KALAKA-3: a database for the recognition of spoken European languages on YouTube audios

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

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

Abstract

This paper describes the main features of KALAKA-3, a speech database specifically designed for the development and evaluation of language recognition systems. The database provides TV broadcast speech for training, and audio data extracted from YouTube videos for tuning and testing. The database was created to support the Albayzin 2012 Language Recognition Evaluation, which featured two language recognition tasks, both dealing with European languages. The first one involved six target languages (Basque, Catalan, English, Galician, Portuguese and Spanish) for which there was plenty of training data, whereas the second one involved four target languages (French, German, Greek and Italian) for which no training data was provided. Two separate sets of YouTube audio files were provided to test the performance of language recognition systems on both tasks. To allow open-set tests, these datasets included speech in 11 additional (Out-Of-Set) European languages. The paper also presents a summary of the results attained in the evaluation, along with the performance of state-of-the-art systems on the four evaluation tracks defined on the database, which demonstrates the extreme difficulty of some of them. As far as we know, this is the first database specifically designed to benchmark spoken language recognition technology on YouTube audios.

Keywords: Spoken Language Recognition, European languages, YouTube audio

1. Introduction

KALAKA-3 was designed and collected to support the Albayzin 2012 Language Recognition Evaluation (LRE) organized by the Spanish Thematic Network on Speech Technologies from May to November 2012 (Rodriguez-Fuentes et al., 2013). This was the third of a series of language recognition evaluations, which started with the Albayzin 2008 LRE (Rodriguez-Fuentes et al., 2010b) and continued with the Albayzin 2010 LRE (Rodriguez-Fuentes et al., 2011).

The Albayzin 2008 LRE used the four official languages in Spain (Basque, Catalan, Galician and Spanish) as target languages, and four European languages (English, French, German and Portuguese) as Out-Of-Set (OOS) languages for open-set tests (OOS languages were not disclosed to participants). Speech data were extracted from wide-band stereo TV broadcast recordings and stored as single-channel 16 kHz 16-bit PCM-encoded WAV files. Despite the high quality of speech recordings, a relatively high confusion among target languages (specially between Galician and Spanish) was found. Note that most speakers of Basque, Catalan and Galician also speak Spanish in their daily lives (Spanish being even the mother language for some of them), making their phoneme inventories, pronunciations, etc. quite close to each other, which would partly explain the high confusion among these languages.

The Albayzin 2010 LRE considered six target languages: Basque, Catalan, English, Galician, Portuguese and Spanish, for which an increased amount of training data was provided, and four OOS languages (Arabic, French, German and Romanian) for open-set tests. Besides studio-quality (clean) TV broadcast speech, a second type of recordings was used, which included background noise/music and/or overlapping conversations. Thus, two different tasks were carried out, the first one on clean speech and the second one on a mix of clean and noisy speech. We found that the average performance on the full set of target languages was better than the average performance on the four official languages in Spain, meaning that error rates for Portuguese and English were lower than the average. This is not surprising, since English and Portuguese utterances come from speakers that do not speak Spanish in their daily lives and thus we may reasonably expect them to be more easily distinguishable from the other four languages. The average performance on the four official languages in Spain was better than that attained in 2008, in part due to improvements in technology but also to the availability of more training data. Finally, we also found that performance degraded remarkably when switching from clean to noisy speech.

Based on these findings, we concluded that language recognition technology should deal with acoustic variability (channel, noise, music, overlapping speakers, etc.), which is inherent to some media (such as the videos posted by people in the Internet), and data availability constraints which may seriously limit the performance of state-of-the-art systems (for a more detailed study, see (Rodriguez-Fuentes et al., 2012b)). Therefore, the Albayzin 2012 LRE was designed with the aim to test the performance of state-of-the-art language recognition systems on unconstrained speech, possibly in a low-resource scenario (i.e. with just a reduced amount of data available for development). These new conditions matched two pre-requisites that, from our point of view, are essential to any technology benchmark: (1) the task must be of practical interest (in this case, indexing multimedia contents with the spoken language); and (2)

the task must be challenging (i.e. difficult) enough in order to foster technological improvements.

With these goals in mind, KALAKA-3 was built by recycling TV broadcast speech (both clean and noisy) from previous evaluations and by collecting new unconstrained speech signals from YouTube videos. TV broadcast speech signals were used for training, whereas audio data extracted from YouTube videos were used for tuning and testing. Two different tasks were defined: (1) *Plenty-of-Training*, which involved six target languages (Basque, Catalan, English, Galician, Portuguese and Spanish) for which a large amount of training data was provided; and (2) *Empty-Training*, which involved four target languages (French, German, Greek and Italian), for which no training data was provided. In both cases, two disjoint sets of YouTube audio files were provided for tuning and testing system performance, respectively. To allow open-set tests, these datasets included speech files in 11 additional OOS languages (Bulgarian, Czech, Croatian, Finnish, Hungarian, Polish, Romanian, Russian, Serbian, Slovak and Ukrainian).

The main features of the database are outlined in Section 2, including a breakdown of the training dataset. Details about the design of the development and evaluation datasets are given in Section 3. Section 4 describes the collection procedure and the validation criteria applied to select YouTube audios. Section 5 presents a summary of the results attained in the Albayzin 2012 LRE, along with the performance of state-of-the-art systems on the four evaluation tracks defined on the database. Finally, conclusions and future work are outlined in Section 6.

2. Database main features

KALAKA-3 consists of three subsets: training, development and evaluation. The training dataset comes entirely from KALAKA-2 (Rodriguez-Fuentes et al., 2012a), the database created to support the Albayzin 2010 LRE, which was an extension of KALAKA (Rodriguez-Fuentes et al., 2010a) (the database that supported the Albayzin 2008 LRE). The training dataset consists of TV broadcast recordings (stored as single-channel 16 kHz 16-bit PCM encoded WAV files), including both planned and spontaneous speech in diverse environment conditions (excluding telephone-channel speech) and multiple speakers. The training dataset is split into two disjoint subsets, consisting of clean speech (around 86 hours) and noisy speech (around 22 hours), respectively (see Table 1).

Table 1: Distribution of training segments per target language for clean and noisy speech in KALAKA-3: number of segments (#) and total duration (T, in minutes).

	Clean speech		**Noisy speech**	
	#	T **(min.)**	**#**	T **(min.)**
Basque	579	860	215	190
Catalan	440	948	209	185
English	322	929	265	220
Galician	675	877	300	227
Portuguese	558	941	295	268
Spanish	486	854	312	318

Noisy-speech segments may include noisy and/or overlapped speech, maybe with short fragments of clean speech. Different and variable types of noise may appear: street, music, cocktail party, laughs, clapping, etc. Most speech overlaps appear in hot spots of informal debates in late night shows, magazines, etc. which, on the other hand, usually feature clean-channel and quiet-background (studio) conditions. In all cases, each training segment contains speech in a single language.

The development and evaluation datasets have been specifically collected for KALAKA-3, and consist of unconstrained YouTube audio signals (originally in different formats and qualities, but all of them stored as single-channel 16 kHz 16-bit PCM encoded WAV files), with the only requirement that a single language is spoken in them. Note that, besides speech produced by possibly multiple speakers, any other sound (music, noise, etc.) could appear in YouTube audios, which makes the task specially challenging. The design of these datasets, the collection procedure and the validation criteria are described in Sections 3 and 4.

In summary, the training dataset amounts to around 108 hours of speech, with 18 hours on average for each one of the 6 target languages considered in the *Plenty-of-Training* task (80% being clean speech and 20% noisy speech). The development and evaluation datasets have the same size (more than 2,000 YouTube audios) and structure, but a different distribution of OOS languages, to avoid overfitting systems to reject specific OOS languages. The whole database amounts to around 200 hours of audio (1.6 times the size of KALAKA-2) and is distributed as a set of downloadable tarballs, after direct request to the authors.

3. Design issues

As noted above, training data were entirely imported from KALAKA-2, so efforts focused on collecting audio from YouTube videos for the development and evaluation datasets. Since language recognition systems were expected to be tuned on the developmet dataset, we kept in mind that the evaluation dataset should be as independent as possible from the development dataset in order to avoid a biased benchmark. This was partially addressed by collecting videos from different YouTube categories, meaning that different topics, different speakers and even different recording conditions appear in both datasets. In previous Albayzin LREs, system performance was measured on three different subsets of speech signals, with nominal durations of 30, 10 and 3 seconds. In the Albayzin 2012 LRE, these nominal duration subsets were not considered anymore. However, to keep things reasonably bounded, YouTube audios were constrained to be between 30 and 120 seconds long, with at least 5 seconds of speech. Also, in order to keep consistency in the database, YouTube audios containing telephone-channel speech were discarded.

4. Collecting YouTube audios

The goal was to collect 300 YouTube audios for each target language (150 for development and 150 for evaluation) and around 100 YouTube audios for each Out-Of-Set language (with different distributions in the development and

evaluation datasets). In this way, each dataset would consist of around 2,000 YouTube audios, which was considered enough for benchmarking language recognition technology.

4.1. Building lists of YouTube videos

After a preliminary study for Spanish, based on a small set of keywords taken from the *aspell*[1] dictionary and considering different YouTube video categories, we chose the six categories most likely to contain speech: *Education*, *News*, *Entertainment*, *Howto*, *Nonprofit* and *Technology*. Then, a large list of YouTube videos was created for each category and each of the 21 languages considered in the database, using the YouTube API[2] to search for language-specific common words in the title, description and other metadata (tags) associated to each video.

The words used for searching video metadata were taken from the *aspell* dictionary of each language, with the following criteria: (1) word inflections and verb tenses were not considered (only *canonical forms* were included); (2) words with less than 6 characters were filtered out; (3) for each language, 2,000 words were randomly chosen among those fulfilling the above mentioned criteria; (4) for any given language, those words that appeared in the *aspell* dictionary of other language were filtered out; and (5) only 1,000 words were retained per language.

Two additional criteria were applied to rank the videos in the list: (1) the first and most important was the existence of any *Creative Commons* license associated to the video; (2) the second was the geographical location (only for those videos containing such information in the metadata): to increase the chances of finding speech in a particular language, priority was given to those videos located within a certain distance (typically, 200 kilometers, though this parameter may vary depending on the size of the country) from a major city where the language of interest was spoken. As a result, we found ourselves with 21 long lists (10 for target languages, 11 for OOS languages) of ranked YouTube videos to validate. It must be noted that our efforts for collecting videos with a *Creative Commons* license were not rewarded at all: we found few of them (it strongly depended on the country) and the resulting lists mostly consisted of videos without any *Creative Commons* license.

4.2. Validating YouTube videos

Each list (a spreadsheet with links to YouTube source pages) was scrolled through to listen to and look at the YouTube videos. The goal was to validate 55 videos per list for target languages, and 17 videos per list for OOS languages (these numbers are slightly higher than needed to account for errors during or after downloading). For target languages, a video was validated only if: (1) there was sufficient amount of speech (around 5 seconds) to make possible the recognition of the target language; (2) there was only speech in the target language; and (3) the environment conditions and the recording quality were good enough for the speech to be intelligible (note that videos including telephone-channel speech were discarded). These criteria were all applied subjectively by the auditors (meaning that e.g. no objective measure of intelligibility was applied). In the case of OOS languages, the second condition was relaxed, so that there could be speech in several languages, provided that none of them was a target language. The balance between categories was respected in all cases except for some languages for which not enough videos were available in some categories.

The validation task was carried out by five auditors (the authors of this paper) and took more than two months. A breakdown of the validated / audited videos per language and category is shown in Table 2.

4.3. Fetching and converting YouTube audios

The validated YouTube videos were automatically downloaded by processing the spreadsheets and applying the *youtube-dl*[3] tool. Then, the *ffmpeg*[4] tool was used to extract the audio layer from the videos and the *SoX*[5] tool was applied to get single-channel 16 kHz 16-bit PCM encoded WAV audio files. Since YouTube contents evolve dynamically (many videos are available only for some months or even weeks before the owner removes them), we made a local copy of all the videos, audios and metadata downloaded from YouTube, strictly for backup purposes. The database does not provide any additional information about the videos, but just the audio and the identity of the spoken language (which is specified in the ground truth files needed to measure system performance).

4.4. Collection of YouTube audios

As a result of the above described procedure, 4,168 YouTube audios were validated out of 21,860 audited videos: 2,059 audios were posted for development (extracted from the *News*, *Education* and *Howto* categories), whereas 2,109 audios were posted for evaluation (extracted from the *Entertainment*, *Nonprofit* and *Tech* categories). There were at least 150 audios per target language in each of the development and evaluation datasets. The OOS languages were distributed as follows: all the audios in Czech, Croatian, Polish and Romanian were posted for development and all the audios in Bulgarian, Finnish, Slovak and Serbian were posted for evaluation, whereas the audios in Hungarian, Russian and Ukrainian were equally distributed in both datasets. A breakdown of the development and evaluation datasets is shown in Table 3. Since some of the validated videos were not available at the time of downloading (typically because the owner removed them), numbers in Table 3 are slightly smaller than those presented in Table 2.

5. Database evaluation

5.1. Albayzin 2012 LRE official results

As noted above, the Albayzin 2012 LRE defined two tasks: *Plenty-of-Training* (involving 6 target languages for which

[1] http://aspell.net/

[2] YouTube API v2.0: https://developers.google.com/youtube/2.0/developers_guide_protocol_audience.

[3] http://rg3.github.io/youtube-dl/

[4] http://www.ffmpeg.org/

[5] http://sox.sourceforge.net/

Table 2: Number of validated YouTube videos (out of: total number of audited videos) per language and category.

		Education	Howto	News	Entertainment	Nonprofit	Tech	TOTAL
Target languages (Plenty)	**Basque**	72 (343)	9 (73)	74 (303)	95 (965)	15 (162)	41 (200)	306 (2,046)
	Catalan	50 (183)	50 (256)	50 (163)	43 (372)	58 (238)	59 (221)	310 (1,422)
	English	50 (111)	50 (159)	50 (130)	51 (277)	58 (194)	55 (186)	314 (1,057)
	Galician	38 (384)	13 (125)	100 (479)	53 (442)	55 (221)	52 (393)	311 (2,044)
	Portuguese	55 (699)	45 (1,021)	60 (156)	50 (542)	55 (395)	60 (576)	325 (3,389)
	Spanish	52 (200)	50 (260)	55 (194)	50 (553)	67 (398)	37 (367)	311 (1,972)
Target languages (Empty)	**French**	50 (146)	50 (162)	51 (116)	53 (269)	53 (208)	52 (192)	309 (1,093)
	German	50 (121)	50 (152)	50 (142)	50 (229)	50 (245)	55 (298)	305 (1,187)
	Greek	50 (196)	53 (297)	54 (199)	55 (252)	55 (239)	55 (287)	322 (1,470)
	Italian	52 (153)	55 (229)	54 (123)	51 (304)	55 (141)	54 (197)	321 (1,147)
OOS languages	**Bulgarian**	15 (61)	15 (65)	15 (37)	19 (101)	19 (65)	16 (94)	99 (423)
	Croatian	15 (52)	15 (54)	15 (35)	15 (59)	15 (47)	15 (131)	90 (378)
	Czech	17 (57)	17 (89)	17 (32)	17 (130)	17 (63)	17 (136)	102 (507)
	Finnish	18 (61)	15 (68)	15 (45)	15 (43)	15 (38)	15 (87)	93 (342)
	Hungarian	17 (54)	17 (139)	17 (57)	17 (147)	17 (59)	17 (117)	102 (573)
	Polish	17 (34)	17 (76)	17 (66)	17 (127)	17 (94)	17 (106)	102 (503)
	Romanian	17 (72)	17 (108)	17 (53)	15 (197)	17 (74)	15 (165)	98 (669)
	Russian	15 (46)	15 (62)	15 (43)	19 (104)	19 (110)	16 (132)	99 (497)
	Serbian	15 (69)	15 (75)	15 (27)	17 (125)	17 (58)	17 (101)	96 (455)
	Slovak	17 (65)	17 (105)	17 (34)	17 (95)	17 (56)	17 (163)	102 (518)
	Ukrainian	15 (35)	15 (85)	15 (30)	19 (99)	19 (79)	17 (100)	100 (428)
TOTAL		697 (3,142)	600 (3,660)	723 (2,464)	738 (5,432)	710 (3,184)	699 (3,978)	4,168 (21,860)

Table 3: Distribution of YouTube audio files per language in the development and evaluation datasets of KALAKA-3.

		Devel	Eval
Target languages (Plenty-of-Training)	**Basque**	154	150
	Catalan	149	158
	English	150	156
	Galician	151	160
	Portuguese	160	163
	Spanish	153	154
Target languages (Empty-Training)	**French**	150	155
	German	146	151
	Greek	155	165
	Italian	158	160
OOS languages	**Bulgarian**	0	98
	Croatian	90	0
	Czech	102	0
	Finnish	0	89
	Hungarian	51	51
	Polish	102	0
	Romanian	98	0
	Russian	45	54
	Serbian	0	91
	Slovak	0	102
	Ukrainian	45	52

plenty of training data was available) and *Empty-Training* (involving 4 target languages for which no training data was provided). Besides, it considered two conditions (closed-set vs. open-set) depending on the presence of OOS languages in the test set. Therefore, four different tracks were defined: *Plenty-Closed* (PC), *Plenty-Open* (PO), *Empty-Closed* (EC) and *Empty-Open* (EO).

A new metric was introduced in this evaluation, based on a calibration-sensitive, multi-class cross-entropy criterion, which measures the information provided by a spoken language recognition system through a set of log-likelihoods and does not require making hard decisions. The performance metric, called *actual relative confusion* (F_{act}), represents the factor by which the system changes the prior confusion (that corresponding to a non-informative system). Good systems will attain relative confusions between 0 and 1 (being 0 only for a perfect system). Under this metric, the task is defined as follows: given a test audio X and assuming N target languages, the system must provide $N + 1$ scores, one per target language plus an additional score for OOS languages, which are interpreted as log-likelihoods. The score for OOS languages is ignored in the closed-set condition. This new metric was developed in collaboration with Niko Brümmer, from Agnitio South Africa, and can be computed by means of a freely available toolkit[6]. For more details on the Albayzin 2012 LRE, see (Rodriguez-Fuentes et al., 2013).

To illustrate the relative difficulty of the four tracks mentioned above, Table 4 presents the F_{act} values for the primary systems (including some late submissions) submitted to the Albayzin 2012 LRE. Best results are marked in boldface. The performance of the most competitive system in the PC track ($F_{act} = 0.071$) was only slightly better than that attained by the same system in the PO track ($F_{act} = 0.085$). This result just reveals that the confusion of OOS languages (most of them Slavic) with target languages (most of them Romance) was quite low. Unfortunately, this can be seen as a design flaw, since the open-set condition would have been harder if OOS languages had been closer to target languages. On the other hand, af-

[6]https://sites.google.com/site/bilbaotoolkit/

ter comparing the results on the *Empty-Training* condition (EC and EO) with those on the *Plenty-of-Training* condition (PC and PO), we conclude that having training data for target languages is key for attaining good performance. Some groups (1 and 6 late) tried to alleviate the lack of training data by using part of the development data to train models for the target languages in the *Empty-Training* condition. This strategy is further explored in Section 5.2.

Table 4: Performance (in terms of the multiclass cross-entropy measure F_{act}) of primary systems (including some late submissions) in the four tracks of the Albayzin 2012 LRE. Best results (lowest F_{act}) are marked in boldface.

Systems	PC	PO	EC	EO
1	**0.071**	**0.085**	–	–
2	0.078	0.120	0.498	0.516
3	0.113	0.114	0.711	0.796
4	0.121	0.160	0.626	0.676
5	0.122	–	–	–
6	0.141	0.184	–	–
7 (late)	0.407	0.216	–	–
1 (late)	–	–	**0.216**	–
6 (late)	–	–	0.310	**0.372**

5.2. Evaluation based on state-of-the-art language recognition systems

To further validate KALAKA-3 as a benchmark for language recognition technology, we have carried out a series of experiments based on two different acoustic systems, both following the Total Variability Factor Analysis (*iVector*) approach as described in (Dehak et al., 2011) and (Martinez et al., 2011). Under this approach, a high-dimensional input vector is mapped to a low-dimensional feature vector (an iVector), hypothetically retaining most of the relevant information. The high dimensional representation usually consists of a supervector stacking the means of a Gaussian Mixture Model (GMM), obtained through Maximum-A-Posteriori adaptation from a Universal Background Model (UBM), based on the feature vectors of the input utterance. Under the iVector approach, an utterance dependent GMM supervector M is decomposed as follows:

$$M = m + Tw$$

where m is the utterance-independent mean supervector, T is the total variability matrix (a low-rank rectangular matrix) and w is the so called iVector (a normally distributed low-dimensional latent vector). The latent vector w can be estimated from its posterior distribution conditioned to the Baum-Welch statistics extracted from the utterance and using a UBM.

Two different feature sets were considered under the iVector approach: (1) the traditional Mel-Filter Cepstral Coefficient / Shifted Delta Cepstrum (MFCC/SDC) features (Torres-Carrasquillo et al., 2002), under a 7-2-3-7 configuration; and (2) the recently introduced log-likelihood ratios of phone posterior probabilities, hereafter called *Phone Log-Likelihood Ratios* (PLLR) (Diez et al., 2012), based on the posteriors provided by the open-software *Temporal Patterns Neural Network* (TRAPs/NN) phone decoders, developed by the Brno University of Technology (BUT) for Czech, Hungarian and Russian (Schwarz, 2008).

Both the MFCC/SDC-iVector and the PLLR-iVector systems performed Voice Activity Detection (VAD) by removing the feature vectors whose highest PLLR value corresponded to a phonetic unit representing non-speech events (silent pauses, short noises, etc.), using the BUT phone decoder for Hungarian. Though this VAD scheme had worked fairly good for detecting speech in telephone audio, in this case we applied it on unconstrained speech with no adaptation, so it could be failing due to e.g. the presence of music or background conversations. A more sophisticated audio classification scheme should be applied instead to avoid these issues.

A gender-independent 1024-mixture UBM was estimated by Maximum Likelihood on the training dataset. The total variability matrix T was estimated according to the procedure defined in (Dehak et al., 2011), but using only data from target languages, as in (Martinez et al., 2011). A generative modeling approach was applied in the iVector feature space, the set of iVectors of each language being modeled by a single Gaussian distribution, as in (Martinez et al., 2011).

When processing an input utterance, our iVector systems provide a score for each target language. A Gaussian backend was estimated, based on the scores obtained for the development set, and applied to the scores obtained for the evaluation set, in order to get log-likelihoods (one per target language). Log-likelihoods were then calibrated and fused according to a discriminative linear model which minimized the so called C_{llr} function on the development set, by means of logistic regression, as explained in (Brummer and van Leeuwen, 2006). To alleviate the lack of training data in the *Empty-Training* condition, one half of the development data was used for training the UBM and the Total Variability matrix, and the other half for estimating the backend and fusion parameters.

Table 5 shows the performance of the iVector systems and their fusion in the four tracks of the Albayzin 2012 LRE. Performance is pretty good in the *Empty-Training* condition when compared to results in Table 4, specially for the fused system. This remarkable performance is probably due to the use of half of the development data for training specific models for target languages. Note that, though this use was not explicitly forbidden in the Albayzin LRE Plan (Rodriguez Fuentes et al., 2012c), the development dataset was designed only for tuning systems, not for training models. Things were arranged this way to force Albayzin 2012 LRE participants to explore alternative ways of performing the task in a low-resource scenario, when no training data was available, i.e. without using specific models for target languages.

In the *Plenty-of-Training* condition, the performance of our systems is still far from the best results in Table 4, specially in the PC track. After a preliminary study of the obtained results, we found a sizeable number of VAD errors which made our systems to produce bad scores. Also, the best systems presented to the Albayzin 2012 LRE resulted from the fusion of several acoustic and phonotactic subsystems.

We reasonably hope to get improved performance by replacing our current VAD by an audio classification module and by fusing our iVector systems with other complementary systems based on phonotactic features (Penagarikano et al., 2011a; Penagarikano et al., 2011b; Penagarikano et al., 2011c). In any case, results in Tables 4 and 5 show that KALAKA-3 provides a challenging benchmark for the development of spoken language recognition technology.

Table 5: Performance (in terms of the multiclass cross-entropy measure F_{act}) of two state-of-the-art systems (iVectors with MFCC features and iVectors with PLLR features) and the fusion of them, in the four tracks of the Albayzin 2012 LRE.

Systems	PC	PO	EC	EO
iVector-MFCC	0.139	0.254	0.238	0.342
iVector-PLLR	0.191	0.294	0.217	0.341
Fusion	0.098	0.128	0.131	0.221

6. Conclusions

In this paper, we have described KALAKA-3, the database created to support the Albayzin 2012 LRE, which explored the performance of state-of-the-art language recognition systems on unconstrained speech, even when no training data were available for target languages. The datasets used for tuning and testing systems consist of YouTube audios, which, as far as we know, are used for the first time to benchmark spoken language recognition technology. Besides describing the collection procedure and the criteria applied to filter and validate YouTube videos, the paper also evaluates the database, by presenting a summary of the results attained in the Albayzin 2012 LRE and two acoustic iVector systems recently developed by our research group, which have not performed as well as expected. We are currently developing a more robust VAD module and several phonotactic systems (known to be complementary to acoustic systems in fusions), that we expect lead to improved performance. KALAKA-3 has been already used for the development of spoken language recognition technology in (D'Haro et al., 2013). Currently, the database is distributed as a set of downloadable tarballs and must be requested directly to the authors. In the future, we plan to license and distribute the development and evaluation datasets of KALAKA-3 through the Linguistic Data Consortium.

7. Acknowledgements

This work has been supported by the University of the Basque Country, under grant GIU13/28, and by the Government of the Basque Country, under program SAIOTEK (projects S-PE12UN055 and S-PE13UN105). Mireia Diez has been supported by a 4-year research fellowship from the Department of Education, University and Research of the Government of the Basque Country.

8. References

Brummer, N. and van Leeuwen, D. (2006). On calibration of language recognition scores. In *Proceedings of Odyssey: The Speaker and Language Recognition Workshop*, pages 1–8.

Dehak, N., Kenny, P., Dehak, R., Dumouchel, P., and Outlet, P. (2011). Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(4):788–798.

D'Haro, L. F., de Córdoba, R., Caraballo, M. A., and Pardo, J. M. (2013). Low-Resource Language Recognition Using a Fusion of Phoneme Posteriorgram Counts, Acoustic and Glottal-Based I-Vectors. In *Proceedings of ICASSP*, pages 6852–6856, Vancouver, Canada.

Diez, M., Varona, A., Penagarikano, M., Rodriguez-Fuentes, L. J., and Bordel, G. (2012). On the Use of Log-Likelihood Ratios as Features in Spoken Language Recognition. In *IEEE Workshop on Spoken Language Technology (SLT)*, Miami, Florida, USA.

Martinez, D., Plchot, O., Burget, L., Glembek, O., and Matejka, P. (2011). Language Recognition in iVectors Space. In *Proceedings of Interspeech*, pages 861–864.

Penagarikano, M., Varona, A., Rodriguez-Fuentes, L., and Bordel, G. (2011a). Dimensionality Reduction for Using High-Order n-grams in SVM-Based Phonotactic Language Recognition. In *Proceedings of Interspeech 2011*, pages 853–856, Florence, Italy, August 28-31.

Penagarikano, M., Varona, A., Rodríguez-Fuentes, L., and Bordel, G. (2011b). A dynamic approach to the selection of high-order n-grams in phonotactic language recognition. In *Proceedings of ICASSP*, pages 4412–4415, Prague, Czech Republic, May 22-27.

Penagarikano, M., Varona, A., Rodríguez-Fuentes, L. J., and Bordel, G. (2011c). Improved Modeling of Cross-Decoder Phone Co-occurrences in SVM-based Phonotactic Language Recognition. *IEEE Transactions on Audio, Speech and Language Processing*, 19(8):2348–2363, November.

Rodriguez-Fuentes, L. J., Penagarikano, M., Bordel, G., Varona, A., and Diez, M. (2010a). 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.

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

Rodriguez-Fuentes, L. J., Penagarikano, M., Varona, A., Diez, M., and Bordel, G. (2011). The Albayzin 2010 Language Recognition Evaluation. In *Proceedings of Interspeech*, pages 1529–1532.

Rodriguez-Fuentes, L. J., Penagarikano, M., Varona, A., Diez, M., and Bordel, G. (2012a). KALAKA-2: a TV broadcast speech database for the recognition of Iberian languages in clean and noisy environments. In *Proceedings of the LREC*, pages 99–105, Istanbul, Turkey.

Rodriguez-Fuentes, L. J., Varona, A., Diez, M., Penagarikano, M., and Bordel, G. (2012b). Evaluation of Spoken Language Recognition Technology Using Broadcast Speech: Performance and Challenges. In *Odyssey 2012: The Speaker and Language Recognition Workshop*, Singapore.

Rodriguez Fuentes, L. J., Brummer, N., Penagarikano, M., Varona, A., Diez, M., and Bordel, G., (2012c). *The Albayzin 2012 Language Recognition Evaluation Plan (Albayzin 2012 LRE)*. URL: http://iberspeech2012.ii.uam.es/images/PDFs/albayzin_lre12_evalplan_v1.3_springer.pdf.

Rodriguez-Fuentes, L. J., Penagarikano, M., Varona, A., Diez, M., and Bordel, G. (2013). The Albayzin 2012 Language Recognition Evaluation. In *Proceedings of Interspeech*, pages 1497–1501.

Schwarz, P. (2008). *Phoneme recognition based on long temporal context*. Ph.D. thesis, Faculty of Information Technology, Brno University of Technology.

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