The Albayzin 2010 Language Recognition Evaluation

Luis Javier Rodriguez-Fuentes, Mikel Penagarikano, Amparo Varona, Mireia Diez, German Bordel

GTTS, Department of Electricity and Electronics
University of the Basque Country, Spain
luisjavier.rodriguez@ehu.es

Abstract

The Albayzin 2010 Language Recognition Evaluation (LRE), carried out from June to October 2010, was the second effort made by the Spanish/Portuguese community for benchmarking language recognition technology. As the Albayzin 2008 LRE, it was coordinated by the Software Technology Working Group of the University of the Basque Country, with the support of the Spanish Thematic Network on Speech Technology. A speech database was created for system development and evaluation. Speech signals were recorded from TV broadcasts, including clean and noisy speech. The task consisted in deciding whether or not a target language was spoken in a test utterance, and involved 6 target languages: English, Portuguese and the four official languages in Spain (Basque, Catalan, Galician and Spanish), other (*Out-Of-Set*) languages being also recorded to allow open-set verification tests. This paper presents the main features of the evaluation, analyses system performance on different conditions, including the confusion among languages, and gives hints for future evaluations.

Index Terms: Language Recognition, Broadcast Speech, Language Resources and Evaluation

1. Introduction

The Albayzin 2010 Language Recognition Evaluation (Albayzin 2010 LRE), carried out from June to October 2010, was the second effort made by the Spanish/Portuguese community for benchmarking language recognition technology. As the Albayzin 2008 LRE, it was coordinated by the Software Technologies Working Group of the University of the Basque Country, with the support of the Spanish Network on Speech Technology [1]. The evaluation aimed to promote creativity, discussion and collaboration between research groups working on automatic language identification and verification, to explore the limits of state-of-the-art technology and eventually to foster research progress and technological developments in this area.

Regarding the task, test conditions and performance measures, the Albayzin 2010 LRE was defined in almost the same terms as the last NIST Language Recognition Evaluations [2, 3], but considering a reduced set of target languages (Spanish, Catalan, Basque, Galician, Portuguese and English) and dealing with speech extracted from multi-speaker TV broadcast recordings. Note that a test segment could contain speech from various speakers. This is a relevant difference with regard to NIST evaluations, whose data were extracted from telephone-channel two-speaker conversations, test segments containing speech from a single speaker.

Test conditions for this evaluation were almost identical to those applied for the Albayzin 2008 LRE [4], adding Portuguese and English as target languages and introducing a new test condition involving noisy and/or overlapped speech. Four different test conditions, depending on the set of non-target languages (closed-set vs. open-set) and the background conditions (clean vs. noisy), and three nominal segment durations (30, 10 and 3 seconds) were considered, leading to 12 different tracks. An award was presented to the system yielding best performance in the CC-30 track (closed-set verification of 30-second segments containing clean-speech), which was mandatory.

The rest of the paper is organized as follows. The language detection task is briefly defined in Section 2. Test conditions, performance measures, and speech data used for development and evaluation are described in Sections 3, 4 and 5, respectively. The evaluation schedule is outlined in Section 6. Results are presented and discussed in Section 7, with special attention to the confusion among languages. Finally, hints for future evaluations are given in Section 8.

2. The language detection task

The language detection task was defined in the same terms as for NIST evaluations [2, 3]: *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 and comparing system decisions with the right ones (stored in a keyfile).

Each trial comprises a segment of audio containing speech in a single language, the identity of the target language of interest and the identities of the languages that might be spoken in the segment (which we will call *non-target* languages). For each trial, the system must output a hard decision (yes/no) about whether or not the target language is spoken in the segment, and 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.

3. Test conditions

3.1. Closed-set vs. open-set verification

Depending on the restrictions imposed to the set languages that might be spoken in the segment, two types of verification tests were defined: (1) *closed-set verification*, where the set of trials is limited to segments containing speech in one of the target languages, and scores are computed based on those trials; and

This work has been supported by the University of the Basque Country under grant GIU10/18, by the Government of the Basque Country under program SAIOTEK (project S-PE10UN87) and by the Spanish MICINN under Plan Nacional de I+D+i (project TIN2009-07446, partially financed by FEDER funds).

(2) *open-set verification*, where scores are computed based on the whole set of trials, including those corresponding to segments containing speech in an *Out-Of-Set* language. This way, systems could be designed specifically for closed-set or open-set verification, and research groups were given the opportunity to submit separate results for each condition. Out-Of-Set languages were not disclosed to participants.

3.2. Clean vs. noisy speech

The development and evaluation datasets consisted of two subsets: (1) *clean* segments, featuring high SNR speech signals, maybe with short fragments of noisy and/or overlapped speech (in a single language); and (2) *noisy* segments, featuring noisy and/or overlapped speech (in a single language), maybe with short fragments of clean speech. The subset of noisy segments might contain different and variable types of noise: street, music, cocktail party, laughs, clapping, etc. Telephone-channel speech signals were not used in any case. Segments containing overlapped speech were extracted from informal debates in late night shows, magazines, etc. which on the other hand, might feature clean-channel and quiet-background (studio) conditions. This condition was introduced with the aim to measure the performance of language verification systems designed to deal with clean speech, when dealing with noisy and/or overlapped speech; and, on the other hand, to measure the performance of language verification systems specifically designed to deal with noisy and/or overlapped speech.

3.3. Duration of speech segments

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. Segments were defined to begin and end at times of non-speech as determined by an automatic speech activity detection algorithm. So, actual segment durations may be slightly longer (but not shorter) than nominal durations. Note that each segment was extracted from an original TV broadcast recording, containing speech in a single language (from one or more speakers) mixed with fragments of non-speech (silence or background noise), so the actual amount of speech was smaller than segment duration. Nominal segment durations were not disclosed to participants (though they could be guessed very easily).

4. Performance measures

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 Out-Of-Set languages), that is, the fraction of trials corresponding to language j that are erroneously accepted as containing language i. In NIST evaluations, the cost model, based on miss and false alarm errors, depends on three application parameters: C_{miss}, C_{fa} and P_{target}. For this evaluation, the same values used in the Albayzin 2008 LRE (the same used in NIST 2007 and 2009 LRE) were applied:

$$C_{miss} = C_{fa} = 1$$
$$P_{target} = 0.5$$

The cost function C_{avg} is computed by averaging costs associated to miss and false alarm errors for all the target languages, as follows:

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

In Equation 1, $P_{non-target}$ is the prior probability of non-target languages (assuming for them a uniform distribution) and P_{OOS} the prior probability of Out-Of-Set languages. In this evaluation, the following values were applied:

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

The average cost C_{avg} was computed separately for each of the four test conditions and for each of the three segment duration categories, and served as the main system performance measure in this evaluation.

4.2. Graphical evaluation: DET curves

Detection Error Tradeoff (DET) curves [5] provide a straightforward way of comparing global performance of different systems for a given test condition and are used in NIST evaluations to support system performance comparisons. In this evaluation, NIST software [6] was used to generate DET curves, including marks for the operation point given by system decisions and the operation point corresponding to the minimum C_{avg}.

5. Data

Participants were allowed to use any available data and subsystems to build their systems. However, for better matching the acoustic conditions of test materials, the organization provided data for system development. In fact, a database (called Kalaka-2) was specifically prepared for this evaluation, including three subsets: train, development and test. Speech signals were extracted from TV broadcast recordings (news, debates, late night shows, etc.), featuring various dialects and/or linguistic competence levels, speech modalities (planned speech, formal conversations, spontaneous speech, etc.), and diverse environment conditions. The sets of TV shows posted to each subset were forced to be disjoint, meaning that any show appearing in one subset did not appear in the other two. This restriction was imposed as an attempt to guarantee speaker independence. Broadcasts were digitally recorded using a Roland Edirol R-09 recorder, audio signals being stored in WAV files (PCM, 16 kHz, single channel, 16 bits/sample).

The database was designed as an extension of Kalaka, the database used for the Albayzin 2008 LRE [7]. To reduce development costs, all the materials of Kalaka were re-used: the train and development datasets of Kalaka were used to build the train dataset of Kalaka-2, and the test dataset of Kalaka was used to build the development dataset of Kalaka-2. Groups that already participated in the 2008 campaign were warned not to use Kalaka, to avoid overtraining. To increase the amount of data, new TV broadcasts were recorded, selected and classified, specially for the two new target languages (Portuguese and English) and for the Out-Of-Set languages. In particular, the test set of Kalaka-2 was entirely extracted from new recordings.

5.1. Train dataset

The train dataset consisted of more than 10 hours of clean speech per target language (for some of them, around 11 hours or even 12 hours were available). Its contents (fragments of variable length) did not all strictly consist of clean speech. Besides some portions of silence, they also featured short fragments containing noisy and/or overlapped speech. In a separate folder, more than 2 hours of noisy/overlapped speech were also provided for each target language (for some of them, more than three hours of noisy speech were provided). No train data were provided containing Out-Of-Set languages. The train dataset amounts to more than 82 hours of speech (80% of the time corresponding to clean speech and 20% to noisy speech).

5.2. Development and test datasets

The development and test datasets had the same size and characteristics, except for the distribution of Out-Of-Set languages and the proportion of clean and noisy speech. Both datasets contained segments with nominal durations of 30, 10 and 3 seconds, with at least 150 speech segments per target language and nominal duration. Each segment contained speech (from one or more speakers) in one of the 6 target languages or in an Out-Of-Set language.

The development set consisted 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 test set consisted 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).

6. Evaluation schedule

The evaluation plan was released and registration opened on May 18, 2010. By June 22, train and development data were sent to registered participants, including a keyfile and a scoring script which allowed to tune system parameters. The scoring script was based on that used for the NIST 2007 and 2009 LRE, with minor changes needed to match the task and to add the identifiers of the 6 target languages considered in this evaluation. A wiki was also activated to improve communication and collaboration between sites and the organizing team.

On September 27, the test dataset was released via web. The deadline for submitting system results (along with system descriptions) was October 17 at 24:00, GMT+1. Results had to be sent in a format similar to that used for NIST evaluations: a text file with a trial per line, each trial consisting of 6 blank-separated fields: background condition (clean/noisy), target language, operation mode (closed-set/open-set), test file, decision and score. Participants could send results for as many systems as they want, but only one primary system per test condition, the remaining systems being *contrastive*.

Preliminary results in all conditions and the keyfile for the test set were released through the wiki on October 25. Only primary systems were taken into account for rankings in all conditions and for the three nominal segment durations, according to C_{avg}, as defined in Section 4.1. Finally, evaluation results and site submissions were presented in a special session of FALA 2010 [8], held in Vigo on November 10-12, 2010.

7. Results

Four teams, two from Spain, one from Portugal and one from Finland, submitted their systems to the Albayzin 2010 LRE. We will briefly refer to them as T1, T2, T3 and T4 (not necessarily in the same order as above). Results (in terms of C_{avg}) for the best primary system in each condition and the best overall result attained in each track are shown in Table 1. Test conditions (CC, OC, CN and ON) are coded so that the first letter refers to the set of non-target languages: C (closed-set) or O (open-set), and the second letter to background conditions: C (clean speech) or N (noisy speech). DET curves corresponding to the best primary systems submitted to the Albayzin 2010 LRE in the four test conditions for the subset of 30-second segments are shown in Figure 1. All sites reported processing times under 1.0×Real-Time. The most competitive systems, applying state-of-the-art language recognition technology, reported processing times of 0.9×RT (T1) and 0.51×RT (T2).

Table 1: Performance (C_{avg}) of the best primary system in each condition, and the best overall result at each track.

		30 sec	**10 sec**	**3 sec**
CC	best primary (T1)	0.0184	0.0418	0.0943
	best overall	0.0181	0.0359	0.0844
OC	best primary (T1)	0.0307	0.0644	0.1202
	best overall	0.0296	0.0445	0.1029
CN	best primary (T2)	0.0316	0.0767	0.1503
	best overall	0.0253	0.0636	0.1217
ON	best primary (T2)	0.0749	0.1092	0.1735
	best overall	0.0475	0.0936	0.1551

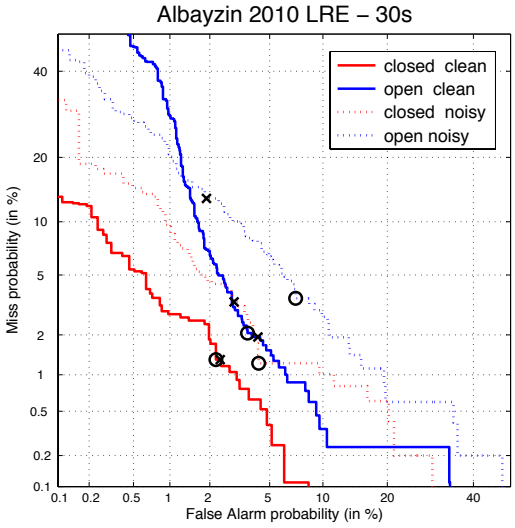

Figure 1: Pooled DET curves for the best primary systems in the four test conditions, on the subset of 30-second segments, including marks for the operation point given by system decisions (X) and that corresponding to the minimum cost (O).

Regarding the mandatory and also the easiest condition (CC), the best primary system yielded $C_{avg} = 0.0184$ on the subset of 30-second segments (thick red curve in Figure 1), whereas the best system overall (a contrastive system by T2) was only slightly better: $C_{avg} = 0.0181$.

Regarding the dependence on segment duration, for the most competitive systems (except for the ON condition, where degradation was smaller) the C_{avg} obtained on the subset of 10-second segments doubled that obtained on the subset of 30-second segments. The same trend was observed for 3-second segments with regard to 10-second segments. This was consistent with previous results for other evaluations.

As illustrated by DET curves in Figure 1, dealing with noisy speech did not lead to catastrophic performance degradations. The increase in cost when moving from clean to noisy speech ranged (depending on the system and condition) from 40% to 80% in most cases, being relatively smaller for short segments. In any case, it seems that fairly good performance can be attained on noisy speech if enough and suitable data are provided to train and calibrate systems.

Finally, when dealing with speech signals in Out-Of-Set languages (open-set condition), the number of false alarms increased and performance degraded. As shown in Table 1, the best primary system in OC-30 yielded $C_{avg} = 0.0307$, meaning around 67% increase in cost with regard to the best primary system in CC-30. A similar figure (63.5%) is obtained when comparing the best overall results in both tracks. As for noisy speech, the increase in cost was less remarkable on the subsets of 10- and 3-second segments.

7.1. Confusion among languages

Table 2 shows P_{miss} (in the main diagonal) and P_{fa} (off-diagonal) computed on test segments containing target languages and Out-Of-Set (OOS) languages, for the best system at the OC-3 track. Qualitatively similar results were obtained for the best system at the ON-3 track. Error probabilities for 30- and 10-second speech segments were very low, though the same trends in language confusion were observed.

Table 2: Confusion among Basque (eu), Catalan (ca), English (en), Galician (gl), Portuguese (pt), Spanish (es) an Out-Of-Set (OOS) languages: P_{miss} in the main diagonal and P_{fa} off-diagonal, for the OC-3 track.

		Target					
		eu	ca	en	gl	pt	es
Segment	eu	0.18	0.07	0.00	0.10	0.00	0.22
	ca	0.01	0.13	0.01	0.18	0.02	0.27
	en	0.01	0.05	0.02	0.00	0.02	0.00
	gl	0.01	0.27	0.00	0.19	0.06	0.62
	pt	0.01	0.08	0.00	0.02	0.01	0.01
	es	0.04	0.27	0.00	0.46	0.03	0.11
	OOS	0.13	0.32	0.17	0.13	0.19	0.15

Regarding the confusion among target languages, note that Romance languages spoken in Spain were highly confusable amongst each other, specially Spanish and Galician. Basque, which is not Romance but whose speakers (most of them) also speak Spanish, was confused with the other languages in Spain, specially Spanish. This may indicate that sharing speakers in a bilingual community makes the two languages more confusable, but other factors (such as mutual influence and common evolution) may be also affecting. Note, for instance, that Portuguese (which is also a Romance language, but less related to the other languages in the Iberian Peninsula) was only marginally confused with Galician and Catalan. Finally, English, which is not Romance and does not share speakers with the other languages, shows almost null confusion rates. Even at the ON-3 track (i.e. for noisy speech, where the confusion was expected to increase), Portuguese and English showed low average false alarm probabilities: 0.0349 and 0.0185, respectively.

When considering OOS segments (which included speech in Arabic, French, German or Romanian), relatively high false alarm rates were found for all the target languages, specially for Catalan, which was mostly confused with Arabic and Romanian. These were also the most confusable OOS languages for Basque, Galician and Spanish. False alarm rates on OOS segments were specially remarkable for Portuguese and English, compared to the low rates observed for them on segments containing target languages. English was mostly confused with German, and Portuguese with French and Romanian. Details are not given here for a lack of space.

8. Hints for future evaluations

Future evaluations should address increasingly challenging tasks that make SLR technology progress. In fact, taking into account the low error rates attained in the CC-30 track (which were further reduced through system fusion), it seems that SLR technology is mature enough to address more challenging tasks. Future evaluations should focus on open-set tests of short segments, containing at most 10 seconds of speech in realistic background conditions (e.g. those appearing in casual broadcast recordings). On the other hand, less related languages can be better discriminated amongst each other, even in challenging conditions (noisy speech, short segments, etc.), as was the case of Portuguese and English in this evaluation, with regard to the languages spoken in Spain (Basque, Catalan, Galician and Spanish). Therefore, future evaluations should define tasks involving highly confusable languages, due either to sociological issues (bilingual communities) or to acoustic, phonetic and lexical similarities (which may happen for languages with the same roots, such as Romance languages). The number of target languages may also affect performance, since the confusion among languages should increase as more target languages were considered. We will take these considerations into account for designing the next Albayzin LRE.

9. Acknowledgements

We thank all the members of the Organizing Committee of FALA 2010 for their help and support. We also thank all the participants for their work and feedback.

10. References

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

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

[3] A. Martin and C. Greenberg, "The 2009 NIST Language Recognition Evaluation," in *Odyssey 2010 - The Speaker and Language Recognition Workshop, paper 030*, 2010.

[4] L. J. Rodriguez-Fuentes, M. Penagarikano, G. Bordel, and A. Varona, "The Albayzin 2008 Language Recognition Evaluation," in *Odyssey 2010: The Speaker and Language Recognition Workshop*, 2010.

[5] 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. 1895–1898.

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

[7] 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 LREC*, 2010, pp. 1678–1685.

[8] *FALA 2010: VI Jornadas en Tecnologia del Habla and II Iberian SLTech Workshop, Vigo (Spain), 10-12 November 2010*, proceedings online at http://fala2010.uvigo.es/.