USING PSEUDO-KEY FOR LANGUAGE RECOGNITION SYSTEM DESIGN

Hanwu Sun, Bin Ma and Haizhou Li

Institute for Infocomm Research

Agency for Science, Technology and Research (A*STAR), Singapore.

{hwsun, mabin, hli}@i2r.a-star.edu.sg

ABSTRACT

In this paper, we present a novel pseudo-key analysis approach for the fusion system of language recognition. The state-of-the-art language recognition systems for the NIST Language Recognition Evaluation (LRE) commonly consist of multiple language classifiers. To avoid the fusion system to be spoiled by one abnormal classifier, pseudo keys are designed to check the integrity of each of the individual classifiers before the system fusion. The scores of individual classifiers are cross-validated based the pseudo keys. The language recognition experiments are conducted on the 2007 NIST LRE corpus based on the Institute for Infocomm Research's submission.

Index term: Language recognition, NIST language recognition evaluation, language, design.

1. INTRODUCTION

Spoken language recognition (SLR) is the process of determining the identity of the language in a spoken utterance. It has become an enabling technology in many applications such as multilingual speech recognition and translation, and spoken document retrieval. In the past decade, the National Institute of Standards and Technology (NIST) has conducted a series of evaluation of SLR technology [1]. The language recognition evaluations (LREs) focus on language and dialect detection in the context of conversational telephony speech. The Institute for Infocomm Research (IIR) team has participated in the 2005 and 2007 NIST LREs and demonstrated the state-of-the-art technologies.

A successful language recognition system should make good use of the discriminative cues for spoken language from multiple resources. These cues mainly come from the acoustic features and phonotactic representations. It is generally agreed upon that the fusion with different cues of discriminative information can improve the performance of language recognition system [2, 3]. Each of these cues can be modeled by a language classifier in the fusion system. Each classifier has a weight in the fusion system according to its performance in the development data set while all the weights sum to 1.0. The weights are estimated through the

development data set, and will be used directly in the fusion system for the evaluation.

In the NIST evaluations, sufficient training and development data with the true keys are available to train the individual language classifiers as well as the fusion weights. In such case, we can guarantee that all the selected language classifiers in the fusion system work well on the development data. However, the true keys of evaluation data are unknown. If a classifier is out of order in processing the evaluation data, the recognition system will be spoiled by including the abnormal classifier in the fusion system with the same weight. So it is necessary to design a strategy to find out the abnormal classifier and prevent the error in the final fusion system.

In this paper, we propose the pseudo-key analysis approach to check the integrity of each of the individual classifier before the system fusion, to avoid a spoiled system for final language recognition. A set of pseudo keys are extracted to cross-validate the scores of individual classifiers. The experiments are conducted on the 2007 NIST Language Recognition Evaluation (LRE) corpus [4].

The paper is organized as follows. Our fusion system for the 2007 NIST LRE submission is briefly described in Section 2. The pseudo-key analysis is introduced in Section 3. Section 4 presents the experiment observations. Finally in Section 5, we make a conclusion.

2. LANGUAGE RECOGNITION SYSTEM

2.1. Individual Language Classifiers

The Institute for Infocomm Research (IIR) system submitted to the 2007 NIST LRE is a fusion of multiple language classifiers. There are 6 classifiers based on phonotactic features and 5 classifiers based on acoustic features, altogether 11 classifiers in the fusion system.

Phonotactic classifiers adopt a set of parallel phone recognizers (PPR) as the front-end to convert a spoken utterance into sequences of phones. PPR-LM [5] classifier uses language model (LM) scores to capture the phonotactic constraints for each target language. PPR-VSM [6] classifier expresses the phone sequences with a high dimensional phonotactic feature vector based on the n-gram counts, and applies the vector space modeling (VSM) for language recognition. TOPT-VSM [7] classifier further improves the

performance of PPR-VSM by recruiting additional phone tokenizers without the need of additional annotated speech samples and additional acoustic modeling efforts. PAD-PPR-VSM [8] makes use of phonetic and acoustic diversification (PAD) to train the PPR with different acoustic model training methods. The above-mentioned four classifiers make use of the seven phone recognizers developed in IIR. Another two phonotactic classifiers, BUT-PPR-LM and BUT-PPR-VSM adopt the phone recognizers developed by the Faculty of Information Technology of the Brno University of Technology (BUT)¹.

The five acoustic feature based language classifiers include ML-GMM using maximum likelihood (ML) training for GMM modeling, MMI-GMM adopting maximum mutual information (MMI) training [9], MFCC-GLDS applying the generalized linear discriminate sequence kernel (GLDS) [10] for the SVM modeling based on MFCC features, LPCC-GLDS applying GLDS SVM based on LPCC features, and PSK-SVM [11] adopting probabilistic sequence kernel (PSK) for SVM modeling.

2.2. System Fusion

The final system was obtained by means of linear fusion of the scores from the 11 individual classifiers:

$$s_i = \sum_{f=1}^N w_f s(f, i) + b, \quad (1)$$

where N is the total number of classifiers and $s(f, i)$ is the score of the i -th language recognition trial from the f -th classifier. The fusion parameters consist of the classifier specific weights w_f and the global bias b . The following minimum equal error rate (EER) objective was used to tune the fusion parameters:

$$(w_f, b)_{\min EER} = \min_{w_f, b} \{ \|P_{miss} + P_{FA}\| \leq P_{miss} = P_{FA} \}, \quad (2)$$

where the miss and false alarm probabilities are given by

$$\begin{aligned} P_{miss} &= \frac{|\{i : i \in \text{True}, s_i < b\}|}{|\{i : i \in \text{True}\}|} \\ P_{FA} &= \frac{|\{i : i \in \text{False}, s_i \geq b\}|}{|\{i : i \in \text{False}\}|} \end{aligned} \quad (3)$$

and $|\{\dots\}|$ denotes the cardinality of the set.

3. PSEUDO KEY

The fusion weights w_f shown in (1) are estimated using the development data, and will be used directly in the fusion system on all the trials of the evaluation data. For each of the individual classifiers in the fusion system, front-end feature extraction, channel and session variability compensation and speaker model training are requested to process the evaluation trials. If errors happen in one of the

procedures, the classifier will be spoiled and definitely bring serious impact to the final language recognition system.

The keys of the evaluation trials are unknown, but we need to find out a way to check the integrity of each individual classifier before the system fusion. We propose to design a set of pseudo keys as if they are the true keys of the evaluation trials. A pseudo-key analysis will be applied to cross-validate the scores between each pair of classifiers, to confirm that all the classifiers are normally conducted on the evaluation data.

A language classifier is normally formulated as a hypothesis test. For each target language l , we build a language detector which consists of two language models $\{l^+, l^-\}$. l^+ may be trained on the data of target language, while l^- may be trained on those data of its competing languages. We can define the confidence of a test sample O belonging to language l^+ as the posterior odds in a hypothesis test under the Bayesian interpretation. We have H_0 , which hypothesizes that O is language l^- , and H_1 , which hypothesizes otherwise. The posterior odd is approximated by the likelihood ratio λ that is used for the final language recognition decision:

$$\lambda = \log \left(\frac{p(O|l^+)}{p(O|l^-)} \right). \quad (4)$$

Suppose that there are M trials for the L target languages. For the close test, the genuine/imposter ratio is given by $1:(L-1)$. From the pool of scores of M trials from each classifier, f , we choose the M/L trials with highest scores as the genuine trials and the remaining trials as impostor trials for the pseudo keys, i.e.,

$$\tilde{k}(f, i) = \begin{cases} \text{True}, & \text{if } s(f, i) \geq \tilde{\lambda}_f \\ \text{False}, & \text{if } s(f, i) < \tilde{\lambda}_f \end{cases}, \quad (5)$$

where $\tilde{k}(f, i)$ denotes the pseudo-key for the i -th trial of the f -th classifier and the threshold $\tilde{\lambda}_f$ is set such that there are M/L trials whose scores are above it. In the above equation, $s(f, i)$ represent the score of the i -th trial from the f -th classifier. Using the pseudo keys from all other classifiers, we can compute the Pseudo EER for the f -th classifier as

$$EER_{\text{pseudo}}(f) = \frac{1}{N-1} \sum_{g=1, g \neq f}^N EER\{s(f, i) | \tilde{k}(g, i)\}, \quad (6)$$

where $EER\{\cdot\}$ is the operator computing the EER, and N is the total number of classifiers.

It is expected there is a bias between the actual EER and Pseudo EER for each classifier. Since we know the exact Pseudo EER and actual EER for the development data, the bias in the development data set can be calculated. If the classifier works well on the evaluation data too, it is expected that such a bias might be kept. The estimated pseudo EER on the evaluation data can be adjusted as:

$$EER_{\text{pseudo}}(f)|_{\text{Eval}} = \frac{1}{N-1} \sum_{g=1, g \neq f}^N EER\{s(f, i) | \tilde{k}(g, i)\} + \Delta EER(f)|_{\text{Dev}}, \quad (7)$$

¹http://www.fit.vutbr.cz/research/groups/speech/index_e.php?id=p-hnrec

and

$$\Delta EER(f)|_{Dev} = EER(f)|_{Dev} - EER_{Pseudo}(f)|_{Dev}.$$

where the subscript ‘Dev’ and ‘Eval’ stand for the development data and evaluation data, respectively. $EER(f)|_{Dev}$ is the actual EER of f -th classifier on the development data set.

4. EXPERIMENTS

We will evaluate the pseudo-key analysis on the 2007 NIST LRE General LR 30-second tasks, both of the closed-test and open-test. The General Language Recognition (LR) task includes 2510 trials in the 14 target languages. We present the pseudo EER using the proposed pseudo-key approach, and study how one abnormal system affects the results on the final system fusion.

4.1. Pseudo EER Results

We applied the pseudo-key approach to analyze the performance of 11 classifiers on the 2007 NIST LRE development data set as well as the evaluation data set. The pseudo EERs were computed using (6) and (7). Figure 1 compares the pseudo EERs and actual EERs for all the 11 classifiers in the closed-test task. It is shown that the pseudo EERs and actual EERs on both the development and evaluation data sets are generally consistent, and the pseudo EERs therefore can provide a good indication of the performance of the classifiers.

Figure 1: Pseudo and Actual EERs Evaluated on the Development and Evaluation Sets of the 2007 NIST LRE (30s General LR close-test condition)

To confirm the reliability of pseudo-key analysis, we carried out the T-test [12] for the confidence level test between the actual EER and Pseudo EER on the 2007 NIST LRE closed-test task. For the development data set, its significance/probability (at 5% significance level two-tailed test) between Pseudo EER and actual EER is 94.29% and its 95% confidence interval on the estimated EER is [-1.0503, 1.0973]. Similarly, we can get the T-test results for the evaluation data set. The significance is 95.55% and its 95% confidence interval on the estimated EER is [-1.6031,

1.5832]. Obviously, we achieve similar significance levels in the T-test on both the development and evaluation data for the closed-test task. Although there are some differences in the 95% confidence intervals of the EER between the development data and evaluation data, it is still within our expected levels.

Similar experiments were conducted for the open-test task. The pseudo EERs and actual EERs are shown in Figure 2 where the similar pattern with that for the closed-test can be observed. For the T-Test on the development data set, the significance/probability (at 5% significance level two-tailed test) between Pseudo EER and actual EER is 95.34% and its 95% confidence interval on the estimated EER is [-1.1188, 1.0693]. For the evaluation data, the significance is 93.55% and its 95% confidence interval on the estimated EER is [-1.4444, 1.4710].

Figure 2: Pseudo and Actual EERs Evaluated on the Development and Evaluation Sets of the 2007 NIST LRE (30s General LR open-test condition)

4.2. Fusion Results

In this section, we study how an abnormal classifier can be traced by pseudo-key analysis and how it affects the final fusion system. The minimum EER objective function in (2) was used to estimate the fusion weights. In order to simulate an abnormal classifier, we randomly chose a classifier and manually reduced the score values of genuine trials and kept the score values of imposter trials unchanged:

$$s(f, i)_{downgrade} = \begin{cases} s(f, i) - C & \text{ith trial is genuine} \\ s(f, i) & \text{otherwise} \end{cases}, \quad (8)$$

where C is a positive constant. Obviously, the larger C , the worse the classifier is. In such a way, we can get the various simulated EERs for this classifier.

We used 7th classifier as the spoiled classifier and kept the remaining 10 classifiers unchanged. While we increased the constant C in (8), the system fusions as in (1) with the same fusion weights were conducted, and their fusion results on the 2007 NIST LRE evaluation data, as well as both of pseudo EERs and actual EERs for 7th classifier are shown in Figure 3.

From Figure 3, we can see that the pseudo EERs are general consistent with the actual EERs at variety of spoiled classifier 7. It also proves that the proposed pseudo-key

analysis can work well for wide range EER conditions. It is not surprised that the final fusion results are decreased as the increased EERs of the classifier 7. One abnormal classifier can greatly affect the final fusion system, so it is important to check the integrity of each of the individual classifiers before the system fusion.

Figure 3: The EERs and Fusion Results under Various Spoiled Level of Classifier 7 (S7) for the 2007 NIST LRE Closed-test (30s General-LR task)

Similarly, the same analysis for the 2007 NIST LRE open-test set was conducted. The results are shown in Figure 4. The similar observations as that for Figure 3 can be obtained.

Figure 4: The EERs and Fusion Results under Various Spoiled Levels of Classifier 7 (S7) for the 2007 NIST LRE Open-test (30s General-LR task)

Table 1: The EER Simulation for the Fusion System under Various Spoiled Levels of Classifier 7

	EER %				
	Baseline	Downgrade simulated Scores			
Close test Classifier 7	3.81	5.75	8.73	23.11	45.95
Close test Fusion	2.19	2.59	3.10	4.22	5.51
Open test Classifier 7	4.87	7.31	11.32	28.72	51.83
Open test Fusion	2.32	2.83	3.29	4.49	6.07

Table 1 shows the simulation for the fusion system under various spoiled levels of classifier 7 by using the

actual keys. The 2007 NIST LRE evaluation data set for both the closed-test and open-test are involved.

5. CONCLUSION

In this paper, a new analysis approach with pseudo keys was proposed to predict the performance of language classifiers in a fusion system of multiple classifiers. From the experimental results on both the closed-test and open-test tasks of the 2007 NIST LRE 30-second general language recognition, the EERs predicted with the pseudo-key analysis are reasonable consistent with the actual EERs. The reliability of pseudo-key analysis has also been confirmed by carrying out the T-test for the confidence level test between the actual EER and Pseudo EER. The simulated abnormal classifier results demonstrated that a spoiled classifier can greatly affect the final fusion system. It is expected that such pseudo key analysis approach can also be applicable for the speaker recognition system.

6. REFERENCE

- [1] <http://www.nist.gov/speech/tests/lang/2007/>.
- [2] E. Singer, P. A. Torres-Carrasquillo, T. P. Gleason, W. M. Campbell and D. A. Reynolds, "Acoustic, Phonetic and Discriminative Approaches to Automatic Language Recognition," in *Proc. Eurospeech*, 2003.
- [3] R. Tong, B. Ma, D. Zhu, H. Li and E. S. Chng, "Integrating acoustic, prosodic and phonotactic features for spoken language identification," in *Proc. ICASSP*, 2006.
- [4] The 2007 NIST Language Recognition Evaluation plan, <http://www.nist.gov/speech/tests/lang/2007/LRE07EvalPlan-v8b.pdf>.
- [5] H. Li, B. Ma, and C.-H. Lee, "A Vector Space Modeling Approach to Spoken Language Identification," *IEEE Transaction on Audio, Speech and Language Processing*, Vol. 15, No. 1, 2007.
- [6] B. Ma, H. Li, and R. Tong, "Spoken Language Recognition Using Ensemble Classifiers," *IEEE Transactions on Audio, Speech and Language Processing*, Vol. 15, issue 7, pp. 2053-2062, 2007.
- [7] R. Tong, B. Ma, H. Li and E. S. Chng, "Target-oriented Phone Tokenizers for Spoken Language Recognition," in *Proc. ICASSP*, 2008.
- [8] K. C. Sim and H. Li, "Fusion of Contrastive Acoustic Models for Parallel Phonotactic Spoken Language Identification," in *Proc. Interspeech* 2007.
- [9] Burget L., Matejka P. and Cernocky J., "Discriminative training techniques for acoustic language identification," *ICASSP*, 2006.
- [10] W. M. Campbell, J. P. Campbell, D. A. Reynolds, E. Singer and P. A. Torres-Carrasquillo "Support vector machines for speaker and language recognition," *Computer Speech and Language*, vol. 20, pp. 210-229, 2006.
- [11] K. A. Lee, C. You, and H. Li, "Spoken language recognition using support vector machines with generative from-end," in *Proc. ICASSP*, 2008,
- [12] R. E. Walpole and R. H. Myers. *Probability and Statistics for Engineers and Scientists*. Macmillan, Inc., New York, 4th edition, 1989.