Multi-site Heterogeneous System Fusions for the Albayzin 2010 Language Recognition Evaluation

Luis Javier Rodríguez-Fuentes [1*], Mikel Penagarikano [1], Amparo Varona [1], Mireia Díez [1], Germán Bordel [1], David Martínez [2], Jesús Villalba [2], Antonio Miguel [2], Alfonso Ortega [2], Eduardo Lleida [2], Alberto Abad [3], Oscar Koller [3], Isabel Trancoso [3,4], Paula Lopez-Otero [5], Laura Docio-Fernandez [5], Carmen Garcia-Mateo [5], Rahim Saeidi [6], Mehdi Soufifar [7], Tomi Kinnunen [6], Torbjørn Svendsen [7], Pasi Fränti [6]

[1] *GTTS, Department of Electricity and Electronics, University of the Basque Country, Spain*
[2] *ViVoLab, Aragon Institute for Engineering Research (I3A), University of Zaragoza, Spain*
[3] *L²F - Spoken Language Systems Lab, INESC-ID Lisboa, Portugal*
[4] *Instituto Superior Técnico, Lisboa, Portugal*
[5] *GTM, Department of Signal Theory and Communications, Universidade de Vigo, Spain*
[6] *School of Computing, University of Eastern Finland (UEF), Joensuu, Finland*
[7] *Department of Electronics and Telecommunications, NTNU, Trondheim, Norway*

[*] luisjavier.rodriguez@ehu.es

***Abstract*—Best language recognition performance is commonly obtained by fusing the scores of several heterogeneous systems. Regardless the fusion approach, it is assumed that different systems may contribute complementary information, either because they are developed on different datasets, or because they use different features or different modeling approaches. Most authors apply fusion as a final resource for improving performance based on an existing set of systems. Though relative performance gains decrease as larger sets of systems are considered, best performance is usually attained by fusing all the available systems, which may lead to high computational costs. In this paper, we aim to discover which technologies combine the best through fusion and to analyse the factors (data, features, modeling methodologies, etc.) that may explain such a good performance. Results are presented and discussed for a number of systems provided by the participating sites and the organizing team of the Albayzin 2010 Language Recognition Evaluation. We hope the conclusions of this work help research groups make better decisions in developing language recognition technology.**

I. Introduction

Best performance in spoken language recognition is commonly attained by combining the scores of different classifiers. Using complementary language cues (spectral, prosodic and phonotactic features) for the development of different systems makes it possible to effectively exploit such complementarity at the score level through discriminative fusion. However, this is just theory. In practice, one cannot predict which system combinations will provide best performance. In fact, the sets of speech signals used to estimate models and calibration parameters may be affecting the scores as much as the features themselves. On the other hand, if fusion parameters are estimated on a suitable (i.e. large enough and representative) set of data, it is expected that the fused system will perform better than the best single system. This explains why, regardless the features and the modeling approaches, the main system is almost always built by fusing *all* the available systems (not just those whose features are expected to be complementary), in an attempt to improve performance as much as possible. The only issue with this *put it all together and shake* approach, is that one ends up building a large number of systems, some of them very close to each other or even identical (differing only in the data they are trained on), expecting that fusion will make the work for us and leaving further analyses aside. Probably, we may have attained the same performance by developing and fusing a reduced (but carefully chosen) set of systems. For instance, previous work has shown that systems based on prosodic features, though yielding higher error rates than systems based on spectral or phonotactic features, can be excellent candidates for fusion [1]. In any case, important lessons can be learnt from studying how much each factor (features, training data, modeling approaches) affects fusion performance and which combinations provide best performance.

This paper aims to enrich the discussion on these issues, by evaluating the performance of many different fusions of heterogeneous subsystems (heterogeneous in terms of features, modeling approaches and/or training data) for a fixed language recognition task, defined on the core condition (closed-set, clean-speech, 30-second segments) of the Albayzin 2010 Language Recognition Evaluation [2]. Sites participating in the evaluation, as well as the organizing team, applied various subsystems to score segments in the development and evaluation datasets. Calibration and fusion were performed by means of *FoCal* [3], a well-known open-source package that is commonly applied for these tasks, for all the possible combinations of k subsystems ($k \in [1, 5]$). Though the conclusions drawn from this work might be somehow task-dependent, we expect that they provide guidelines (and time savings) for future developments.

The rest of the paper is organized as follows. Section II summarizes the most relevant information about the evaluation.

Section III describes the post-evaluation activity proposed to participants, including details about the calibration and fusion approaches. The main features of the subsystems submitted by each site are given in Section IV. Language recognition performance for the best fusions of k subsystems ($k \in [1, 5]$) is presented and discussed in Section V. Finally, conclusions are given in Section VI.

II. The Albayzin 2010 Language Recognition Evaluation

The Albayzin 2010 Language Recognition Evaluation (Albayzin 2010 LRE) was the second effort made by the Spanish/Portuguese speech and language processing communities for benchmarking language recognition technology, after a first edition in 2008 [4]. The task, test conditions and performance measures were defined in almost the same terms as for the last NIST LRE [5][6]. Six target languages were considered (Spanish, Catalan, Basque, Galician, Portuguese and English) and speech signals were extracted from multi-speaker TV broadcast recordings. Note that a test segment could contain speech from various speakers. This is a relevant difference with regard to NIST evaluations, whose data were extracted from telephone-channel two-speaker conversations, test segments containing speech from a single speaker.

Four different test conditions, depending on the set of non-target languages (closed-set vs. open-set) and the background conditions (clean vs. noisy speech), and three nominal segment durations (30, 10 and 3 seconds) were considered, leading to 12 different tracks. A core mandatory condition was defined: closed-set verification of 30-second segments containing clean-speech, on which the work presented in this paper focuses.

A. Performance measures

Language recognition performance is commonly computed by presenting the system a set of trials (each involving a speech segment and a target language) and comparing system decisions with the right ones (stored in a keyfile). For each trial, the system must output a hard decision (yes/no) about whether or not the target language is spoken in the segment, and a score (a real number) 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.

Performance was primarily measured by means of the average cost across target languages, C_{avg}, defined in the same way as for NIST evaluations. In fact, the NIST LRE scoring script was used, with minor changes needed to match the task and to add the identifiers of the 6 target languages considered in this evaluation (see [2] for details). The cost model consisted of three application parameters: C_{miss}, C_{fa} and P_{target}. The same values used in NIST 2007 and 2009 LRE were applied:

$$C_{\text{miss}} = C_{\text{fa}} = 1$$
$$P_{\text{target}} = 0.5$$

B. Data

Participants were allowed to use any available data and subsystems to build their systems. However, for better matching the acoustic conditions of test data, the organization provided two additional datasets, *train* and *development*, designed for training language models and tuning system parameters, respectively. Speech signals were extracted from TV broadcast recordings. The TV shows posted to train, development and test sets were forced to be disjoint, as an attempt to guarantee speaker independence. Audio signals were stored in WAV files (PCM, 16 kHz, single channel, 16 bits/sample).

The train dataset consisted of more than 10 hours of clean speech per target language (in some cases, almost 12 hours) and more than 2 hours (in some cases, more than 3 hours) of noisy/overlapped speech per target language, amounting to more than 82 hours (roughly, 80% corresponding to clean speech and 20% to noisy speech). The development and test datasets, each amounting to more than 21 hours (roughly, 70% corresponding to clean speech and 30% to noisy speech), contained speech segments with nominal durations of 30, 10 and 3 seconds, with at least 150 speech segments per target language and nominal duration.

C. Participation

Four teams submitted their systems to the evaluation:

- GTC-VIVOLAB from the University of Zaragoza (Spain),
- L^2F (Spoken Language Systems Lab) from INESC-ID Lisboa (Portugal),
- GTM from the University of Vigo (Spain), and
- UEF-NTNU from the University of Eastern Finland (UEF) at Joensuu (Finland) and the Norwegian University of Science and Technology (NTNU) at Trondheim (Norway).

A wiki was activated to improve communication and collaboration between sites and the organizing team. A more detailed description of the evaluation conditions (rules, file formats, schedule, etc.) and a brief discussion of the results can be found in [2].

III. The post-evaluation (fusion) activity

Just after releasing evaluation results, in the interim before the evaluation workshop, the fusion activity was proposed to participants, the information being provided through the wiki, which was also used to upload the required scores and to publish the results. Somehow inspired by [7], the post-evaluation activity aimed to analyse the fusion of heterogeneous systems for a language recognition task, trying to find performance dependence on different factors (features, train/development data, modeling approach, etc.). In other words, our aim was to provide evidence (if any) about which factors affected fusion performance in language recognition, to help future developments. A well-established fusion approach was chosen, based on the *FoCal* Multiclass toolkit [3], developed by Niko Brümmer and widely used in the research community for the

calibration and fusion of speaker and language recognition systems.

Submissions to the Albayzin 2010 LRE were, in most cases, the result of fusing various subsystems. For this study, we asked each site for the scores (preferably log-likelihoods) yielded by subsystems on two fixed datasets: (1) the subset of the development dataset corresponding to closed-set trials involving 30-second clean-speech segments (i.e. the core condition); and (2) the test dataset. Sites also provided details about features, training data and the basic modeling methodology applied for each subsystem.

The submitted scores, which were expected to be (and were interpreted as) log-likelihoods, were not applied any normalization nor backend before using them for fusions. Let us consider the score corresponding to subsystem j for the input utterance x and target language t: $s_j(x,t)$. Calibration and fusion of k subsystems can be simultaneously performed by computing the fused scores $s_f(x,t)$ as a linear combination of subsystem scores, as follows [7]:

$$s_f(x,t) = \sum_{j=1}^{k} \alpha_j s_j(x,t) + \beta_t \qquad (1)$$

The weights α_j ($j \in [1,k]$) and β_t ($t \in [1,L]$) are estimated in a discriminative way, using *FoCal*, applying linear logistic regression to minimize the so called C_{LLR} function, which measures the information provided by the scores with regard to taking decisions based on the prior alone, so that $C_{LLR} = 0$ means a perfect system (no information loss), $C_{LLR} = \log_2 L$ corresponds to a well calibrated system whose scores provide no information, and intermediate values (the most common situation) represent a useful recognizer whose scores do help taking good decisions.

Note that Equation 1 is useful not only for calibrating and fusing the scores of two or more subsystems, but also for calibrating the scores of a single subsystem. Therefore, Equation 1 is applied for all k (including $k = 1$). Fusion parameters were optimized on the scores submitted for the subset of segments corresponding to the core condition in the development set.

IV. Sub-systems

The participating sites submitted scores for 20 subsystems. Scores for three additional subsystems, developed according to the evaluation setup, were added by the organizing team (GTTS, University of the Basque Country) to enrich the analyses. In the following paragraphs, the most relevant features of all the subsystems are outlined.

A. GTC-VIVOLAB subsystems

GTC-VIVOLAB sent log-likelihood ratios for three acoustic subsystems, the first one based on Maximum Likelihood (ML) 2048-mixture Gaussian Mixture Models (GMM), the second on Maximum Mutual Information (MMI) 2048-mixture GMM, and the third on Joint Factor Analysis (JFA, following [8][9], featuring channel compensation and linear scoring) applied on a high-dimensional GMM supervector space, obtained by Maximum-A-Posteriori (MAP) adaptation of a Universal Background Model (UBM). Acoustic features were, in all cases, Mel-Frequency Cepstral Coefficients (MFCC) with cepstral mean normalization, concatenated to their Shifted Delta Cepstra Coefficients (SDC) under a 7-1-3-7 configuration [10], yielding a 56-dimensional acoustic vector.

GTC-VIVOLAB also sent scores for five Phone Recognizer followed by Language Model (PRLM) subsystems, using a Spanish phone recognizer developed by themselves, based on GMM/HMM and conventional MFCC, and the Brno University of Technology (BUT) phone recognizers for English, Czech, Hungarian and Russian, based on ANN/HMM and Temporal Patterns (TRAPS) with Split Temporal Context (STC) [11]. The recognized phone sequences/lattices were used to train an n-gram (4-gram in most cases) language model for each target language (by means of the SRILM tool [12]) and to score test utterances.

Training and development data were limited to those provided for the Albayzin 2010 LRE, except for the phone recognizers, which were trained on different (ortographically transcribed) databases. A more detailed description of GTC-VIVOLAB subsystems and the results obtained in the Albayzin 2010 LRE can be found in [13].

B. L^2F subsystems

The L^2F team sent scores for five PRLM subsystems with n-gram language modeling, using five different phone recognizers for African Portuguese (AF), Brazilian Potuguese (BR), American English (EN), European Spanish (ES) and European Portuguese (PT), all of them based on multistream MultiLayer Perceptron (MLP) phone classifiers. MLPs were trained on different amounts of Broadcast News (BN) transcribed data. Acoustic features included Perceptual Linear Prediction features (PLP, 13 static + first derivative), PLP with log-RelAtive SpecTrAl speech processing features (PLP-RASTA, 13 static + first derivative) and Modulation SpectroGram features (MSG, 28 static). The phonotactics of each target language for each type of speech (clean/noisy) was modeled with a 3-gram backoff model (using Witten-Bell discounting), by means of the SRILM toolkit [12].

L^2F also sent scores for a Gaussian supervector (GSV) subsystem [14], using SDC features under a 7-1-3-7 configuration, based on PLP features with RASTA filtering and mean and variance normalization. A 256-mixture UBM was trained with approximately 9 hours of randomly selected clean speech (including 1.5 hours per target language). Support Vector Machines (SVM) with linear kernel were trained as target language models using the LibLinear implementation of the libSVM tool [15]. For each target language and each type of speech (clean/noisy), all the training segments of that language were used as positive samples and the remaining segments as negative samples.

In all cases, language models for clean and noisy speech were trained, and a Gaussian backend was trained and applied to 12-dimensional score vectors to get 7 log-likelihoods: 6

for target languages and one for out-of-set languages. In this work, only the six log-likelihood values corresponding to target languages were used.

C. GTM subsystems

The original submission of GTM to the Albayzin 2010 LRE already consisted of two individual subsystems based on two techniques, *Fishervoices* and *NMFvoices*, that had been previously applied to speaker identification [16]. Both techniques aimed to get a reduced and discriminative subspace through linear transformations of the original feature space. In both cases, each utterance was initially represented by a matrix of dimension $m \times n$, containing the means of the GMM obtained by Bayesian adaptation of a UBM to that utterance, m being the number of Gaussians and n the dimension of the feature space. Acoustic features consisted of 12 MFCC, the normalized log-energy and their delta and acceleration coefficients, with mean and variance normalization. Both for Fishervoices and NMFvoices, the decision about the target language in a test utterance was taken by searching for the closest training utterance (i.e. that yielding the lowest euclidean distance) in the reduced subspace. System development relied exclusively on the data provided for the Albayzin 2010 LRE.

D. UEF-NTNU subsystems

The UEF-NTNU team sent the scores for one phonotactic and three acoustic subsystems. The phonotactic subsystem was based on the BUT English phone recognizer (16 kHz, trained on TIMIT), and applied a Phone Recognizer followed by Vector Space Modeling (PR-VSM) approach [17]. Counts of unigrams and bigrams were used, and 300 dimensions were retained after SVD. For classification, two GMMs were used for target and non-target scores.

The acoustic subsystems were developed under three different approaches. The Generalized Linear Discriminant Sequence (GLDS) - SVM subsystem, based on [18], used 49-dimensional SDC features, a third order polynomial expansion and a neural network with one hidden layer as language classifier. The MMI-GMM subsystem, based on [19], used 49-dimensional SDC features and 256-mixture MMI trained (3 iterations) language-dependent GMMs. Finally, the HMM-VSM approach, an alternative to Gaussian tokenizers commonly used under the VSM approach [17], trained a HMM on the sequences of labels obtained using a 128-mixture UBM as tokenizer. The HMM was then used as event recognizer before applying a VSM classifier.

The PR-VSM and HMM-VSM subsystems were applied a Gaussian backend, whereas a neural network was employed as backend for the GLDS-SVM and MMI-GMM subsystems.

E. GTTS subsystems

The organizing team provided the scores for 3 phonotactic PR-SVM subsystems, following [18]. BUT phone recognizers for Czech, Hungarian and Russian, along with HTK [20], were used to produce phone lattices. Lattices, which encode multiple hypotheses with acoustic likelihoods, were then used to produce expected counts of phone n-grams (up to 3-grams), which fed a discriminative classifier based on SVM, using LibLinear [15]. Except for the phone recognizers (trained on different databases), model estimation and parameter tuning were performed on the data provided for the Albayzin 2010 LRE. A more detailed description of GTTS phonotactic systems (applied to a different task) can be found in [21].

V. Results

First we analyse the performance of individual subsystems. As explained in Section III, subsystem scores were calibrated by means of *FoCal*. As shown in Table I, the C_{avg} ranged from 0.0207 to 0.5033, most subsystems yielding values lower than 0.1. Note that C_{LLR} and C_{avg} on the test set were highly correlated, with few differences in ranking, whereas the C_{LLR} on the test set did not correlate well with that obtained on the development set. Also, though calibration parameters were adjusted in order to minimize C_{LLR} on the development set, subsystems yielding the lowest C_{LLR} on the development set (L2F-PRLM-ES, GTC-JFA and L2F-GSV) did not perform as well on the test set (the case of L2F-GSV is remarkable). This may reveal overtraining on the development set for some subsystems.

TABLE I
PERFORMANCE OF SUBSYSTEMS ON THE CORE CONDITION (CLOSED-SET, CLEAN-SPEECH, 30-SECOND SEGMENTS), IN TERMS OF C_{LLR} ON THE DEVELOPMENT AND TEST SETS, AND C_{avg} ON THE TEST SET. SUBSYSTEMS ARE RANKED ACCORDING TO C_{avg} (BEST FIRST).

id	Subsystem	C_{LLR} (dev)	C_{LLR} (eval)	C_{avg} (eval)
1	GTTS-PRSVM-CZ	0.23853	0.20643	0.0207
2	GTTS-PRSVM-HU	0.18495	0.22897	0.0233
3	GTC-JFA	0.11387	0.23281	0.0244
4	GTTS-PRSVM-RU	0.30046	0.23346	0.0265
5	L2F-PRLM-ES	0.07893	0.29615	0.0273
6	L2F-PRLM-EN	0.25598	0.29991	0.0326
7	L2F-PRLM-BR	0.24360	0.33552	0.0365
8	L2F-PRLM-PT	0.20721	0.32241	0.0389
9	GTC-PRLM-HU	0.45970	0.36321	0.0423
10	GTC-PRLM-RU	0.47957	0.41525	0.0477
11	GTC-PRLM-CZ	0.53160	0.41928	0.0495
12	L2F-GSV	0.16201	0.47611	0.0496
13	GTC-MMI	0.32480	0.49063	0.0497
14	L2F-PRLM-AF	0.24453	0.51002	0.0539
15	GTC-ML	0.47741	0.62041	0.0705
16	UEF-NTNU-GLDS	0.77615	0.88410	0.0969
17	UEF-NTNU-MMI	0.89888	0.88936	0.0976
18	UEF-NTNU-HMM	0.90744	0.84839	0.1130
19	GTC-PRLM-ES	0.80570	1.15025	0.1324
20	GTM-Fisher	1.28786	1.31241	0.1696
21	GTC-PRLM-EN	1.19591	1.98024	0.2375
22	GTM-NMF	1.67096	2.03993	0.2680
23	UEF-NTNU-PRVSM-EN	0.88437	5.87520	0.5033

Most subsystems followed the phonotactic approach and were based on hybrid ANN/HMM phone decoders, trained on different amounts and types of data, using MFCC and/or PLP-RASTA features. Only one of the 10 best subsystems followed an acoustic approach: the Joint Factor Analysis subsystem by GTC-VIVOLAB (GTC-JFA) —the best of those originally submitted to the evaluation—, based on SDC features and trained only on the data provided for the Albayzin 2010 LRE.

Table II shows the 5 fusions of k subsystems yielding the lowest C_{avg} on the test set, for $k \in [2, 5]$. C_{LLR} performance on the development and test sets is shown too. First, note that best fusions are built around best subsystems (those appearing in the 5 first positions in Table I), specially for $k = 2$ and $k = 3$. This result was expected, since the fused system should perform better than component subsystems, at least on the development set (this was not always the case on the test set).

TABLE II
RANKING OF THE 5 BEST FUSIONS OF k SUBSYSTEMS, FOR $k \in [2, 5]$, IN TERMS OF C_{avg} FOR THE CORE CONDITION ON THE TEST SET. NUMBERS IN PARENTHESES REFER TO SUBSYSTEM IDENTIFIERS IN TABLE I.

k	**Fusion**	C_{LLR} **(dev)**	C_{LLR} **(eval)**	C_{avg} **(eval)**
2	(3)+(5)	0.02662	0.12151	0.0094
	(1)+(5)	0.03970	0.12304	0.0095
	(1)+(4)	0.15605	0.14120	0.0117
	(1)+(3)	0.06293	0.13204	0.0120
	(2)+(3)	0.05516	0.11722	0.0120
3	(3)+(5)+(6)	0.02066	0.10831	0.0066
	(1)+(5)+(13)	0.03219	0.10546	0.0074
	(1)+(3)+(5)	0.02043	0.09043	0.0078
	(3)+(5)+(10)	0.02031	0.12137	0.0078
	(3)+(5)+(15)	0.01887	0.15120	0.0081
4	(1)+(5)+(10)+(13)	0.02707	0.11011	0.0059
	(2)+(3)+(5)+(15)	0.01550	0.11466	0.0059
	(3)+(5)+(6)+(9)	0.01926	0.09759	0.0065
	(1)+(5)+(7)+(13)	0.03193	0.10255	0.0066
	(3)+(5)+(6)+(8)	0.02065	0.10780	0.0066
5	(2)+(3)+(5)+(9)+(15)	0.01430	0.09723	0.0054
	(1)+(5)+(6)+(10)+(13)	0.02573	0.10273	0.0057
	(3)+(5)+(9)+(13)+(15)	0.01557	0.11427	0.0057
	(1)+(2)+(3)+(5)+(15)	0.01254	0.08527	0.0058
	(1)+(5)+(10)+(11)+(13)	0.02482	0.09978	0.0058

Note that *all the possible fusion combinations* have been estimated and evaluated, which is quite tedious. To get the best fusion of k subsystems, an incremental procedure may have been applied by adding a new subsystem to the best fusion of $k - 1$ subsystems. However, attending to Table II, this would generally lead to suboptimal fusions: (1)+(5) for $k = 2$, (1)+(5)+(13) for $k = 3$, (1)+(5)+(10)+(13) for $k = 4$ —this is optimal—, and (1)+(5)+(6)+(10)+(13). In any case, fusions obtained this way would be close to optimal and the procedure would be much faster than an exhaustive search. Alternatively, if we added a new subsystem to, say, the 5 best fusions of $k-1$ subsystems, the computational cost would not increase dramatically and we would very likely find the best fusion of k subsystems (at least, that would be the case in Table II).

With regard to the modeling approaches involved in best fusions, the PRLM-ES subsystem by L2F, in spite of being rank 5, appears in 17 (out of 20) fusions in Table II, thus revealing as the most *useful* subsystem. The Joint Factor Analysis subsystem by GTC-VIVOLAB (with 13 appearances) and the PRSVM-CZ subsystem by GTTS (with 10 appearances) can be also regarded as *useful* subsystems. Note that for $k = 2$, three of the 5 best fusions (including the best one) involve one acoustic and one phonotactic subsystem. For $k = 3$, the 4 best fusions involve one acoustic and two phonotactic subsystems and the fifth involves two acoustic and one phonotactic subsystems. Except for $k = 2$, all the fusions in Table II include at least one acoustic subsystem, which is quite relevant taking into account the prevalence of phonotactic subsystems in Table I. These results confirm the intuition that acoustic and phonotactic subsystems contribute complementary information. In particular, the tandem GTC-JFA + L2F-PRLM-ES ((3)+(5)) seems to be key in many configurations.

A more detailed analysis reveals that not only the use of different modeling approaches but also other factors may provide complementary information and lead to a high performance fusion. Let us focus on the second best fusion for $k = 2$: the PRSVM-CZ subsystem by GTTS and the PRLM-ES subsystem by L^2F follow very similar approaches, both based on phonotactic features, but they complement quite well. Besides the use of different classifiers (n-gram language models vs. SVM), other factors come to explain this result: the acoustic features on which phone recognizers rely, the phone inventory —which is language-dependent and may cover different (complementary) sounds— and the data used to train phone recognizers. The third best fusion for $k = 2$, involving the PRSVM-CZ and PRSVM-RU subsystems by GTTS, is even more illustrative, since both subsystems are identical except for the phone inventory and the data used to train phone recognizers.

Obviously, low-ranked subsystems are less likely to be part of best fusions, except for those that complement well some of the best subsystems. That is the case of GMM-ML (rank 15) and GMM-MMI (rank 13) subsystems by GTC, which appear in many of the best fusions for $k = 3, 4, 5$. The use of different features and different modeling approaches (acoustic systems in an ecosystem full of phonotactic approaches) may be providing complementary and useful information. Regarding this, it would have been nice to deal with acoustic subsystems not based on spectral features but on prosodic and/or articulatory features.

Up to this point, we have analysed a handful of fusions (those yielding best performance), trying to find evidences that explain why some configurations work better than others, but we still find it difficult to translate fusion performance into a rank of *useful* (good for fusion) subsystems. To that end, given a set of subsystems Γ, we now define the *average k-fusion performance* of subsystem $S \in \Gamma$, $\bar{C}_{LLR}(S, \Gamma, k)$, as the average C_{LLR} performance of fusions of k subsystems including S. The best subsystem according to this metric would be that providing, on average, the best fusions on Γ. Table III shows the 5 best subsystems according to $\bar{C}_{LLR}(S, \Gamma, k)$ computed on the test set, for $k \in [1, 5]$. Note that $\bar{C}_{LLR}(S, \Gamma, 1) = C_{LLR}(S)$.

Attending to Table III, the best subsystems match almost exactly those yielding best performance individually. The PRSVM-CZ subsystem by GTTS seems to be the more robust in providing high performance fusions. Note that the PRLM-ES subsystem by L^2F, which is part of the 5 best fusions for $k = 3, 4, 5$, does not even appear in Table III (it ranks sixth

TABLE III
RANKING OF SUBSYSTEMS ACCORDING TO THE AVERAGE k-FUSION PERFORMANCE ON THE TEST SET, FOR $k \in [1, 5]$.

(subsystem id) : $C_{LLR}(S, \Gamma, k)$				
$k = 1$	$k = 2$	$k = 3$	$k = 4$	$k = 5$
(1) : 0.2064	(1) : 0.1712	(1) : 0.1481	(1) : 0.1348	(1) : 0.1300
(2) : 0.2290	(2) : 0.1875	(3) : 0.1622	(3) : 0.1484	(2) : 0.1434
(3) : 0.2328	(3) : 0.1880	(2) : 0.1627	(2) : 0.1490	(3) : 0.1504
(4) : 0.2335	(4) : 0.2069	(6) : 0.1787	(6) : 0.1598	(6) : 0.1529
(5) : 0.2962	(6) : 0.2159	(4) : 0.1814	(4) : 0.1650	(4) : 0.1561

for $k = 2, 3, 4$ and eleventh for $k = 5$). This only means that its performance in fusion is quite sensitive to the subsystems it is combined with. As a result, the proposed measure seems useless, since the average behaviour it describes tells nothing about the maximum achievable performance, which is the key issue when deciding which subsystems would be worth developing and fusing.

VI. CONCLUSIONS

In this work, we have studied the performance of fusions of heterogeneous multi-site language recognition subsystems on the core task defined for the Albayzin 2010 LRE. Fusions of k subsystems (for $k \in [1, 5]$) have been estimated and applied by means of the well-established *FoCal* package. As expected, best fusions involved the best individual subsystems. In this study, an acoustic subsystem based on JFA and various phonotactic subsystems populated most of the best configurations. Despite the prevalence of phonotactic approaches (or maybe *because of that*), the JFA subsystem and two other acoustic subsystems revealed as key elements to get high-performance fusions. It was observed that subsystems based on different classifiers and/or different features may significantly contribute to fusion performance. Even low-ranked subsystems may take part in high-performance fusions if they contribute new complementary information to best individual subsystems. An average k-fusion performance measure was defined with the aim to rank subsystems according to their *usefulness* in fusions, but it did not identify some of the subsystems involved in best fusions. More work is needed to define suitable measures and effective ways of determining which kind of subsystems (features, modeling approaches, classifiers, etc.) would be worth developing to get optimal fusions. For now, developing a diverse ecosystem of subsystems (including acoustic and phonotactic approaches, short-time spectral and segmental features, different classifiers, etc.) seems to be the most secure way to high-performance fusions.

ACKNOWLEDGEMENTS

The work of GTTS was supported by the University of the Basque Country under grant GIU10/18, by the Government of the Basque Country under program SAIOTEK (project S-PE10UN87, partially financed by FEDER funds) and by the Spanish MICINN under Plan Nacional de I+D+i (project TIN2009-07446, partially financed by FEDER funds). The works of T. Kinnunen and R. Saeidi were supported by the Academy of Finland and Nokia Foundation, respectively.

REFERENCES

[1] R. W. Ng, C.-C. Leung, T. Lee, B. Ma, and H. Li, "Prosodic attribute model for spoken language identification," in *Proceedings of IEEE ICASSP*, Dallas, Texas, USA, March 14-19 2010, pp. 5022–5025.

[2] L. J. Rodriguez-Fuentes, M. Penagarikano, A. Varona, M. Diez, and G. Bordel, "The Albayzin 2010 Language Recognition Evaluation," in *Proceedings of Interspeech*, Firenze, Italia, August 28-31 2011.

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

[4] L. J. Rodriguez-Fuentes, M. Penagarikano, G. Bordel, and A. Varona, "The Albayzin 2008 Language Recognition Evaluation," in *Odyssey 2010: The Speaker and Language Recognition Workshop*, 2010.

[5] A. F. Martin and A. N. Le, "NIST 2007 Language Recognition Evaluation," in *Odyssey 2008 - The Speaker and Language Recognition Workshop, paper 016*, 2008.

[6] A. Martin and C. Greenberg, "The 2009 NIST Language Recognition Evaluation," in *Odyssey 2010 - The Speaker and Language Recognition Workshop, paper 030*, 2010.

[7] N. Brümmer, L. Burget, J. Cernocky, O. Glembek, F. Grezl, M. Karafiat, D. van Leeuwen, P. Matejka, P. Schwarz, and A. Strasheim, "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*, vol. 15, no. 7, pp. 2072–2084, 2007.

[8] P. Kenny, "Joint factor analysis of speaker and session variability: Theory and algorithms," CRIM, Tech. Rep. Technical Report CRIM-06/08-13, 2005, [Online]. Available: http://www.crim.ca/perso/patrick.kenny/.

[9] O. Glembek, L. Burget, N. Dehak, N. Brümmer, and P. Kenny, "Comparison of Scoring Methods Used in Speaker Recognition with Joint Factor Analysis," in *Proceedings of IEEE ICASSP*, Taipei, April 2009, pp. 4057–4060.

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

[11] P. Schwarz, "Phoneme recognition based on long temporal context," Ph.D. dissertation, Faculty of Information Technology, Brno University of Technology, http://www.fit.vutbr.cz/, Brno, Czech Republic, 2008.

[12] A. Stolcke, "SRILM - an extensible language modeling toolkit," in *Proceedings of ICSLP (Interspeech)*, November 2002, pp. 257–286.

[13] D. Martinez, J. Villalba, A. Miguel, A. Ortega, and E. Lleida, "I3A Language Recognition System for Albayzin 2010 LRE," in *Proceedings of Interspeech*, Firenze, Italy, 28-31 August 2011.

[14] W. M. Campbell, D. E. Sturim, and D. A. Reynolds, "Support vector machines using GMM supervectors for speaker verification," *IEEE Signal Processing Letters*, vol. 13, no. 5, pp. 308–311, 2006.

[15] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "LIBLINEAR: A library for large linear classification," *Journal of Machine Learning Research*, vol. 9, pp. 1871–1874, 2008, software available at http://www.csie.ntu.edu.tw/∼cjlin/liblinear.

[16] P. Lopez-Otero, L. Docio-Fernandez, and C. Garcia-Mateo, "A Fishervoice-based Speaker Identification System," in *Proceedings of FALA 2010: VI Jornadas en Tecnologia del Habla and II Iberian SLTech Workshop*, Vigo (Spain), 10-12 November 2010.

[17] H. Li, B. Ma, and C.-H. Lee, "A Vector Space Modeling Approach to Spoken Language Identification," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 1, pp. 271–284, January 2007.

[18] W. M. Campbell, J. P. Campbell, D. A. Reynolds, E. Singer, and P. A. Torres-Carrasquillo, "Support vector machines for speaker and language recognition," *Computer Speech and Language*, vol. 20, no. 2-3, pp. 210–229, April 2006.

[19] P. Matejka, L. Burget, P. Schwarz, and J. Cernocky, "Brno University of Technology System for NIST 2005 Language Recognition Evaluation," in *Proceedings of Odyssey 2006 - The Speaker and Language Recognition Workshop*, San Juan, Puerto Rico, 2006, pp. 57–64.

[20] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Lui, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK Book (for HTK Version 3.4)*, Cambridge, UK, 2006.

[21] M. Penagarikano, A. Varona, L. J. Rodriguez-Fuentes, and G. Bordel, "Dimensionality reduction for using high-order n-grams in SVM-based phonotactic language recognition," in *Proceedings of Interspeech*, Firenze, Italy, August 28-31 2011.