Evaluation of Spoken Language Recognition Technology Using Broadcast Speech: Performance and Challenges

Luis J. Rodríguez-Fuentes, Amparo Varona, Mireia Diez, Mikel Penagarikano, Germán Bordel

Software Technologies Working Group (http://gtts.ehu.es)
Department of Electricity and Electronics, University of the Basque Country UPV/EHU
Barrio Sarriena s/n, 48940 Leioa, Spain
luisjavier.rodriguez@ehu.es

Abstract

Spoken Language Recognition (SLR) technology has remarkably improved in the last years, partly thanks to NIST Language Recognition Evaluations (LRE), which have become standard benchmarks for testing new approaches. NIST evaluations focus on narrow-band conversational telephone speech and deal with some specific target languages. Recent efforts to expand the scope of SLR technology assessment include the Albayzin 2008 and 2010 LRE, which deal with wide-band TV broadcast speech. In this work, a SLR system based on state-of-the-art approaches is developed and evaluated on the Albayzin 2008 and 2010 LRE datasets, looking to identify those conditions that make the task challenging and eventually to guide the design of future evaluations using the same kind of data. We present and analyse system performance under different conditions, regarding: (1) the set of target languages (including details about the confusion of languages with each other) and the amount of data available to estimate models; and (3) the presence of background noise.

1. Introduction

The development of Spoken Language Recognition (SLR) technology has been largely boosted by NIST Language Recognition Evaluations (LRE) [1], held in 1996 and every two years since 2003. As a result, the datasets produced and distributed for such evaluations have become standard benchmarks to prove the usefulness of new approaches. NIST LRE datasets consist of narrow-band (8 kHz) conversational telephone speech (in all LRE) and narrow-band (mostly telephone speech) segments from worldwide Voice of America broadcasts (only in 2009 and 2011 LRE). The number of target languages ranges from 7 in NIST 2005 LRE to 24 in NIST 2011 LRE. There seems to be a symbiotic relationship between the research community that provides the algorithms and the government agencies that support the production of data. This fruitful collaboration also features some issues: (1) NIST LRE focus on telephone speech for a specific type of applications (large-scale verification of telephone conversations in some interesting target languages); (2) NIST LREs have undoubtly helped improve SLR technology also for wide-band speech (16 kHz and above), but there is a lack of resources to objectively assess such improvement; and (3) by developing technology based on NIST LRE datasets, we may be addressing challenges specific to that kind of signals (narrow-band, single-speaker, etc.) and limiting potential improvements that may be accomplished by using other datasets.

Recently, aiming to expand the scope of SLR technology assessment, we organized the Albayzin 2008 and 2010 Language Recognition Evaluations [2, 3], both supported by the Spanish Thematic Network on Speech Technologies [4] and the ISCA Special Interest Group on Iberian Languages (SIG-IL). These evaluations were inspired by the NIST 2007 LRE [5] (same task definition, test procedures, performance measures, file formats, etc.), but featured a number of differences: (1) speech signals were extracted from wide-band (16 kHz) TV broadcasts involving multiple speakers; (2) the set of target languages was relatively small (compared to the sets used in the last NIST LREs) though quite challenging due to acoustic, phonetic and lexical similarities; and (3) the target application was Spoken Document Retrieval (SDR). It is worth noting that the Albayzin 2010 LRE dataset has been recently used as benchmark [6][7].

In this work, a spoken language recognition system based on state-of-the-art approaches is developed and evaluated on the Albayzin 2008 and 2010 LRE datasets, looking to identify those conditions that make the task challenging and eventually to guide the design of future evaluations using the same kind of data. System performance is compared with regard to the set of target languages and the amount of training data available, including details about the confusion of languages with each other. Performance degradation as a consequence of the presence of background noise is also evaluated. By the way, since the same system was previously applied to the NIST 2011 LRE [8], yielding high performance, results reported in this paper support the use of Albayzin LRE datasets as alternative or complementary benchmarks for the assessment of SLR technology, specially when dealing with wide-band speech for SDR applications.

The rest of the paper is organized as follows. The main features of the Albayzin LREs are briefly outlined in Section 2. Section 3 describes the Albayzin LRE datasets used for the experiments reported in this paper. Section 4 describes the acoustic and phonotactic subsystems and the backend and fusion strategy applied to get the final (fused) scores. System performance on different tracks, either within a single evaluation or across the two evaluations, is presented and discussed in Section 5, including a detailed analysis on the confusion of target languages with each other. Finally, the most challenging conditions according to the obtained results and different setups for future evaluations are discussed in Section 6.

This work has been partially supported by the University of the Basque Country, under grant GIU10/18 and project US11/06, by the Government of the Basque Country, under program SAIOTEK (project S-PE11UN065), and the Spanish MICINN, under *Plan Nacional de I+D+i* (project TIN2009-07446, partially financed by FEDER funds). Mireia Diez is supported by a research fellowship from the Department of Education, Universities and Research of the Government of the Basque Country.

2. The Albayzin Language Recognition Evaluations: An Overview

As for NIST evaluations [5, 9], the task defined for the Albayzin LRE consisted on deciding (by computational means) whether or not a target language was spoken in a test utterance, which is usually known as *language detection* or *language verification*. Performance was computed by presenting the system a set of trials and comparing system decisions with the right ones (stored in a keyfile). Each trial comprised a segment of audio containing speech in a single language and the identity of the target language. For each trial, the system was required to output a hard decision about whether the target language was spoken in the segment, and a score such that the highest the score the most likely the target language was spoken in the segment.

The four official languages spoken in Spain: Basque, Catalan, Galician and Spanish, were used as target languages in the Albayzin 2008 LRE. Though small, this set was expected to be challenging, due to the potential confusability of languages with each other. The set of Iberian languages was completed in the Albayzin 2010 LRE by adding Portuguese as target language. Due to its international relevance and its pervasiveness in broadcast news, English was also added as target language in the Albayzin 2010 LRE (it was long after the evaluation that we realized that English can be also regarded as Iberian, since it is the official language in Gibraltar). In both evaluations, speech segments in other (Out-Of-Set, OOS) languages were also included to allow open-set verification trials. In fact, systems could be specifically tuned for either closed-set or open-set verification, since separate tracks were defined to evaluate performance under both conditions. Following the NIST LRE protocol, separate tracks were also defined to evaluate system performance on speech segments of three nominal durations (30, 10 and 3 seconds).

In the Albayzin 2008 LRE, two separate tracks were defined depending on the materials used for training: (1) *restricted development*, for which only the data provided for the evaluation could be used to train models; and *free development*, for which any available data could be used to train models. By restricting system development to the training materials provided for the evaluation, we wanted to remove the relative advantage of some groups having many available data for training, to focus the challenge on the modeling and classification approaches and also to measure the dependence of system performance on the availability of training data.

In the Albayzin 2010 LRE there was no limitation regarding the training materials, but two additional tracks were also defined, depending on the presence of background noise. The first one, defined for reference (though somewhat unrealistic), considered only clean-speech segments, whereas the second one considered all (clean-speech and noisy-speech) segments, aiming to reflect more closely the kind of resources that SDR applications must commonly deal with. Besides clean-speech data, speech segments featuring background noise, music and/or conversations (overlapped speech) were separately provided in the Albayzin 2010 LRE for training, development and evaluation. Each segment contained a single language, which also applied to segments with background conversations, except for the case of segments in OOS languages, which might contain speech in two or more languages, provided that none of them were target languages. By providing noisy speech data, systems could be trained, tuned and evaluated also for the noisy-speech condition. This was a relevant move with regard to the Albayzin 2008 LRE, where only clean-speech segments were processed.

In both evaluations, system performance was primarily measured by means of the well-known average cost C_{avg} (pooled across target languages), which is a combination of two basic error rates: 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}):

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

where L is the number of target languages and C_{miss}, C_{fa}, P_{target}, $P_{non-target}$ and P_{OOS} are cost model (application dependent) parameters. For these evaluations, the same values used in NIST 2007 and 2009 LRE were applied:

$$\begin{aligned} C_{miss} &= C_{fa} = 1 \\ P_{target} &= 0.5 \\ P_{OOS} &= \begin{cases} 0.0 & \text{closed-set condition} \\ 0.2 & \text{open-set condition} \end{cases} \\ P_{non-target} &= \frac{1 - P_{target} - P_{OOS}}{L-1} \end{aligned}$$

Detection Error Tradeoff (DET) curves [10] were also used to compare the global performance of different systems for a given test condition. NIST software [11] was used to generate DET curves, including marks for the operation point given by system decisions (actual C_{avg}) and the operation point corresponding to the optimal threshold (minimum C_{avg}).

3. The Albayzin LRE Datasets

Two databases, KALAKA and KALAKA-2, were created to support the Albayzin 2008 and 2010 LRE, respectively. Both included separate subsets of speech segments for training, development and evaluation. In this section we just provide their most relevant features (for further details, see [12][13]).

3.1. Shared Features

All the speech signals of KALAKA and KALAKA-2 were digitally recorded from TV broadcast shows (news, debates, interviews, talk shows, etc.) using a Roland Edirol R-09 recorder. Most recordings involved several speakers, sometimes featuring various dialects, linguistic competence levels and/or speech modalities (planned speech, spontaneous speech, etc.) and diverse environment conditions. Audio signals were stored in WAV files (uncompressed PCM, 16 kHz, single channel, 16 bits/sample).

In both cases, the sets of TV shows posted to training, development and evaluation were forced to be disjoint, meaning that any show appearing in one set did not appear in the other two. This restriction was imposed as an attempt to achieve speaker independence.

3.2. The Albayzin 2008 LRE Datasets

To build KALAKA, which featured 4 target languages (Basque, Catalan, Galician and Spanish), TV broadcast recordings were audited in order to filter out segments containing background noise, music, speech overlaps, etc. Clean-speech segments of a wide range of lengths (each spoken in a single language by one or more speakers) were collected this way. No further processing was applied to speech segments posted to the training set (see Table 1). Segments posted to the development and evaluation sets (featuring both target and OOS languages) were taken as source to automatically extract segments of fixed durations (30, 10 and 3 seconds).

Table 1: Distribution of training segments per target language in the Albayzin 2008 LRE: 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

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 in OOS languages. The evaluation dataset has the same structure, except for the distribution of OOS languages (see Table 2).

Table 2: Distribution of segments (the same for each duration) for OOS languages in the development and evaluation datasets of the Albayzin 2008 LRE.

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

KALAKA amounts to around 50 hours of speech, 36 for training (around 9 hours per target language), 7.7 hours for development and 7.7 hours for evaluation (both distributed the same way: more than 90 minutes of speech per target language and more than 90 minutes of speech for OOS languages all together).

3.3. The Albayzin 2010 LRE Datasets

KALAKA-2 was designed as an extension of KALAKA, including two additional target languages (Portuguese and English) and extended datasets. To reduce development costs, all the materials of KALAKA were recycled for KALAKA-2. New TV broadcasts were also recorded, selected and classified, specially for Portuguese and English and for OOS languages. In particular, the evaluation dataset of KALAKA-2 was completely new and independent of KALAKA.

The train dataset of KALAKA-2 contains at least 10 hours (in most cases, around 11 hours) of clean speech and at least 2 hours (in some cases, more than 3 hours) of noisy speech per target language, amounting to around 82 hours of speech. The distribution of training data is shown in Table 3.

The development and evaluation datasets are identical in size and characteristics, except for the distribution of OOS languages and the proportion of clean and noisy speech. Both datasets contain segments with nominal durations of 30, 10 and 3 seconds, with at least 150 speech segments per target language and nominal duration. Clean-speech segments were extracted by completely automatic means, as for KALAKA. In the case of noisy speech, segments lasting from 30 to 35 seconds were manually selected by experts. Then, 10- and 3-second noisy-speech segments were automatically extracted from them, the same way as for clean speech.

Table 3: Distribution of training segments per target language for clean and noisy speech in the Albayzin 2010 LRE: number of segments (#) and total duration (T, in minutes).

	Clean speech		**Noisy speech**	
	#	T **(minutes)**	**#**	T **(minutes)**
Basque	406	644	112	135
Catalan	341	687	107	131
English	249	731	136	152
Galician	464	644	125	134
Portuguese	387	665	160	197
Spanish	342	625	133	222

The development set consists of 4950 speech segments, 3492 containing clean speech and 1458 containing noisy speech, their total duration being 21.24 hours (70% of the time corresponding to clean speech and 30% to noisy speech). The evaluation set consists of 4992 speech segments, 3345 containing clean speech and 1647 containing noisy speech, their total duration being 21.43 hours (67% of the time corresponding to clean speech and 33% to noisy speech). The distribution of segments per language is shown in Table 4.

Table 4: Distribution of segments per language (the same for each duration) in the development and evaluation datasets of the Albayzin 2010 LRE.

		Devel		**Eval**	
		clean	**noisy**	**clean**	**noisy**
Target languages	**Basque**	146	29	130	74
	Catalan	120	47	149	55
	English	133	60	135	69
	Galician	137	60	121	83
	Portuguese	164	77	146	58
	Spanish	136	83	125	79
OOS languages	**Arabic**	100	25	115	22
	French	120	32	70	34
	German	108	73	13	32
	Romanian	0	0	111	43

4. The SLR System

The spoken language recognition system developed for this work resembles almost exactly that presented by our research group to the NIST 2011 LRE (under different backend/fusion configurations) [8], yielding very competitive performance (fourth best primary system): $C_{avg} = 0.0892$ for the 24 worst performing language pairs and $C_{avg} = 0.0169$ when the average was computed over all the pairs. In the following paragraphs, we provide a brief description of the component subsystems and the backend and fusion configuration.

4.1. Acoustic Subsystems

For the acoustic subsystems, the concatenation of 7 Mel-Frequency Cepstral Coefficients (MFCC) and the Shifted Delta Cepstrum (SDC) coefficients under a 7-2-3-7 configuration, were used as acoustic features. A gender independent 1024-mixture GMM (Universal Background Model, UBM) was estimated by Maximum Likelihood on the training dataset, using binary mixture splitting, orphan mixture discarding and variance flooring. Finally, zero-order and centered and normalized first-order Baum-Welch statistics were computed for each input utterance.

4.1.1. Dot-Scoring Subsystem

The Linearized Eigenchannel GMM (LE-GMM) subsystem, that we briefly call *Dot-Scoring* subsystem, makes use of a linearized procedure to score test segments against target models [14]. The log-likelihood ratio between the target model and the UBM used for scoring can be approximated as follows:

$$score\left(f,l\right) = \log \frac{P\left(f|\lambda_l\right)}{P\left(f|\lambda_{ubm}\right)} \approx \hat{m}_l^t \cdot \hat{x}_f \qquad (2)$$

where $\hat{m}_l$ denotes the centered and normalized channel-compensated MAP-means corresponding to language l, computed as follows:

$$\hat{m}_l = \left(\tau I + diag(n_l)\right)^{-1} \hat{x}_l \qquad (3)$$

where $\tau = 16$ is the relevance factor, n_l are the zero-order statistics for language l and $\hat{x}_l$ and $\hat{x}_f$ are the channel-compensated first-order statistics corresponding to language l and target signal f, respectively. Channel compensation was performed by using Niko Brümmer's recipe [15]. The channel matrix was estimated using only data from target languages.

4.1.2. iVector Generative Subsystem

The estimation of the total variability matrix T, the computation of iVectors and the estimation of the generative Gaussian models were performed as in [16]. The total variability matrix was estimated using only data from target languages. The iVector scores were computed as follows:

$$score\left(f,l\right) = \mathcal{N}\left(w_f;\mu_l,\Sigma\right) \qquad (4)$$

where w_f is the iVector for target signal f, μ_l is the mean iVector for language l and Σ is a common (shared by all languages) within-class covariance matrix.

4.2. Phonotactic Subsystems

Three phonotactic subsystems were developed under a phone-lattice-SVM approach. Given an input signal, an energy-based voice activity detector was applied in first place, which split and removed long-duration non-speech segments. Then, the Temporal Patterns Neural Network (TRAPs/NN) phone decoders developed by the Brno University of Technology (BUT) for Czech (CZ), Hungarian (HU) and Russian (RU) [17], were applied to perform phone tokenization. Regarding channel compensation, noise reduction, etc. the three subsystems relied on the acoustic front-end provided by BUT decoders.

BUT decoders were configured to produce phone posteriors that were converted to phone lattices by means of HTK [18] along with the BUT recipe, on which expected counts of phone n-grams were computed using the *lattice-tool* of SRILM [19]. Finally, a Support Vector Machine (SVM) classifier was applied, SVM vectors consisting of counts of features representing the phonotactics of an input utterance. In this work, phone n-grams up to $n = 3$ were used, weighted as in [20]. L2-regularized L1-loss support vector classification was applied, by means of LIBLINEAR [21], whose source code was slightly modified to get regression values.

4.3. Backend and Fusion

Backend and fusion parameters were optimized in preliminary experiments on the development set of the Albayzin 2010 LRE, and then applied for the experiments in both the Albayzin 2008 and 2010 LRE evaluation datasets. In particular, it was found that applying a backend to subsystem scores improved performance only in the open-set verification condition. So, for the closed-set verification experiments, the raw subsystem scores were used. Regarding the backend approach, best results were found when using a (generative) Gaussian backend.

Therefore, in the open-set condition a Gaussian backend was applied to the L scores provided by each subsystem, and $L + 1$ log-likelihoods were output (one per target language plus an additional log-likelihood for OOS languages, estimated based on the scores for the L target languages). The resulting $5 \times (L+1)$ log-likelihood values were fused by applying linear logistic regression under a multiclass paradigm, obtaining $L+1$ calibrated scores. Finally, a minimum expected cost Bayes decision was made based on these scores, according to application-dependent language priors and costs. The same procedure was applied in the closed-set condition, but using $5 \times L$ raw scores. The *FoCal* toolkit was used to estimate and apply the backend and calibration/fusion models [22, 23].

5. Performance: Analysis and Discussion

In this section, we present system performance under two different setups. In the first one, the Albayzin LRE tasks are compared on the free-development clean-speech test condition, which is common to both evaluations. Results are compared with regard to the set of target languages and the available amount of training and development data on the closed-set test condition, with a detailed analysis of the confusion of target languages with each other. Results are then presented for the open-set test condition, comparing performance degradation on the 2008 and 2010 evaluation datasets and analysing the confusion of OOS languages with target languages and of target languages with each other on the open-set condition of the Albayzin 2010 LRE. The second set of experiments aims to evaluate performance degradation when dealing with noisy speech, based on the Albayzin 2010 LRE datasets.

5.1. Comparing Albayzin LRE Tasks on Clean Speech

Table 5 presents system performance for the free-development clean-speech closed-set (CC) test condition on the evaluation datasets of Albayzin 2008 and 2010 LRE: *eval2008* and *eval2010*. DET curves for the CC-30s condition are shown in Figure 1. In both cases, to allow a more detailed study of the factors that may explain performance results, a subset of the Albayzin 2010 evaluation dataset has been defined (*eval2010 (4L)*), which includes only the trials corresponding to the 4 target languages of the Albayzin 2008 LRE. This way, we can compare the suitability of the training sets defined in 2008 and 2010 for those languages, and on the other hand, get an estimate of the difficulty of the 2008 and 2010 tasks, by using the same models (estimated on the 2008 training set) to process eval2008 and eval2010 (4L).

Results suggest that the 2008 task is more difficult than the 2010 task. As shown in the first and fourth lines of Table 5, C_{avg} in the 2008 task is around 8, 3 and 2 times higher than that found in the 2010 task, for the subsets of 30-, 10- and 3-second segments, respectively. Note also how the red (2008) and blue (2010) DET curves in Figure 1 are remarkably far apart from

Table 5: Performance (C_{avg}) for the Albayzin 2008 and 2010 LRE in the clean-speech closed-set (CC) test condition. To allow significant comparisons, the performance on the evaluation set of Albayzin 2010 LRE using only the trials corresponding to the 4 target languages of Albayzin 2008 LRE (eval2010 (4L)) is also shown.

	CC-30s	**CC-10s**	**CC-3s**
train2008 + eval2008	0.0514	0.0761	0.1722
train2008 + eval2010 (4L)	0.0275	0.0552	0.1535
train2010 + eval2010 (4L)	0.0133	0.0506	0.1466
train2010 + eval2010	0.0063	0.0263	0.0888

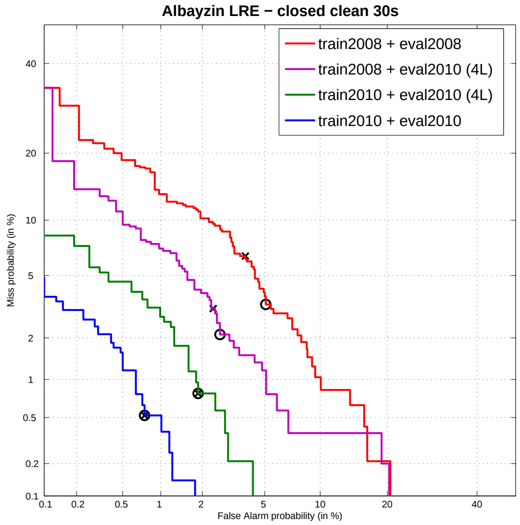

Figure 1: Pooled DET curves for the Albayzin 2008 and 2010 LRE in the clean-speech closed-set 30-second test condition (CC-30s). To allow significant comparisons, DET curves on the evaluation set of Albayzin 2010 LRE using only the trials corresponding to the 4 target languages of Albayzin 2008 LRE (eval2010 (4L)) are also shown.

each other. This may be due in part to the different amount of training data used to estimate models: in KALAKA there are less than 9 hours per target language, whereas in KALAKA-2 all the target languages have at least 10 hours, and some of them more than 12 hours of training data. It could be also explained by the two additional target languages introduced in 2010: Portuguese and English may be less confused with other languages than the average and make the pooled C_{avg} fall down. Finally, the 2008 task (i.e. the set of evaluation segments supplied for the Albayzin 2008 LRE) could be intrinsically more difficult than the 2010 task, for different reasons: use of regional dialects, high acoustic variability, lack of coverage for some speakers, etc. In particular, the criteria applied to filter out noisy speech segments were not so strict in 2008 as in 2010, so there could be significant differences in this regard.

Probably, as performance results in Table 5 and DET curves in Figure 1 suggest, a mix of the above arguments may explain the differences. The different amount of training data is the argument behind the difference between the purple and green DET curves, both computed on the eval2010 (4L) dataset, but using models estimated on the 2008 and 2010 training datasets, respectively. The corresponding costs are shown in the second and third lines of Table 5. Note that, though there is a large difference in the CC-30s condition (around 50% cost reduction, from 0.0275 when using train2008 to 0.0133 when using train2010), differences are much smaller for the CC-10s and CC-3s conditions (8% and 4.5% cost reductions, respectively).

The new target languages introduced in 2010 (Portuguese and English) explain the difference between the green (4 target languages) and blue (6 target languages) DET curves (see the corresponding costs in the third and fourth lines of Table 5), since the same system (built on the Albayzin 2010 LRE training and development datasets) is being applied to two sets of segments: eval2010 and eval2010 (4L), that are identical except for the fact that eval2010 (4L) excludes segments corresponding to Portuguese and English.

Finally, the instrinsic difficulty of the 2008 task comes to explain the difference between the red (eval2008) and purple (eval2010 (4L)) DET curves, since the same system (built on the Albayzin 2008 LRE training and development sets) is applied to process both datasets. Note that on the CC-30s condition, the C_{avg} for eval2008 is almost twice the C_{avg} for eval2010 (4L). Again, differences are smaller on the CC-10s and CC-3s conditions.

The confusion of target languages with each other (miss and false alarm probabilities) on the CC-3s condition of Albayzin 2010 LRE is shown in Table 6. Clearly, system performance was not homogeneous when disaggregated for all the target languages. The lowest error rates were obtained for Portuguese, English and Basque. On the other hand, Spanish and Galician were highly confused between each other, also showing significant miss rates, which was also observed for the Albayzin 2008 LRE [2]. These results may be partly explained by the fact that most of the target languages (Catalan, Galician, Portuguese and Spanish) are Romance languages, whereas English is a Germanic language and Basque, though influenced by Romance languages (specially by Spanish and French), has completely different roots and its lexicon is quite different from those of the other languages.

Table 6: Error probabilities per target language: Basque (eu), Catalan (ca), English (en), Galician (gl), Portuguese (pt) and Spanish (es), for the closed-set clean-speech 3-second test condition (CC-3s) of the Albayzin 2010 LRE. *Miss probability* is shown in the diagonal and *false alarm probability* out of the diagonal.

		Target					
		eu	**ca**	**en**	**gl**	**pt**	**es**
Segment	**eu**	0.054	0.046	0.015	0.139	0.000	0.162
	ca	0.107	0.060	0.013	0.181	0.107	0.195
	en	0.015	0.037	0.015	0.000	0.052	0.022
	gl	0.099	0.198	0.033	0.207	0.083	0.397
	pt	0.027	0.075	0.034	0.055	0.027	0.055
	es	0.112	0.152	0.024	0.336	0.016	0.144

The low confusion rates for Portuguese, compared to the high confusion of the other Romance languages with each other, may probably come from the coexistence of speakers of Catalan, Galician and Spanish, specially in the last century (nowadays, most Catalan and Galician speakers also speak Spanish in their normal life), whereas Portuguese speakers have historically had little contact with speakers of the other languages. In any case, it is surprising the low confusion between Portuguese and Galician, despite being quite close in many of their features. A reasonable explanation for this and, by the way, for the high confusion between Galician and Spanish, could be that many of the Galician speakers were in fact Spanish speakers having Galician as their second language.

For the open-set clean speech (OC) condition, the system developed and evaluated using the Albayzin 2010 LRE datasets achieved better performance than that developed and evaluated using the Albayzin 2008 LRE datasets. Relative cost reductions of 77%, 64% and 45% were obtained on the OC-30s, OC-10s and OC-3s conditions, respectively (see Table 7). Again, these results may be explained in various ways, including a larger training dataset, less confusable target languages (on average), etc.

Table 7: Performance (C_{avg}) for the Albayzin 2008 and 2010 LRE in the open-set clean-speech (OC) test condition.

	OC-30s	**OC-10s**	**OC-3s**
Albayzin 2008 LRE	0.0759	0.1211	0.2004
Albayzin 2010 LRE	0.0171	0.0437	0.1094

The confusion of languages with each other (miss and false alarm probabilities) on the OC-3s condition of the Albayzin 2010 LRE is shown in Table 8. The presence of impostor trials with OOS languages (see the last line of Table 8) had a strong impact on the false alarm rates for all the target languages. The origin of OOS languages is diverse: Romanian and French are Romance languages whereas German is a Germanic language and Arabic is Semitic. This diversity may be behind the high confusion rates of OOS languages with all the target languages. In relative terms, the impact was more noticeable for Portuguese and English, compared to the closed-set condition shown in Table 6, where they accumulated low false alarm probabilities. In absolute terms, the highest confusion with OOS trials was found for Catalan (0.304) and Spanish (0.210). Overall, best performance (the lowest error probabilities) was found for English, Portuguese and Basque. As for the CC-3s condition, the highest confusion was found between Spanish and Galician, with false alarm probabilities of 0.616 and 0.587.

Table 8: Error probabilities per target language: Basque (eu), Catalan (ca), English (en), Galician (gl), Portuguese (pt) and Spanish (es), for the closed-set clean-speech 3-second test condition (OC-3s) of the Albayzin 2010 LRE. *Miss probability* is shown in the diagonal and *false alarm probability* out of the diagonal. Error probabilities for OOS segments are shown too.

		Target					
		eu	**ca**	**en**	**gl**	**pt**	**es**
Segment	**eu**	0.062	0.062	0.000	0.146	0.000	0.231
	ca	0.094	0.107	0.000	0.201	0.074	0.201
	en	0.000	0.007	0.052	0.000	0.007	0.000
	gl	0.116	0.223	0.000	0.141	0.074	0.587
	pt	0.000	0.027	0.014	0.048	0.041	0.041
	es	0.136	0.208	0.000	0.616	0.008	0.112
	OOS	0.149	0.304	0.123	0.113	0.159	0.210

To provide a complete picture of performance on clean speech, Figure 2 shows DET curves on the CC-30s and OC-30s conditions for systems developed and evaluated on the Albayzin 2008 and 2010 LRE datasets. Note that the difference in performance between the closed-set and open-set conditions is similar for both datasets. Note also that the difference in performance for equivalent tasks defined on the Albayzin 2008 and 2010 LRE datasets, using the same state-of-the-art SLR system, is around 5 points in terms of Equal Error Rate.

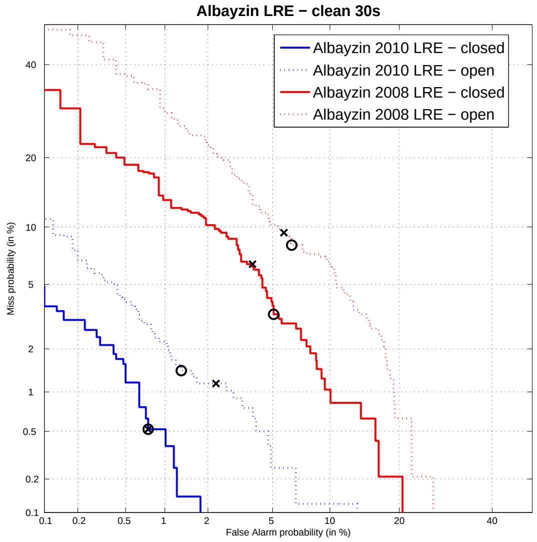

Figure 2: Pooled DET curves computed on the CC-30s and OC-30s test conditions, for systems developed and evaluated on the Albayzin 2008 and 2010 LRE datasets.

5.2. Performance on Noisy Speech

A SLR system was built based on the training and development (clean and noisy) speech signals provided for the Albayzin 2010 LRE. Note that a system built this way is not specially optimized to deal with noisy speech, but just enabled to deal with *any kind* of speech signals. This system was applied on the closed-set noisy-speech (CN) and open-set noisy-speech (ON) conditions of the Albayzin 2010 LRE, to check performance degradation when the evaluation set includes not only clean but also noisy speech segments. Results are shown in Table 9.

Table 9: Performance (C_{avg}) for the Albayzin 2010 LRE on the closed-set noisy-speech (CN) and open-set noisy-speech (ON) test conditions.

	CN-30s	**CN-10s**	**CN-3s**
Albayzin 2010 LRE	0.0177	0.0599	0.1476
	ON-30s	**ON-10s**	**ON-3s**
	0.0390	0.0867	0.1740

Performance for the noisy-speech condition was far worse than that found for the clean-speech condition. In particular, the C_{avg} for the CN condition was 2.81, 2.28 and 1.66 times higher than that reported for the CC condition on 30-, 10- and 3-second segments, respectively (see Table 5). By the way, it is worth noting that noisy speech produced higher degradation than OOS trials (open-set condition, see Table 7). This is graphically shown in Figure 3, with DET curves corresponding to two systems, built on clean speech (CC-30s and OC-30s conditions)

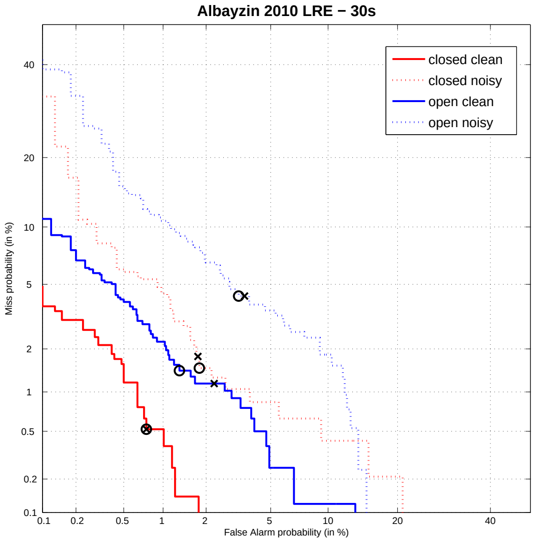

Figure 3: Pooled DET curves for two systems built on the Albayzin 2010 LRE datasets: a system built on clean speech (CC-30s and OC-30s conditions) and a system built on both clean and noisy speech (CN-30s and ON-30s conditions).

and both clean and noisy speech (CN-30s and ON-30s conditions) from the Albayzin 2010 LRE datasets. Finally, the C_{avg} for the ON condition was 6.19, 3.30 and 1.96 times higher than that reported for the CC condition on 30-, 10- and 3-second segments, respectively. As expected, the worst performance in all experiments on the Albayzin 2010 LRE was achieved for the ON-3s condition, with $C_{avg} = 0.1740$.

6. Conclusions and Challenges for Future Evaluations

The work presented in this paper involved the development and evaluation of a state-of-the-art spoken language recognition system on the Albayzin 2008 and 2010 LRE datasets, with the purpose of identifying the most challenging conditions, which may support design decisions in future evaluations.

It was found that the tasks defined for the Albayzin 2008 LRE were more challenging than those defined for the Albayzin 2010 LRE. Based on results on clean-speech data, three possible explanations were proposed: (1) the different amount of training and development data available to estimate models and tune system parameters; (2) the newly added target languages in 2010 (Portuguese and English), which were less confusable with other languages than the average and might push the average cost down; and (3) intrinsic features of the evaluation datasets, probably related to the presence of background noise in speech data tagged as clean that were provided for the Albayzin 2008 LRE.

A detailed analysis of the confusion of target languages with each other was performed, both in closed-set and open-set conditions, revealing that closely related languages (e.g. Romance languages in Spain) tend to be the most confused. When including OOS trials, target languages for which low confusion rates had been found in the closed-set condition, featured much higher error rates, due to the confusion with some similar OOS languages.

Finally, though reasonably good performance was attained even on noisy speech by using all the available data (clean and noisy speech) to train and calibrate systems, the highest degradation was found when dealing with noisy speech.

Future evaluations should address increasingly challenging tasks that make SLR technology progress and be useful in realistic applications. Attending to the results obtained using the Albayzin 2008 and 2010 LRE datasets, the most challenging conditions for a SLR system are related to: (1) the presence of background noise, music and/or conversations (which is common in many realistic aplications); (2) the acoustic, phonetic and lexical similarity of target languages (due to common roots or close evolution) and the need to reject OOS languages, which in some cases may be similar to target languages; and (3) the amount of speech available to make decisions (short segments). Besides them, the lack of training and development data could be also regarded as a challenging condition, as is frequently the case of low-resource target languages.

Taking these considerations into account, we propose three possible setups for future language recognition evaluations (not necessarily carried out by us):

- *Dialect recognition*. This task is intrinsically difficult, since dialects are variants of the *same* language (e.g. Spanish dialects). Dialect recognition has become an interesting application in the last years. In fact, the NIST 2011 LRE already addressed it as a pairwise language detection task. OOS languages should be provided to avoid the system to erroneously recognize a foreign language as a dialect. Speech segments of different durations and with diverse background conditions should be considered to match a realistic application, though each of them would probably involve a single speaker.
- *Large-scale European language recognition*. This task would involve many (30-50) European languages, as well as other OOS languages, and would basically extend the concept applied to design the Albayzin LREs so far. The high number of target languages is a challenging condition, since the confusion will probably increase as more target languages are considered. Speech data would be preferably recorded from the media (broadcast TV, internet TV, etc.), with the purpose of supporting language recognition for multilingual spoken document indexing and retrieval in multimedia resources. Speech data would include segments of different durations, featuring various speakers, diverse environment conditions, etc. Data collection would require the collaboration of research groups throughout Europe, and previous work should be done to define the protocols, copyright issues, data distribution, etc.
- *Language recognition in the wild*. This task would process uncontrolled resources in the internet, such as youtube videos, and would involve a small set of target languages, for which there could be many, few or no training data at all. Audio files would be required to include a minimum amount of speech (thus having a minimum duration), but they may also include music, animal sounds and whatever other non-speech sounds, and their quality would be diverse (from clean studio quality to outside far-field microphone recordings).

7. References

[1] *NIST LRE*, http://www.itl.nist.gov/iad/mig/tests/lre/.

[2] L. J. Rodriguez-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, pp. 172–179.

[3] Luis Javier Rodriguez-Fuentes, Mikel Penagarikano, Amparo Varona, Mireia Diez, and German Bordel, "The Albayzin 2010 Language Recognition Evaluation," in *Proceedings of Interspeech*, Firenze, Italia, August 28-31 2011, pp. 1529–1532.

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

[5] Alvin F. Martin and Audrey N. Le, "NIST 2007 language recognition evaluation," in *Proceedings of Odyssey 2008 - The Speaker and Language Recognition Workshop, paper 016*, 2008, p. 1.

[6] D. Martínez, J. Villalba, A. Miguel, A. Ortega, and E. Lleida, "I3A Language Recognition System for Albayzin 2010 LRE," in *Proceedings of Interspeech*, Firenze, Italy, 2011, pp. 849–852.

[7] Luis Javier Rodríguez-Fuentes, Mikel Penagarikano, Amparo Varona, Mirieia Díez, Germán Bordel, David Martínez, Jesús Villalba, Antonio Miguel, Alfonso Ortega, Eduardo Lleida, Alberto Abad, Oscar Koller, Isabel Trancoso, Paula Lopez-Otero, Laura Docio-Fernandez, Carmen Garcia-Mateo, Rahim Saeidi, Mehdi Soufifar, Tomi Kinnunen, Torbjørn Svendsen, and Pasi Frånti, "Multi-site Heterogeneous System Fusions for the Albayzin 2010 Language Recognition Evaluation," in *Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop (ASRU) 2011*, Hawaii, USA, December 2011.

[8] Mikel Penagarikano, Amparo Varona, Luis J. Rodriguez-Fuentes, Mireia Diez, and German Bordel, "University of the Basque Country (EHU) Systems for the 2011 NIST Language Recognition Evaluation," in *Proceedings of the NIST 2011 Language Recognition Evaluation (LRE) Workshop*, Atlanta (USA), 6-7 december 2011.

[9] Alvin Martin and Craig Greenberg, "The 2009 NIST language recognition evaluation," in *Odyssey 2010 - The Speaker and Language Recognition Workshop, paper 030*, Brno, Czech Republic, 2010, pp. 165–171.

[10] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The DET Curve in Assessment of Detection Task Performance," in *Proceedings of Eurospeech*, 1997, pp. 1985–1988.

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

[12] L. J. Rodriguez-Fuentes, M. Penagarikano, G. Bordel, A. Varona, and M. 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, pp. 1678–1685.

[13] L. J. Rodriguez-Fuentes, M. Penagarikano, A. Varona, M. Diez, and G. Bordel, "KALAKA-2: a TV broadcast speech database for the recognition of Iberian languages in clean and noisy environments," in *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*, Istanbul, Turkey, 23-25 May 2012.

[14] Albert Strasheim and Niko Brümmer, "SUNSDV system description: NIST SRE 2008," in *NIST Speaker Recognition Evaluation Workshop Booklet*, 2008.

[15] N. Brümmer, A. Strasheim, V. Hubeika, P. Matejka, L. Burget, and O. Glembek, "Discriminative Acoustic Language Recognition via Channel-Compensated GMM Statistics," in *Proceedings of Interspeech*, Brighton, UK, 2009, pp. 2187–2190.

[16] D. Martínez, O. Plchot, L. Burget, O. Glembek, and P. Matejka, "Language Recognition in iVectors Space," in *Proceedings of Interspeech*, Firenze, Italy, 2011, pp. 861–864.

[17] Petr Schwarz, *Phoneme recognition based on long temporal context*, Ph.D. thesis, Faculty of Information Technology, Brno University of Technology, http://www.fit.vutbr.cz/, Brno, Czech Republic, 2008.

[18] Steve Young, Gunnar Evermann, Mark Gales, Thomas Hain, Dan Kershaw, Xunying Lui, Gareth Moore, Julian Odell, Dave Ollason, Dan Povey, Valtcho Valtchev, and Phil Woodland, *The HTK Book (for HTK Versión 3.4)*, Entropic, Ltd., Cambridge, UK, 2006.

[19] Andreas Stolcke, "SRILM - an extensible language modeling toolkit," in *Proceedings of Interspeech*, November 2002, pp. 257–286.

[20] F. Richardson and W. Campbell, "Language recognition with discriminative keyword selection," in *Proceedings of ICASSP*, 2008, pp. 4145–4148.

[21] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin, "LIBLINEAR: A library for large linear classification," *Journal of Machine Learning Research*, vol. 9, pp. 1871–1874, 2008, Software available at http://www.csie.ntu.edu.tw/∼cjlin/liblinear.

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

[23] N. Brümmer, L. Burget, J. Cernocky, O. Glembek, F. Grezl, M. Karafiat, D. van Leeuwen, P. Matejka, P. Schwarz, and A. Strasheim, "Fusion of heterogeneous speaker recognition systems in the STBU submission for the NIST speaker recognition evaluation 2006," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 7, pp. 2072–2084, 2007.