interspeech 2008 incorporating SST 08

Context-Dependent Phone models and Models Adaptation for Phonotactic Language Recognition

Mohamed Faouzi BenZeghiba, Jean-luc Gauvain, Lori Lamel

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

Abstract

The performance of a PPRLM language recognition system depends on the quality and the consistency of phone decoders. To improve the performance of the decoders, this paper investigates the use of context-dependent instead of context-independent phone models, and the use of CMLLR for model adaptation. This paper also discusses several improvements to the LIMSI 2007 NIST LRE system, including the use of a 4-gram language model, score calibration and fusion using the FoCal Multi-class toolkit (with large development data) and better decoding parameters such as phone insertion penalty. The improved system is evaluated on the NIST LRE-2005 and the LRE-2007 evaluation data sets. Despite its simplicity, the system achieves for the $30s$ condition a C_{avg} of 2.4% and 1.6% on these data sets, respectively.

Index Terms: Context-dependent, Phone lattice, CMLLR adaptation, Language recognition.

1. Introduction

Language recognition (LR) is the task of identifying the language of a given speech segment. Two approaches are widely used, phonotactic and acoustic. Phonotactic approaches [1] require a phone decoder such as a standard Hidden Markov Model (HMM) [1], an Artificial neural network [2] or a binary tree [3]. Acoustic approaches such as Gaussian Mixture Models (GMM) [4] or Support Vector Machines [5] do not require any decoder. State-of-the-art LR systems often make use of these two approaches.

The paper focuses on the Parallel Phone decoders followed by Language Modeling (PPRLM) phonotactic approach [1]. We think that there are still ways to improve this approach, in particular by improving the underlying models. In this approach, multiple speech decoders (each specific to a given language), are used to map speech segments spoken in any language to a phone sequence or phone lattice [6]. Usually, the mapping is performed without any phonotactic constraints. Phone sequences or phone lattices are then used to estimate *n*-gram statistics and to generate the *n*-gram language models. The performance of a PPRLM based LR system depends highly on the performance and the consistency of the phone decoders. Better phone recognition leads to better *n*-gram estimates, which in turn leads to better LR results. A straightforward method to improve the decoders is to use more training data[2, 7].

The consistency of the decoder can be affected by different types of variability. Among them, the inherent variability of speech due to speaker, context, and the channel variability. It is well known that the local context, that is the preceding and following phones, can have a large influence on the acoustic realization of a phone. A consistent decoder should be able to deal with this kind of variability either by trying to remove it or by explicitly modeling it. This paper investigates the use of Context-Dependent (CD) phone models for LR. Such models have been shown to be more effective than Context-Independent (CI) phone models, when a sufficient large training data is available [8]. Channel variability can dramatically degrade the performance of a decoder, leading to a poor LR results. This was demonstrated in the 2005 NIST LRE. To reduce this effect, this paper investigates the use of Constrained Maximum Likelihood Linear Regression (CMLLR) adaptation technique [9, 10, 11]

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

2. Language model training

In a PPRLM system, each target language is represented by multiple *n*-gram language models (LM), depending on the number of decoders. The generation of these *n*-gram LMs is done in two steps, phone lattice decoding followed by phone *n*-gram probability estimation.

Phone Lattice decoding:
A phone lattice is a rich and compact representation of the phone hypotheses. Phone lattice decoding has been proposed as an alternative to the best phone sequence decoding [6]. The language likelihood estimation is done by taking the summation over the phone sequences present in the phone lattice instead of just using the best sequence, and results in a significant improvement in performance. In this work, prior to the phone lattice decoding, a CMLLR adaptation procedure is performed. Given some data, the parameters of one or several transformations (depending on the number of classes) are estimated to transform the means and covariance matrices of the Gaussian components of the HMM model, in such a way that the likelihood of the data is maximized. That is:

$$\hat{\mu} = A_c \mu + b_c \quad (1)$$

$$\hat{\Sigma} = A_c \Sigma {A_c}^T \quad (2)$$

where μ and Σ are the original mean and covariance matrix parameters, $\hat{\mu}$ and $\hat{\Sigma}$ are the transformed mean and covariance matrix parameters, and A_c is the transformation matrix of class c (a full transformation in our case). The parameters of these transformations are estimated iteratively using EM algorithm. The way in which CMLLR adaptation has been applied can be described as follows:

During training, for each audio file of a given language, a Viterbi decoding pass is first performed to find the best phone

segmentation which is then used to estimate the transformation $[A_c, b_c]$ using the maximum likelihood criterion (6 iterations). Then a new set of features are generated as follows:

$$\hat{o}_t = A_c o_t + b_c \quad (3)$$

where o_t is the observation vector at time t. The transformed features $\hat{o}_t$ are then used in a second decoding pass to generate phone lattices from which a back-off *n*-gram language model is estimated. The decoding is done without any phonotactic constraints (i.e., no grammar is used). This procedure is valid for both training and test, but during test, CMLLR adaptation is performed only for segments longer than $6s$ based on the number of frames after speech/non-speech segmentation.

N-gram LM generation:
The *n*-gram probabilities are estimated by computing the expected *n*-gram frequencies from the phone lattices. Back-off 3-gram and 4-gram models are generated with Witten-Bell discounting using the SRILM toolkit[1].

3. Experimental set-up

3.1. Phone decoders

During system development, different phone decoders were explored, using context-independent and context-dependent phone models with a varying number of contexts (ranging from 200 to 5000). Phone decoders for 3 languages (English, French and Spanish) were used in this work. For each of these languages, both CI and CD decoders were trained on the same amounts of data. The decoders were trained on 25 hours for Spanish, 116 hours for French and 1760 hours for English. The CD decoders used in the experiments reported here are word-position independent models, covering about 3000 contexts with 3000 tied states. The Spanish, French and English decoders have 27, 36 and 48 phones, respectively. Each phone is modeled by a 3-state, left-to-right HMM with a mixture of 32 Gaussians per state. Silence is modeled by a single state with a mixture of 1024 Gaussians.

3.2. Training and development data

In this work, we make use of the same data that MIT Lincoln Labs used to develop their NIST LRE 2007 system [12]. Briefly, the training data was created from LRE-96 train and dev sets, the NIST LRE-07 train set, and samples randomly selected from the Callhome, Fisher and Mixer databases. Depending on the language, the amount of training data is varying from about 2.5 (for Bengali) to about 71 (for English) hours of speech. For development data, we pooled together the MIT Lincoln Labs dev and test sets. This data is created by including all the previous NIST LRE eval sets, the NIST LRE-07 dev set and samples from Callhome, OGI-22 and Mixer databases.

3.3. Evaluation data sets

The performance of the different PPRLM systems are evaluated on the NIST LRE-05 and 07 eval sets[2],[3] containing speech segments with three nominal durations (3s, 10s and 30s). In the NIST LRE 2005, there are 7 target languages. Some of the speech segments are extracted from the Mixer and Fisher databases, but most of them are from the OHSU database. There are 3662 speech segments for each duration. In the NIST LRE 2007, there are 14 languages. There are about 2530 speech segments for each duration and they are extracted mainly from the Fisher, Mixer, Callfriend and OGI databases.

3.4. Pre-processing

Standard 12 PLP coefficients with energy are extracted very 10 ms over a 30 ms window, applying cepstral mean and variance normalization. These features are augmented by their delta and delta-delta, resulting in a 39 dimension feature vector. Speech activity detection was carried out using Gaussian mixture models to segment the audio signal int speech/non-speech regions. Two Gaussian mixtures, one for speech and one for non-speech with 2048 and 512 mixtures, respectively, were used.

3.5. Tasks and performance measure

In this work, both closed set language identification and detection tasks are considered. Language scores estimated by each PRLM (corresponding to the use of one decoder) are first stacked into one vector and then mapped by a duration dependent backend consisting of LDA followed by a linear logistic regression (LLR). These backends are trained on the development data using the FoCal Multi-class toolkit[4]. The language likelihood outputs are then used to estimate the Language Error Rate (LER) using a *maximum log-likelihood* criterion or to estimate the *detection log likelihood ratio* llr [13]:

$$llr(s|\ell) = \log\left[\frac{P(\ell)p(s|\ell)}{\sum_{k\neq\ell} P(k)p(s|k)}\right] \quad (4)$$

where $p(s|\ell)$ is the likelihood of the test segment s given the target language ℓ and $P(\ell)$ is the a priori probability of the language ℓ. This llr is then compared to a threshold Δ to make a decision, and to compute the C_{avg}[5]. The threshold is estimated as follows:

$$\Delta = \log(C_{fa}/C_{miss}) - \log(P_{tar}/(1-P_{tar})) \quad (5)$$

where C_{fa} and C_{miss} are the costs of false acceptance and false rejection, respectively, and P_{tar} is the *a prior* probability of the target language. These three parameters are given by NIST. Additionally, the $C_{llr\ avg}$ measure is reported.

4. Experimental results

To demonstrate the effectiveness of the use of CD phone models and CMLLR adaptation, several experiments were conducted. Before discussing the results, it is worth giving a short description of the LIMSI system submitted to the NIST LRE 2007 evaluation.

4.1. The LIMSI 2007 NIST LRE system

The system has the following characteristics:

- A PPRLM with 4 phone decoders, three with CD phone models (French, Spanish and English) and one with CI phone models (Arabic)
- CMLLR adaptation

[1] http://www.speech.sri.com/projects/srilm/
[2] http://www.nist.gov/speech/tests/lang/2005/
[3] http://www.nist.gov/speech/tests/lang/2007/
[4] http://niko.brummer.googlepages.com/focalmulticlass
[5] http://www.nist.gov/speech/tests/lang/2007/LRE07EvalPlan-v8b.pdf

- Phone lattice decoding
- Back-off 3-gram phonotactic models
- Training data was composed of the LDC CallFriend corpora, the LRE 2005 (OHSU database) and the NIST LRE 2007 databases
- Development data created by regrouping all the previous LRE eval sets and the LRE-07 dev set
- Detection decision was based on the log likelihood ratio No score calibration or advanced fusion techniques used

This system achieved a C_{avg} of 4.7%, 8.6% and 18.1% on the $30s$, $10s$ and $3s$ conditions, respectively. This system was improved by using better decoding parameters, 4-gram language models and additional development data with score calibration and fusion techniques.

4.2. Improved system with 3-gram LM

The first experiment compares the performance of different PPRLMs on the $30s$ condition, using 3-gram phonotatic models. Figures 1 and 2 show DET curves for 4 different systems. Tables 1 and 2 report the performance in terms of C_{avg}, $C_{llr\ avg}$, and LER. The following discussion focuses on the C_{avg}, but the findings are true for the other measures.

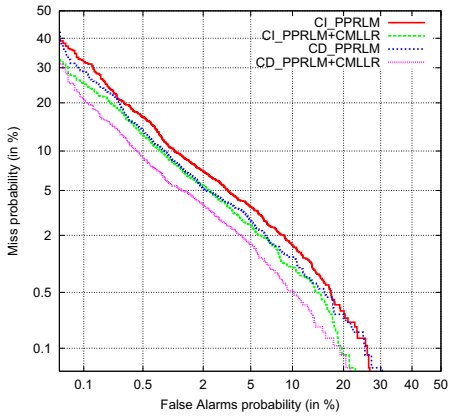

Figure 1: *Det curves for different PPRLM systems on the* $30s$ *condition of the NIST LRE* 2005 *eval set.*

SYSTEM	C_{avg}[%]	$C_{llr\ avg}$	LER[%]
CI PPRLM	4.8	0.16	8.8
CI PPRLM+CMLLR	3.9	0.13	7.6
CD PPRLM	3.7	0.13	7.6
CD PPRLM+CMLLR	**2.8**	**0.10**	**5.8**

Table 1: *Performance of different systems in terms of* C_{avg}, $C_{llr\ avg}$ *and LER on the* $30s$ *condition of the NIST LRE* 2005.

The following observations can be made: Using CD phone models with CMLLR adaptation gives the best results for all three measures on both of the NIST LRE evaluation sets. Compared to the submitted system, a substantial improvement is seen, with a relative gain of 42.5%.

On the LRE-07 eval data, the contribution of CMLLR adaptation is almost negligible compared to the contribution of CD

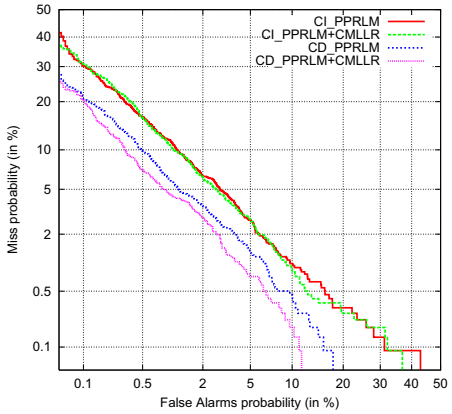

Figure 2: *Det curves for different PPRLM systems on the* $30s$ *condition of the NIST LRE* 2007 *eval set.*

SYSTEM	C_{avg}[%]	$C_{llr\ avg}$	LER[%]
CI PPRLM	3.2	0.13	11.9
CI PPRLM+CMLLR	2.9	0.12	11.8
CD PPRLM	2.1	0.08	8.4
CD PPRLM+CMLLR	**2.0**	**0.07**	**7.0**

Table 2: *Performance of different systems in terms of* C_{avg}, $C_{llr\ avg}$ *and LER on the* $30s$ *condition of the NIST LRE* 2007.

models. The relative improvements in the C_{avg} using CMLLR adaptation are 9% and 5% with CI and CD phone models, respectively. This is not surprising, since the test speech segments are extracted from databases that are used during the training and development steps. So the effect of channel mismatch is minimal. However, the use of CD phone models compared to CI phone models, gives a substantial improvement (about 31% relative) with and without CMLLR. It is worth mentioning that both the CI and CD decoders of the same language are trained with the same amount of data. So, the improvement comes from the fact that CD decoders are more consistent than their corresponding CI ones.

With respect to the LRE-05 data, the contribution of the CD phone models and CMLLR adaptation are equally important and additive. The relative improvements in the C_{avg} using CMLLR are 19%, and 24% with the CI and CD phone models, respectively. In this data, a large portion of speech segments are selected from the OHSU database that contains a lot of cellular data. This type of data was not used during system development, which in part explains the relatively poor performance. CMLLR adaptation significantly reduces this effect. Further significant improvement (28.2%) was achieved when CMLLR adaptation is used with CD phone models.

4.3. Using 4-gram LMs

In the above experiments, 3-gram phone LMs were used. Better performance can be expected by using a higher order *n*-gram, if the estimation of the *n*-gram frequencies is accurate. We conducted another experiment using 4-gram LMs, and compared the gain obtained by the CD PPRLM (with and without CMLLR) to that obtained by the best CI PPRLM (i.e., CI PPRLM + CMLLR). The obtained results are shown in Table 3 and 4.

SYSTEM	C_{avg}[%]	$C_{llr\ avg}$	LER[%]
CI PPRLM+CMLLR	3.4	0.12	7.2
CD PPRLM	2.9	0.10	6.4
CD PPRLM+CMLLR	**2.4**	**0.08**	**5.0**

Table 3: *Performance of different systems using 4-gram LMs on the* $30s$ *condition of the NIST LRE* 2005.

SYSTEM	C_{avg}[%]	$C_{llr\ avg}$	LER[%]
CI PPRLM+CMLLR	2.5	0.12	11.3
CD PPRLM	1.8	0.08	7.0
CD PPRLM+CMLLR	**1.6**	**0.07**	**6.0**

Table 4: *Performances of different systems using 4-gram LM on the* $30s$ *condition of the NIST LRE* 2007.

As can be expected, the 4-gram LM outperforms the 3-gram LM for all systems. But interestingly, the improvement on the LRE-05 set using CD PPRLM without adaptation (21.6%) is much higher than that obtained by the CI PPRLM with adaptation (12.8%). That is, despite the effect due to channel mismatch of the test data, the 4-gram statistics are better estimated with the CD decoder than with the adapted CI decoder.

Table 5 reports the C_{avg} measure of the CD PPRLM with CMLLR adaptation for segments with the three nominal durations ($30s$, $10s$ and $3s$). Although results with the $10s$ segments were not reported above, we observed that only small improvements (compared to the CD PPRLM) are obtained. This is can be explained by the fact that $10s$ of speech is too small for effective adaptation. It is worth mentioning that during the test, CMLLR adaptation is performed only for segments longer than $6s$.

EVAL. SET	$30s$	$10s$	$3s$
LRE-05	2.4%	6.3%	15.1%
LRE-07	1.6%	5.7%	15.9%

Table 5: *Performance in terms of* C_{avg} *of the CD PPRLM with CMLLR adaptation on the three conditions using 4-gram LMS.*

If we compare the results of this system to the one submitted to NIST LRE 2007, the relative improvements for $30s$, $10s$ and $3s$ segments are 66%, 36%, and 12.1%, respectively.

5. Conclusion

Our goal was to improve the quality and the consistency of phone decoders in a PPRLM language recognition system. We have shown that a consistent and substantial gain can be obtained by using context-dependent instead of context-independent phone models. The gain from using CMLLR adaptation is subject to the degree of mismatch between the test and train conditions. Several improvements were made to the LIMSI NIST LRE 2007 system. We found the FoCal Multiclass toolkit to be very useful, in particular when a large amount of development data is available. State of the art language recognition systems are a combination of several acoustic and phonotactic approaches. This is the case for the best performing systems in the 2007 NIST LRE [14]. With respect to those systems, the improved system described here achieves competitive results despite its simplicity.

6. Acknowledgements

We would like to thank MIT Lincoln Labs for sharing their training and development data lists with us. We also thank Abdel Messaoudi for his guidance and fruitful discussions.

7. References

[1] 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.

[2] P. Matejka, P. Shwarz, J. Cernocky and P. Chytil "Phonotactic Language Identification using High Quality Phoneme Recognition", *Proceedings of Interspeech'05* pp. 2237-2240.

[3] J. Navratil "Recent advances in phonotactic language recognition using binary decision trees" *proceedings of ICSLP'06*, pp. 421-424, Pittsburgh, USA.

[4] L. Burget, P. Matejka and J. Cernocky, "Discriminative Training Techniques for Acoustic Language Identification", *proceeding of ICASSP'06*, Vol. 1, pp. 197-200, Toulouse, 2006

[5] W. J. Campbell, D. Reynolds, E. Singer, P. Torres "Support Vector Machines for Speaker and Language Recognition", Computer Speech and Language, Vol. 20, No. 2-3, pp. 210-229, April 2006.

[6] J. L. Gauvain, A. Messaoudi and H. Schwenk, "Language Recognition Using Phone Lattices", Proceedings of ICSLP 2004.

[7] Doroteo T. Toledano et al., "Improved language Recognition Using Better Phonetic Decoders and Fusion with MFCC and SDC Features", *proceedings of InterSpeech'07*, pp. 194-197, Antwerp, Belgium.

[8] D. Zhu, M. Adda-Decker, and F. Antoine, "Different Size Multilingual Phone Inventories and Context-Dependent Acoustic Models for Language Identification:, *proceedings of InterSpeech'05*, pp. 2833-2836, Lisbon, Portugal.

[9] V. V. Digalakis, D. Rtischev and L. G. Neumeyer, "Speaker Adaptation using Constrained estimation of Gaussain mixtures", IEEE Trans. Speech and Audio Proc, 3(5):357-366, 1995.

[10] M. J. F. Gales "Maximum Likelihood Linear Transformation For HMM-based Speech Recognition", Computer Speech and Language 12, pp. 75-98, 1998.

[11] W. Shen and D. Reynolds. "Improved Phonotactic language recognition with acoustic adaptation" *Proceedings. of InterSpeech'07*, pp. 358-361, Antwerp, Belgium.

[12] D. Reynolds et al, "MITLL 2007 Language Recognition Evaluation System Description" *2007 NIST Language Recognition Workshop*, Orlando, Florida. December 11-12, 2007

[13] N. Brummer and D. A. van Leeuwen, "On Calibration of language recognition scores" *Proceedings of 2006 IEEE Odessey- The Speaker and Language Recognition Workshop*, pp. 1-8.

[14] A. Martin and A. Le "2007 NIST Language Recognition Evaluation: Evaluation overview & Results" *2007 NIST LRE Workshop*, Orlando, Florida. Dec. 11-12, 2007.