# KALAKA-2: a TV Broadcast Speech Database for the Recognition of Iberian Languages in Clean and Noisy Environments

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

Grupo de Trabajo en Tecnologas Software (GTTS, http://gtts.ehu.es)
Departamento de Electricidad y Electrnica, University of the Basque Country UPV/EHU
Barrio Sarriena s/n, 48940 Leioa, Spain
luisjavier.rodriguez@ehu.es

**Abstract**

This paper presents the main features (design issues, recording setup, etc.) of KALAKA-2, a TV broadcast speech database specifically designed for the development and evaluation of language recognition systems in clean and noisy environments. KALAKA-2 was created to support the Albayzin 2010 Language Recognition Evaluation (LRE), organized by the Spanish Network on Speech Technologies from June to November 2010. The database features 6 target languages: Basque, Catalan, English, Galician, Portuguese and Spanish, and includes segments in other (Out-Of-Set) languages, which allow to perform open-set verification tests. The best performance attained in the Albayzin 2010 LRE is presented and briefly discussed. The performance of a state-of-the-art system in various tasks defined on the database is also presented. In both cases, results highlight the suitability of KALAKA-2 as a benchmark for the development and evaluation of language recognition technology.

**Keywords:** Spoken Language Recognition, Broadcast Speech, Iberian Languages

## 1. Introduction

A TV broadcast speech database, named KALAKA-2, was designed, collected and built with the purpose of supporting the Albayzin 2010 Language Recognition Evaluation (LRE) organized by the Spanish Thematic Network on Speech Technologies from May to November 2010 (Rodriguez-Fuentes et al., 2011). This was the second of a series of language recognition evaluations, which started with the Albayzin 2008 LRE (Rodriguez-Fuentes et al., 2010a). In fact, KALAKA-2 is a major update of KALAKA (Rodriguez-Fuentes et al., 2010b), the database created to support the Albayzin 2008 LRE, which consisted of wide-band TV broadcast speech recordings and featured 4 target languages: Basque, Catalan, Galician and Spanish. The update involves the addition of Portuguese and English as target languages, the addition of new Out-Of-Set languages and the use of noisy and/or overlapped speech for a new test condition. Besides recycling all the materials of KALAKA, new TV broadcast shows were recorded, including both planned and spontaneous speech in diverse environment conditions (excluding telephone-channel speech) and multiple speakers.

The database consists of three subsets: training, development and test, which allow to build and evaluate language recognition systems for six target languages: Basque, Catalan, English, Galician, Portuguese and Spanish. English has been included for its leading role as a world interchange language (note that English is also the official language in Gibraltar). The remaining languages have jointly evolved in the Iberian peninsula during centuries, most of them sharing a common origin in Latin, so the recognition task could be specially challenging, as performance results in the Albayzin 2008 LRE already revealed. The KALAKA-2 datasets further increase the difficulty of the task by extending the challenge to noisy/overlapped speech.

The development and test datasets include not only target languages but also Out-Of-Set (OOS) languages, so that open-set evaluations can be carried out. OOS languages (Arabic, French, German and Romanian) have been chosen based on the availability of TV channels, but also taking into account their similarity to target languages. In this regard, both French and Romanian are Romance languages, as four of the target languages; on the other hand, Arabic had a strong influence on Spanish (specially at the lexical level) and a moderate influence on Portuguese, Catalan and Galician; finally, German belongs to the same family of languages as English.

The train dataset amounts to more than 82 hours of speech, with more than 10 hours (in some cases, more than 12 hours) of clean speech per target language and more than 2 hours (in some cases, more than 3 hours) of noisy/overlapped speech per target language. The development and test datasets have the same size (around 21 hours of speech) but a different distribution of OOS languages. The whole database amounts to around 125 hours of speech (2.5 times the size of KALAKA) and is distributed in five DVD, after direct request to the authors. In the future, we plan to license the database through a distribution agency such as LDC or ELRA.

The rest of the paper is organized as follows. The design of the database and the recording setup are addressed in Sections 2 and 3, respectively. Section 4 describes how the recorded materials were processed and organized, including the recycling of KALAKA, the classification of recordings, the selection of speech materials, the extraction of fixed (nominal) length segments and the encoding of filenames. Section 5 summarizes the results obtained in the Albayzin 2010 LRE and presents a state-of-the-art language recognition system developed and evaluated on KALAKA-2. Finally, conclusions are given in Section 6.

## 2. Design Issues

KALAKA-2 was designed as an extension of KALAKA with the purpose of: (1) adding two new target languages (English and Portuguese), and (2) allowing the evaluation of systems under a new test condition for noisy and/or overlapped speech.

The materials produced for KALAKA were fully recycled for KALAKA-2, as follows: the train and development sets of KALAKA were posted to the train set of KALAKA-2, and the test set of KALAKA was posted to the development set of KALAKA-2.

New TV broadcasts were recorded, selected and classified, specially for the two newly added target languages (Portuguese and English), for the OOS languages and for the noisy condition in all languages, taking care of including as much diversity as possible regarding speakers, speech modalities, etc. Also, as an attempt to avoid unintentional overfitting to a given speaker or environment and to make the evaluation as robust as possible, disjoint subsets of TV shows were posted to the training, development and test datasets. It must be noted that the test set of KALAKA-2 was entirely built on new recordings, thus being completely independent of KALAKA.

No constraints were imposed to training segments regarding duration, whereas development and test segments of three nominal durations (30, 10 and 3 seconds) were produced, to measure language recognition performance as a function of the available amount of speech.

## 3. Recording Setup

To keep consistency, new recordings were done under the same setup (cable TV provider, devices, connectors, audio conversions, etc.) used for KALAKA. CD quality (16 bit / 44.1 kHz / stereo) recordings were done through a home connection to cable TV, by means of a Roland Edirol R-09 ultra-light digital audio recorder (http://www.roland.com/products/en/R-09). Audio signals were downsampled to 16 kHz, left and right channels being averaged into one single channel, by means of *SoX* (http://sox.sourceforge.net). This way, storage requirements were reduced in a factor of 5.51, while keeping an acceptable (wide-band) quality for speech processing applications. The resulting signals were stored in WAV files.

KALAKA-2 recordings were made at three different times: October-November 2008 (Arabic, Romanian and English), April-May 2010 (Arabic, German, French, Romanian, English and Portuguese) and August-September 2010 (Basque, Catalan, Galician and Spanish). Table 1 shows the TV channels and the recorded time for each language. The recorded time for all languages amounts to around 257 hours, which is more than two times the size of KALAKA-2.

Table 1: TV channels and recorded time (in minutes) for each language in KALAKA-2.

| ***Language*** | ***TV Channels*** | ***Recorded time*** |
|---|---|---|
| Basque | ETB1, ETBSat | 1996 |
| Catalan | TVCi | 1842 |
| English | DWTV, BBCWorld, CNN, Bloomberg | 2705 |
| Galician | TVG | 2240 |
| Portuguese | RTPi | 2608 |
| Spanish | TVE1, La 2, La Sexta, Cuatro, Tele5, Antena3, ETB2, TV Canaria Sat, AndalucíaTV, TeleMadrid, ExtremaduraTV, CNNPlus | 2090 |
| Arabic | Al Jazeera | 497 |
| French | TV5Monde Europe | 499 |
| German | DWTV | 431 |
| Romanian | PROTV | 552 |

## 4. Database Construction

### 4.1. Classification of Recordings

Side information was gathered for each recording: language, show type (news, documentary, talk show, debate, etc.), duration, environment conditions, rate of speech overlaps, etc. This information was used to distribute TV shows into the training, development and test datasets, keeping in mind that the three datasets should contain similar proportions of show types, and that all the recordings of a given TV show should be posted to the same dataset. To avoid tuning systems to reject specific OOS languages, different proportions of OOS languages were posted to the development and test datasets (see Table 3 for details).

### 4.2. Selection of Speech Fragments

This task was performed by listening to and looking at audio signals. The selected fragments may contain speech from two or more speakers, but only a single language. Two types of fragments were discarded for further use: (1) narrow-band (telephone-channel) speech fragments, and (2) fragments with background speech or speech overlaps using a different language than that used in the foreground (the nominal language).

The remaining materials were cut into clean and noisy speech fragments. Clean speech fragments were allowed to have any length greater than 30 seconds. Noisy speech fragments were forced to be between 30 and 35 seconds long. Since finding long fragments under a single background condition was not easy, the purity constraint was relaxed. In the case of clean speech, besides some portions of silence, relatively short portions of noisy and/or overlapped speech were also allowed. In the case of noisy speech, relatively short portions of clean speech were allowed.

This task was performed from May to June 2010 (training and development datasets) and in September 2010 (evaluation dataset) by members of the research team. After discussing and determining the selection criteria for the resulting sets of segments to be as homogeneous as possible, each member worked in a fully autonomous way and the resulting speech fragments (of indefinite duration) were pooled together.

No further processing was applied to speech fragments posted to the training dataset, which consists of two separate subsets, the first one containing more than 10 hours (in some cases, more than 12 hours) of clean speech per target language, and the second one containing more than 2 hours (in some cases, more than 3 hours) of noisy/overlapped speech for each target language. No training data are provided for OOS languages. The distribution of training data is shown in Table 2.

Table 2: Distribution of training segments per target language in KALAKA-2, 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 |

### 4.3. Automatic Extraction of Fixed-Length Segments

Clean-speech fragments posted to the development and test datasets were taken as source to automatically extract segments of fixed duration (30, 10 and 3 seconds), using a greedy algorithm which aimed to catch natural segments (i.e. speech segments surrounded by low-energy regions) with small (always positive) deviations from nominal durations (see (Rodriguez-Fuentes et al., 2010b) for details).

Noisy-speech fragments posted to the development and test datasets were stored as 30-second segments, since their duration ranged from 30 to 35 seconds. Then, a greedy algorithm similar to that used for clean speech was applied to automatically extract 10- and 3-second noisy-speech segments.

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 at least 150 speech segments per target language and nominal duration. Each segment contains speech from one or more speakers in one of the 6 target languages or in an OOS language.

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% 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% corresponding to clean speech and 33% to noisy speech). The distribution of segments per language is shown in Table 3.

### 4.4. Filename Encoding

The speech files of KALAKA-2 were initially stored with conventional names, according to the same protocol defined for KALAKA: a sequence LLCDDXXX.wav, where LL is the international language code (ca, eu, gl, en, es, pt, ar, de, fr, ro), C is the dataset identifier (t, d, e), DD is the duration code (00: undefined, 03: 3 seconds, 10: 10 seconds, 30: 30 seconds), and XXX is a three-digit number which identifies each segment under each category. Then, with the purpose of keeping language content undisclosed, new filenames consisting of a seemingly random string of 8 hexadecimal digits (followed by the .wav extension) were produced. To that end, the encoding/decoding algorithm defined for KALAKA was applied, based on a password, SHA-1 hashing of file contents and a XOR scheme (see (Rodriguez-Fuentes et al., 2010b) for details). The database is distributed with encoded filenames and a keyfile describing file contents.

Table 3: Distribution of segments per language (the same for each duration) in the development and evaluation datasets of KALAKA-2.

| | | **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 |

## 5. Database Evaluation

### 5.1. The Albayzin 2010 LRE

#### 5.1.1. Task and Conditions

Following NIST evaluations (see e.g. (Martin and Greenberg, 2010)), the task defined for the Albayzin 2010 LRE consisted on deciding by computational means whether or not a target language was spoken in a test utterance. 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 and the set of *non-target* languages (those that may appear in the segment instead of the target language). For each trial, the system is required to output: (1) a hard decision about whether the target language is spoken in the segment; and (2) a score, such that the higher the score the greater the confidence that the segment contains the target language.

The Albayzin 2010 LRE involved independent language verification trials for a set of 6 target languages: Basque, Catalan, English, Galician, Portuguese and Spanish. Three segment durations (30, 10 and 3 seconds), two evaluation modes (closed-set vs. open-set) and two environment conditions (clean vs. noisy speech) were defined, leading to 12 evaluation tracks. Remind that closed-set evaluation assumes that only target languages can be spoken in test utterances, whereas open-set evaluation relaxes that assumption by allowing any language (i.e. also OOS languages) to be spoken in test utterances.

#### 5.1.2. Performance Measures

System performance was primarily measured in terms of the well-known *average cost* $C_{avg}$ (Martin and Le, 2008), which is a combination of the miss and false alarm error rates ($P_{miss}$ and $P_{fa}$) obtained by the system at a given operation point (threshold), pooled across target languages. The $C_{avg}$ measure depends on language priors ($P_{target}$, $P_{non-target}$ and $P_{OOS}$) and application dependent costs ($C_{miss}$ and $C_{fa}$). Details about the applied values can be found in (Rodriguez-Fuentes et al., 2011). Detection Error Tradeoff (DET) curves (Martin et al., 1997) were also computed (using NIST software[1]) to visualize and compare the global performance of systems, including marks for the actual and minimum $C_{avg}$ operation points.

#### 5.1.3. Results

The best $C_{avg}$ performance attained at each track in the Albayzin 2010 LRE is shown in Table 4. Test conditions (CC, OC, CN and ON) are coded so that the first letter refers to closed-set (C) or open-set (O) evaluation and the second letter refers to clean-speech (C) or noisy speech (N) test segments. DET curves corresponding to the best primary systems in the four test conditions for the subset of 30-second segments are shown in Figure 1.

Table 4: Best performance ($C_{avg}$) attained at each track in the Albayzin 2010 LRE.

| | **30s** | **10s** | **3s** |
|---|---|---|---|
| **CC** | 0.0181 | 0.0359 | 0.0844 |
| **OC** | 0.0296 | 0.0445 | 0.1029 |
| **CN** | 0.0253 | 0.0636 | 0.1217 |
| **ON** | 0.0475 | 0.0936 | 0.1551 |

For the easiest condition (CC-30s), the best system yielded $C_{avg} = 0.0181$, which is comparable to the performance attained by state-of-the-art language recognition systems on NIST LRE datasets. Note also that this performance is much better than that attained for a similar task in the Albayzin 2008 LRE ($C_{avg} = 0.0552$, see (Rodriguez-Fuentes et al., 2010a)). Such an important difference may be due in part to technology improvements from 2008 to 2010, but also to the availability of more training data for target languages, and to the introduction of English and Portuguese as target languages, which makes the task easier on average, since most systems manage to discriminate them from the other target languages, as the low false alarm probabilities of English and Portuguese (compared to those of the other target languages) suggest (see (Rodriguez-Fuentes et al., 2011) for details).

Regarding the dependence on the nominal duration of test segments, the $C_{avg}$ obtained on 10-second segments is around twice that obtained on 30-second segments, and the same trend is observed for 3-second segments with regard to 10-second segments. This is consistent with previous results for other evaluations.

[1] http://www.itl.nist.gov/iad/mig/tools/DETware_v2.1.targz.htm

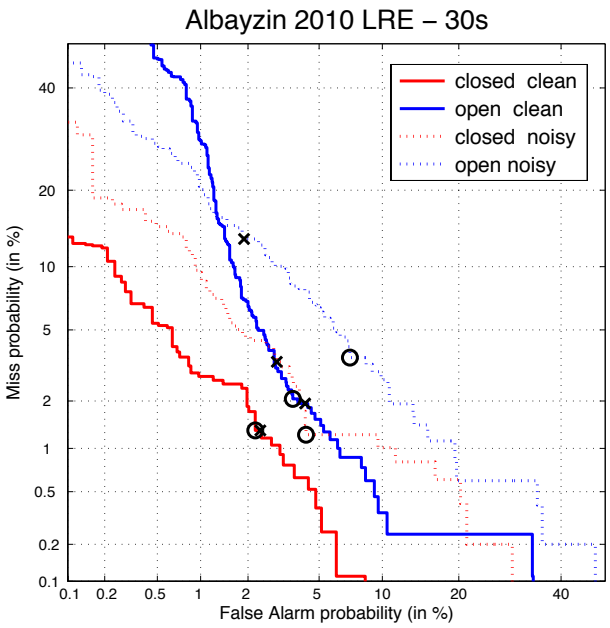


Figure 1: DET curves for the best primary systems submitted to the Albayzin 2010 LRE in all tracks (30-second segments). Operation points corresponding to actual ($\times$) and minimum ($\circ$) $C_{avg}$ are marked on the curves.

Performance degraded when moving from the closed-set to the open-set evaluation, due to a higher number of false alarms related to OOS languages. As shown in Table 4, the best system in the OC-30s condition yielded $C_{avg} = 0.0296$, meaning around 63.5% increase in cost with regard to the best system in the CC-30s condition. The increase in cost was less noticeable for 10- and 3-second segments.

Finally, as illustrated in Figure 1, dealing with noisy speech led to the highest (but not catastrophic) performance degradations. The increase in cost when moving from clean to noisy speech ranged from 40% to 80%, being relatively smaller for short segments.

### 5.2. State-of-the-Art System Development and Evaluation Based on KALAKA-2

The development of KALAKA and KALAKA-2 was motivated by the lack of multilingual wide-band speech databases (specially including Iberian languages) for the development and evaluation of language recognition systems in applications not suitably covered by NIST LRE datasets, such as multilingual spoken document retrieval on wide-band broadcast speech resources. In particular, KALAKA-2 allows for the development and evaluation of language recognition systems for wide-band broadcast speech in both quiet and noisy background environments, which is a relevant step towards more realistic conditions with regard to KALAKA, where only clean speech was considered. A state-of-the-art language recognition system has been developed and evaluated based on the datasets of KALAKA-2. As will be shown below, the relatively low performance attained by this system in some tasks, specially those dealing with noisy speech, highlights the suitability of KALAKA-2 as a benchmark for the development and evaluation of new approaches.

The system presented in this section resembles almost exactly that submitted by our research group to the 2011 NIST LRE (Penagarikano et al., 2011), which fused two acoustic and three phonotactic subsystems and demonstrated very competitive performance (fourth best primary system in the NIST 2011 LRE core condition): $C_{avg} = 0.0892$ for the 24 worst performing language pairs and $C_{avg} = 0.0169$ when the average was computed over all the pairs.

All the models (except for the phone decoders applied in the phonotactic subsystems) have been trained *exclusively* on the training dataset of KALAKA-2. Two sets of models have been estimated: the first one, used for the clean-speech tracks, was trained on clean speech; the second one, used for the noisy-speech tracks, was trained on the whole training dataset, including both clean and noisy speech. Backend and fusion parameters have been estimated using the development dataset of KALAKA-2, under two configurations: (1) closed-set, for which only segments containing target languages were used; and (2) open-set, for which all the segments were used. In the following paragraphs, we provide a brief description of the component subsystems and the backend and fusion approaches.

#### 5.2.1. Acoustic Subsystems

Acoustic features consist of the concatenation of 7 Mel-Frequency Cepstral Coefficients and the Shifted Delta Cepstrum coefficients (Torres-Carrasquillo et al., 2002) under a 7-2-3-7 configuration, a gender independent 1024-mixture Gaussian Mixture Model (GMM) is used as Universal Background Model (UBM) and zero-order and centered and normalized first-order Baum-Welch statistics were computed for each input utterance.

The first acoustic subsystem follows the Linearized Eigenchannel GMM (LE-GMM) approach (also known as *Dot-Scoring*), which makes use of a linearized, channel compensated and normalized approximation of the likelihood ratio in the GMM-UBM approach to score test segments against target models (Strasheim and Brümmer, 2008) (Brümmer et al., September 2009). The second acoustic subsystem follows the Total Variability generative iVector approach, as described in (Martínez et al., 2011).

#### 5.2.2. Phonotactic Subsystems

Three phonotactic sub-systems were developed under a phone-lattice Support Vector Machine (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, Hungarian and Russian (Schwarz, 2008), were applied to perform phone tokenization. Regarding channel compensation, noise reduction, etc. the three sub-systems 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 (Young et al., 2006) along with the BUT recipe, on which expected counts of phone n-grams were computed using the *lattice-tool* of SRILM (Stolcke, 2002). Finally, a 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 (Richardson and Campbell, 2008). L2-regularized L1-loss support vector classification was applied, by means of LIBLINEAR (Fan et al., 2008), whose source code was slightly modified to get regression values.

#### 5.2.3. Backend and Fusion

When processing an input utterance, each subsystem provides a score for each target language. In this work, a generative Gaussian backend is estimated for each subsystem (based on the scores obtained for the development set) and applied to get log-likelihoods for target languages (under the open-set configuration, an additional log-likelihood for OOS languages is also computed). Log-likelihods are then fused according to a discriminative linear model which minimizes the so called $C_{LLR}$ function on the development set, by means of logistic regression under a multiclass paradigm (Brümmer and van Leeuwen, 2006). After the fusion model is applied, well-calibrated scores are obtained, for which a minimum expected cost Bayes decision threshold is applied, according to application-dependent language priors and costs. Backend and fusion parameters have been separately estimated for each nominal duration on the development set, and then applied to the corresponding segments in the evaluation set. The *FoCal* toolkit has been used to estimate and apply the backend and fusion models (FoCal, 2008).

#### 5.2.4. Results

Table 5 shows the $C_{avg}$ performance attained by the state-of-the-art language recognition system described above in all the tracks of the Albayzin 2010 LRE. Figure 2 shows the corresponding DET curves for the tracks involving 30-second segments. Remind that two different systems have been developed, the first one built and evaluated on clean speech (CC and OC tracks, solid DET curves) and the second one built and evaluated on a mix of clean and noisy speech (CN and ON tracks, dotted DET curves).

Table 5: $C_{avg}$ performance attained at each track of the Albayzin 2010 LRE by a sate-of-the-art language recognition system developed on KALAKA-2.

| | **30s** | **10s** | **3s** |
|---|---|---|---|
| **CC** | 0.0063 | 0.0263 | 0.0888 |
| **OC** | 0.0171 | 0.0437 | 0.1094 |
| **CN** | 0.0177 | 0.0599 | 0.1476 |
| **ON** | 0.0390 | 0.0867 | 0.1740 |

System performance was consistently better than that attained in the Albayzin 2010 LRE (see Table 4 and Figure 1) in all conditions, except for 3-second segments. This could be due to a lack of robustness in the estimation of backend and fusion parameters when using extremely short segments (specially when dealing with noisy speech).

The average cost was extremely low for the CC-30s track ($C_{avg} = 0.0063$), meaning that, quite probably, this database would not support technology improvements for that condition (closed-set evaluation, clean speech, 30-second segments), since differences in performance would

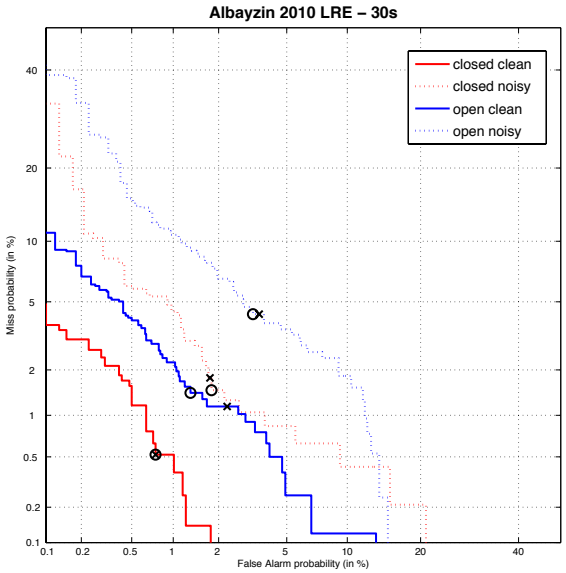


Figure 2: DET curves for a state-of-the-art language recognition system developed on KALAKA-2, in the four tracks of the Albayzin 2010 LRE involving 30-second segments. Operation points corresponding to actual ($\times$) and minimum ($\circ$) $C_{avg}$ are marked on the curves.

be too small and not significant. In the remaining tracks, the average costs were high enough to allow statistically significant performance improvements. In particular, performance for the noisy-speech condition was far worse than that found for the clean-speech condition. It is worth noting that moving from clean to noisy speech produced higher degradation than moving from closed-set to open-set evaluation. In other words, keeping the test closed to target languages but using noisy speech seems to be more challenging than expanding the test with OOS languages but using clean speech. This conclusion is graphically supported by DET curves in Figure 2: performance in the closed-set noisy-speech condition (dotted red curve) was consistenly worse than that in the open-set clean-speech condition (solid blue curve). Finally, performance degraded as less speech (i.e. less information) was available to make decisions (short segments), following the same pattern observed above for the Albayzin 2010 LRE.

## 6. Conclusions

In this paper, we address the design, data collection, construction and evaluation of KALAKA-2, a database containing wide-band (16 kHz) clean and noisy speech signals recorded from TV broadcasts. KALAKA-2 was created and used specifically for the Albayzin 2010 Language Recognition Evaluation, but can be freely requested to authors. It amounts to around 125 hours of speech and consists of three datasets: training, development and evaluation, which allow to build and evaluate language recognition systems for six target languages: Basque, Catalan, English, Galician, Portuguese and Spanish (Iberian languages + English). The database also includes speech signals in other languages, to allow open-set verification trials.

The best performance attained in all the tracks of the Albayzin 2010 LRE has been presented as a means of evaluating the database. The best performance in the core condition (closed-set, clean-speech, 30-second segments) was much better than the best performance for the same condition in the Albayzin 2008 LRE, due to several reasons, including the availability of more training data and the introduction of two target languages (English and Portuguese) feturing extremely low confusion rates with the other target languages. Significant degradation was observed as less speech was available: roughly, the cost doubled when moving from 30-second to 10-second segments, and doubled again when moving from 10-second to 3-second segments. Moving from closed-set to open-set tests (i.e. allowing for test segments with OOS languages) also led to degraded performance, but the highest degradation was found when dealing with noisy speech.

A second evaluation has been carried out, using the datasets of KALAKA-2 to build and evaluate a state-of-the-art language recognition system. Performance using this system was better than that attained in the Albayzin 2010 LRE in all conditions except for 3-second segments (probably due to unreliable estimations of backend and fusion models), but the same trends are observed (e.g. the highest degradation is produced by noisy speech) and the same conclusions can be drawn. In brief, the average costs were high enough to allow statistically significant performance improvements in all the tracks except for the easiest one (closed-set, clean-speech, 30-second segments). The most challenging condition involves extremely short (3-second) noisy speech segments in open-set tests, for which the best performance reported in this paper is $C_{avg} = 0.1551$. Therefore, KALAKA-2 provides plenty of margin to support further language recognition technology developments for wide-band broadcast speech in quiet and noisy background environments.

## 7. Acknowledgements

This work has been supported by the University of the Basque Country UPV/EHU, 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 the Department of Education, Universities and Research of the Government of the Basque Country, under a 4-year research fellowship.

## 8. References

N. Brümmer and D.A. van Leeuwen. 2006. On calibration of language recognition scores. In *Proceedings of Odyssey - The Speaker and Language Recognition Workshop*, pages 1–8.

N. Brümmer, A. Strasheim, V. Hubeika, P. Matejka, L. Burget, and O. Glembek. September 2009. Discriminative Acoustic Language Recognition via Channel-Compensated GMM Statistics. In *Proceedings of Interspeech*, pages 2187–2190, Brighton, UK.

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

FoCal, 2008. *Toolkit for Evaluation, Fusion and Calibration of statistical pattern recognizers*. http://sites.google.com/site/nikobrummer/focal.

Alvin Martin and Craig Greenberg. 2010. The 2009 NIST Language Recognition Evaluation. In *Odyssey 2010 - The Speaker and Language Recognition Workshop*, pages 165–171, Brno, Czech Republic.

Alvin F. Martin and Audrey N. Le. 2008. NIST 2007 Language Recognition Evaluation. In *Proceedings of Odyssey 2008 - The Speaker and Language Recognition Workshop, paper 016*.

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

David Martínez, Oldrich Plchot, Lukás Burget, Ondrej Glembek, and Pavel Matejka. 2011. Language Recognition in iVectors Space. In *Proceedings of the Interspeech*, pages 861–864, Firenze, Italy.

Mikel Penagarikano, Amparo Varona, Luis J. Rodriguez-Fuentes, Mireia Diez, and German Bordel. 2011. 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).

F. Richardson and W. Campbell. 2008. Language recognition with discriminative keyword selection. In *Proceedings of ICASSP*, pages 4145–4148.

L. J. Rodriguez-Fuentes, M. Penagarikano, G. Bordel, and A. Varona. 2010a. The Albayzin 2008 Language Recognition Evaluation. In *Proceedings of Odyssey 2010: The Speaker and Language Recognition Workshop*, pages 172–179, Brno, Czech Republic.

L. J. Rodriguez-Fuentes, M. Penagarikano, G. Bordel, A. Varona, and M. Diez. 2010b. 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)*, pages 1678–1685, Valleta, Malta.

Luis Javier Rodriguez-Fuentes, Mikel Penagarikano, Amparo Varona, Mireia Diez, and German Bordel. 2011. The Albayzin 2010 Language Recognition Evaluation. In *Proceedings of Interspeech*, pages 1529–1532, Firenze, Italia.

Petr Schwarz. 2008. *Phoneme recognition based on long temporal context*. Ph.D. thesis, Faculty of Information Technology, Brno University of Technology, Brno, Czech Republic.

Andreas Stolcke. 2002. SRILM - an extensible language modeling toolkit. In *Proceedings of Interspeech*, pages 257–286.

Albert Strasheim and Niko Brümmer. 2008. SUNSDV system description: NIST SRE 2008. In *NIST 2008 Speaker Recognition Evaluation Workshop Booklet*.

P. A. Torres-Carrasquillo, E. Singer, M. A. Kohler, R. J. Greene, D. A. Reynolds, and J. R. Deller. 2002. Approaches to language identification using Gaussian mixture models and Shifted Delta Cepstral features. In *Proceedings of ICSLP*, pages 89–92.

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. 2006. *The HTK Book (for HTK Versin 3.4)*. Entropic, Ltd., Cambridge, UK.