International Symposium on Chinese Spoken
Language Processing (ISCSLP 2004)
Hong Kong
December 15-18, 2004

LANGUAGE IDENTIFICATION USING DISCRIMINATIVE WEIGHTED LANGUAGE MODELS

Shizhen WANG, Jia LIU, Runsheng LIU
Department of Electronic Engineering, Tsinghua University, Beijing
wangsz02@mails.tsinghua.edu.cn

ABSTRACT

In this paper, discriminative weighted language models are proposed to better distinguish between similar languages. Through Parallel Phone Recognizers followed by Language Modeling (PPRLM) system in the first stage, two best candidates are hypothesized and then processed using discriminative language models. Experimental results show that, compared with the traditional one-pass language identification (LID) systems, the proposed two-pass method can greatly improve the performance without considerably increasing the computational costs. Tested on the evaluation set of CallFriend corpus, the final system achieved an error rate of 14.90% on the 30s 12-way close-set task.

1. INTRODUCTION

With the increasing demand on multilingual services as a result of the large spread of human-machine interfaces and the explosion of international telecommunications, there has been a great deal of research in the field of automatic language identification (LID) [1, 2,], which plays an integral role in many multilingual speech-based systems. Much of the research so far has placed its focus on the approach of Parallel Phone Recognizers followed by Language Modeling (PPRLM), which directly employ phonotactics information to hypothesize the target language in one pass. In this paper, we proposed a somewhat different strategy which exploits PPRLM system in the first stage to choose the two best candidates, and then discriminative weighted language models are used to make the final decision.

The paper is organized as follows. A brief review of the PPRLM system is given in section 2. Then in section 3 we present our proposed system and discuss discriminative weighted language models. The experimental results are shown in section 4. The conclusion is given in section 5.

2. BASELINE PPRLM SYSTEM

The phone based approach, PPRLM system identifies the language of an utterance based on the statistical characteristics of phone sequences [2, 3]. As shown in Figure 1, it mainly consists of two components: the front-end is language-independent phone recognizers running in parallel, which is used to convert speech utterances into phone sequences; the back-end is N-Gram interpolated language models which are trained to learn the phonotactics rules of the languages of interest. During recognition, the phone recognizers tokenize the test utterances into phone sequences, which are then scored against each language model. The final likelihood scores are calculated as the average of the individual log likelihoods emanating from the corresponding language models. The language of the model with the highest score is hypothesized.

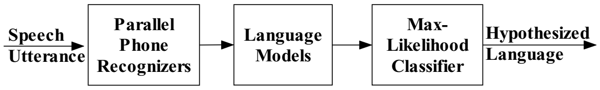

Figure 1. Block diagram of the baseline PPRLM system

3. PROPOSED LID SYSTEM

The block diagram of the proposed LID system is given in Figure 2. The system is composed of the following four components.

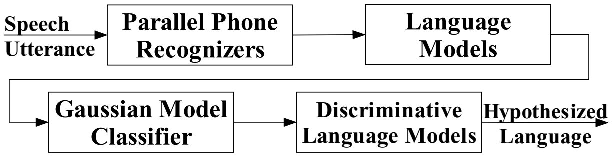
Speech
Utterance
Parallel Phone Recognizers
Language Models
Gaussian Model Classifier
Discriminative Language Models
Hypothesized
Language

Figure 2. Block diagram of the proposed LID system

3.1 Parallel Phone Recognizers

This component tokenizes the incoming speech utterance into phone sequences. In this paper, HMM based phone recognizers were trained using a phonetically labeled subset of the OGI training speech in each of the following six languages: English, German, Hindi, Japanese, Mandarin, and Spanish.

3.2 Language Models

Language models take as input the phone sequences and calculate the log likelihood scores for each language to be identified. An interpolated bigram language model[4] is used in our experiments, which is

$$\tilde{P}(w_t \mid w_{t-1}) = \alpha_2 P(w_t \mid w_{t-1}) + \alpha_1 P(w_t) + \alpha_0$$

where w_{t-1} and w_t are consecutive symbols observed in the phone stream. The *P*'s are ratios of counts observed in the training data, e.g.

$$P(w_t \mid w_{t-1}) = C(w_{t-1}, w_t) / C(w_{t-1})$$

where $C(w_{t-1}, w_t)$ is the number of times symbol w_{t-1} is followed by w_t, and $C(w_{t-1})$ is the number of occurrences of symbols w_{t-1}. The α 's are estimated iteratively using the E-M algorithm so as to minimize perplexity, and in our experiments $\alpha_2 = 0.666$, $\alpha_1 = 0.333$, $\alpha_0 = 0.001$.

During recognition, the test utterance is first passed through the phone recognizer, producing a phone sequence, $W = \{w_1, w_2, \ldots w_T\}$. The log likelihood, L, that the interpolated bigram language model for language l, λ_l produced the phone sequence W, is

$$L(W \mid \lambda_l) = \frac{1}{T} \sum_{t=1}^{T} \log \tilde{P}(w_t \mid w_{t-1}, \lambda_l)$$

where w_0 represents the symbol of starting of a sentence.

3.3 Gaussian Model Classifier

Different from the max-likelihood classifier used in the common PPRLM system, we considered the language model scores of each phone recognizer as elements of a feature vector, and used Gaussian model (GM) with multi-dimension mean and covariance to capture its statistical distribution characteristics. We trained a GM classifier for each of the six front-end phone recognizers, and the final likelihood score equaled the average of six GM scores in log domain. The classifier made a maximum decision based on the total language scores. Instead of hypothesizing the language of the model with the highest score, the classifier output the two best candidates for later processing in a second stage as described in section 3.4.

3.4 Discriminative Language Models

As a post-classification method, discriminative weighted language models are applied to process the two best hypotheses. The basic idea is to improve the significance of individual phone-pairs whose probabilities differ among the languages, while suppressing those having similar bigram values [5]. The discriminative weight γ between languages m and n , which could be regarded as the dissimilarity measure of the two languages, was defined as:

$$\gamma^{mn}_{w_i w_j} = \max\left\{ \frac{\tilde{P}(w_i \mid w_j, \lambda_m)}{\tilde{P}(w_i \mid w_j, \lambda_n)}, \frac{\tilde{P}(w_i \mid w_j, \lambda_n)}{\tilde{P}(w_i \mid w_j, \lambda_m)} \right\}$$

and used to weight the log-probabilities in the score for each of the best hypothesized languages $l, \lambda_l \in \{\lambda_m, \lambda_n\}$:

$$L^*(W \mid \lambda_l) = \frac{1}{T} \sum_{t=1}^{T} \gamma^{mn}_{w_t w_{t-1}} * \log \tilde{P}(w_t \mid w_{t-1}, \lambda_l)$$

The final decision was made based on the new scores. Because the discriminative weights can be computed offline and stored, the additional computational costs coming along with the second stage classification are negligible.

4. EXPERIMENTS AND RESULTS

4.1 Speech Corpus

The data source used to evaluate our system was the CallFriend corpus of conversational telephone speech collected by the Linguistic Data Consortium (LDC) [6]. The corpus consists of telephone conversations between friends in the following 12 languages: Arabic, English, Farsi, French, German, Hindi, Japanese, Korean, Mandarin, Spanish, Tamil, and Vietnamese. The training set was used to train the language models, the development partition to train the Gaussian model classifiers, and the 30s segments from the evaluation set for testing the final system.

4.2 First-Pass Performance of LID Systems

These experiments aimed to compare the N-Best performance of the baseline PPRLM systems using max-likelihood classifier (ML) with that of systems using GM classifier (GM). The results are summarized in Table 1 in terms of error rates (percent).

From table 1 we can see that the use of GM classifier can greatly improve the performance and that the performance increase is relatively small when the candidates number is larger than 2. That’s why the GM classifier output just the two best hypotheses in our experiments, given that the increase in the number of candidates can considerably increase the processing difficulty in the second stage.

Classifier	1-Best	2-Best	3-Best	4-Best	5-Best
ML	24.98	14.22	9.39	6.46	4.67
GM	17.69	9.17	5.36	3.40	1.98

Table 1 First-Pass N-Best performance of LID systems

4.3 Combining Acoustic Scores and Language Model Scores

To further improve the first-pass performance, we use acoustic scores from the parallel phone recognizers as another feature and combine them with language model scores as new feature vectors. They are normalized using linear discriminant analysis (LDA)[7] to decorrelate and decrease the dimension. Table 2 shows the error rates of N-Best candidates where LMS refers to using language model scores as GM feature, and LMS/AS refers to combining language model scores and acoustic scores. The use of acoustic scores further improves the performance.

GM Features	1-Best	2-Best	3-Best	4-Best	5-Best
LMS	17.69	9.17	5.36	3.40	1.98
LMS/AS	17.10	8.35	4.76	2.92	1.67

Table 2 First-Pass N-Best performance of LID systems using combined features

4.3 Second-Pass Performance of LID Systems

Table 3 shows the second-pass performance of the proposed LID systems. From table 3 we can see that discriminative weighted language models can clearly enhance the performance and that the LID system with combined features outperforms the system that only uses language model scores.

GM Features	1st Pass 1-Best	1st Pass 2-Best	2nd Pass Final	Error Rate Decrease
LMS	17.69	9.17	15.92	10.01
LMS/AS	17.10	8.35	14.90	12.87

Table 3 Second-Pass Performance of LID Systems

5. CONCLUSION

Different from the common one-pass LID systems, we propose a system using two-pass strategy, which exploits PPRLM system to hypothesize the two best candidates in the first pass, and applies discriminative weighted language models to make the final decision in the second stage. The purpose of discriminative weighted language models is to improve the significance of the discriminatory phone-pairs whose probabilities differ between languages and suppress the indistinguishable phone pairs whose probabilities are similar between languages. To improve the first-pass performance, language model scores are combined with acoustic scores to make new features and LDA is used to process the features, which are then used in the Gaussian model classifier. Experimental results show that the two-pass strategy can greatly improve the performance of the LID system, while does not considerably increase the computational costs. Evaluated on the CallFriend 30s 12-way close-set task, we obtained an error rate of 14.90%, a decrease of 12.87% compared to the one-pass system.

6. ACKNOWLEDGEMENT

This project is supported by the National Natural Science Foundation of China (NSFC) (No. 60272016).

7. REFERENCE

1. Zissman, Marc A., Berkling, Kay M., “Automatic language identification”, Speech Communication, 35(1-2), pp. 115-124, August, 2001

2. Zissman, M.A., “Comparison of four approaches to automatic language identification of telephone speech”, IEEE Trans. Speech and Audio Proc., SAP-4(1), pp. 31-44, 1996

3. Zissman, M.A., “Language identification using phoneme recognition and phonotactic language modeling”, Proc. ICASSP’95, Vol. 5, pp. 3503-3506, 1995

4. F. Jelinek, “Self-organized language modeling for speech recognition”, Readings in Speech Recognition, A. Waibel and K.-F. Lee, Eds. Palo Alto, CA: Morgan Kaufmann, pp. 450-506, 1990

5. J. Navrátil, W. Zühlke, “Double-bigram decoding in phonotactics language identification”, Proc. ICASSP’97, Vol. 2, pp. 1115-1118, 1997

6. Linguistic Data Consortium, Philadelphia, PA, 1996, http://www.ldc.upenn.edu/Catalog/byType.jsp#speech.telephone, LDC96S46-LDC96S60.

7. M. Schafföner, “Improved robustness of automatic speech recognition using linear discriminative analysis”, Master’s thesis, Otto-von-Guericke University, Magdeburg, March 2003.