Odyssey 2012
The Speaker and Language Recognition Workshop
25-28 June 2012, Singapore

Fusing Language Information from Diverse Data Sources for Phonotactic Language Recognition

Mohamed Faouzi BenZeghiba, Jean-Luc Gauvain and Lori Lamel

Spoken Language Processing Group
LIMSI - CNRS B.P. 133 91403 ORSAY CEDEX FRANCE

Abstract

The baseline approach in building phonotactic language recognition systems is to characterize each language by a single phonotactic model generated from all the available language-specific training data. When several data sources are available for a given target language, system performance can be improved using language source-dependent phonotactic models. In this case, the common practice is to fuse language source information (i.e., the phonotactic scores for each language/source) early (at the input) to the backend. This paper proposes to postpone the fusion to the end (at the output) of the backend. In this case, the language recognition score can be estimated from well-calibrated language source scores.

Experiments were conducted using the NIST LRE 2007 and the NIST LRE 2009 evaluation data sets with the $30s$ condition. On the NIST LRE 2007 eval data, a C_{avg} of 0.9% is obtained for the closed-set task and 2.5% for the open-set task. Compared to the common practice of early fusion, these results represent relative improvements of 18% and 11%, for the closed-set and open-set tasks, respectively. Initial tests on the NIST LRE 2009 eval data gave no improvement on the closed-set task. Moreover, the C_{llr} measure indicates that language recognition scores estimated by the proposed approach are better calibrated than the common practice (early fusion).

1. Introduction

The baseline approach in phonotactic language recognition is to build a phonotactic model for each target language in the application. The phonotactic model is generated from phone n-gram statistics (counts) that represent the phonotactic characteristics which are considered to be language specific. These statistics are estimated from phone lattices generated from language training data using one or several phone recognizer(s). The key issue in this approach is the estimation of the phone n-gram statistics. If they are accurately estimated, a phonotactic system with state-of-the-art performances can be built. The accuracy of the estimation of the phone n-gram statistics can be improved using various techniques, including the use of more adequate train data [1], the use of better phone recognizers, better optimization of the phone lattice decoding, and parameter tuning for count estimation [3].

In practice, phonotactic language models are generated using different data sources. The data sources can be distinguished using different criteria, such as the nature of speech (conversational telephone speech or broadcast data), gender (male or female), dialect (e,g; Indian or American for the English language) and channel type. For an open-set task, the data from each language in the out-of-set languages can be considered as a different data source.

When several data sources are available for a specific language, the baseline approach is to merge all phone n-gram statistics from the different sources for a given language and generate a single phonotactic model characterizing the language. In this case, the fusion of the language source information is performed at the *phonotactic level*. It has been reported that for such training conditions, using source-dependent phonotactic models can improve system performance [4] [5]. In this latter approach, fusing language sources information is performed at the beginning of the back-end fusion module. This kind of fusion will be referred to as *within-language* fusion.

When the phonotactic system makes use of the Parallel Phone decoders followed by Language Modeling (PPRLM) approach, the language recognition score is a combination of the language/source scores estimated by each individual PRLM. This kind of language information fusion will be referred to as *between-prlm* fusion. Depending on the way the integration of the *within-language* and the *between-prlm* fusion techniques is performed, several language information fusion configurations can be envisaged.

This paper proposes a new language information fusion technique when multiple data sources are available for some target languages. In this technique, each language source is considered as a separate language and the final language source recognition score is estimated at the output of the fusion module. The final score for each target language is the simple average of the scores of its language sources. Language information is integrated using first *between-prlm* fusion followed by *within-language* fusion. This paper proposes the use of *an unsupervised labeling* technique to assign each segment in the language development data to its most likely source.

The rest of this paper is organized as follows: Section 2 describes the PPRLM system used in this work. Section 3 describes several language information fusion techniques. Section 4 describes the experimental set-up and provides an analysis of the obtained results.

2. System description

The language recognition system makes use of the Parallel Phone Recognizer followed by Language Modeling (PPRLM)i approach [6]. A block diagram of a baseline PPRLM system is shown in Figure (1).

The PPRLM system uses 3 context-dependent phone recognizers, for English, Spanish and French. They have 38, 36 and 27 phones, respectively. These recognizers are trained using Conversational Telephone Speech (CTS). The acoustic models are word-position independent, and trained on 25 hours

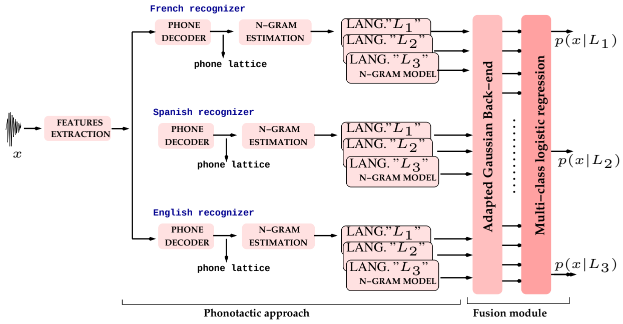

Figure 1: Block diagram of a baseline PPRLM system using three phone recognizers and three target languages.

for Spanish, 116 hours for French, and 1760 hours for English. Each model covers about 2000 phone contexts, with 2000 tied states and a mixture of 32 Gaussians per state. Silence is modeled by a single state, with a mixture of 1024 Gaussians. Constrained Maximum Likelihood Linear Regression (CMLLR) adaptation procedure is performed, prior the phone lattices decoding [2]. Phone lattice decoding [7] is done without any phonotactic constraints. The expected phone *n*-gram estimated from phone lattices are used to generate an interpolated Back-off *4*-gram phonotactic models with the Witten-Bell discounting using the SRILM toolkit.[1] The standard approach is for each individual PRLM and each target language, one phonotactic model is generated. Alternatively, a separate phonotactic model can be generated for each language source. In this case, each target language will be represented by one or multiple phonotactic models, depending on the available data sources. Both approaches are investigated in this work.

3. Language information fusion techniques

Language information fusion can be performed at different levels in a PPRLM system: at the *phonotactic*, the *within-language* and the *between-prlm* levels. This section describes several configurations to integrate these fusion techniques and estimate the language decision score. In all cases, language/source score fusion and calibration is performed using the widely used Gaussian back-end followed by logistic regression techniques [8]. These are implemented in the FoCal Multiclass toolkit[2]. For completeness these two fusion techniques are briefly overviewed before presenting the different configurations considered in this work.

3.1. Gaussian Backend (GB)

In this technique, language/source phonotactic scores of a given speech segment, are stacked in a phonotactic score vector. The dimension of this vector is equal to the total number of phonotactic models in the PPRLM system. The set of phonotactic score vectors associated with a given language/source are used to train a language/source dependent multivariate normal distribution $N(\mu_c, \Sigma)$ (one Gaussian). All Gaussians share a common full covariance matrix.

In [9], Gaussian back-end performances are significantly improved using *maximum a posteriori* (MAP) adaptation [10]. The mean vector of the class-dependent Gaussian was adapted from the mean vector of a background Gaussian:

$$\hat{\mu}_\ell = \alpha_\ell \mu_\ell + (1 - \alpha_\ell)\bar{\mu} \tag{1}$$

where μ_ℓ and $\hat{\mu}_\ell$ are the mean of the class-dependent Gaussian before and after adaptation, the vector $\bar{\mu}$ is the mean of the background Gaussian estimated from the mean of the other language/source dependent Gaussians. The parameter α_ℓ is defined as follows:

$$\alpha_\ell = \frac{n_\ell}{n_\ell + \tau} \tag{2}$$

where n_ℓ is the number of examples for the language/source ℓ and τ is the relevant adaptation factor optimized using k-fold cross-validation technique. Adapted Gaussian backend outperforms significantly conventional Gaussian backend in particular when the amount of development data is small.

3.2. Multi-class Logistic Regression (MLR)

As reported in [9] and will be shown in the results, if the amount of adequate development data is large enough then, further improvements can be obtained by calibrating language/source dependent Gaussian likelihoods (estimated by the Gaussian backend) using the discriminative multi-class logistic regression (MLR). This scheme for language score fusion and calibration is also used in [12].

3.3. Configurations for language information fusion

The fusion module in the complete PPRLM system consists of a GB sub-module followed by the MLR sub-module. The number of inputs and outputs of the MLR are equal to the size of the GB (i.e, the number of languages/sources represented in the GB). Depending on the modeling approach and the size of the GB, three language information fusion configurations were explored. Figure (2) shows these configurations for a simplified PPRLM system.

3.3.1. Configuration A

As shown in Figure (2), this configuration corresponds to the case where each target language is represented by a single phonotactic model for each individual PRLM component. This model is generated using phone *n*-gram counts merged from all language data sources, thereby fusing the language source information at the phonotactic level. In the PPRLM fusion module, each language is represented by one Gaussian and one output in the GB and the MLR fusion sub-modules, respectively.

In this configuration, the language recognition score is estimated by first fusing the phonotactic information followed by fusing language phonotactic scores (i.e, *between-prlms*).

3.3.2. Configuration B

This configuration corresponds to the case where each target language is represented by multiple source-dependent phonotactic models for each individual PRLM component as shown in Figure (2). However in the PPRLM fusion module, each language is represented by one Gaussian and one output in the GB and the MLR fusion sub-modules, respectively. Therefore configurations A and B have the same size of the Gaussian backend (i.e, the number of target languages) but they differ by the dimension of the phonotactic score vector.

In this configuration the language recognition score is estimated by simultaneously integrating the *within-language* and the *between-prlm* fusion techniques.

[1] http://www.speech.sri.com/projects/srilm/

[2] http://niko.brummer.googlepages.com/focalmulticlass

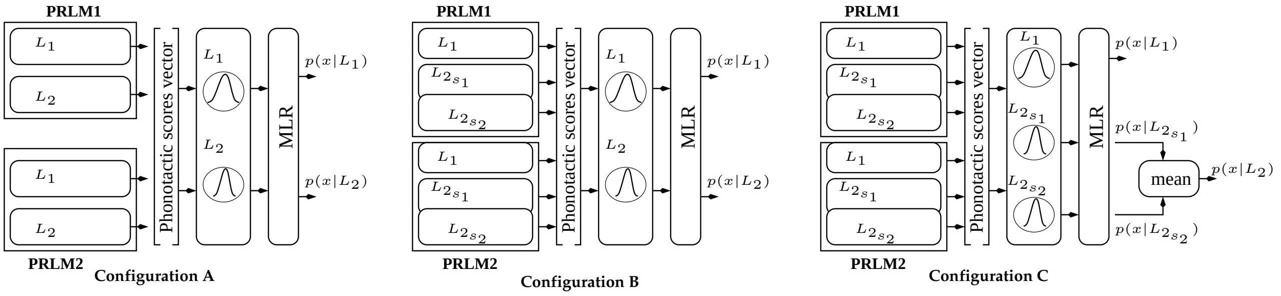

Figure 2: Three fusion module configurations compared in this work. The PPRLM is composed of two PRLMs (the corresponding phone recognizers are not shown for clarity purposes) and the set of target languages contains two languages L_1 and L_2. Language L_1 has one data source, so only one phonotactic model is generated for that language. Language L_2 has two data sources denoted $L_{2_{s_1}}$ and $L_{2_{s_2}}$. So, L_2 is represented by one or two phonotactic models depending on the modeling approach. These configurations differ by the size of the Gaussian backend (2 for configurations A and B, and 3 for configuration C) and the dimension of the phonotactic score vectors which is equal to 4 in configuration A and 6 in configurations B and C.

3.3.3. Configuration C

The third configuration shown in Figure (2) corresponds to the case where each target language is represented by one or multiple language source-dependent phonotactic models in each individual PRLM component. In the PPRLM fusion module, each language source is represented by one Gaussian and one output in the GB and the MLR fusion sub-modules, respectively. The language score $p(x|\ell)$ of a speech segment x can be estimated from language-source scores $p(y|\ell_s)$ as follows:

$$p(x|\ell) \simeq \sum_{s \in S_\ell} \lambda_s p(x|\ell_s) \tag{3}$$

where S_ℓ is the set of effectively modeled sources[3] for the language ℓ, and λ_s is the weight (representing prior knowledge) of the source ℓ_s.

In this configuration, the language recognition score is estimated by first using *between-prlm* fusion to estimate the language-source recognition scores followed by *within-language* fusion to produce the final language score.

From practical point of view, configuration C is very attractive, as it allows the integration of prior knowledge about language sources in the estimation of language recognition score. These priors are application dependent and can be set before the use of the system. For example if language sources correspond to dialects, the dialect scores can be weighted according to the origin of the audio files. Alternatively, these priors can be estimated dynamically during system use. In this case the weight λ_s will depend on the test segment. In this work, no prior knowledge is used so the weight λ_s is set to be equal to $\frac{1}{\|S_\ell\|}$. That is the language recognition score in (3) is equal to the average of the language source recognition scores.

The implementation of the configuration C requires the availability of sufficient amounts of development data for each modeled source. If development data for a given source is not available, then a phonotactic model for this source can be always built and its scores included in the phonotactic score vector, but this source will not be explicitly represented in the fusion module.

In practical situations, the languages of the development data are known but often the source is not known or may not represented in the training data. If the set of development languages are known, but the speech segments are not labeled, one option is to automatically assign labels to the data. This paper proposed to use *unsupervised data labeling* to address this issue as described in Section 4.2.3.

The difference between configurations B and C, is that in B all the dev data for a given target language is used to train one language-dependent Gaussian, while in C, the same amount of data is distributed over the modeled language sources to train multiple language source-dependent Gaussians (i.e, the number of parameters in the fusion module is increased). This splitting might affect system performance, in particular when the amount of dev data attributed to a language source is small.

4. Experimental set-up and results

The performance obtained with the three different language information configurations was evaluated using the $30s$ condition of the NIST LRE 2007[4] (lid07e1) and the NIST LRE 2009 (lid09e1) eval data sets. For the NIST LRE 2007, experiments are conducted for both closed-set and open-set tasks. For the NIST LRE 2009, only closed-set task is evaluated.

4.1. Task definition

The task of the interest is language detection. Given a speech segment x, the detection decision is made based on the language log likelihood ratio estimated as follows:

$$llr(x|\ell_k) \simeq \log \left[\frac{P_{\text{tar}}.p(x|\ell_k)}{P_{\text{oos}}.p(x|\ell_{\text{oos}}) + \sum\limits_{\substack{\ell_i \in L_T \\ \ell_i \neq \ell_k}} P_{\text{non-tar}}.p(x|\ell_i)} \right] \tag{4}$$

where $p(x|\ell)$ is the likelihood of x given the target language ℓ. It can be the output of the GB or the MLR. L_T is the set of target languages and oos represents the out-of-set languages. The target language *prior* P_{tar} is equal to 0.5. The out-of-set language *prior* P_{oos} is equal to 0.0 for the closed-set task and 0.2 for the open-set task[5]. The $P_{\text{non-tar}}$ is equal to:

$$P_{\text{non-tar}} = (1 - P_{\text{tar}} - P_{\text{oos}})/(\|L_T\|) \tag{5}$$

[3] Sources with sufficient amount of train and development data

[4] http://www.nist.gov/speech/tests/lang/2007/

[5] These values are given by NIST.

The *llrs* scores are then compared to the theoretical threshold $\Delta = 0$. Performance is reported in terms of C_{avg} as defined by NIST and the multi-class C_{llr} measure[6].

4.2. Experiments on the NIST LRE 2007

The training datasets used in these experiments are those provided by NIST. No external sources were used.

4.2.1. Evaluation data

In the NIST LRE 2007, there are 14 target languages, and about 2509 speech segments, of which 352 segments belong to one of the 5 out-of-set languages (French, Italian, Punjabi, Tagalog, Indonesian). Speech segments were mainly from the Fisher, Mixer, Callfriend and OGI corpora.

4.2.2. Training data

Three data sources are used to train language phonotactic models: the NIST LRE-96 train dataset (part of Callfriend CF database), the LRE-07 train data set and the OHSU database. Table (1) reports the amount of training data available per language and source. These amounts are computed after removing non-speech segments detected automatically by a speech activity detector.

4.2.3. Development data

The development data includes the NIST LRE 1996, 2003, 2005 dev and eval data, the NIST LRE-07 dev (lid07d1) data and the MITDev[7] data. In these data sets, the language sources of most segments are given and they match training data sources. For other segments, language source information was either missing or did not match one of the training data sources. To label these latter segments (i.e, associate each segment in the dev data to one of the train data sources), an *unsupervised labeling* procedure is applied. For each unlabeled segment x of a given language ℓ, a language source recognition score is estimated as the average language source phonotactic scores generated by each individual PRLM. The segment x is labeled with the source $\ell_s^\star$ having the highest score. Formally,

$$\ell_s^\star = \underset{\ell_s}{\text{argmax}} \frac{1}{N_d} \sum_{d=1}^{N_d} p_d(x|\ell_s) \tag{6}$$

where N_d is the number of PRLMs. Because this unsupervised labeling is not applied to all dev data segments, it will be referred to as *partially-unsupervised*.

As a contrast condition it was assumed that the language source information was unavailable for all dev segments (i.e. the provided language source information was not used). The labels were assigned automatically to all dev data segments using Equation (6). This labeling will be referred to as *fully-unsupervised*.

Table (2) reports the number of dev segments for each language source obtained using *partially* (left number) and *fully* (right number) unsupervised procedures.

For open-set task, additional out-of-set (OOS) data comprises 895 segments from 8 OGI-22 languages and French segments from LRE-96 and LRE-03 eval data sets are used. Three of these languages are also in the eval data. For this OOS data, the language source of the segment is the language itself of the segment.

4.3. Results and discussion

The configurations described in Section 3.3 were evaluated on a language detection task for closed- and open-set conditions.

4.3.1. Closed-set task

For this task, the decision score is estimated according to (4) with $P_{oos} = 0.0$. Table (3) reports the results for the 3 different fusion configurations. The baseline system corresponds to configuration A.

CONFIG	FUSION	$C_{avg}\%$	C_{llr}
A	ADAPTED GB	2.1	0.1222
	+ MLR	1.5	0.0690
B	ADAPTED GB	1.5	0.0845
	+ MLR	1.1	0.0528
C	ADAPTED GB	1.6	0.0826
(Partially-sup)	+ MLR	0.9	0.0437
C	ADAPTED GB	1.6	0.0842
(Fully-sup)	+ MLR	0.9	0.0428

Table 3: *System performance for different language information fusion configurations in terms of* $100 \times C_{avg}$ *and* C_{llr} *obtained on the NIST LRE 2007 for the closed-set task.*

It can be observed that configuration B significantly outperforms configuration A by 27% relative. This result indicates that modeling the training data sources separately might improve the performance of the PPRLM system. The PPRLM systems corresponding to these two configurations have the same fusion module architecture (same size GB and same number of classes in the MLR) but they differ by the information provided to this module. In configuration A, the dimension of the phonotactic score vectors is $\|L_T\| \times N_d$ (i.e, $14 \times 3 = 42$) while in configuration B, this dimension is equal to $N_s \times N_d$ (N_s is the total number of modeled sources, in these experiments $N_s = 25$). The information provided by the phonotactic score vector in configuration B is richer, and the fusion parameter estimation is expected to be better. This is particularly true when enough dev data segments are available.

Further improvements can be obtained with configuration C (18% relative, compared to configuration B). In configuration B, the *within-language* fusion was done earlier in the fusion module (at the input to the GB), while in configuration C, it is delayed to the end of the fusion module (at the output of the MLR). This result suggests that the language recognition score (providing there is enough dev data for each source) can be better estimated by fusing the language source recognition scores at the decision level rather than earlier in the system. The C_{avg} obtained with configuration C is equal to 0.9%. These results obtained with a purely phonotactic system are competitive with the best published results ($C_{avg} = 0.87\%$) obtained with a language recognition system [11] using both acoustic and phonotactic sub-systems.

The same trend can be also observed with the C_{llr} measure, indicating that language scores estimated by configuration C are better calibrated than those estimated by configurations A and B. Finally, no difference in performance was observed between the *Partially* and *Fully* unsupervised data labeling.

[6] http://www.nist.gov/speech/tests/lang/2007/LRE07EvalPlan-v8b.pdf

[7] Defined by MIT Lincoln Labs for their NIST LRE 2007 system.

Language	LRE96-TR	OHSU	LRE07-TR	Language	LRE96-TR	OHSU	LRE07-TR
Arabic	8.3	–	3.5	Bengali	–	–	3.6
Chinese	17.4	29	11.1	English	18.7	22.9	–
Farsi	8.9	–	–	German	8.8	3.8	–
Hindustani	9.2	6.2	3.4	Japanese	8.6	16.4	–
Korean	7.9	17.5	–	Russian	–	–	3.4
Spanish	18.5	11.2	–	Tamil	7.4	8.6	–
Thai	–	–	3.5	Vietnamese	10.6	–	–

Table 1: *The amount of training data (in hours) for each target language and data source after removing the automatically detected non-speech segments.*

Language	LRE-TR	OHSU	LRE07-TR	Language	LRE96-TR	OHSU	LRE07-TR
Arabic	635/634	–	276/277	Bengali	–	–	115/115
Chinese	567/433	743/839	462/467	English	1004/982	1726/1748	–
Farsi	337/337	–	–	German	450/447	105/108	–
Hindustani	243/232	199/209	158/159	Japanese	252/213	689/728	–
Korean	196/162	452/486	–	Russian	–	–	447/447
Spanish	857/837	472/492	–	Tamil	264/238	249/275	–
Thai	–	–	80/80	Vietnamese	326/326	–	–

Table 2: *Number of dev segments for each language source using an unsupervised labeling procedure. (Left: partially-unsupervised/ Right: fully-unsupervised)*

4.3.2. Open-set task

For the open-set task, the standard approach [12] is to use development data from out-of-set (OOS) languages (i.e; languages that are different from the set of target languages). For each segment in the OOS dev data, a phonotactic score vector is estimated using the phonotactic models. The phonotactic score vectors are then used to train an OOS Gaussian to be added to the GB. This approach was used to compare the three configurations on an open-set task. The OOS dev data consist of segments of $30s$ long from 9 Languages. The language detection score is estimated according to (4) with $P_{\text{oos}} = 0.2$. Results are reported in Table (4). Before analyzing the results, it is worth mentioning that in each configuration, the target language score is estimated as in the closed-set task. The focus in this section is on modeling the OOS languages and estimating their scores. For configuration C, the partially-unsupervised labeled dev data is used.

Config	Fusion	$C_{avg}[\%]$	C_{llr}
A	Adapted GB	4.3	0.2190
	+ MLR	3.5	0.1472
B	Adapted GB	3.4	0.1768
	+ MLR	2.8	0.1242
C (Fully-unsup)	Adapted GB	3.7	0.2024
	+ MLR	2.5	0.1309

Table 4: *System performances in terms of* $(100 \times C_{avg})$ *for different language information fusion configurations obtained on the NIST LRE 2007 for the open-set task.*

Configuration A is the standard open-set approach. The target languages are represented by a single phonotactic model for each PRLM, and the OOS phonotactic score vectors are used to train a single OOS Gaussian.

Configuration B is basically the standard open-set approach but each target language is represented by one or multiple source-dependent phonotactic models. The OOS phonotactic score vectors are used to train a single OOS Gaussian. This configuration improves system performances by 20% relative compared to configuration A. It can also be seen that the result obtained with the adapted Gaussians backend ($C_{avg} = 3.4\%$) is comparable to the best result obtained with configuration A ($C_{avg} = 3.5\%$).

In configuration C, each language in the OOS languages is considered as a different source and represented by a separate Gaussian in the GB. This means that the number of Gaussians representing the OOS languages is 9 (i.e, the number of OOS languages in the dev data). In this configuration, the target language score is estimated according to (3). The OOS recognition score $p(x|\ell_{\text{oos}})$ is estimated as follows:

$$p(x|\ell_{\text{oos}}) \simeq \sum_{s \in L_{\text{OOS}}} \beta_s p(x|\ell_{\text{oos}_\text{s}}) \qquad (7)$$

where, L_{OOS} is the set of OOS languages, $\ell_{\text{oos}_\text{s}}$ is one of the OOS languages and β_s is the OOS source weight that can be set differently depending on the application or can be estimated dynamically. In this work, β_s is set to be equal to $\frac{1}{\|L_{\text{OOS}}\|}$. A similar modeling approach was proposed in [12] but the scoring was done differently.

With this configuration, a further 11% relative improvement in the C_{avg} is obtained. The obtained results ($C_{avg} = 2.5\%$) is close to the best published result (2.39%) [11].

4.4. Experiments on the NIST LRE 2009

For the NIST LRE 2009 data, initial experiments were conducted only for the closed-set task. The training and dev data sets are those provided by NIST and no external sources are used. There are 23 target languages and the eval data is composed of both CTS and VOA (Voice-Of-America) data segments.

4.4.1. Training data

The training data consists of data described previously, with adional data from the mixer3 corpus and the VOA data, both

Language	CTS	OHSU	VOA	Language	CTS	OHSU	VOA
Amharic	–	–	9.7	Bosnian	–	–	5.2
Cantonese	7.3	–	4.3	Creole	–	–	9.87
Croatian	–	–	8.0	Dari	–	–	9.8
English_American	3.8	15.2	–	English_Indian	3.9	7.7	–
Farsi	12.8	–	10.0	French	9.6	–	9.9
Georgian	–	–	4.6	Haussa	–	–	9.7
Hindi	12.8	6.2	9.8	Korean	11.5	17.5	9.6
Mandarin	24.9	29	9.7	Pashto	–	–	9.9
Portuguese	–	–	9.8	Russian	6.7	–	9.9
Spanish	18.5	11.2	9.7	Turkish	–	–	7.4
Ukrainian	–	–	3.9	Urdu	6.9	–	9.9
Vietnamese	14.2	–	9.9	-	–	–	–

Table 5: *The amount of training data (in hours) for each target language and data source after removing the automatically detected non-speech segments.*

Language	CTS	OHSU	VOA	Language	CTS	OHSU	VOA
Amharic	–	–	300/300	Bosnian	–	–	100/100
Cantonese	80/106	–	90/64	Creole	–	–	300/300
Croatian	–	–	168/168	Dari	–	–	300/300
English_American	557/111	338/784	–	English_Indian	160/103	215/272	–
Farsi	318/356	–	300/262	French	316/318	–	300/298
Georgian	–	–	132/132	Hausa	–	–	300/300
Hindi	393/383	143/200	300/356	Korean	314/267	314/399	241/276
Mandarin	877/585	644/1008	300/228	Pashto	–	–	300/300
Portuguese	–	–	300/300	Russian	320/333	–	300/287
Spanish	976/576	259/526	300/253	Turkish	–	–	186/186
Ukrainian	–	–	78/78	Urdu	160/175	–	300/285
Vietnamese	391/426	–	258/223	-	–	–	–

Table 6: *Number of dev segments for each language source with both supervised and unsupervised labeling. (Left: supervised/ Right: unsupervised)*

provided by NIST. The VOA data was further processed to select telephone segments with high inter-speaker variability. This processing is done via the use of an *incremental open-set speaker verification* procedure. The result of this procedure is a set of speakers with their associated speech segments. The set of speakers is divided into two non-overlapping sub-sets. The VOA train data was selected from one subset and the dev data from the other. For each target language, a maximum of 10 hours of VOA speech was selected for training. To increase the speaker variability and the generalization capabilities of the VOA phonotactic models a maximum of 15 minutes of speech data per speaker was used.

The entire training data was separated in three subsets: CTS, OHSU and VOA sources. Although OHSU is a CTS data type, it has different characteristics compared to the other CTS data sets provided by NIST and was therefore not combined with the other CTS data. The total number of modeled sources for all target languages was 39. Table (5) reports the amount of data (in hours) for each source and target language.

4.4.2. Development data

The development data is composed of CTS, OHSU and VOA data segments. The CTS segments are extracted from eval and dev data sets of the NIST LRE 1996, 2003, 2005 and 2007. The OHSU segments are extracted from the eval and dev data sets of the NIST LRE 2005. The speakers in the VOA segments are from a separated set than those used for training, with a maximum of 300 segments per target language is used.

For these experiments, supervised (the source -CTS, OHSU or VOA- for each segment in the dev data is already given) and unsupervised (as described in section 4.2.3) labeling are compared. Table (6) reports the number of dev segments per data source and language for both supervised and unsupervised labeling.

4.5. Results and Discussion

Table (7) reports the obtained results.

Config	Fusion	C_{avg}[%]	C_{llr}
A	Adapted GB	2.99	0.1854
	+ MLR	2.1	0.0879
B	Adapted GB	2.99	0.1904
	+ MLR	1.99	0.0851
C	Adapted GB	2.99	0.2005
(Supervised)	+ MLR	1.98	0.0819
C	Adapted GB	2.99	0.2046
(Unsupervised)	+ MLR	1.99	0.0858

Table 7: *System performance in terms of* (100 × C_{avg}) *and* C_{llr} *for different language information fusion configurations obtained on the NIST LRE 2009 for closed-set task.*

It can be observed that configurations B and C outperform configuration A in terms of both C_{avg} (small improvements) and C_{llr} measures. There is no difference between configuration B and C in terms of C_{avg} measure, but the language score is better calibrated with configuration C when supervised labeling of the dev segments is used.

It should be mentioned here that configuration B and C use the same set of phonotactic score vectors (same dev data) to train the fusion module. But in configuration C, the fusion module has more parameters than in B. In configuration C there are 39 Gaussians in the GB (compared to 23 in configuration B), and the same holds for the number of MLR classes. In both configurations, the dimension of the phonotactic score vector is $39*3 = 117$. In configuration C and for a given target language, the phonotactic score vectors are split over the modeled target language sources. As a result, the number of dev segments attributed to one or more sources may not be large enough (see Table 6 for Cantonese) to better estimate the source-dependent Gaussian. This might affect the effectiveness of configuration C. One possible simple solution to this problem is to model only sources with a number of dev data segments that is higher than a fixed threshold.

5. Conclusions

This paper has investigated several techniques to fuse language information when multiple training data sources for all or some target languages are available. Experimental results on the NIST LRE 2007 and NIST LRE 2009 data sets suggest that instead of merging all data sources and creating a single phonotactic model, better performance can be obtained by modeling language sources separately, provided that enough dev data are available for each source. When multiple source-specific phonotactic models are used, the results show that the language recognition score can be better estimated by fusing well calibrated language source scores (configuration C), rather than at the input to the fusion model (configuration B). These experiments indicate that unsupervised labeling of the dev segments according to language source performs as well as known language source labels.

In the proposed technique, the language recognition score is estimated by taking the average of the language source scores (Equation 3). That is, all sources are equally important. Better optimization of the weights can be expected to improve system performances.

For a language with a relatively high number of data sources, modeling each data source separately might not be possible since there may be insufficient data for each source. In this case, two or more data sources can be merged, however what criteria should be used to decide which data sources to merge is an open question that needs to be addressed.

6. Acknowledgments

This work has been partially supported by OSEO, French State agency for innovation, under the Quaero program.

7. References

[1] P. Matejka, P. Shwarz, J. Cernocky and P. Chytil, ”Phonotactic Language Identification using High Quality Phoneme Recognition”, *Proceedings of Eurospeech’05*

[2] M.F. BenZeghiba, J.L. Gauvain and L. Lamel, ”Context-Dependent Phone models and Models Adaptation for Phonotactic Language Recognition”, *Interspeech’08*, pp. 313-316.

[3] M.F. BenZeghiba, J.L. Gauvain and L. Lamel, Improved N-Improved N-gram Phonotactic Models For Language Recognition *Interspeech’10*, Makuhari, Japan, September 2010.

[4] O. Glembek, P. Matejka, L. Burget and T. Mikolov ”Advances in Phonotactic Language Recognition” *Interspeech’08*, 743-746.

[5] M.F. BenZeghiba, J.L. Gauvain and L. Lamel, ”Gaussian Back-end Design for Open-set Language Detection” *ICASSP’09*, pp. 4349-4352.

[6] M. A. Zissman, ”Comparison of Four Approaches to Automatic Language Identification of Telephone Speech”, IEEE Trans. Speech and Audio Proc., 4(1):31-44, 1996.

[7] J. L. Gauvain, A. Messaoudi and H. Schwenk, ”Language Recognition Using Phone Lattices”, *ICSLP’04*

[8] N. Brummer and D.A. van Leeuwen, ”On Calibration of language recognition scores” *2006 IEEE Odyssey: The Speaker and Language Recognition Workshop*, 1-8.

[9] M.F. BenZeghiba, J.L. Gauvain and L. Lamel, ”Language Score Calibration using Adapted Gaussian Back-end”, *interspeech’09.*

[10] J. L Gauvain and C. H. Lee ”Maximum a-posteriori estimation for multivariate Gaussian mixture observations of Markov chain” IEEE Trans. Speech and Audio Proc., 2:31-44, 1994.

[11] A. McCree, F. Richardson, E. Singer and D. Reynolds ”Beyond Frame Independence: Parametric Modeling of Time Duration In Speaker and Language Recognition” *Interspeech’08*, pp.767-770

[12] P. A. Torres-Carrasquillo et al. ”The MITLL NIST LRE 2007 Language Recognition system” *Interspeech’08* 719-722