GAUSSIAN BACKEND DESIGN FOR OPEN-SET LANGUAGE DETECTION

Mohamed Faouzi BenZeghiba, Jean-Luc Gauvain and Lori Lamel

Spoken Language Processing Group
LIMSI - CNRS B.P. 133 91403 ORSAY CEDEX FRANCE
{*mfb,gauvain,lamel*}@*limsi.fr*

ABSTRACT

This paper proposes a new approach to the challenging open-set language detection task. Most state-of-the-art approaches make use of data sources with several out-of-set languages to model such languages. In the proposed approach, no additional data from out-of-set languages is required, only date from the target languages is used. Experiments are conducted using the LRE-05 and the LRE-07 evaluation data sets with the $30s$ condition. A C_{avg} of 4.5% and 3.4% is obtained on these data set, respectively. These results are comparable with other reported results.

Index Terms— Language recognition, Open-set, Phonotactic approach, Gaussian Backend, Adaptation

1. INTRODUCTION

Language recognition is the task of automatically determining the language of a given speech segment. This is achieved by extracting information that conveys language characteristics from the speech signal. The general task of language recognition can be divided into several sub-tasks including language detection.

Language detection is a binary decision of whether the language of a speech segment corresponds to a specific language from a set of target languages. When the language of the test segment is constrained to be one of the target languages, the task is known as *closed-set* language detection, otherwise it is known as *open-set* language detection. The main difference between the two tasks is that in the former case, the system has *a priori* knowledge about the possible languages of the speech segment, while such knowledge is not available in an open-set task. State-of-the-art language detection systems achieve high performance on the closed-set task, but performance often degrades substantially on an open-set task. Successful approaches to the open-set task make use of speech data in languages different from the set of target languages [1, 2]. The characteristics of these out-of-set (OOS) languages are incorporated in the system in order to improve detection of such languages. Acquiring the additional data can be costly and time consuming, which may be a reason why only a few participants in previous NIST Language Recognition Evaluations (LRE) submitted results for the open-set task.

In this paper a simple technique to deal with the open-set task that is proposed does not rely on the use of additional data from OOS languages, but rather only needs data from target languages. The proposed technique is evaluated on the LRE-05 and LRE-07 eval data, and compared with other state-of-the-art approaches.

2. SYSTEM DESCRIPTION

The language recognition system makes use of the Parallel Phone Recognizer followed by Language Modeling (PPRLM) approach [3] where each target language is represented by multiple *n*-gram language models (LM), generated using phone recognizers trained in different languages. This work uses phone lattice decoding [4] which has been demonstrated to outperform one-best phone decoding. The PPRLM system uses 3 context-dependent phone decoders for English, French and Spanish. The acoustic models are word-position independent, and trained on 25 hours for Spanish, 116 hours for French and 1760 hours for English. Each model covers about 3000 phone contexts, with 3000 tied states and a mixture of 32 Gaussians per state. The Spanish, French and English decoders have 27, 36 and 48 phones, respectively. Silence is modeled by a single state, with a mixture of 1024 Gaussians.

Prior to the phone lattice decoding, Constrained Maximum Likelihood Linear Regression (CMLLR) adaptation procedure is performed. During training a Viterbi decoding pass is first performed to find the best phone segmentation for each audio file of a given target language, which is then used to estimate the CMLLR transforms using the maximum likelihood criterion. These transforms are used to generate new features which are then used in a second decoding pass to generate phone lattices. Phone decoding is done without any phonotactic constraints (i.e., no grammar is used). This procedure is applied both during training and test (only for segments longer than $6s$).

The *n*-gram probabilities are estimated by computing the expected *n*-gram frequencies from the phone lattices. Back-off *4*-gram phonotactic models are generated with Witten-Bell discounting using the SRILM toolkit.[1] The standard approach is for each individual phone decoder to generate one language model for each target language. When the training data contains several data sources, the robustness of the phonotactic approach can be enhanced by generating multiple LMs, one for each decoder and target language per available sources [5].This approach has been adopted here. The effectiveness of using context-dependent phone models with CMLLR adaptation was demonstrated in [6].

3. EXPERIMENTAL SET-UP

3.1. Training and development data

In this work experiments were carried out using the training and development data sets defined by MIT Lincoln Labs when developing their NIST LRE 2007 system [7]. The training portion is comprised of the LRE-96 train and dev sets, the NIST LRE-07 train set, and

This work was in part supported by OSEO under the Quaero program.

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

randomly selected samples from the Callhome, Fisher and Mixer corpora. The amount of training data per language varies from about 2.5 hours (for Bengali) to about 71 hours (for English) of speech. The development data includes all segments from the LRE-96 and LRE-03 evaluation sets, half of the data in the LRE-07 dev set and segments from Callhome, Fisher and Mixer. The validation data includes segments from the OGI-22 and Mixer corpora, the second part of the LRE-07 dev set. [2]

In order to compare the proposed approach with the more widely used ones, 895 segments from 8 OGI-22 languages and French LRE-96 and LRE-03 eval set are used as out-of-set development data.

3.2. Evaluation data sets

The performances of different techniques for open-set language detection are evaluated using the $30s$ segments in the NIST LRE-05[3] and LRE-07[4] evaluation sets.

There are 7 target languages in the LRE-05 eval data with some speech segments from the Mixer and Fisher corpus, but the majority from the OHSU corpus. Of the total of 3662 speech segments, only 86 correspond to the one out-of-set language (German). For the LRE-07 data, there are 14 target languages, and about 2509 speech segments, of which 352 belong to one of the 5 out-of-set languages (French, Italian, Punjabi, Tagalog, Indonesian). Speech segments were extracted mainly from the Fisher, Mixer, Callfriend and OGI corpora.

3.3. Pre-processing

Standard 12 PLP coefficients with energy are extracted every 10 ms, with a 30 ms window. Cepstral mean removal and variance normalization are applied to each segment. These features are augmented by their first and second derivatives, resulting in a 39 dimensional feature vector. Speech activity detection was carried out using Gaussian mixture models to segment the audio signal into speech/non-speech regions. Two Gaussian mixtures, one for speech and one for non-speech with 2048 and 512 mixtures, respectively, were used.

3.4. Post-processing

Language scores estimated by the individual decoders are fused as shown in Figure 1 to estimate the final detection scores. The parameters of the fusion module are optimized on the development and the validation data using the FoCal Multi-class toolkit.[5]

It was observed that the dynamic range of the language scores is decoder-dependent, where decoders with fewer phones give higher language likelihoods. To reduce this effect, mean normalization (MN) is applied to the score for each individual segment and the normalized scores are stacked in a feature vector. The set of feature vectors associated with a given target language are used to train a language dependent Gaussian or a multivariate normal distribution. All Gaussians share a common full covariance matrix and form what is called *Gaussian Backend* GB.

For this case, the decision function based on the normal density $N(\mathbf{x}|\Sigma_\ell, \mu_\ell)$ for language ℓ can be simplified to:

$$\delta_\ell(\mathbf{x}) = (\mathbf{x} - \mu_\ell)^{\mathbf{t}} \mathbf{\Sigma}^{-1} (\mathbf{x} - \mu_\ell) \tag{1}$$

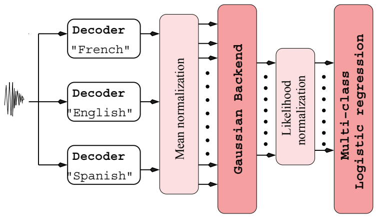

Fig. 1. *A block diagram of the fusion module*

where μ_ℓ is the mean vector, and Σ is the common covariance matrix. Developing and removing terms that are independent of ℓ, the decision function can be expressed as follows:

$$d_\ell(\mathbf{x}) = (\mathbf{\Sigma^{-1}} \mu_\ell)^{\mathbf{t}} \mathbf{x} - \frac{\mathbf{1}}{\mathbf{2}} \mu_\ell^{\mathbf{t}} \mathbf{\Sigma^{-1}} \mu_\ell \tag{2}$$

As explained in [9], the Gaussian backend can be seen as an affine transform. The *linear* part of the transform is the same as a linear discriminant analysis transform (LDA) and the *translation* part corresponds to the calibration task for setting the "language dependent threshold". If the parameters of the GB are well estimated, then it performs both score calibration and fusion. It has been reported that the output of the GB can be further calibrated using a discriminative Logistic Regression (LR) [7, 8, 6]. The outputs of the GB is converted to a *log likelihood ratio* (LLR) by normalizing each language likelihood with respect to the other likelihoods. The LLRs are then used to train the Multi-class LR.

3.5. Tasks and performance measure

The task of interest is that of open-set language detection. The detection decision is made based on the *detection log likelihood ratio* llr [10]:

$$llr(s|\ell) = \log \left[\frac{P_{tar}.p(s|\ell)}{P_{oos}.p(s|o) + \sum_{L_T,\ k \neq \ell} P_{non-tar}.p(s|k)} \right] \tag{3}$$

where $p(s|\ell)$ is the likelihood of the test segment s given the target language ℓ. It can be the outputs of the GB or the MLR. L_T is the set of target languages and o represents the OOS languages.

The target language *prior* P_{tar} is equal to 0.5. The OOS language *prior* P_{oos} is equal to 0.0 for the closed-set task and 0.2 for the open-set task.[6] The $P_{non-tar}$ is equal to:

$$P_{non-tar} = (1 - P_{tar} - P_{oos})/(N_L - 1) \tag{4}$$

where N_L is the number of target languages. The $llrs$ in (3) are then compared to the theoretical threshold $\Delta = 0$ to make a decision. The performance measure is the C_{avg}[7] estimated as follows:

$$C_{avg} = \frac{1}{N_L} \sum_{L_T} \left\{ \begin{array}{l} C_{miss}.P_{tar}.P_{miss}(L_T) \\ + \sum_{L_N} C_{fa}.P_{non-tar}.P_{fa}(L_T, L_N) \\ + C_{fa}.P_{oos}.P_{fa}(L_T, L_o) \end{array} \right\} \tag{5}$$

where P_{miss} and P_{fa} are the error probabilities computed from the decision results. and $C_{fa} = C_{miss} = 1.0$ are the costs of the false acceptance and false rejection, respectively, and were specified by NIST.

[2] The LRE-05 eval set was also included in the validation data when evaluating on the LRE-07 eval set.

[3] http://www.nist.gov/speech/tests/lang/2005/

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

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

[6] These values are given by NIST.

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

4. RESULTS AND DISCUSSION

4.1. Closed-set detection results

To fairly asses the contribution of each step in the fusion module, a closed-set language detection experiment was conducted. The results for the $30s$ condition in terms of C_{avg} are given in Table 1. The C_{avg} is computed by replacing P_{oos} in the above equations with 0.

MN		√		√		√
GB	√	√	√	√	√	√
LLR					√	√
MLR			√	√	√	√
$LRE-05$	2.23	2.16	2.12	2.15	2.13	**2.03**
$LRE-07$	2.70	2.64	1.43	1.45	1.40	**1.28**

Table 1. *Performance of the system in terms of* ($100 \times C_{avg}$) *for a closed-set detection task with* $30s$ *test segments. MN: mean normalization, GB: Gaussian Backend, LLR: Log-likelihood ratio and MLR: Multi-class logistic regression*

Several observations can be made. First, adding the MLR is more beneficial for the LRE-07 eval data than for LRE-05 eval data. A possible explanation is that the LRE-05 GB has fewer parameters ($d = 33$) compared to the LRE-07 GB ($d = 78$)[8]. So the GB for LRE-05 is better estimated, resulting in well calibrated scores. Second, the MN and LLR are more effective when used together than individually as can be seen by comparing the last 3 columns of Table 1. A small improvement is obtained with the GB when the scores are mean normalized (comparing columns 1 and 2), but this improvement is not carried over when used with MLR, as can be seen by comparing columns 3 and 4. Third, the system described in this paper achieves a relative improvement of 13% for both eval sets compared to the system described in [6]. This previous system in which only one LM was generated per decoder for each language, and the fusion was effectuated with the GB followed by MLR had a C_{avg} of 2.4% and 1.6%, respectively, for the LRE-05 and LRE-07 eval data.

4.2. Open-set detection results

4.2.1. Baseline approach

As introduced earlier, a characteristic of the open-set language detection task is that no *a priori* knowledge about the out-of-set languages is available to the system. Speech segments belonging to one of these languages should be detected and rejected. A simple approach to the open-set task is to ignore the problem and rely on the robustness of the system designed for the closed-set task (Section 4.1). That is, no additional information characterizing the OOS languages is integrated within the system. This approach will be referred to as the *baseline* approach. In the other approaches studied here, additional knowledge is incorporated at the fusion level.

4.2.2. Gaussian backend design for open-set

The basic idea is to train an additional Gaussian to represent the OOS languages. One approach tested in [2], consisted of using scores obtained for segments belonging to languages that are not in the target language list to train the additional Gaussian. These scores are estimated using only LMs for the target languages. While this approach generally gives better results than the *baseline* approach, it requires the availability of additional data from several OOS languages. Searching for audio files for these languages that match some specific conditions and pre-processing them is time consuming. It has also been observed that performance is better on test data on languages for which some data happened to be included in the training data for the OOS Gaussian. In this work, an OOS Gaussian was trained on the OOS development data described in Section 3.1. This approach will be referred to as the *OOS-GB* approach.

The new approach proposed in this paper uses the same basic idea but, makes the assumption that no additional data from OOS languages are available. The issue here then becomes how to design the OOS Gaussian when only data from target languages are available? First a Target-Independent (TI) Gaussian was trained using the development data of all target languages. This TI Gaussian is used to represent the OOS languages. This approach will be referred to as the *TI-GB* approach. Results of these three approaches in terms of C_{avg} estimated using equation (3) applied at the Gaussian backend outputs are reported in Table 2. The OOS-GB approach performs consistently better than the other approaches.

	LRE-05	LRE-07
BASELINE	6.30%	4.87%
OOS-GB	5.50%	4.74%
TI-GB	5.65%	5.54%

Table 2. *Detection Performance in terms of* C_{avg} *for different approaches on an open-set task for the LRE-05 and LRE-07 eval sets.*

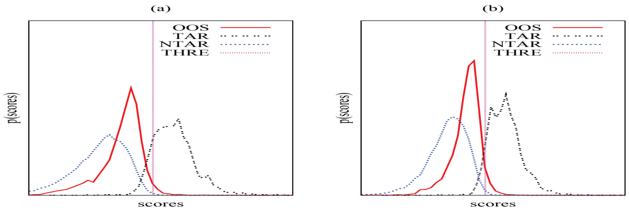

Fig. 2. *Distribution of different types of scores with the TI-GB approach for the LRE-07 eval set. (a) before adaptation and (b) after adaptation*

Figure 2(left) shows the distribution of scores (log likelihood detection ratio) for target (TAR), non-target (NTAR) and OOS languages using the TI-GB approach on the LRE-07 eval set (similar plots are obtained for the LRE-05 eval data). The vertical line represents the detection threshold. It can be observed that a relatively large portion of target segments are rejected. This is also true for the OOS-GB approach. Explicitly modeling some of the characteristics of OOS languages reduces significantly the false acceptance rate, but at the same time it increases the false rejection rate P_{miss}. The values of P_{fa} and P_{miss} combined with their costs and priors determine the final cost, which can be better or worse than the *baseline* approach. The TI-GB was trained using only data from target languages, so it models some of the target language specific information, increasing the confusability between the target independent and the target dependent Gaussians. We have also observed that when a

[8] If only one LM is generated per target language for each decoder, then d is 21 and 42 for LRE-05 and LRE-07, respectively.

target segment is correctly detected (accepted), most of the time, the difference between its score and the best non-target/OOS score is relatively high. But when it is missed, this difference is rather small.

Based on these observations and inspired by the work described in [11], we explored the effect of slightly perturbing the mean of the target-dependent Gaussian using the target-independent Gaussian.

$$\mu_\ell^{new} = \alpha\mu^{TI-GB} + (1-\alpha)\mu_\ell^{old} \qquad (6)$$

This can be seen as a simplified form of adaptation, where α is optimized under the closed-set condition, in order to remain compatible with the assumption that no additional data from OOS languages is available. The training of the GB is done as follows: First, the dev data is used to train a primary GB and the validation data is used to optimize the value of α. The dev and validation data were then pooled to train the final GB. It is worth mentioning here, that for closed-set task, the TI Gaussian is used only as an *a priori* model for adaptation, but not for score estimation (in Equation (3) $P_{oos} = 0$). The obtained results are reported in Table 3 for the closed and open-set tasks.The new distributions of scores are shown in Figure 2 (right).

	LRE-05 $(\alpha = 0.2)$	LRE-07 $(\alpha = 0.3)$
CLOSED-SET TASK	2.12%	1.61%
OPEN-SET TASK	4.53%	3.37%

Table 3. *Performances in terms of C_{avg} for open and closed set tasks using the adapted GB on the LRE-05 and LRE-07 eval data.*

It can be observed that adapting the GB has the effect of shifting the score distribution slightly to the right. The amount of this shift is controlled by the parameter α. As a consequence, the P_{miss} gets reduced, while the P_{fa} is increased. However, the reduction in P_{miss} is larger than the increase in P_{fa}, and in terms of C_{avg}, there is a relative improvement of 20% and 40%, respectively, on the LRE-05 and LRE-07 eval data. These results are better than the results with the other approaches reported in Table 2. It is interesting to note that there is even an improvement for closed-set task with respect to the baseline system (column 2 of Table 1) in particular for the LRE-07.

4.2.3. Combination with logistic regression

The use of the multi-class logistic regression with the adapted Gaussian backend is not an easy task. The MLR is a supervised discriminative technique, requiring some representative examples for each class. In other words, some segments for the OOS class should be available, which is in contradiction with our basic assumption. To overcome this problem, the TI Gaussian was used to generate several random examples as a representative of the OOS class. Results of combining GB with MLR for the three approaches described above are reported in Table 4. Experiments with the adapted GB are run several times. The variance (± 0.1) is due to the randomness of the generated features.

The results show that, in contrast to the OOS-GB, MLR does not bring any additional improvement to the adapted GB. A possible explanation is that the scores estimated by the adapted GB are reasonably well calibrated and the use of MLR cannot bring further improvement. However 3 of the OOS languages in the LRE-07 eval data were also in the dev data. Removing these languages from the dev data, the C_{avg} on the LRE-07 data with the OOS-GB approach increases to 3.9%. In this case, the adapted GB approach outperforms the OOS-GB approach.

	BASELINE	OOS-GB	ADAPTED GB
LRE-05	6.9%	4.3%	$4.5\% \pm 0.1$
LRE-07	4.7%	3.4%/3.9*%	$3.4\% \pm 0.1$

Table 4. *Performance in terms of C_{avg} on the open-set task for the full system with different design approaches for the Gaussian backend. *The 3 OOS languages in eval data are removed from dev.*

5. CONCLUSION

This paper proposes a new approach to open-set language detection with the assumption that no data from out-of-set languages is required. This approach is based on the adaptation of a target-independent Gaussian, trained with data from all target languages. Results are comparable with the best reported state-of-the-art approaches using additional data to model possible OOS languages. The proposed approach can be further improved by fine optimization (language-dependent) of the adaptation factor α.

6. ACKNOWLEDGMENTS

We would like to thank MIT Lincoln Labs for sharing their training and development data lists with us.

REFERENCES

[1] E. Singer, et al. ”Acoustic, Phonetic, and Discriminative Approaches to Automatic Language Identification”, *Interspeech’03:* 1345-1348.

[2] Campbell, W. et al. ”Advanced Language Recognition using Cepstra and Phonotactics: MITLL System Performance on the NIST 2005 Language Recognition Evaluation” *Speaker and Language Recognition Workshop*, 1-8, 2006.

[3] 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.

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

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

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

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

[8] P. Matejka et al. ”BUT Language Recognition System for NIST 2007 Evaluations” *Interspeech’08*, 739-742.

[9] D.A. van Leeuwen and N. Brummer, ”Channel-dependent GMM and Multi-class Logistic Regression models for language Recognition” *2006 IEEE Odyssey: The Speaker and Language Recognition Workshop*.

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

[11] M. Hebert and S. Douglas Peters ”Improved Normalization Without Recourse to an Impostor Database For Speaker Verification” *ICASSP’00*, **2**:1213-1216.