Odyssey 2010
The Speaker and Language Recognition Workshop
28 June – 1 July 2010, Brno, Czech Republic

PARALLEL ACOUSTIC MODEL ADAPTATION FOR IMPROVING PHONOTACTIC LANGUAGE RECOGNITION

Cheung-Chi Leung [1], *Bin Ma* [1] *and Haizhou Li* [1,2]

[1] Institute for Infocomm Research, A*STAR, Singapore 138632
[2] Department of Computer Science and Statistics, University of Eastern Finland, Finland
{ccleung,mabin,hli}@i2r.a-star.edu.sg

ABSTRACT

In phonotactic language recognition systems, the use of acoustic model adaptation prior to phone lattice decoding has been proposed to deal with the mismatch between training and test conditions. In this paper, a novel approach using diversified phonotactic features from parallel acoustic model adaptation is proposed. Specifically, the parallel model adaptation involves independent mean-only and variance-only MLLR adaptation. A quantitative method to measure the diversity between two sets of high-dimensional phonotactic features is introduced. Our experiment shows that this novel approach achieves an EER of 3.07% in the 30-second condition of the 2007 NIST Language Recognition Evaluation (LRE) tasks. It brings a 17.3% relative improvement in EER over the baseline system using a SAT phone model and CMLLR for model adaptation.

***Index Terms*—** MLLR adaptation, phone recognizer, phone lattice, spoken language recognition

1. INTRODUCTION

Automatic spoken language recognition is a task to determine the identity of the language corresponding to a given spoken utterance. Most automatic spoken language recognition systems can be classified into two main categories, namely phonotactic and acoustic approaches. We are interested in the phonotactic approach in this paper.

In a typical phonotactic system, a phone recognizer (PR) or a parallel phone recognizer (PPR) frontend performs phonotactic information extraction and a backend classifier discriminates between target languages using the extracted phonotactic information. Recently various advanced techniques have been studied in phonotactic systems. For instance, Constrained Maximum Likelihood Linear Regression (CMLLR) adaptation [1,2], Speaker Adaptive Training (SAT) [2] and discriminative training [3] in phone modelling, and the use of phone lattice [4] instead of single best phone sequence have been proposed for the phone recognizer frontends. Moreover, vector space modelling (VSM) approach [5] (A similar approach was referred to as the PPR-SVM in the MIT Lincoln Lab's language recognition system [6]) and anti-model training [7] have been studied for the backend classifiers.

The idea of phonotactic feature diversification has been well adopted in phonotactic systems. The rationale for diversified features is that we expect the incorrect decisions (false acceptance and false rejection) made by different subsystems using different types of useful features occur in different test trials and so the incorrect decisions made by each subsystem can be corrected by other subsystems through fusion (in feature-level, model-level, score-level and etc). The use of parallel phone recognizers can be considered as a way to utilize feature diversification. Of course, a number of phone recognizers can have a larger phonetic coverage and thus a better phonotactic analysis. Moreover, the phone transcription made by each individual phone recognizer is inevitably error-prone and the phonotactic features provided by different phone recognizers can provide complementary information. Instead of using 1-best hypotheses, multiply hypotheses in phone lattice provide another way of feature diversification.

The use of MLLR adaptation prior to phone lattice decoding and SAT in phone recognizer frontend has been proposed to deal with the mismatch between training and test conditions, and showed a satisfactory result [1,2]. In this paper, a novel approach using diversified phonotactic features from parallel acoustic model adaptation is proposed. Specifically, in our experiments, the parallel model adaptation is implemented by a score-level fusion of systems using mean-only and variance-only MLLR adaptation respectively. A quantitative method to measure the diversity between two sets of phonotactic features is introduced. Various types of phone models and MLLR adaptation techniques are also tested in our phonotactic system. The remaining of this paper is organized as follows: Section 2 describes the model adaptation and phonotactic feature extraction in a VSM-based phonotactic system. Section 3 and 4 describes the experimental setup and results respectively. Finally we conclude in Section 5.

2. PHONOTACTIC LANGUAGE RECOGNITION

In a state-of-the-art phonotactic system with a parallel phone recognizer frontend, the language identification task can be viewed as a two-step optimization as follows:

$$\hat{l} = \arg\max_{l} \left[\sum_{f=1}^{F} P(f) \log P(\hat{T}_f \mid \lambda_{f,l}^{LM}) \right] \qquad (1)$$

where $\hat{T}_f = \arg\max_{T \in B_f} \left[\log P(O \mid T, \lambda_f^{AM}) \right].$ (2)

Eq. (2) represents the phone recognizer frontend in which most possible phone sequences $\hat{T}_f$ for phone recognizer f are decoded using its acoustic model λ_f^{AM} and the sequence of feature vectors O for the spoken utterance. B_f is a set of multiply hypotheses represented in phone lattice.

Eq. (1) represents the system backend in which each candidate spoken language l is represented by a set of language models $\lambda_{f,l}^{LM}$. VSM [5] is adopted as language models in this paper. $P(f)$ is the prior probability of phone recognizer f and it can be considered as a combination weight of the phone recognizer towards the overall a posteriori probability. F is the number of phone recognizers used. To focus on parallel model adaptation, only one phone recognizer is used in our experiments. Our proposed parallel model adaptation takes acoustic model adaptation a into account and modifies the computation of Eq. (1) as:

$$\hat{l} = \arg\max_{l} \left[\sum_{f=1}^{F} \sum_{a=1}^{A} P(f,a) \log P(\hat{T}_f \mid \lambda_{f,a,l}^{LM}) \right] \qquad (3)$$

where $\hat{T}_f = \arg\max_{T \in B_f} \left[\log P(O \mid T, \lambda_{f,a}^{AM}) \right].$ (4)

Similarly, $P(f,a)$ is the prior probability of phone recognizer f using acoustic model adaptation a. An equal combination weight is used in our experiments. To evaluate our proposed method, a detection task in the 2007 NIST Language Recognition Evaluation (LRE) [10] is used. The detection decision based on log likelihood ratio llr can be computed as:

$$llr = \sum_{f=1}^{F} \sum_{a=1}^{A} P(f,a) \log \frac{P(\hat{T}_f \mid \lambda_{f,a,l+}^{LM})}{P(\hat{T}_f \mid \lambda_{f,a,l-}^{LM})} \qquad (5)$$

where $\lambda_{f,a,l+}^{LM}$ represents a hypothesized language model and $\lambda_{f,a,l-}^{LM}$ represents a language model formed by other competing languages.

2.1. Model Adaptation in Phone Recognizer Frontend

Prior to the phone lattice generation, a set of linear transformation for the mean and variance parameters of a Gaussian mixture HMM system can be computed so as to reduce the mismatch between the phone model and the adaptation data (i.e. the estimate of phone sequence in the first-pass decoding of training or test data). Given the adaptation data, the transformation is as follows:

$$\hat{\mu} = A_c \mu + b_c \qquad (6)$$

$$\hat{\Sigma} = H_c \Sigma H_c^{T} \qquad (7)$$

where μ and $\hat{\mu}$ are the original and transformed Gaussian mean vectors, $\sum$ and $\hat{\sum}$ are the original and transformed Gaussian covariance matrices (the original covariance matrix is usually diagonal in implementation), A_c and H_c are the transformation matrices, and b_c is the bias vector of class *c*. Single class/global transform and full transformation matrices are used in this paper. In our preliminary experiments, iterative CMLLR adaptation [8] and CMLLR adaptation with multiple regression classes were tested and they could not bring any improvement in language recognition, so they are not considered in this paper.

In CMLLR adaptation, the mean and its corresponding variance parameters share the same transform (i.e. $A_c = H_c$). Instead of applying the same set of transformation to the mean and variance parameters, unconstrained transforms can be used. Mean-only, variance-only and mean-and-variance MLLR transforms are considered in this paper. In the mean-and-variance MLLR adaptation, the mean transform is firstly estimated from the phone model and the adaptation data. Then the variance transform is estimated given the mean transform, the phone model and the same adaptation data.

Instead of using the speaker independent (SI) phone model, a SAT model with less speaker- or session-induced effect can be used in the phone recognizer frontend. In the SAT model training, a set of single-class CMLLR transforms (one per speaker) is generated using the training utterance and the SI model. The resultant SAT model is formed using the training features with their speaker-specific CMLLR transforms.

2.2 Diversified Phonotactic Features with Parallel MLLR Adaptation

In the VSM approach, language classification is performed using a high dimensional vector space from phone *n*-gram statistics [5]. Suppose that p is the size of phone set in a phone recognizer. The phone sequence/phone lattice generated by the phone recognizer from a training/test utterance can be transformed to a high dimensional phonotactic feature $c = \{c_1, c_2 ..., c_i, ..., c_s\}$. If an *n*-gram order of 3 is considered, c_i can be a phone unigram, bigram or trigram statistics. The dimension of this phonotactic feature, which depends on the *n*-gram order *n* and the phone set size p, equals $s = \sum_{i=1}^{n} p^i$.

In unconstrained mean-and-variance MLLR adaptation, it was reported that the variance adaptation can further increase the likelihood of the adaptation data:

$$\Pr(O \mid \lambda'') \geq \Pr(O \mid \lambda') \geq \Pr(O \mid \lambda) \qquad (8)$$

where λ is the original model, λ' is the model adapted by a mean transform, λ'' is the model adapted by a mean and a variance transforms and O is the observation sequence of the adaptation data. However, this does not necessarily bring a further word error rate reduction in ASR tasks [9]. This motivates us to generate different phonotactic features using different adapted models (from independent mean-only and variance-only MLLR transforms) and investigate the language recognition performance when the diversified features are combined in the same way as combining diversified features from parallel phone recognizers.

To study the relationship between the diversity of phonotactic features from different adapted models and the language recognition performance, we assume c^a and c^b are two phonotactic features extracted from an utterance using two different adapted models. We define the diversity between the two phonotactic features using the Euclidean distance between their "seen" phone *n*-gram statistics as:

$$d(c^a, c^b) = \frac{\sqrt{\sum_{i \in U} (c_i^a - c_i^b)^2}}{n_U} \qquad (9)$$

where c_i^a and c_i^b are the i-th *n*-gram statistics from the two phonotactic features, U is the set of *n*-gram statistics which are non-zero in both or either of the phonotactic features, and n_U is the number of elements in the set U.

3. EXPERIMENTAL SETUP

We conducted our experiments on the 30-sec test trials of the 2007 NIST Language Recognition Evaluation (LRE) tasks [10]. Given a speech trial, the system decided whether the target language was spoken. There were 14 target languages in the evaluation task. In the experiments, we reported the results on the 30-second closed-set trials. Equal error rate (EER) was used to evaluate system performance. To ensure each target language have an equal contribution to the EER, all the trials are grouped according to their target languages and an average EER is calculated from the EER of each target language group. The training data of the target languages was obtained from CallFriend corpus, OHSU corpus and the training dataset of NIST LRE-07.

In the frontend, an English phone recognizer was trained using around 15 hour utterance in Switchboard I Cellular corpus. A standard three-state left-to-right HMM topology was adopted. 32 Gaussian components per state were used in each phone model. Twelve Mel-frequency cepstral coefficients (MFCC) with the energy coefficient as well as the first and second derivatives (i.e. 39-dimension feature) were used and normalized using Histogram equalization [11].

Two-pass decoding (Three-pass in certain systems using the SAT model) for the training data of each target language/dialect and test speech data was performed. A monophone loop grammar was used. In the first-pass decoding, single best phone sequences were generated. In the second-pass decoding with the model adaptation (the third-pass in certain systems using the SAT model), phone lattices were generated and then converted to expected *n*-gram statistics (*n*≤3) for forming high dimensional phonotactic features in the VSM. To optimize for the language recognition task, a phone insertion penalty of 0 was chosen in phone/phone lattice decoding.

4. EXPERIMENTAL RESULTS

4.1. Comparison of Model Adaptation Techniques

In the first experiment, we compared the performance of different model adaptation techniques with the SI and SAT models. Their performance in the PR-VSM system is shown in Table 1. When SI model was used, different model adaptation techniques provided a similar system improvement. CMLLR and mean-only MLLR adaptation performed the best, improving around 10% relatively in EER over the system without adaptation. Although mean-and-variance MLLR could further increase the likelihood of the adaptation data as described in Eq. (8), it did not provide a further improvement in language recognition.

Table 1. Performance of PR-VSM systems with different adaptation techniques and different phone model types

Phone Model Type	System ID	Adaptation Technique	EER (%)
SI	A1	No	4.39
	A2	CMLLR	3.97
	A3	Mean-only MLLR	3.97
	A4	Variance-only MLLR	4.18
	A5	Mean-and-variance MLLR	4.14
SAT	B1	CMLLR	3.71
	B2	CMLLR => Mean-only MLLR	3.94
	B3	CMLLR => Variance-only MLLR	3.85

Generally a further performance gain could be obtained by model adaptation when the SAT model was used. Since mean-and-variance MLLR did not improve the language recognition performance, it was not further tested with the SAT model. Moreover, our experiment found that when mean-only and variance-only MLLR were used with the SAT model, a better performance could be achieved if CMLLR was applied before applying the corresponding transforms. This was the reason why CMLLR was always used first when decoding with the SAT model. The best individual configuration, which uses the SAT model with CMLLR adaptation, provides an overall 15% relative improvement over the system using SI model and without adaptation.

4.2. System Fusion with Different Model Adaptation Techniques

Secondly, we tested whether the phonotactic systems with different types of adapted models provided complement information to each other and whether the corresponding system fusion could provide a further system improvement. The system fusion is performed using an equal-weight combination of the trial scores from different systems.

By considering the eight phonotactic systems listed in Table 1, 28 possible ways of two-system fusion were made. For each fused system, the average diversity between the two sets of phonotactic features was measured using Eq. (9) and the EER was calculated. A graph showing the relationship between the average feature diversity and the corresponding EERs is given in Figure 1. A trend can be observed that when two systems with more diversified phonotactic features are fused, better performance can be achieved in the fused system.

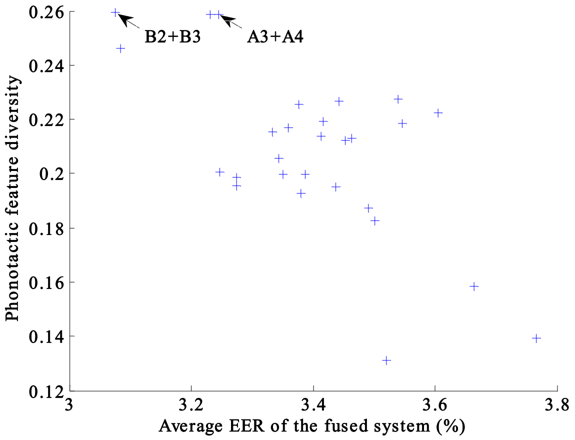

Figure 1. Phonotactic feature diversity and EER(%) in fusion of any two PR-VSM systems

In the last experiment, we fused the system using mean-only MLLR (system A3 and B2) and variance-only MLLR (system A4 and B3). In the diversity test, we found that these two kinds of adaptation provide relatively high phonotactic feature diversity between each other (see system fusion A3+A4 and B2+B3 in Figure 1). Our preliminary test showed that more gain could be made if more systems with different model types and adaptation techniques are fused. However, this involves many passes of decoding (and with different models) and makes the resultant system inefficient. It is worth noting that our major interest is to investigate whether incorporating different adapted models leading to highly diversified phonotactic features can effectively improve the language recognition performance. The fused system is obtained by an equal-weight combination of the trial scores from different systems, each with different adaptation techniques, and the corresponding result is shown in Table 2.

Table 2. Fusion of PR-VSM systems using mean-only and variance-only MLLR adaptation

Phone Model Type	System Fusion	EER(%)	Relative Improvement in EER (%)*
SI	A3 + A4	3.25	12.4
SAT	B2 + B3	3.07	17.3

* Comparison based on system B1 (EER: 3.71% as shown in Table 1)

The fusion of systems with diversified phontactic features showed a substantial performance improvement. Even when the SI model was used, the fusion of systems using mean-only and variance-only MLLR adaptation could outperform the system using the SAT model and CMLLR adaptation. This fusion still provided substantial improvement when the SAT model was used. The fused system using the SAT model obtained the EER of 3.07%, which represents a 30% relative improvement over the system using the SI model and without adaptation. The improvement should be comparable with the performance gain obtained by a parallel phone recognizer frontend.

5. CONCLUSIONS

In this paper, various types of CMLLR/MLLR adaptation techniques prior to the phone lattice decoding are studied in a phonotactic language recognition system. Apart from phonotactic feature diversification provided by parallel phone recognizers, diversification using parallel acoustic model adaptation is illustrated. Generally CMLLR or mean-only MLLR adaptation is used in phone lattice decoding. However, it is showed that an independent variance-only MLLR adaptation can give another set of phonotactic features, which provides complementary information to the original one.

Our experiment shows that parallel model adaptation can provide a substantial language recognition improvement even when SAT model and CMLLR adaptation are adopted. The parallel model adaptation effectively reduces the mismatch between training and test data (e.g. speaker and session variation). This performance gain is comparable to the one obtained by a parallel phone recognizer frontend. In our future work, the interaction with a parallel phone recognizer frontend will be studied.

6. REFERENCES

[1] M. F. BenZeghiba., J. L. Gauvain, L. Lamel, “Context-Dependent Phone Models and Models Adaptation for Phonotactic Language Recognition,” Interpseech 2008, pp.313-316.

[2] W. Shen, D. Reynolds, “Improving Phonotactic Language Recognition with Acoustic Adaptation,” Interspeech 2007, pp. 358-361.

[3] K. C. Sim, H. Li, "On Acoustic Diversification Front-end for Spoken Language Identification," IEEE Transactions on Audio, Speech and Language Processing, Vol 16, No. 5, pp.1029-1037, 2008.

[4] J.L. Gauvain, A. Messaoudi, and H. Schwenk, “Language recognition using phone lattices,” In Proc. ICSLP, 2004.

[5] H. Li, B. Ma, and C.-H. Lee, “A vector space modeling approach to spoken language identification,” IEEE Trans. Audio, Speech, Lang. Process, vol. 15, no. 1, pp. 271–284, Jan. 2007.

[6] W. M. Campbell et al., “Advanced language recognition using cepstra and phonotactics: MITLL system performance on the NIST 2005 language recognition evaluation,” in Proc. IEEE Odyssey: The Speaker and Language Recognition Workshop, pp.1-8, 2006.

[7] P. Matejka, P. Schwarz, L. Burget, and J. Cernocky, “Use of antimodels to further improve state-of-the-art PRLM language recognition system,” in IEEE Proc. Int. Conf. Acoust., Speech, Signal Process., 2006, pp. I-197–I-200.

[8] P. C. Woodland, D. Pye, M. J. F. Gales, “Iterative Unsupervised Adaptation Using Maximum Likelihood Linear Regression,” In Proc. ICSLP, pp. 1133-1136, 1996.

[9] M. J. F. Gales and P. C. Woodland, “Mean and Variance Adaptation within the MLLR Framework,” Computer, Speech & Language, vol. 10, pp. 249-264.

[10] NIST Language Recognition Evaluation 2007, http://www.itl.nist.gov/iad/mig/tests/lre/2007/LRE07EvalPlan-v8b.pdf.

[11] A. De la Torre et al., “Histogram equalization of the speech representation for robust speech recognition,” IEEE Trans. on Audio and Speech processing, vol. 13, no. 3, pp. 355–366, 2005.