Verification of the four Spanish official languages on TV show recordings *

Verificación de las cuatro lenguas oficiales españolas en grabaciones de programas de televisión

A. Varona, M. Penagarikano, L.J. Rodriguez-Fuentes, M. Diez, G. Bordel
University of the Basque Country, UPV/EHU
GTTS, Department of Electricity and Electronics
amparo.varona@ehu.es

Resumen: En este trabajo se presentan resultados de verificación sobre las cuatro lenguas oficiales españolas: castellano, catalán, euskera y gallego. Se analizan los resultados obtenidos en tests cerrados y abiertos (estos últimos incluyendo segmentos en frances, portugés, aleman o inglés) y considerando segmentos de voz de 30 segundos. Se realiza también un estudio detallado del rendimiento del sistema por cada lengua objetivo. Se usa la base de datos *KALAKA* creada especialmente para la *Evaluación Albayzín 2008 de sistemas de verificación de la lengua*.
El sistema de verificación principal resulta de la fusión de un sistema acústico y 6 subsistemas fonotácticos. El sistema acústico toma información de las características espectrales de la señal de audio, mientras que los sistemas fonotácticos utilizan secuencias de fonemas producidas por varios decodificadores acústicos. En este trabajo se alcanza una tasa EER= 3,58 % y un coste C_{LLR}= 0.30 en test cerrado, lo que implica una mejora relativa del 24,5 % con respecto a los mejores resultados obtenidos en la evaluación *Albayzin 2008 VL*.
Palabras clave: Verificación de la lengua, Gaussian Mixture Models, Support Vector Machines

Abstract: This paper presents language recognition results obtained for the four official Spanish languages: Spanish, Catalan, Basque and Galician. Results were obtained in closed and open tests (these latter including segments in French, Portuguese, German or English) on a subset of 30 second segments. A detailed study per target language is also included. Experiments were carried out on the *KALAKA* database, especially recorded for *The Albayzin 2008 Language Recognition Evaluation*.
The main verification system resulted from the fusion of an acoustic system and 6 phonotactic subsystems. To model the target language, the acoustic subsystem takes information from the spectral characteristics of the audio signal, whereas phonotactic subsystems use sequences of phones produced by several acoustic-phonetic decoders. The best fused system attained a 3,58 % EER and C_{LLR}= 0.30 in closed tests, which means 24,5 % improvement with regard to the best result obtained in the *Albayzin 2008 LRE*.
Keywords: Language Verification/Recognition, Gaussian Mixture Models, Support Vector Machines

1. Introducction

As for the National Institute of Standards and Technology (NIST) evaluations (Martin and Le, 2008), the language detection task can be stated as follows: *given a segment of speech and a language of interest (target language), determine whether or not that language is spoken in the segment, based on an automated analysis of the data contained in the segment*. Performance is computed by presenting the system a set of trials. Each trial comprises the following elements: (1) a segment of audio containing speech in a single language; (2) the target language; and (3) the non-target languages, that is, those languages that may be spoken in the segment. For each trial, the system must output: (1) a

* 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)

hard decision (yes/no) about whether or not the target language is spoken in the segment; and (2) a score indicating how likely is for the system that the target language is spoken in the segment, the higher the score the greater the confidence that the segment contains the target language.

In NIST Language Recognition Evaluations (LRE), test data included narrow-band (8kHz) segments for conventional recorded conversational telephone speech (in all LRE) as well as narrow-band segments from worldwide Voice of America broadcast (only in 2009 LRE). The number of target languages was 14 in NIST 2007 LRE and 23 in NIST 2009 LRE. In both evaluations, best results were obtained by the systems submitted by the Massachusetts Institute of Technology (MIT) Lincoln Laboratory (Torres-Carrasquillo et al., 2008), (Torres-Carrasquillo et al., 2010). Our experience in those evaluations has been very positive (Penagarikano et al., 2007) (Penagarikano et al., 2009). Further developments made on the NIST 2007 LRE have been published in (Penagarikano et al., 2010b) and (Penagarikano et al., 2010a).

But we are mainly interested in recognizing the Spanish official languages. In 2008, we organized and coordinated *the Albayzin 2008 Language Recognition Evaluation* which was held as part of the *5th Biennial Workshop on Speech Technology* (JTH, 2008). The Albayzin 2008 LRE was inspired by NIST 2007 LRE, but with only 4 target languages (Spanish, Catalan, Basque and Galician) and using wide-band audio signals (16kHz). For Albayzin 2008 LRE, the *KALAKA* database was recorded from TV shows (Rodriguez-Fuentes et al., 2010b). In this competition there were 4 participant groups and preliminary results showed the difficulty of the task despite having only four target languages (Rodriguez-Fuentes et al., 2010a).

In this work, a language recognition/verification system has been built based on the train and development sets of KALAKA and the materials implicitly used to built phone decoders (see subsection 2.2). The system consists of a hierarchical fusion of 7 individual subsystems. An acoustic ("low-level") subsystem and 6 phonotactic ("high-level") subsystems. To model each target language, the acoustic subsystem takes information from the spectral characteristics of the audio signal, whereas phonotactic subsystems use sequences of phones produced by three acoustic-phonetic decoders, developed by the Brno University of Thechnology (BUT) for Czech, Hungarian and Russian (Schwarz, 2008). As we shall see, these two types of language recognition systems (acoustic and phonotactic) provides complementary information and their fusion leads to best results.

The rest of the paper is organized as follows. Section 2 presents the language recognition/verification systems used in this work. Section 3 defines the measures used to evaluate language verification system performance. The KALAKA database is described in Section 4. Section 5 summarizes results attained on KALAKA in closed-set and open-set tests. Finally, conclusions are outlined in Section 6.

2. Language recognition/ verification technology

The main system consists of a hierarchical fusion of 7 individual subsystems, an acoustic subsystem, and 6 phonotactic subsystems.

2.1. The acoustic subsystem

For the acoustic subsystem, 7-2-3-7 SDC-MFCC were used as acoustic parameters and Gaussian Mixture Models (GMM) were used as acoustic models by means of the *Sautrela* toolbox (Penagarikano and Bordel, 2005). Then a Support Vector Machine (SVM) classifier is applied on the vector space defined by 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 using UBM parameters and stacked to construct the so called GMM supervectors which feed the SVM classifier (Campbell et al., 2006). The SVM was developed using *SVMtorch* (Collobert and Bengio, 2001)

2.2. The phonotactic subsystems

Phonotactic language recognizers exploit the ability of phone decoders to convert a speech utterance into a sequence of symbols containing acoustic, phonetic and phonological information. Models for target languages are built by decoding hundreds or even

thousands of training utterances and using the phone-sequence (or phone-lattice) statistics (typically, counts of n-grams) in different ways. The most common phonotactic approaches are the so called PPRLM (Parallel Phone Recognizers followed by Language Models) (Zissman, 1996), referred to as Phone-LM in this paper, and Phone-SVM (Support Vector Machines applied on counts of phone n-grams) (Campbell et al., 2006). In both cases, N phone decoders are applied to the input utterance, and each output i ($i \in [1, N]$) is scored for each target language j ($j \in [1, L]$), by applying the model $\lambda(i, j)$ (estimated using the outputs of the phone decoder i for the subset of the training database corresponding to language j). Scores for the subsystem i are calibrated, typically by means of a Gaussian backend. A t-norm (Auckenthaler, Carey, and Lloyd-Thomas, 2000) is applied before calibration. Finally, $N \times L$ calibrated scores are fused using the *FoCal* toolkit (FoCal, 2008). A linear logistic regression was applied, to get L final scores for which a minimum expected cost Bayes decision is taken, according to application-dependent language priors and costs (see (Brümmer and du Preez, 2006) for details). Figure 1 shows the structure of a phonotactic language recognizer.

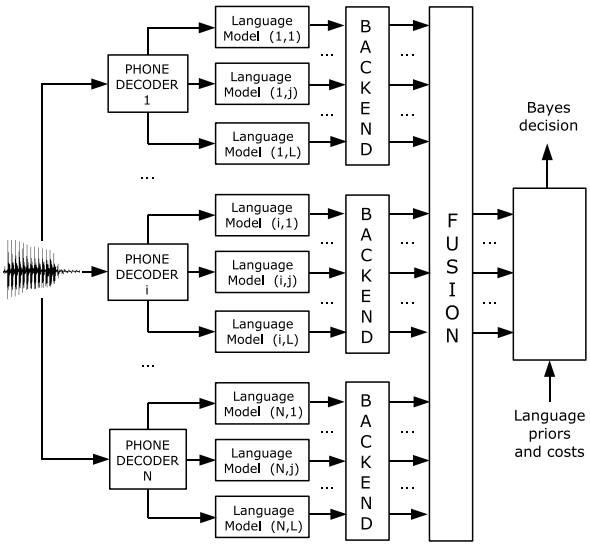

Figure 1: A phonotactic language recognition system.

In this work, 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.

BUT 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 LMs with Witten-Bell smoothing. It was used the SRI Language Model toolkit (Stolcke, 2002) to estimate phone sequence n-gram models.
- *Phone-SVM*: SVM built on bag-of-N-gram vectors (including up to 4-grams), weighted as proposed in (Richardson and Campbell, 2008). It was used the *libLinear* (Fan et al., 2008)

3. Performance measures

The language recognition task defined in this evaluation considers two types of errors: (1) *misses*, those for which the correct answer is *yes* but the system says *no*; and (2) *false alarms*, those for which the correct answer is *no* but the system says *yes*. Therefore, for any test condition the corresponding error rates can be computed as the fraction of target trials that are rejected (*miss rate*, P_{miss}) and the fraction of impostor trials that are accepted (*false alarm rate*, P_{fa}), and suitable cost functions can be defined as combinations of these basic error rates.

3.1. Graphical evaluation: DET curves

Detection Error Tradeoff (DET) curves (Martin et al., 1997) provide a straightforward way of comparing global performance

of different systems for a given test condition. A DET curve is generated by computing P_{miss} and P_{fa} for a wide range of operation points (thresholds), based on the scores yielded by the analyzed system for a given test set. DET curves are used in NIST evaluations to support system performance comparisons. In this work, DET curves were generated by means of NIST software.

3.2. Equal Error Rates

The most common performance measure is the Equal Error Rate (EER), which reports system performance when the false acceptance probability (P_{miss}) is equal to the missed detection probability (P_{fa}). EER is a very simple measure, useful in many context but it does not allow to compare the global performance of two systems.

3.3. Log-Likelihood Ratio average cost C_{LLR}

When scores represent (or can be interpreted) as log-likelihood ratios, it is possible to evaluate systems also in terms of the so called C_{LLR} (Brümmer and du Preez, 2006), which is used as an alternative performance measure in NIST evaluations. C_{LLR} shows two important features: (1) it allows to evaluate system performance globally by means of a single numerical value, which is somehow related to the area below the DET curve, provided that scores can be interpreted as log-likelihood ratios; and (2) C_{LLR} does not depend on application costs; instead, it depends on the calibration of scores, an important feature of detection systems. To compute C_{LLR}, the *FoCal* toolkit can be used (FoCal, 2008).

Let $LR(X, i)$ be the *likelihood ratio* corresponding to segment X and target language i. The likelihood ratio can be expressed in terms of the conditional probabilities of X with regard to the alternative target and non-target hypotheses, as follows:

$$LR(X, i) = \frac{prob(X|i)}{prob(X|\neg i)} \tag{1}$$

Let consider an evaluation set E, consisting of the union of $L + 1$ disjoint subsets: E_j ($j \in [1, L]$) containing segments in the target language j, and E_0 containing segments in *unknown* languages. Pairwise costs $C_{LLR}(i, j)$, for $i \in [1, L]$ and $j \in [0, L]$, are defined as follows:

$$C_{LLR}(i,j) = \begin{cases} \frac{1}{|E_i|} \sum_{X \in E_i} \log_2(1 + LR(X,i)^{-1}) & j = i \\ \frac{1}{|E_j|} \sum_{X \in E_j} \log_2(1 + LR(X,i)) & j \neq i \end{cases} \tag{2}$$

Finally, the average cost C_{LLR} is computed by adding the pairwise costs for all the combinations of target and non-target (including Out-Of-Set) languages, as follows:

$$\begin{aligned} C_{LLR} &= \frac{1}{L} \sum_{i=1}^{L} \{P_{target} \cdot C_{LLR}(i,i) \\ &+ \sum_{\substack{j=1 \\ j \neq i}}^{L} P_{non-target} \cdot C_{LLR}(i,j) \\ &+ P_{OOS} \cdot C_{LLR}(i,0)\} \end{aligned} \tag{3}$$

where P_{target} is the prior probability of target languages, $P_{non-target}$ is the prior probability of non-target languages and P_{OOS} is the prior probability of *unknown* (Out-Of-Set) languages. In this work, the same values used in the two last NIST LRE (2007 and 2009) are applied:

$$\begin{aligned} P_{OOS} &= \begin{cases} 0{,}0 & \text{closed-set} \\ 0{,}2 & \text{open-set} \end{cases} \\ P_{target} &= 0{,}5 \\ P_{non-target} &= \frac{1 - P_{target} - P_{OOS}}{L - 1} \end{aligned}$$

The cost function C_{LLR} returns an unbounded non-negative value which can be interpreted as information bits, with lower values representing better performance, the value 0 corresponding to a perfect system and the value $\log_2(L)$ corresponding to a system which just relies on (uniform) priors, thus providing no information to decide a trial. Further details about the reasons for using and the interpretation of C_{LLR} can be found in (Brümmer and du Preez, 2006; Brümmer and van Leeuwen, 2006).

4. The KALAKA database

The KALAKA speech database (Rodriguez-Fuentes et al., 2010b) 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. Due to the interaction between these languages, the

task of distinguishing them can be more difficult than expected.

KALAKA consists of wide-band (16kHz) segments extracted from 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 training set contains around 9 hours of speech per target language, which amounts to around 36 hours of training data.

Both development and evaluation data include utterances in target and *unknown* languages, so that closed-set and open-set evaluations can be carried out.

- The development dataset consists of 1800 speech segments, distributed in three subsets, each containing 600 segments with nominal durations of 30, 10 and 3 seconds, respectively. Each subset consists of 120 segments per target language and 120 additional segments from *unknown* languages (70 for French, 10 for Portuguese and 40 for Enghish).
- The evaluation dataset has the same structure, except for the distribution of non-target languages (10 for French, 70 for Portuguese and 40 for German).

Development and evaluation sets contains around 7.7 hours of speech each: more than 90 minutes of speech per target language and more than 90 minutes of speech for *unknown* languages all together.

5. Results

In this work, closed-set and open-set tests were carried out on the subset of 30-second speech segments of KALAKA. In closed-set verification, the set of trials is limited to segments containing speech in one of the target languages, and scores are computed based on those trials. In open-set verification, scores are computed based on the whole set of trials for a given test, including those corresponding to segments containing speech in an *unknown* (Out-Of-Set) language.

5.1. Closed-set evaluation

Table 1 shows results (EER and C_{LLR}) using various single language verification subsystems and systems resulting from different fusions. 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 subsystems) are shown in Figure 2.

Table 1: EER and C_{LLR} of single and fused language recognition systems on the closed-set evaluation subset of 30-second speech segments.

		EER	**C_{LLR}**
GMM-SVM (A)		16.11 %	0.96
Phone-LM (B)	CZ	16,08 %	0,94
	HU	13,19 %	0,80
	RU	14,17 %	0,86
	Fusion	7,53 %	0,49
Phone-SVM (C)	CZ	7,95 %	0,58
	HU	8,44 %	0,59
	RU	10,10 %	0,68
	Fusion	5,45 %	0,42
Partial Fusions	(A+B)	5,52 %	0,38
	(A+C)	4,06 %	0,35
	(B+C)	4,83 %	0,38
Fusion (A+B+C)		3,58 %	0,30

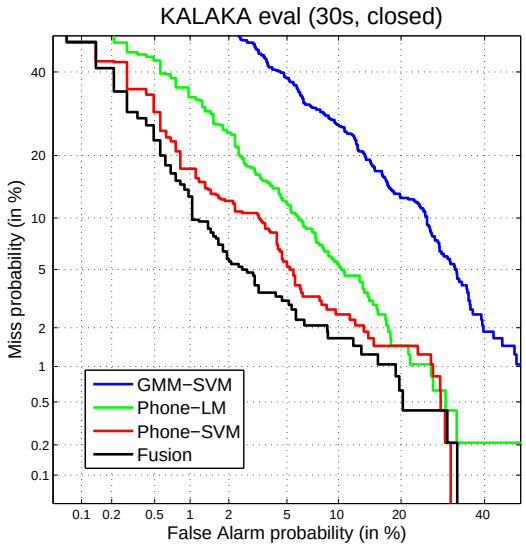

Figure 2: Pooled DET curves of various subsystems: GMM-SVM, Phone-LM, Phone-SVM and the fusion of all of them, on the closed-set evaluation subset of 30-second speech segments.

The low-level (acoustic) subsystem is clearly worse than the high-level (phonotactic) subsystems. But it can be seen that the fusion of both types of language recognition systems takes advantage from complementary information and leads to the best results. These are the main comments on these results:

- The acoustic GMM-SVM subsystem yields the worst result.
- Individual phone-LM subsystems yield quite poor results but their fusion was quite successful.
- Individual phone-SVM subsystems yield quite good results (specially CZ) and their fusion was very successful.
- The partial fusion of the acoustic subsystem with phone-ML provides around 26,7 % and 24,4 % of improvement in terms of EER and C_{LLR} respectively, with regard to results obtained exclusively with phone-ML.
- The partial fusion of the acoustic subsystem with phone-SVM provides around 25,5 % and 19 % of improvement in terms of EER and C_{LLR} respectively, with regard to results obtained with phone-SVM.
- The partial fusion between Phone-LM and Phone-SVM provides around 11,5 % of improvement with regard to the results obtained with phone-SVM.

The best result is obtained when the three systems are fused: EER= 3,58 % and C_{LLR}=0,30 (around 12 % of improvement with regard to the best partial fusion). The performance of this system is remarkably better than those of the most competitive systems submitted to the Albayzin 2008 LRE, yielding around 24,5 % relative improvements in terms of EER (Rodriguez-Fuentes et al., 2010a).

EER and DET curves are similar to those attained in NIST LRE, which deals with much more data and target languages (Penagarikano et al., 2010b). Since we are not comparing the *same systems* on two different tasks, but different systems on different tasks, we cannot extract conclusions. Anyway, these results may indicate that this task is, in fact, more difficult than expected, taking into account that we are dealing with wideband (good quality) speech signals and just 4 target languages. This difficulty may be due to the presence of various sources of variability (speakers, environment, channel, etc.) but more probably to the acoustic and lexical similarity among the target languages, which evolved jointly in different regions of the Iberian Peninsula, being Castilian Spanish the shared and most influential language.

Table 2 shows a detailed analysis of the behaviour of the main fused system for each target language (see DET curves per language in Figure 3). Note that, since the number of segments for each target language is 120, an error of 0,83 % means that only one segment is missed.

Table 2: Percentage of errors per target language (*miss probability* in the diagonal and *false alarm probability* out of the diagonal) for the closed-set tests, using the main fused system.

		Target			
		Spanish	Catalan	Basque	Galician
Segment	Spanish	8,33	1,67	5,00	15,00
	Catalan	0,83	0,83	1,67	0,83
	Basque	0,83	0.00	0,83	0,83
	Galician	12,50	4,17	0.00	4,17

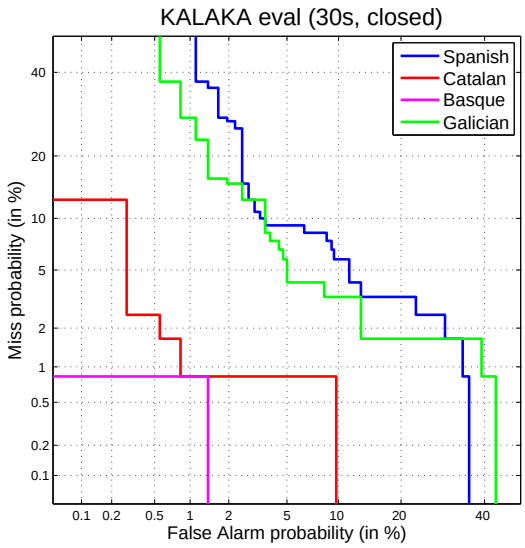

Figure 3: DET curves for target languages, using the main fused system in the closed-set test condition.

Clearly, system performance was not homogeneous when disaggregated for all the target languages. The best recognition performances were obtained for Basque and Catalan. In both cases, only one of the 120 target segments was missed and few segments were taken as false alarms for other target languages. On the other hand, it appears that Spanish and Galician were confused each other, also showing a significant miss rate.

5.2. Open-set evaluation

Results in open-set tests are presented in Table 3 and Figure 4. System performance is remarkably worse in open-set than in closed-set verification tests. But the relative behavior of single and fused systems is similar to that observed in closed-set verification tests.

- The acoustic GMM-SVM subsystem yields the worst result.
- Individual phone-LM subsystems yield quite poor results but their fusion was successful.
- Individual phone-SVM subsystems yield quite good results (specially CZ) and their fusion was quite successful.
- The partial fusion of the acoustic subsystem with phone-ML provides around 18 % and 11 % of EER and C_{LLR} improvements, respectively.
- The partial fusion of the acoustic subsystem with phone-SVM provides around 7 % and 5 %of EER and C_{LLR} improvements, respectively.

As in the closed-set tests experiments, low-level and high-level language recognition systems provide complementary information and their fusion leads to the best results.

The performance of the main fused system (EER= 9,23 %) is similar to those of the most competitive systems submitted to the Albayzin 2008 LRE (Rodriguez-Fuentes et al., 2010a). The increase of EER with regard to the best partial fusion (EER=8,33 %) is due to a local effect (as can be observed in the shape of the DET curves). But, as noted in subsection 3.3, C_{LLR} allows us to evaluate the system globally and it reflects a 1,3 % improvement with regard to the best partial fusion.

Table 4 shows the detailed analysis of the behavior of the main fused system for each target language (see DET curves in Figure 5). The best recognition performance was obtained for Basque, whereas performance was quite worse for the three other target languages.

In the open-set tests, the presence of impostor trials with *unknown* languages had little impact in the performance for Basque (which yielded again the best performance), whereas the impact was quite remarkable for

Table 3: Performance (EER and C_{LLR}) of single and fused language recognition systems on the open-set evaluation subset of 30-second speech segments

		EER	C_{LLR}
GMM-SVM (A)		20,04 %	1,51
Phone-LM (B)	CZ	18,81 %	1,42
	HU	18,35 %	1,33
	RU	17,48 %	1,44
	Fusion	13,75 %	1,10
Phone-SVM (C)	CZ	11,83 %	1,03
	HU	12,71 %	1,08
	RU	14,23 %	1,12
	Fusion	8,96 %	0,82
Partial Fusions	(A+B)	11,27 %	0,98
	(A+C)	8,33 %	0,78
	(B+C)	9,48 %	0,82
Fusion (A+B+C)		9,23 %	0,77

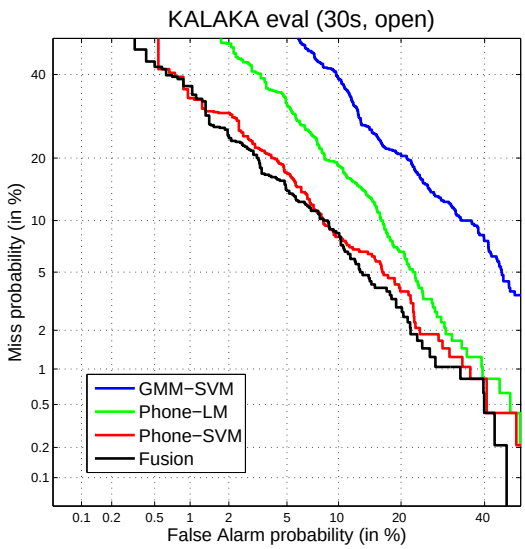

Figure 4: Pooled DET curves of various systems: GMM-SVM, Phone-LM, Phone-SVM and the system fusing all of them, on the open-set evaluation subset of 30-second speech segments.

Catalan. Note the high cost associated to Catalan for impostor trials with *unknown* languages in the open-set tests.

The high performance for Basque may be due to the different origins of Basque with regard to the other target languages (which are Romance languages). Basque has been influenced by Romance languages (specially by Spanish and French), but has completely different roots, and its lexicon is quite different from those of the other languages appearing in KALAKA. On the other hand, the high confusion of Catalan (and at a lower degree,

also of Galician) with the *unknown* languages may be due to its similarity to French or Portuguese (note that all of them are Romance languages).

Table 4: Percentage of errors per target language (*miss probability* in the diagonal and *false alarm probability* out of the diagonal) for the open-set tests, using the main fused system.

		Target			
		Spanish	Catala	Basque	Galician
Segment	Spanish	18,33	1,67	2,50	10,83
	Catalan	0.00	11,67	0.00	0.83
	Basque	0.00	0.00	1,67	0.00
	Galician	15,83	2,50	0.0	6,67
	Unknown	3,33	35,00	11,67	21,67

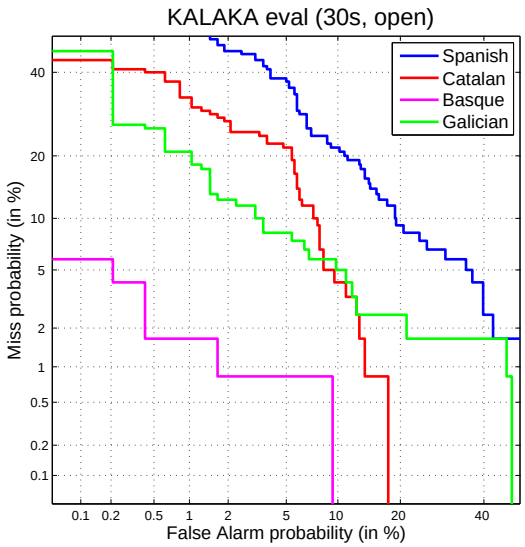

Figure 5: DET curves for target languages, using the main fused system in the open-set test condition.

6. Conclusions

In this work, various systems were evaluated on a language verification task defined on the four Spanish official languages (Spanish, Catalan, Basque and Galician. Closed-set and open-set tests (these latter including segments in French, Portuguese, German or English) were performed on subsets of 30-seconds speech segments obtained from TV shows.

The main verification/recogniton system resulted from the hierarchical fusion of an acoustic subsystem and 6 phonotactic subsystems. To model each target language, the acoustic subsystem used spectral characteristics of the audio signal whereas phonotactic subsystems used sequences of phones produced by acoustic-phonetic decoders. The best fused system attained 3,58 % EER and C_{LLR}= 0.30 in closed-set tests (a 24,5 % improvement with regard to previous results). Performance was remarkably worse (9,23 % EER and C_{LLR}= 0.77) in open-set tests.

Finally, system performance was not homogeneous when disaggregated for target languages. In particular, the best recognition performance was obtained for Basque, which may be due to the different origins of Basque with regard to the other target languages.

References

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

Brümmer, N. and J. A. du Preez. 2006. Application-independent evaluation of speaker detection. *Computer Speech & Language*, 20(2-3):230–275.

Brümmer, N. 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.

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

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

Fan, Rong-En, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. LIBLINEAR – A Library for Large Linear Classification. *The Journal of Machine Learning Research*, 9:1871–1874, June.

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

JTH, 2008. *5th Biennial Workshop on Speech Technology*. Bilbao, Spain, 12-14 November. http://jth2008.ehu.es/en/index.html.

Martin, A., 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.

Martin, A.F. 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.

Matejka, P., 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.

Penagarikano, M., A.Varona, M. Zamalloa, L.J. Rodriguez, G. Bordel, and J. P. Uribe. 2009. University of the Basque Country + Ikerlan System for NIST 2009 Language Recognition Evaluation. In *2009 NIST Language Recognition Evaluation (LRE) Workshop*, Baltimore, MD, USA.

Penagarikano, M. 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.

Penagarikano, M., G. Bordel, L.J. Rodriguez, and J. P. Uribe. 2007. University of the Basque Country + Ikerlan System for NIST 2007 Language Recognition Evaluation. In *2007 NIST Language Recognition Evaluation (LRE) Workshop*, Orlando, Florida, USA.

Penagarikano, M., A. Varona, L.J. Rodriguez-Fuentes, and G. Bordel. 2010a. Improved Modeling of Cross-Decoder Phone Co-occurrences in SVM-based Phonotactic Language Recognition. In *Odyssey 2010: The Speaker and Language Recognition Workshop*, Brno, Czech Republic.

Penagarikano, M., A. Varona, L.J. Rodriguez-Fuentes, and G. Bordel. 2010b. Using cross-decoder phone co-ocurrences in phonotactic language recognition. In *Proceedings of the 35th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2010)*, pages 5034–5037, Dallas, Texas (USA).

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

Rodriguez-Fuentes, L. J., M. Penagarikano, G. Bordel, and A. Varona. 2010a. The Albayzin 2008 Language Recognition Evaluation. In *Odyssey 2010: The Speaker and Language Recognition Workshop*, Brno, Czech Republic, 28 June - 1 July.

Rodriguez-Fuentes, L. J., 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)*, Valleta, Malta, 17-23 May.

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

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

Torres-Carrasquillo, P.A., 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.

Torres-Carrasquillo, P.A., E. Singer, T. Gleason, A. McCree, D.A. Reynolds, F. Richardson, and D.E. Sturim. 2010. The MITLL NIST LRE 2009 language recognition system. In *Proceedings of ICASS 2010*, pages 4994–4997.

Zissman, M.A. 1996. Comparison of four approaches to automatic language identification of telephone speech. *IEEE Transactions on Speech and Audio Processing*, 4(1):31–44, January.