Overview of the Albayzin 2010 Language Recognition Evaluation: database design, evaluation plan and preliminary analysis of results

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

This paper presents an overview of the Albayzin 2010 Language Recognition Evaluation, carried out from June to October 2010, organized by the Spanish Thematic Network on Speech Technology and coordinated by the Speech Technology Working Group of the University of the Basque Country. The evaluation was designed according to the test procedures, protocols and performance measures used in the last NIST Language Recognition Evaluations. Development and evaluation data were extracted from KALAKA-2, a database including clean and noisy speech in various languages, recorded from TV broadcasts and stored in single-channel 16-bit 16 kHz audio files. The task consisted in deciding whether or not a target language was spoken in a test utterance. Four different conditions were defined: closed-set/clean-speech, closed-set/noisy speech, open-set/clean-speech and open-set/noisy speech. Evaluation was performed on three subsets of test segments, with nominal durations of 30, 10 and 3 seconds, respectively. The task involved 6 target languages: English, Portuguese and the four official languages spoken in Spain (Basque, Catalan, Galician and Spanish), other (*unknown*) languages being also recorded to allow open-set verification tests. Four teams (2 from Spanish universities, one from a Portuguese research center and one from a Finnish university) presented their systems to this evaluation. The best primary system in the closed-set/clean-speech condition on the subset of 30-second segments yielded $C_{avg} = 0.0184$ (around 2% EER).

Index Terms: Language Recognition Evaluation, KALAKA-2, Spanish Thematic Network on Speech Technology

1. Introduction

The Albayzin 2010 Language Recognition Evaluation (Albayzin 2010 LRE), 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], aimed to promote creativity, discussion and collaboration between research groups (specially from Spain and Portugal, though worldwide participation was welcome) 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], with three important changes:

- Portuguese and English were added as target languages,
- The so-called *restricted development* condition was not considered anymore, and
- a new test condition involving noisy and/or overlapped speech was introduced.

Therefore, four different test conditions, depending on the operation mode (closed-set vs. open-set) and the background conditions (clean vs. noisy), were defined. Also, 3 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. The test conditions and the measures used to evaluate system performance are described in Sections 3 and 4, respectively. Section 5 describes the database and Section 6 addresses issues related to the organization of Albayzin 2010 LRE. Results are presented and briefly discussed in Section 7, with special attention to the closed-set clean-speech condition (which was mandatory), and devoting some space to a special activity carried out after evaluation results were submitted. Finally, conclusions and future work are outlined 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

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).

and comparing system decisions with the right ones (stored in a keyfile).

Each trial comprises the following elements:

- 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:

- 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. This means that, for each trial, non-target languages are limited to all the target languages except for the target language of interest in that trial.
- In *open-set verification*, scores are computed based on the whole set of trials, including those corresponding to segments containing speech in an *unknown* language. This means that, for each trial, non-target languages are all the possible languages except for the target language of interest in that trial.

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. The set of *unknown* languages were not disclosed to participants.

3.2. Clean vs. noisy speech

The development and evaluation datasets consisted of two subsets:

- *clean* segments, featuring high SNR speech signals, maybe with short fragments of noisy and/or overlapped speech (in a single language), and
- *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 be 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. As noted above, each segment contains speech in a single language, which also applies to overlaps and fragments with background speech, except for the case of segments in unknown languages, which might contain speech in two or more languages, provided that none of them are target languages.

This condition was introduced with two objectives:

- to measure the performance of language verification systems designed to deal with clean speech, when dealing with noisy and/or overlapped speech, and
- 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

The language verification task defined for 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* languages), 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:

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

The cost model depends on three application parameters: C_{miss}, C_{fa} and P_{target}. For this evaluation, the same values used in the Albayzin 2008 LRE (which are also the same used in NIST 2007 and 2009 LRE) were applied:

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

Finally, an average cost is defined by adding the contributions for all the combinations of target and non-target languages, as follows:

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

where $P_{non-target}$ is the prior probability of non-target languages (assuming for them a uniform distribution) and P_{OOS} the prior probability of *unknown* (*Out-Of-Set*) languages. In this evaluation, the following values were 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} 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. Log-Likelihood Ratio (LLR) average cost

Sites may specify that their scores could be interpreted as log-likelihood ratios. In such cases, detection results were also evaluated in terms of the so called C_{LLR} [5], which is commonly 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; and (2) it does not depend on application costs.

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:

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

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:

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

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:

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

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 [7] provide a straightforward way of comparing global 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, NIST software [8] 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

The database used for this evaluation, called KALAKA-2, was organized in three subsets: train, development and evaluation. Speech signals were extracted from TV broadcast recordings (news, documentaries, debates, interviews, reportages, magazines, 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. 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). We strongly recommended to prepare systems starting from the materials provided for this evaluation, but participants were allowed to use any available data and subsystems. 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.

The database was designed as an extension of KALAKA, the database created ad-hoc for the Albayzin 2008 LRE [9]. To reduce development costs, all the materials of KALAKA were re-used for KALAKA-2, as follows:

- The train and development datasets of KALAKA were used to build the train dataset of KALAKA-2.
- The evaluation dataset of KALAKA was used to build the development dataset of KALAKA-2.

To complete the datasets of KALAKA-2, new TV broadcasts were recorded, selected and classified, specially for the two new target languages (Portuguese and English) and for the *unknown* (Out-Of-Set) languages. In particular, the evaluation dataset was completely new and independent of KALAKA.

5.1. Training data

The train dataset consisted of more than 10 hours of clean speech per target language. 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. No data were provided containing *un-*

known (Out-Of-Set) languages. The distribution of training data is shown in Table 1.

Table 1: Distribution of training segments per target language for clean and noisy speech: 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

5.2. Development and evaluation data

The development and evaluation datasets had the same size and characteristics, except for the distribution of unknown 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 *unknown* (Out-Of-Set) language. Speech segments were given random names, so that languages and durations appeared in a random sequence.

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 evaluation 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). The distribution of segments per language is shown in Table 2.

In the case of clean speech, speech segments of 30, 10 and 3 seconds were automatically extracted from fragments of clean speech 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.

In the case of noisy speech, segments from 30 to 35 seconds were manually extracted from recordings, and then segments with nominal durations of 10 and 3 seconds were automatically extracted from the former, according to the same criteria applied for clean speech.

Table 2: Distribution of segments per language (the same for each duration) in the development and evaluation datasets.

		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
Unknown languages	**Arabic**	100	25	115	22
	French	120	32	70	34
	German	108	73	13	32
	Romanian	0	0	111	43

6. Rules and schedule

All the registered participants received three DVD containing train speech for the six target languages, plus an additional DVD with development data, a keyfile and a scoring script which allowed to tune system parameters (such as verification thresholds, fusion weights, etc.). 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. The evaluation dataset was released via web (restricted to registered participants) and eventually distributed in a single DVD at FALA 2010.

Registration involved the commitment to use data exclusively for research purposes, distribution being allowed only with explicit permission. After the evaluation, registered participants were allowed to use the data to develop or evaluate their own systems, provided that they acknowledged that use by means of a suitable reference:

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

Four test conditions were defined: CC (closed-set, clean-speech), CN (closed-set, noisy-speech), OC (open-set, clean-speech) and ON (open-set, noisy-speech). The ranking of systems in all conditions and for the three nominal segment durations was determined by taking into account the average cost C_{avg}, as defined in Section 4.1. Participants could send results for as many systems as they want, but only one primary system per test condition, the remaining systems being *contrastive*. Only primary systems were taken into account for rankings. The CC condition was mandatory: an award was presented for the best system (i.e. that yielding the least C_{avg}) in the CC condition on the subset of 30-second segments.

Detection 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. Since multiple systems could be submitted, a naming protocol was established, consisting of a site identifier, a test condition identifier and a system identifier (primary, contrastive1, contrastive2, etc.). Each participant 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.).

The Evaluation Schedule was as follows:

- *May 18, 2010*
  - The evaluation plan is released through the website of FALA 2010.
  - Registration for Albayzin 2010 LRE opens.
  - An online registration form is made available through the website of FALA 2010.
- *June 22, 2010.*
  - Train and development datasets are sent to registered sites via courier.
  - A *wiki* is activated to improve communication and collaboration between the registered participants and the organizing team.
- *July 15, 2010.*
  - Registration for Albayzin 2010 LRE closes.
- *September 27, 2010.*
  - The evaluation dataset is released via web (restricted to registered participants).
  - System submission (via e-mail) opens.
- *October 17, 2010.*
  - System submission deadline (24:00, GMT+1).
- *October 25, 2010.*
  - Preliminary results in all conditions and the keyfile for the evaluation dataset are released to participants through the wiki.
- *November 2, 2010.*
  - Deadline for submitting final system descriptions (including analysis of results).
- *November 10-12, 2010 (FALA 2010, Vigo, Spain).*
  - Albayzin 2010 LRE Workshop: delivery of a DVD including documentation and evaluation data to registered participants, poster presentations and discussion.
  - Plenary session: summary of results and awards.

7. Results

Four teams, two from Spain, one from Portugal and one from Finland, submitted their systems to the Albayzin 2010 LRE (see Table 3). Full descriptions of the submitted systems can be found as regular papers in the proceedings of FALA 2010. Results (in terms of C_{avg}) in the four test conditions and for the three segment durations are shown in Tables 4, 5, 6 and 7.

Regarding the mandatory condition, the best primary system on the subset of 30-second segments was submitted by GTC-VIVOLAB (thus the award winner), yielding $C_{avg} = 0.0184$ (around 2% EER). Note, however, that the best system in this condition was the second contrastive system submitted by L^2F, with $C_{avg} = 0.0181$. The postkey submissions by L^2F outperformed the replaced systems, but not their second contrastive system nor the primary system by GTC-VIVOLAB. The DET curves for all the primary (thick lines) and contrastive (thin lines) systems submitted to this condition (CC-30) are shown in Figure 1. The actual and the minimum achievable costs (marked in the DET curves of primary systems with X and O, respectively) are shown in Figure 2, revealing calibration losses for some systems.

Table 3: Teams participating in the Albayzin 2010 Language Recognition Evaluation.

Team ID	**Research institution**	**Submitted conditions**
GTC-VIVOLAB	University of Zaragoza	CC, OC
L2F	L^2F INESC-ID Lisboa	All
UEF_NTNU	Univ. of Eastern Finland	CC
GTM	University of Vigo	CC, CN

Table 4: Performance (C_{avg}) of primary and contrastive systems submitted to Albayzin 2010 LRE in the CC test condition. Systems submitted after the keyfile release are shown too (though they do not count for rankings).

	CC-30	**CC-10**	**CC-3**
VIVOLAB_UZ_CC_pri	**0.0184**	0.0418	0.0943
VIVOLAB_UZ_CC_alt1	0.0238	0.0498	0.1087
L2F_CC_pri	0.0320	0.0513	0.1034
L2F_CC_pri_postkey	0.0223	0.0359	0.0853
L2F_CC_alt1	0.0910	0.0540	0.1065
L2F_CC_alt1_postkey	0.0219	0.0363	0.0844
L2F_CC_alt2	0.0181	0.0459	0.1055
UEF-NTNU_CC_pri	0.1636	0.3035	0.3799
UVIGO-GTM_CC_pri	0.1916	0.2934	0.4447
UVIGO-GTM_CC_alt1	0.2888	0.3181	0.3956

Table 5: Performance (C_{avg}) of primary and contrastive systems submitted to Albayzin 2010 LRE in the CN test condition. Systems submitted after the keyfile release are shown too (though they do not count for rankings).

	CN-30	**CN-10**	**CN-3**
L2F_CN_pri	0.0316	0.0767	0.1503
L2F_CN_pri_postkey	0.0416	0.0810	0.1273
L2F_CN_alt1	0.3556	0.0892	0.2080
L2F_CN_alt1_postkey	0.0403	0.0754	0.1217
L2F_CN_alt2	0.0253	0.0636	0.1342
UVIGO-GTM_CN_pri	0.2744	0.3534	0.4476
UVIGO-GTM_CN_alt1	0.2978	0.3412	0.4309

Regarding the dependence on the available amount of speech, for the most competitive systems 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 (e.g. see results for the best primary system in the CC condition). This was consistent with previous results for other evaluations. The following analyses will focus on the subset of 30-second segments.

As may be expected, the performance degraded in open-

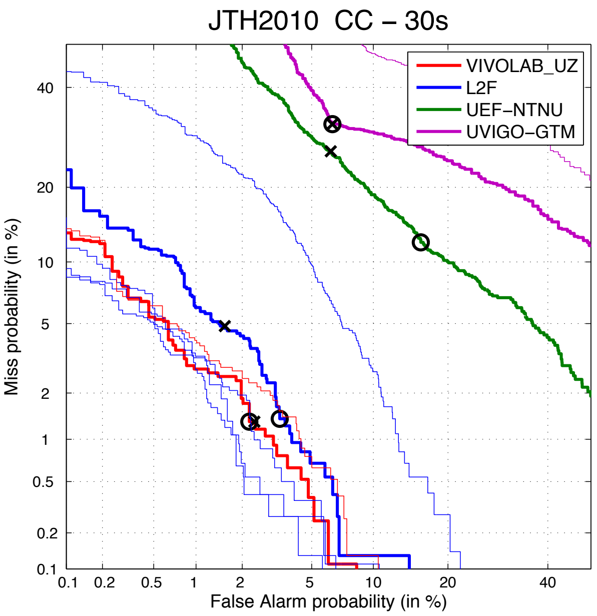

Figure 1: Pooled DET curves of systems submitted to the Albayzin 2010 LRE in the CC condition for the subset of 30-second segments.

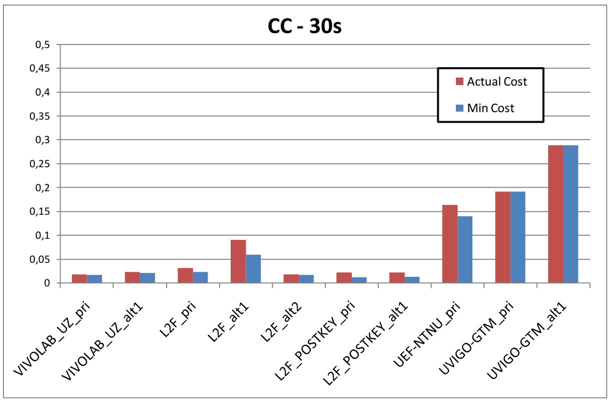

Figure 2: Actual and minimum achievable costs of systems submitted to the Albayzin 2010 LRE in the CC condition for the subset of 30-second segments.

set tests, due to the increase of false alarms in trials involving speech signals in *unknown* languages (e.g. see Tables 4 and 6). For instance, the primary system of GTC-VIVOLAB yielded $C_{avg} = 0.0307$ in the OC-30 condition, which means around 67% increase in cost with regard to the CC-30 condition. Similar figures were observed for other systems in the same conditions: 49% and 88% cost increases for the primary and second contrastive L^2F systems, respectively. A detailed study of the confusion among languages is not included here for a lack of space.

Finally, a new condition was introduced in this evaluation with the aim to test how much the performance of language recognition systems degraded when dealing with noisy and/or overlapped speech. Not all the sites submitted results for the CN and ON conditions. In fact, only L^2F submitted results for all the conditions, so the analysis will focus on the primary and second contrastive L^2F systems (which are competitive systems with a consistent behavior across all conditions). Performance degradation was not so catastrophic as we expected. In fact, when comparing CN-30 with CC-30 results, the primary L^2F system surprisingly showed a slight improvement, whereas the second contrastive L^2F system showed *only* a 40% cost increase. The latter result is quite representative, since the increase in cost ranges from 30% to 50%, depending on the system and condition. In any case, it seems that good performance can be attained even on noisy speech if data are provided to train and calibrate systems.

Table 6: Performance (C_{avg}) of primary and contrastive systems submitted to Albayzin 2010 LRE in the OC test condition. Systems submitted after the keyfile release are shown too (though they do not count for rankings).

	OC-30	**OC-10**	**OC-3**
VIVOLAB_UZ_OC_pri	0.0307	0.0644	0.1202
VIVOLAB_UZ_OC_alt1	0.0373	0.0635	0.1309
L2F_OC_pri	0.0478	0.0750	0.1297
L2F_OC_pri_postkey	0.0296	0.0468	0.1073
L2F_OC_alt1	0.1416	0.1225	0.1460
L2F_OC_alt1_postkey	0.0309	0.0445	0.1029
L2F_OC_alt2	0.0341	0.0611	0.1289

Table 7: Performance (C_{avg}) of primary and contrastive systems submitted to Albayzin 2010 LRE in the ON test condition. Systems submitted after the keyfile release are shown too (though they do not count for rankings).

	ON-30	**ON-10**	**ON-3**
L2F_ON_pri	0.0749	0.1092	0.1735
L2F_ON_pri_postkey	0.0700	0.0981	0.1551
L2F_ON_alt1	0.3778	0.1311	0.2328
L2F_ON_alt1_postkey	0.0839	0.0948	0.1609
L2F_ON_alt2	0.0475	0.0936	0.1654

7.1. Processing times

Processing times for the submitted systems, in terms of real-time factor (×RT), along with the CPU and memory specifications of the servers used to run the experiments, are shown in Table 8 (only data provided by the participating teams are shown). All the systems are reported to run under 1×RT, but on servers with very different computational power. The most competitive systems have reported processing times of 0.9 (GTC-VIVOLAB) and 0.51 (L^2F).

Table 8: Processing time (×RT) for the submitted systems.

Systems	**CPU-RAM**	**×RT**
GTC-VIVOLAB	–	0.9
L2F	2xQuad Xeon E5530 2.4GHz, 48 GB	0.51
UEF_NTNU	Xeon X5450 3.0GHz	0.051
GTM (p)	Xeon E5620 2.4 GHz,18 GB	0.0288
GTM (c)	Xeon E5620 2.4 GHz,18 GB	0.0533

7.2. Exploring cross-site fusions

We proposed to participants an interesting way of collaboration: to investigate which subsystems combined better under a FoCal-based fusion paradigm, which may help future developments of language recognition systems (and potential collaborations). We focused on the core condition (closed-set, clean speech, 30-second segments). To accomplish that objective, we asked for them to submit log-likelihoods for their subsystems, giving details of the applied methodology. This way, previously unexplored cross-site fusions may give valuable cues of which kind of systems would be worth developing and combining.

Three sites submitted the log-likelihoods for the six target languages produced by their subsystems on the CC-30 evaluation subset. Note that this information had not been previously disclosed, since each team studied and optimized the fusion of subsystem scores, and what they called *system* was in fact the fusion of various subsystems. Additionally, the organizing team (GTTS) included the log-likelihoods of its own subsystems: three phonotactic-SVM subsystems, for Czech, Hungarian and Russian BUT decoders, using expected n-gram counts (up to 3-grams) computed on phone-lattices.

All the information was uploaded and results presented through the wiki created for this evaluation. For a lack of space, we do not include results here. Only note that the best cross-site fusion (including 5 subsystems from GTTS, GTC-VIVOLAB and L^2F) yielded $C_{avg} = 0.0054$, almost three times lower than that obtained by the best system in the CC-30 condition.

8. Conclusions

In this paper, the main features of the Albayzin 2010 Language Recognition Evaluation have been described, and results obtained by the submitted systems have been presented and briefly discussed. The evaluation involved six target languages: the four official languages spoken in Spain (Basque, Catalan, Galician and Spanish) plus Portuguese and English. A new database, KALAKA-2, was created for the evaluation, including clean and noisy speech in various languages, recorded from TV broadcasts and stored in single-channel 16-bit 16 kHz audio files.

In closed-set clean-speech verification tests on the evaluation subset of 30-second segments, the best primary system, employing state-of-the-art technology, yielded $C_{avg} = 0,0184$. This reveals a remarkable technology improvement with regard to the previous Albayzin 2008 LRE, where the best system yielded $C_{avg} = 0,0552$ on a similar task.

A new condition has been introduced in this evaluation, with the aim to evaluate performance degradation when dealing with noisy and/or overlapped speech. The increase in cost observed in noisy-speech tests (with regard to clean-speech tests) ranged from 30% to 50%, depending on the system and condition. This reveals that reasonably good performance can be attained even on noisy speech if enough and suitable data are available to train and calibrate systems.

Finally, a post-eval activity was organized which tried to investigate which subsystems combined better under a FoCal-based fusion paradigm. Starting from the log-likelihoods for the six target languages produced by the subsystems of various teams, we discovered that cross-site fusion may provide great performance improvements (the best 5-subsystem fusion yielding $C_{avg} = 0.0054$).

9. Acknowledgements

We thank the Organizing Committee of FALA 2010, specially Laura Docio, coordinator of the Albayzin 2010 Evaluations, for their help and support. We also thank all the participants, for their work and feedback, which were essential for the success of this evaluation.

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*, (Brno, Czech Republic), 28 June - 1 July 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*, (Brno, Czech Republic), 28 June - 1 July 2010.

[5] N. Brümmer and J. du Preez, “Application-Independent Evaluation of Speaker Detection,” *Computer, Speech and Language*, vol. 20, pp. 230–275, April-July 2006.

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

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

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

[9] 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.