg}) in open-set than in closed-set verification tests. Just by inspecting DET curves, we realized that performance consistently worsened as the available amount of speech reduced from 30 to 10, and from 10 to 3 seconds. As may be expected, restricting the development conditions to using only those materials provided for Albayzin 2008 LRE (instead of any available materials) led to higher costs. The best system, with development restrictions, yielded $C_{avg} = 0,0778$ in closed-set verification tests on the set of 30-second segments (a 40% relative increase in cost with regard to the system without development restrictions). Finally, system performance was not homogeneous when disaggregated for target languages. In particular, the best recognition performance was obtained for Basque, which may be due to the different origins of Basque with regard to the other target languages (which are Romance languages).

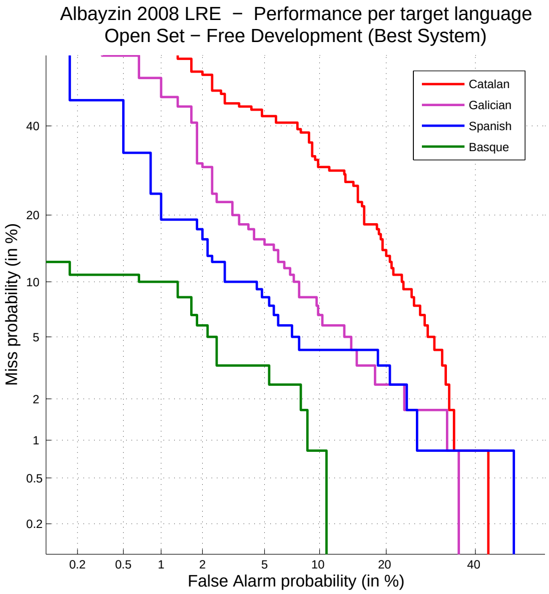

Figure 5: DET curves for target languages, using the best system in the OF test condition.

We plan to carry out a second evaluation this year, the Albayzin 2010 LRE, using again 16 kHz TV broadcast speech signals, but including also Portuguese and English as target languages and renewing the set of *unknown* languages. This new feature might make the evaluation more appealing for research teams from outside Spain. The evaluation would be held from June to October 2010 and results would be presented at the 6th Biennial Workshop on Speech Technology, to be held in Vigo (Spain) in November 2010. The evaluation plan will be posted through ISCA.

8. Acknowledgements

We thank specially Rubén San Segundo, president of the Spanish Network on Speech Technologies, and Inma Hernáez, president of the Organizing Committee of the 5th Biennial Workshop on Speech Technology, held in Bilbao in November 2008, for their help and support. We also thank all the participants, for their work, their feedback and kindness, which were essential for the success of the evaluation.

9. References

[1] *5th Biennial Workshop on Speech Technology*, Bilbao, Spain, 12-14 November 2008, http://jth2008.ehu.es/en/index.html.

[2] *Spanish Network on Speech Technology*, Web (in Spanish): http://lorien.die.upm.es/∼lapiz/rtth/.

[3] Luis Javier Rodriguez-Fuentes, Mikel Penagarikano, German Bordel, Amparo Varona, and Mireia Diez, "KALAKA: A TV broadcast speech database for the evaluation of language recognition systems," in *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010)*, Valleta, Malta, 17-23 May 2010.

[4] Alvin F. Martin and Audrey N. Le, "NIST 2007 Language Recognition Evaluation," in *Odyssey 2008 - The Speaker and Language Recognition Workshop, paper 016*, 2008.

[5] Niko Brümmer and Johan A. du Preez, "Application-independent evaluation of speaker detection," *Computer Speech & Language*, vol. 20, no. 2-3, pp. 230–275, 2006.

[6] N. Brümmer and D.A. van Leeuwen, "On calibration of language recognition scores," in *Proceedings of Odyssey 2006 - The Speaker and Language Recognition Workshop*, 2006, pp. 1–8.

[7] *NIST Language Recognition Evaluations*, http://www.itl.nist.gov/iad/mig/tests/lre.

[8] *NIST 2007 LRE scoring software*, http://www.itl.nist.gov/iad/mig/tests/lre/2007/score_lre07.v01d.tgz.

[9] *Toolkit for Evaluation, Fusion and Calibration of statistical pattern recognizers*, http://sites.google.com/site/nikobrummer/focal.

[10] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The DET curve in assessment of detection task performance," in *Proceedings of the 5th European Conference on Speech Communication and Technology (Eurospeech'97)*, Rhodes, Greece, 22-25 September 1997, pp. 1895–1898.

[11] *NIST DET-Curve Plotting software for use with MATLAB*, http://www.itl.nist.gov/iad/mig/tools/DETware_v2.1.targz.htm.