KALAKA: a TV Broadcast Speech Database for the Evaluation of Language Recognition Systems

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

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

Abstract

A speech database, named KALAKA, was created to support the Albayzin 2008 Evaluation of Language Recognition Systems, organized by the Spanish Network on Speech Technologies from May to November 2008. This evaluation, designed according to the criteria and methodology applied in the NIST Language Recognition Evaluations, involved four target languages: Basque, Catalan, Galician and Spanish (official languages in Spain), and included speech signals in other (*unknown*) languages to allow open-set verification trials. In this paper, the process of designing, collecting data and building the train, development and evaluation datasets of KALAKA is described. Results attained in the Albayzin 2008 LRE are presented as a means of evaluating the database. The performance of a state-of-the-art language recognition system on a closed-set evaluation task is also presented for reference. Future work includes extending KALAKA by adding Portuguese and English as target languages and renewing the set of *unknown* languages needed to carry out open-set evaluations.

1. Introduction

A speech database, named KALAKA, was designed, collected and built with the aim to support the Albayzin 2008 Evaluation of Language Recognition Systems (http://jth2008.ehu.es/en/albayzin.html) organized by the Spanish Network on Speech Technologies from May to November 2008. Hereafter, we will refer to this evaluation as Albayzin 2008 LRE. This evaluation was designed according to the criteria and methodology applied in NIST Language Recognition Evaluations (http://www.itl.nist.gov/iad/mig/tests/lre/). In particular, the Evaluation Plan for the 2007 NIST LRE was taken as reference (Martin and Le, 2008). There is, however, a significant difference: NIST LRE materials were extracted from spontaneous conversations through telephone (narrow-band) channels involving two speakers, whereas those of KALAKA were extracted from (wide-band) TV shows, including both planned and spontaneous speech in diverse environment conditions involving a varying number of speakers. Various types of TV shows were recorded, with prevalence of broadcast news, talk shows and debates. The database was named after the talk show *Kalaka* (which could be translated to English as *offensive* or *annoying* talk), broadcast by the Basque channel ETB1.

Training data provided in KALAKA allows to build language recognition systems with four target languages: Basque, Catalan, Galician and Spanish. These are all official languages in Spain, though only Spanish is spoken in the whole territory, whereas the other three are spoken (with different usage levels) in specific regions. In any case, remarkable differences have been observed between planned speech produced by professional speakers in broadcast news and spontaneous speech produced by people in interviews. In particular, Spanish features several regional dialects, some of them reflecting features (pronunciation, intonation, words, syntactic forms, etc.) inherited from yet extinct Iberian languages, and others reflecting features imported from Basque, Catalan or Galician, which at the same time have historically received a strong influence from Spanish. So, the task of recognizing these four target languages could be more challenging than expected. In fact, one of the goals of the evaluation was to measure the accuracy that state-of-the-art language recognition systems could attain for this task.

Development and evaluation data include utterances not only in target languages but also in other languages (unknown from the point of view of the application), so that open-set evaluations can be carried out. Those *unknown* languages (English, French, Portuguese and German) have been chosen attending to the availability of TV channels, with a higher presence of French and Portuguese, which may increase task difficulty, since these two languages are assumed to share some features with Catalan and Galician, respectively.

The training set contains around 9 hours of speech per target language, which amounts to around 36 hours of training data. The development and evaluation sets contains around 7.7 hours of speech each one, with the same distribution: more than 90 minutes of speech per target language and more than 90 minutes of speech in other (*unknown*) languages. The whole database amounts to around 50 hours of speech and is distributed (after direct request to the authors) in three DVD.

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 classification of recordings, selection of speech materials, extraction of fixed (nominal) length segments and encoding of filenames. Section 5 summarizes results attained in Albayzin 2008 LRE and presents a state-of-the-

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).

art language recognition system developed and evaluated on KALAKA. Finally, conclusions and future work are outlined in Section 6.

2. Design issues

We started from the following two considerations:

1. In order to avoid undesired variabilities, recording conditions (channel and audio processing) should be the same for all the languages, meaning that a single recording setup (devices, connectors, audio conversions, etc.) should be defined.

2. With regard to other sources of variability (environment, speaker, etc.), each language should contain as much diversity as possible, thus minimizing (or at least, averaging) the effect of such factors in the evaluation.

We chose cable TV, in particular, that provided by *Euskaltel* (http://www.euskaltel.com) in the Basque Country, because it gives easy access to audio in different languages: Spanish from several generalist and regional channels; Basque, Catalan and Galician from the corresponding regional channels in the Basque Country, Catalonia, Valencia (a region in eastern Spain where a variation of catalan is spoken) and Galicia; and English, German, French, Portuguese, etc. from international channels.

In order to foster data independence and make the evaluation as robust as possible, disjoint subsets of TV shows were assigned to train, development and evaluation. This way, each subset still contains a diverse choice of speakers, but the probability of finding the same speaker in two subsets (and therefore, the risk of modeling, optimizing for or matching the speaker and not the language) is very low.

Training materials had no constraints regarding duration, whereas development and evaluation data followed the guidelines of NIST LRE, by defining three evaluation subsets according to the nominal duration of speech segments: 30, 10 and 3 seconds, respectively. This allowed to measure the recognition performance as a function of the available amount of speech. Obviously, it was expected that the shorter the speech segment, the lower the accuracy in recognizing the spoken language.

3. Recording setup

Recordings were done through a home connection to cable TV, by means of a digital audio recorder. A Roland Edirol R-09 ultra-light audio recorder was chosen, with the following features (see http://www.roland.com/products/en/R-09 for further details): up to 24 bit / 48 kHz linear PCM and up to 320 kbps MP3 recording, SD card direct storage, built-in stereo microphone, mic and line audio inputs and high speed file transfer through USB 2.0. CD quality (16 bit / 44.1 kHz / stereo) recordings were done by connecting the audio output of the cable TV decoder to the line input of the R-09. The resulting files were stored in WAV format for further processing.

Audio signals were downsampled to 16 kHz, left and right channels being averaged into one single channel, by means of *SoX* (Sound eXchange: 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.

Filenames were given according to the following pattern:

<TVshow>[TVchannel]_<date>_<language>.wav

Date consisted of a sequence of numbers of the form yyyymmdd (year, month and day), left padded with zeros if necessary (for instance, 20080503 represents May 3rd, 2008). International codes were used to denote language: es (Spanish), ca (Catalan), eu (Basque), gl (Galician), en (English), de (German), fr (French) and pt (Portuguese). The TV channel was added only when, for a given language, TV shows from different channels were recorded, or when the name of the show was not descriptive enough (as in the case of news). Here we present some examples:

NoticiasTVCanaria_20080319_es.wav
ElTiempoAndaluciaTV_20080421_es.wav
ElTiempoTeleMadrid_20080423_es.wav
EuromaxxDWTV_20080331_en.wav
HardTalkBBCWorld_20080317_en.wav

Most recordings were done from April 18th to May 2nd 2008. After that time, in order to complete the evaluation dataset, a few additional recordings were done from August 15th to September 13th 2008. As explained in next section, recordings were filtered and many segments discarded because of high noise levels, speech overlaps, etc. So, the size of recorded materials was around 3 times the size of speech segments finally used in KALAKA. Table 1 shows the TV channels used for the recordings and the recorded time for each language. Recorded time for all languages amounts to 138 hours.

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

Language	***TV Channels***	***Recorded time***
Spanish	TVE1, La 2, La Sexta, Cuatro, Tele5, Antena3, ETB2, TV Canaria Sat, AndalucíaTV, TeleMadrid	1818
Catalan	TVCi	1777
Basque	ETB1	1905
Galician	TVG	1731
German	DWTV	275
French	TV5Monde Europe	320
English	DWTV, BBCWorld	257
Portuguese	RTPi	218

Table 2: Recorded time, absolute (minutes) and relative (%), of the six types of TV shows for the target languages.

	Spanish	***Catalan***	***Basque***	***Galician***
Debates and interviews	495 - 27.23	499 - 28.08	631 - 33.12	515 - 29.75
Talk-shows	500 - 27.50	428 - 24.09	498 - 26.14	642 - 37.09
News	353 - 19.42	336 - 18.91	341 - 17.90	405 - 23.40
Sports	126 - 6.93	120 - 6.75	120 - 6.30	17 - 0.98
Entertaining	230 - 12.65	249 - 14.01	153 - 8.03	83 - 4.79
Documentaries	114 - 6.27	145 - 8.16	162 - 8.50	69 - 3.99
Total	1818 - 100.00	1777 - 100.00	1905 - 100.00	1731 - 100.00

4. Creating the database

Audio files were processed in four steps: (1) classification (according to contents); (2) selection of speech segments; (3) automatic extraction of 30-, 10- and 3-second speech segments (needed for the development and evaluation datasets); and (4) generation of encoded filenames (hiding language information). The last step was necessary because information about language (present in conventional filenames) had to be hidden to sites participating in the Albayzin 2008 LRE.

4.1. Classification of recordings

The recording process included taking notes about each TV show: type (news, talk show, debate, etc.), duration, environment conditions, rate of speech overlaps, etc. This information was used to distribute TV shows into train, development and evaluation datasets, keeping in mind the diversity and independence conditions: (1) the three datasets should contain similar proportions of show types; and (2) all the recordings of a given TV show should be posted to the same dataset.

TV shows were classified in six categories: (1) debates and interviews; (2) talk-shows; (3) news; (4) sports; (5) entertaining (contests, reality shows, etc.); and (6) documentaries. Recorded time (absolute and relative) of the six types of TV shows for the target languages is presented in Table 2. To allow a good characterization of target languages, debates and interviews (which feature a high rate of clean non-overlapped speech from many speakers) were most of them posted to the train dataset.

As noted above, speech data were recorded not only for the four target languages, but also for other languages (*unknown* from the point of view of the application), with the only aim to carry out open-set evaluations, not to train models for them. Obviously, training models on other languages would improve acoustic coverage and help verification, but we cannot assume that such data will be available in practice. On the other hand, training (and using) models for languages actually appearing in the evaluation dataset would violate the assumption that they were unknown. And finally, training (and using) models for languages not appearing in the evaluation dataset may help but may also damage verification. So, TV shows in English, German, French and Portuguese were posted only to the development and evaluation datasets. Their relative distribution was designed according to the percentages given in Table 3.

Table 3: Planned distribution of data (%) for *unknown* languages.

	Dev	***Eval***	***Total***
German	0.00	16.67	16.67
French	29.17	4.16	33.33
English	16.67	0.00	16.67
Portuguese	4.16	29.17	33.33
Total	50.00	50.00	100.00

Proportions of *unknown* languages were deliberately different for development and evaluation, to avoid tuning systems to reject specific languages. On the other hand, in order to make things even more difficult, due to the relative proximity of French to Catalan, and Portuguese to Galician, the proportion of these languages was twice the proportion of English and German.

4.2. Selection of speech segments

Fragments containing noisy speech, music, speech overlaps, etc. were discarded. Only speech segments with a low level of background noise were validated for KALAKA. This task was performed by means of *Wavesurfer* (Sjolander and Beskow, 2000) (http://www.speech.kth.se/wavesurfer/), which allows listening to and looking at audio signals, selecting segments and storing them. As a result, speech segments of indefinite length (each segment spoken in a single language) were extracted from recorded materials and stored in WAV files.

The main part of this task was performed by three members of the research team, from May to July 2008. Additional materials were also processed by one of the researchers in September 2008. After discussing and determining the selection criteria (for the resulting segments to be as homogeneous as possible), each researcher selected materials in a fully autonomous way, and the resulting files were pooled for further processing. For intermediate storage, filenames were generated by adding a three-digit number to the name of the source file. This way, the first speech segment extracted from the file *LaNitAlDia_20080317_ca.wav* was named *LaNitAlDia_20080317_ca_001.wav*, the second speech segment was named *LaNitAlDia_20080317_ca_002.wav*, etc.

No further processing was applied to speech segments posted to the train dataset. The number of training segments per target language, as well as their total approximate du-

Table 4: Number and total duration of training segments for the target languages in KALAKA.

	Spanish	***Catalan***	***Basque***	***Galician***	***All***
Number of segments	282	278	342	401	1303
Total duration (minutes)	529	538	531	532	2130

ration (in minutes), are shown in Table 4. Speech segments posted to development and evaluation datasets were taken as source to extract spech segments of fixed (nominal) durations of 30, 10 and 3 seconds, according to the criteria given in Section 4.3.

The resulting WAV files were stored in the corresponding folders (train, devel and eval), with conventional names consisting of the sequence LLCDDXXX.wav, where LL is the international language code (es, ca, eu, gl, de, fr, en, pt), 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. This way, *cat00023.wav* represents the 23th speech segment of undefined duration in Catalan in the train dataset; *ptd30011.wav* represents the 11th 30-second speech segment in Portuguese in the development dataset; and *eue10143.wav* represents the 143rd 10-second speech segment in Basque in the evaluation dataset.

4.3. Automatic extraction of 30-, 10- and 3-second segments

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

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

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

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

A single-pass greedy approach was applied, which looked for 30-second segments fulfilling the three conditions given above. First, the whole file was taken as input and non-overlaped 30-second speech segments were searched and stored for further processing. Next, each 30-second segment found in the above search was validated, as follows: (1) non-overlapped 10-second segments inside the 30-second segment were searched and stored; (2) each 10-second speech segment was processed the same way, by looking for non-overlapped 3-second speech segments; and (3) as soon as a 3-second speech segment was found inside a 10-second segment, the validation procedure ended, all the intermediate files were deleted and time marks of the 30-, 10- and 3-second segments were stored.

The search for d-second speech segments works as follows: (1) the input signal is processed in overlapped frames of 100 milliseconds, with a frame step (time resolution) of 10 milliseconds; (2) frame energies are computed and stored; (3) low-energy fragments are implicitly defined by two heuristically fixed energy thresholds (for start and end) and are required to last more than 100 milliseconds; and (4) a greedy search is applied which extracts non-overlapped segments lasting from d to $d + k$ seconds, forced to begin and end at low-energy fragments, k being a tolerance parameter.

On average, this algorithm retrieved 65% of the input speech. Note that two additional files of 10 and 3 seconds were produced for each 30-second segment located by the algorithm: each 3-second segment was part of a 10-second segment, which in turn was part of a 30-second segment. Since 30-, 10- and 3-second evaluation subsets were built on the same materials, performance differences measured on these subsets should be attributed, almost exclusively, to the varying amount of available speech.

The development dataset consists of 1800 speech segments, distributed in three subsets, each containing 600 segments of 30, 10 and 3 seconds, respectively. Each subset consists of 120 segments per target language and 120 additional segments spoken in *unknown* languages, following the distribution shown in Table 3: 70 segments spoken in French, 10 in Portuguese and 40 in English. The evaluation dataset has the same structure, except for the distribution of *unknown* languages, which, according to Table 3, consists of 10 segments spoken in French, 70 in Portuguese and 40 in German.

4.4. Encoding filenames

For the sake of completeness, we briefly address here the algorithm applied to encode conventional filenames. The algorithm was designed according to the following conditions:

1. Encoding must be reversible: the conventional filename must be recoverable from the encoded filename.

2. The encoded filename will be generated from three data: the conventional filename, file contents and a password.

3. The conventional filename will be recovered from three data: the encoded filename, file contents and the same password used to produce the encoded filename.

4. The conventional filename must match the structure described in section 4.2.

5. The encoded filename will consist of a seemingly random string of 8 hexadecimal digits (followed by the .wav extension).

All the speech files of KALAKA (around 5000), also including those of the train dataset, were given an encoded filename. Encoding consisted on applying an exclusive or (XOR) to a 4-byte string derived from the conventional filename (through a bidirectional translation table), the password and a SHA-1 hash computed on file contents. The same algorithm was applied to recover the conventional filename, taking advantage of the reversibility property of the XOR (if c = a XOR b, then b = a XOR c) and inverting the correspondence between 4-byte strings and conventional filenames. Using a SHA-1 hash implies that there is no one-to-one correspondence between conventional and encoded filenames, i.e. two conventional filenames may produce the same encoded filename. However, the probability of such an event is very low in a set of 5000 filenames, since there are $(2^4)^8 = 2^{32}$ potential encodings. In any case, different passwords could be applied until no collision was found. In practice, no collision was detected when applying this algorithm with various passwords.

5. Using the database

5.1. The Albayzin 2008 LRE

Following NIST evaluations (Martin and Le, 2008), the Albayzin 2008 LRE involved independent language verification trials for a set of 4 target languages: Basque, Catalan, Galician and Spanish. Given a test utterance S and a target language L, the task consisted on deciding whether or not L was actually spoken in S. Besides the decision, a score should be provided for each trial, the higher the score the greater the confidence that the target language was spoken in the segment.

Three evaluation subsets of 30-, 10- and 3-second speech segments, two development conditions (restricted vs. free) and two evaluation modes (closed-set vs. open-set) were considered. Restricted development implied that only those materials provided in KALAKA could be used to build the system, and external materials could be used neither directly nor indirectly. For instance, acoustic models trained on an external acoustic database were not allowed. Free development allowed using any kind and amount of materials. Closed-set evaluation assumed that only target languages could be spoken in test utterances. Open-set evaluation relaxed that assumption by allowing any (known or unknown) language to be spoken in test utterances.

System performance was measured by presenting a set of trials, each trial consisting of a pair (test utterance, target language), and then computing the C_{avg} cost function (Martin and Le, 2008), which depends on the miss and false alarm error rates, language priors (P_{target}, $P_{non-target}$ and $P_{Out-Of-Set}$) and application dependent costs (C_{miss} and C_{fa}). The C_{avg} function was computed separately for the three evaluation subsets of 30-, 10- and 3-second segments, for the restricted and free development conditions and for the closed-set and open-set evaluation modes. For those sites indicating that their scores could be interpreted as log-likelihood ratios, an alternative peformance measure was also computed, the so called C_{LLR} (Brümmer and du Preez, 2006), which does not depend on application costs. Finally, Detection Error Tradeoff (DET) curves (Martin et al., 1997) were computed (using NIST software[1]) to visualize and compare global performance of systems, including both actual and minimum C_{avg} operation points.

The Albayzin 2008 LRE presented an award for the system yielding the least C_{avg} in the restricted-condition closed-set evaluation on the subset of 30-second speech segments. DET curves of four primary (continuous line) and one contrastive (dotted line) systems participating in that competition are shown in Figure 1. The C_{avg} attained by the most competitive systems (corresponding to two undisclosed participants) are shown in Table 5. Note that, though state-of-the-art technology was employed, results reveal that the proposed task was more challenging than expected, the best systems yielding a C_{avg} of around 0.05 (roughly corresponding to 5% EER) in the free-development closed-set evaluation on 30-second segments, and around 0.09 (roughly corresponding to 9% EER) in the free-development open-set evaluation on 30-second segments.

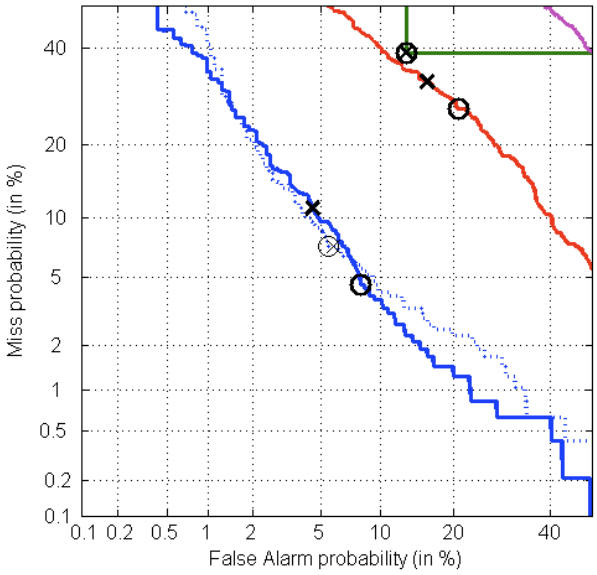

Figure 1: Pooled DET curves of systems participating in the restricted-development closed-set evaluation on 30-second speech segments. Operation points corresponding to actual (X) and minimum (O) C_{avg} are marked on the curves.

5.2. Developing language recognition technology

NIST evaluations have promoted the creation and use of large multilingual narrow-band (telephone channel) speech databases, supporting the development of language recognition technology as a preprocessing step for the automatic transcription of telephone conversations in some interesting languages. Few multilingual wide-band speech databases are available, and none of them includes the official languages in Spain. Creating KALAKA was moti-

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

Table 5: C_{avg} yielded by the most competitive systems presented to Albayzin 2008 LRE.

	C_{avg} (30-second segments) F: free development, R: restricted development, O: open-set, C: closed-set					
Competition	**FO**	**RO**		**FC**	**RC**	
System	*primary*	*primary*	*contrastive*	*primary*	*primary*	*contrastive*
Site A	0.0946	0.1313	0.1110	0.0552	0.0778	0.0656
Site B	0.1204	0.2787	–	0.0556	0.2420	–

vated just by the lack of a multilingual speech database featuring the official languages in Spain. Using wide-band (16 kHz, single channel) broadcast recordings made sense since we were primarily interested in building a language recognition module for the backend of an audio indexing and retrieval system dealing with wide-band broadcast news in Spanish and Basque (which is likely to be extended to broadcast news in Catalan and Galician) (Bordel et al., 2009). Open-set language recognition was needed because broadcast news often include fragments in foreign languages (French, English, Arabic, etc.) that must be discarded for automatic transcription and indexing. Therefore, KALAKA was designed and is being used in our research group for both basic and applied research on language recognition.

5.2.1. The main language recognition system

A language recognition system has been built starting from the train and development sets of KALAKA and the materials implicitly used to built phone decoders (see below). Performance has been measured, in terms of C_{avg} and DET curves, on the evaluation set of KALAKA. The system consists of a hierarchical fusion of 7 individual subsystems: an acoustic GMM-SVM subsystem using 7-2-3-7 SDC-MFCC, three Phone-SVM subsystems and three Phone-LM subsystems (see descriptions below). In order to make it easier for other researchers to verify our results, open software resources were used to build all the subsystems.

For the GMM-SVM subsystem, acoustic models were estimated using the *Sautrela* toolbox (Penagarikano and Bordel, 2005). The phonotactic systems were based on the phone decoders developed and made available by the Brno University of Thechnology (BUT) for Czech, Hungarian and Russian (Schwarz, 2008). BUT decoders have been previously used by other groups −besides BUT (Matejka et al., 2007), the MIT Lincoln Laboratory (Torres-Carrasquillo et al., 2008)− as the backend for phonotactic language recognition, yielding high recognition accuracies. Each BUT decoder runs its own acoustic front-end, so it can be seen as a black box which takes a speech signal as input and gives the 1-best phone decoding as output. The Phone-LM subsystems applied the SRI Language Model toolkit (Stolcke, 2002) to estimate phone sequence n-gram models. Finally, all the subsystems based on Support Vector Machines (SVM) (Campbell et al., 2006a) (Campbell et al., 2006b) were developed using either *SVMTorch* (Collobert and Bengio, 2001) or *libSVM* (Chang and Lin, 2001), for dense and sparse vectors, respectively.

Scores produced by language recognition subsystems were first normalized, by means of a *t-norm* (Auckenthaler et al., 2000), and then calibrated, by means of a Gaussian backend. Finally, normalized and calibrated scores were fused by applying linear logistic regression optimization. A minimum expected cost Bayes decision threshold was then established, according to the application-dependent language priors and costs −see (Brümmer and van Leeuwen, 2006) and (Brümmer et al., 2007) for details.

5.2.2. The GMM-SVM subsystem

The GMM-SVM subsystem applies a SVM classifier on the vector space defined by Gaussian Mixture Model (GMM) parameters. The GMM corresponding to a target language is constructed by using training samples of that language to adapt the means of a Universal Background Model (UBM) consisting of 1024 mixture components. Maximum A Posteriori (MAP) adaptation is performed using a relevance factor of $\tau = 16$. The adapted means are normalized and stacked to construct the so called GMM supervectors which feed the SVM classifier (Campbell et al., 2006b).

5.2.3. Phonotactic subsystems

As noted above, the phonotactic subsystems were based on the Brno University of Technology (BUT) TRAPS/NN decoders for Czech, Hungarian and Russian. These decoders were designed to process 8 kHz raw PCM signals. Therefore, the original 16 kHz signals were downsampled to 8 kHz. Prior to phone tokenization, an energy based Voice Activity Detector (VAD) was used to split and remove low-energy (presumably non-speech) segments from the signals. Non-phonetic units appearing in phone sequences were all mapped to silence, leading to inventories of 43, 59 and 49 phonetic units for Czech, Hungarian and Russian, respectively.

Two different phone sequence modeling techniques were applied:

- *Phone-LM*: 4-gram language models with Witten-Bell smoothing.
- *Phone-SVM*: SVM (with a linear kernel), built on bag-of-N-gram vectors (including up to 3-grams), weighted as proposed in (Richardson and Campbell, 2008).

5.2.4. Results

Table 6 shows the performance (C_{avg}) of single and fused systems on the closed-set evaluation subset of 30-second speech segments. DET curves for the GMM-SVM subsystem, the Phone-LM fused system, the Phone-SVM fused system and the main system (fusing all the previous systems) are shown in Figure 2. The performance of the main

Table 6: Performance (C_{avg}) of single and fused language recognition systems on the closed-set evaluation subset of 30-second speech segments of KALAKA.

		C_{avg}
Single systems	GMM-SVM	0.1611
	PHONE (CH) - LM	0.1545
	PHONE (HU) - LM	0.1427
	PHONE (RU) - LM	0.1305
	PHONE (CH) - SVM	0.0940
	PHONE (HU) - SVM	0.1017
	PHONE (RU) - SVM	0.1215
Fused systems	PHONE - LM	0.0892
	PHONE - SVM	0.0774
	PHONE	0.0691
	ALL	0.0576

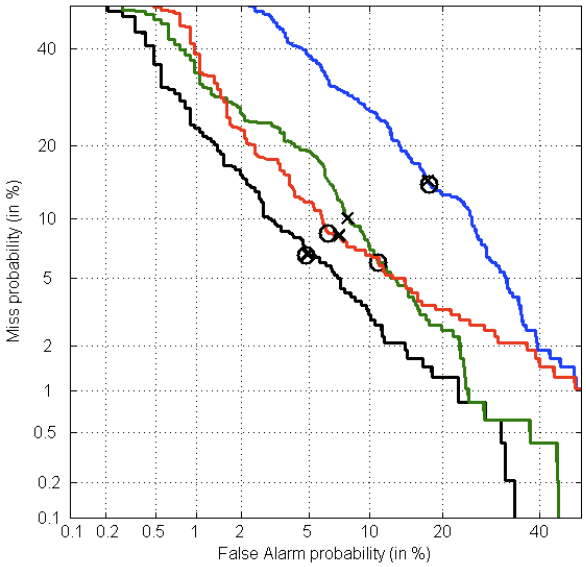

Figure 2: Pooled DET curves of various systems: GMM-SVM (blue), Phone-LM (green), Phone-SVM (red) and the system fusing all of them (black), on the closed-set evaluation subset of 30-second speech segments. Operation points corresponding to actual (X) and minimum (O) C_{avg} are marked on the curves.

system ($C_{avg} = 0.0576$) is similar to that of the most competitive systems submitted to the Albayzin 2008 LRE (in the free-development condition). In any case, taking into account that state-of-the-art technology has been applied, these results confirm that the task defined on KALAKA is quite challenging and may help further developments in language recognition technology.

6. Conclusions and future work

In this paper, we have addressed the design, data collection and evaluation of KALAKA, a database consisting of wide-band (16 kHz) audio signals taken from TV broadcasts, created and used specifically for the Albayzin 2008 Language Recognition Evaluation, which was carried out from May to November 2008. The database includes train, development and evaluation materials for four target languages: Basque, Catalan, Galician and Spanish (official languages in Spain). It also includes speech signals in other languages to allow open-set verification trials.

Results attained in the Albayzin 2008 LRE have been presented as a means of evaluating the database. Preliminary results using various state-of-the-art language recognition sub-systems and the system resulting from their fusion have been also presented to provide more evidences of the difficulty of the task. Taking into account the performance attained in both cases, we can conclude that tasks defined on KALAKA can be challenging enough to support further developments in language recognition technology.

Future work will focus on preparing an extended version of KALAKA to support a second evaluation this year, the Albayzin 2010 LRE, using again wide-band (16 kHz) TV broadcast speech signals, but including also Portuguese and English as target languages and renewing the set of *unknown* languages. We hope this new feature will make the evaluation more appealing for research teams from outside Spain. The evaluation would be held from June to October 2010 and results would be presented at the 6th Biennial Workshop on Speech Technology, to be held in Vigo (Spain) in November 2010. If things go as we expect, the evaluation plan would be posted through ISCA in June 2010.

7. References

R. Auckenthaler, M. Carey, and H. Lloyd-Thomas. 2000. Score normalization for text-independent speaker verification systems. *Digital Signal Processing*, 10(1-3):42–54, January.

G. Bordel, M. Diez, I. Landera, S. Nieto, M. Penagarikano, L.J. Rodriguez-Fuentes, A. Varona, and M. Zamalloa. 2009. Hearch: A Multilingual Spoken Document Retrieval System. In *IEEE Automatic Speech Recognition and Understanding Workshop (demo)*.

N. Brümmer and J. du Preez. 2006. Application-Independent Evaluation of Speaker Detection. *Computer, Speech and Language*, 20(2-3):230–275, April-July.

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, L. Burget, J. Cernocky, O. Glembek, F. Grezl, M. Karafiat, D. van Leeuwen, P. Matejka, P. Schwarz, and A. Strasheim. 2007. Fusion of heterogeneous speaker recognition systems in the STBU submission for the NIST speaker recognition evaluation 2006. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(7):2072–2084.

W. M. Campbell, J. P. Campbell, D. A. Reynolds, E. Singer, and P. A. Torres-Carrasquillo. 2006a. Support vector machines for speaker and language recognition. *Computer Speech and Language*, 20(2-3):210–229.

W.M. Campbell, D.E. Sturim, D.A. Reynolds, and A. Solomonoff. 2006b. SVM Based Speaker Verification Using a GMM Supervector Kernel and NAP Variability Compensation. In *Proceedings of ICASSP*, volume 1, pages 97–100.

C.C. Chang and C.J. Lin, 2001. *LIBSVM: a library for support vector machines*. http://www.csie.ntu.edu.tw/∼cjlin/libsvm.

R. Collobert and S. Bengio. 2001. SVM Torch: Support Vector Machines for Large-Scale Regression Problems. *The Journal of Machine Learning Research*, 1:143–160.

A.F. Martin and A.N. Le. 2008. NIST 2007 Language Recognition Evaluation. In *Proceedings of Odyssey 2008 - The Speaker and Language Recognition Workshop, paper 16*, Stellenbosch, South Africa.

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.

P. Matejka, L. Burget, O. Glembek, P. Schwarz, V. Hubeika, M. Fapso, T. Mikolov, and O. Plchot. 2007. BUT system description for NIST LRE 2007. In *Proc. 2007 NIST Language Recognition Evaluation Workshop*, pages 1–5, Orlando, US. National Institute of Standards and Technology.

M. Penagarikano and G. Bordel. 2005. Sautrela: A Highly Modular Open Source Speech Recognition Framework. In *Proceedings of the ASRU Workshop*, pages 386–391, San Juan, Puerto Rico, December.

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

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

Kare Sjolander and Jonas Beskow. 2000. Wavesurfer - An Open Source Speech Tool. In *Proceedings of ICSLP 2000*, volume 4, pages 464–467, Beijing, China.

A. Stolcke. 2002. SRILM - an extensible language modeling toolkit. In *Proc. Intl. Conf. on Spoken Language Processing*, pages 257–286, November.

P.A. Torres-Carrasquillo, E. Singer, W.M. Campbell, T. Gleason, A. McCree, D.A. Reynolds, F. Richardson, W. Shen, and D.E. Sturim. 2008. The MITLL NIST LRE 2007 language recognition system. In *Proceedings of Interspeech 2008*, pages 719–722.