AN APPROACH TO LANGUAGE IDENTIFICATION WITH ENHANCED LANGUAGE MODEL

Yonghong Yan *Etienne Barnard*

Center for Spoken Language Understanding
Oregon Graduate Institute of Science and Technology
P.O.Box 91000, Portland, OR 97291-1000
Phone: (503)690-1121 ext. 1637, FAX (503)690-1334
yan@cse.ogi.edu barnard@cse.ogi.edu

ABSTRACT

An approach to Language Identification (LID) based on language-dependent phone recognition is presented. This LID system is designed to exploit varying phonotactic constraints of different languages. Based on the output of language-dependent phone recognizers, various LID features are extracted. Two methods are proposed to enhance the language modeling accuracy, (1) language models based on forward and backward bigrams, and (2) backpropagation based language model optimization. The system was evaluated on a standard 11-language task and a standard nine-language task. The results(correct rate) reached 87.6% for 45-second long utterances and 73.6% for 10-second long utterances for the 11-language task, and reached 87.8% and 74.0% respectively on the nine-language task. By adding channel normalization, the performance of our best systems was further improved to 90.8% and 77.1% for the 11-language task, and 91.1% and 77.5% on the nine-language task.

1. Introduction

This paper presents extensions of our previous work[1,2], which is a language dependent phone recognition based approach to language identification(LID). Our baseline LID system is composed of three parts: (1) language-dependent phone recognizers, (2) LID score generators, and (3) a final language classifier. The baseline system architecture is similar to that in [5,6].

In [2] we introduced a forward and backward bigram based language model and evaluated our system on a six-language task. Recently, our system has been generalized for an 11-language task; a novel language model optimization method based on error backpropagation is introduced in this paper. The methods proposed were evaluated on a standard 11-language task. Comparative experiments with traditional bigram-based language models were also conducted. The results show that the enhanced language model increases the discrimination between different languages. The rest of paper is organized as follows: Section 2 provides an overview of our LID system, Section 3 presents the methods used to enhance language modeling accuracy, Section 4 describes the database, Section 5 presents the experiments and results, and a discussion is presented in Section 6.

2. Overview of the LID System

Our LID system is composed of *three* parts: language dependent phone recognizers, LID score generators and the final classifier. A general system architecture for an N-language task is given in Figure 1. For an N-language task, M($N \geq M$) language-dependent phone recognizers were implemented. These general-purpose recognizers run in parallel, and independently decode the input speech vectors into phone strings. The score generator of each recognizer takes the decoded phone string, estimates various LID feature scores and sends them to the final classifier. As shown in the figure, for each language in the task, a score generator contains a set of LID models, described in terms of the phone inventory of the language-dependent recognizer associated with this score generator . The scores calculated by the generator were normalized by the number of phones in the decoded path of the test utterance before sending to the final classifier. The final classifier identifies the language spoken based on the output of all the score generators. Two sets of LID models were used in our baseline system, (1) a bigram $b(p_i|p_{i-1})$

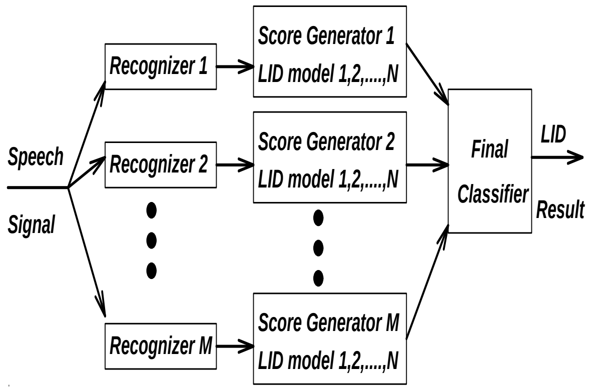

Figure 1: General Structure of the LID system

based language model(b_{LF}), and (2) a duration model. The language model used in our baseline system is the commonly used bigram based language model:

$$b_{LF} = \sum_{i=1}^{T} log(\alpha b(p_i|p_{i-1}) + \beta u(p_i)), \quad (1)$$

where $u(p_i)$ is the unigram term, and p_i is the *ith* phone in decoded path.

The duration model is a generalized left context-dependent model, represented by a set of histograms. For each phone, six duration models are estimated depending on whether its preceding phone is a *vowel, fricative, stop, nasal, affricative* or *glide*. The duration models used in this paper are given as:

$$P_D = \sum_{i=1}^{T} log((1-\alpha)P(O_i|O_{i-1} \in S) + \alpha P(O_i)) \quad (2)$$

where $P(O_i|O_{i-1} \in S)$ is the context-dependent model, and S is one of the six broad categories. $P(O_i)$ is the original context independent monophone duration model, which is used here as a smoothing factor with weight α.

3. Two Methods used to Enhance the Language Model

The Language model is the central part of our system. In this section we will discuss the two major methods used to enhance the modeling accuracy.

3.1. Two-bigram-based Language model

Our language model is designed to exploit sequential constraints on phonemes, which is a natural reflection of the phonotactic differences between different languages. The commonly used bigram-based($b(p_i|p_{i-1})$) language model exploits left context information; thus only the forward information is captured. With a limited amount of training data, we propose using two bigram models b($p_i|p_{i-1}$) (forward) and b($p_i|p_{i+1}$) (backward) in the language model. One possible way to combine these two bigram models into a language model is:

$$b_{LFB} = \sum_{i=1}^{T} log(\alpha b(p_i|p_{i-1}) + \beta b(p_i|p_{i+1}) + \gamma u(p_i)) \quad (3)$$

Compared with a trigram-based language model, which has approximately N^3 parameters, the number of parameters in our model is order N^2, which decreases the requirement of training data size while still providing a good approximation of left and right context information.

3.2. Optimization of the language model

A cost function based on the maximum likelihood LID scores calculated by a score generator is proposed in this paper. For any score generator in the system, the cost function for a particular input utterance is defined as:

$$E = (S_T - S_M)^2 \quad (4)$$

where

$$S_M = \max_i S_i, \quad i = 1, ..., N \quad (5)$$

S_i is the LID score for language i in an N-language task, S_T is the score for the expected target language($S_T \in S_i$). When $T = M$, *i.e.*, the correct language has the highest score, $E = 0$, and when $T \neq M$, E provides a quantitative measurement for the error in the language model. By back propagating the error, optimization of the language model for this score generator will be achieved.

Here we give an example to illustrate how the optimization is realized; this example is for the optimization of the bigram term in the forward language model. Other terms in the score generator can be optimized similarly.

By using (1) and (4),

$$\begin{aligned} E &= (S_T - S_M)^2 \\ &= (\sum_{i=1}^{T} log(\alpha b(p_i|p_{i-1}) + \beta u(p_i)) - S_M)^2 \end{aligned}$$

$$\frac{\partial E}{\partial b(p_i|p_{i-1})} = 2(S_T - S_M)\frac{\partial S_T}{\partial b(p_i|p_{i-1})}$$

$b(p_i|p_{i-1})$ is updated as,

$$\Delta b(p_i|p_{i-1}) = -\eta \frac{\partial E}{\partial b(p_i|p_{i-1})} \quad (6)$$

where η is the learning rate. *i.e.,*

$$b_{new}(p_i|p_{i-1}) = b(p_i|p_{i-1}) + \Delta b(p_i|p_{i-1}) \quad (7)$$

After one iteration, re-normalization of the bigram probability results in a new bigram term for the language model. A cross validation procedure is performed, and the criterion for stopping the optimization procedure is based on the system performance on the cross validation data set using the new model.

4. Database

The Oregon Graduate Institute Multi_Language Telephone Speech Corpus (OGI_TS)[4] was used. The 11 languages in the task are: English, Farsi, French, German, Hindi, Japanese, Korean, Mandarin, Spanish, Tamil, Vietnamese. Six language-dependent continuous HMM-based phone recognizers were implemented, for English, German, Hindi, Japanese, Mandarin and Spanish(which have been phonetically labeled). The model for each context-independent phone is a three-state left-to-right HMM, with 3 Gaussian mixtures on each state. The phone accuracy of the recognizers are around 45% to 55%.

In our implementation, Speech wave data are parameterized every 20ms with 10ms overlap between contiguous frames. For each frame a feature vector with 26 dimensions is calculated: 12 LPC_cepstra, 12 delta cepstra, normalized energy and delta energy.

Four parts of the utterances in the database were used: "story-before-the-tone" (story-bt), "story-after-the-tone" (story-at), "rooms"(room) and "numbers"(num). The training and development sets had 1785 utterances total, which were further divided into following three sets.

- Training set. There were 716 utterances in this subset.
- Development set 1. There were 631 utterances in this subset.
- Development set 2. There were 438 utterances in this subset.

The numbers of utterances in each language were approximately balanced.

Training set were used to train all the LID models. All the sets were used to train and develop the final classifier. When the proposed optimization method used, Development set 2 was used for cross-validation. The test set used was the same test set used by the National Institute of Standards and Technology(NIST) in their March 1994 evaluation for these 11 languages. It contained 195 whole utterances(45-second long) and 625 ten-second utterances, which were also part of OGI _TS database.

The nine-language NIST'95 evaluation test set was also used to evaluate our best systems. The nine languages are: English, French, German, Hindi, Japanese, Mandarin, Spanish, Tamil and Vietnamese. The test set contains 180 whole utterances and 653 ten-second long utterances.

In both test sets, the numbers of utterances in each language were also balanced.

None of above sets overlapped.

5. Experiments

In order to test the LID performance of the enhanced language model, five experiments were carried out. The baseline system has a linear classifier as the final classifier. The results of each experiment on the 11-language task are given in Table 1, where **BS** denotes the baseline system(with forward bigram based language model and duration model being used). Four enhanced systems are also listed in Table 1. The systems listed as **BS_NN**, **BS_BK**, and **BS_OP** each contain one improvement with respect to the **BS** system. **BS_NN** contains a feed-forward neural-network with conjugate gradient optimization[3] being used as the final classifier, in **BS_BK** both the forward and backward bigram based language models are used, and **BS_OP** includes optimization of the language models. Finally, **BS_EN** denotes our final system with the optimized forward bigram and backward bigram based language model and a neural-net as the final classifier.

Training set and development set 1 were involved in language optimization, and development set 2 was used for cross validation to prevent over-fitting. A small learning rate was used in (7). After three iterations, the optimization procedure stops.

The results of our best systems achieved on the nine language task are given in Table 2. In the table, the results of **BS_EN** were achieved during the actual evaluation.

In the table, **BS_UD** denotes the system which is currently still under development. Here, **BS_UD** is an enhanced version of **BS_EN**. Cepstral mean subtraction[7,8] is used for channel normalization. Performance(phone accuracy) of each phone recognizer achieved 1-10% error reduction compared with the performance of those used in BS_EN. A more detailed report on this improvement will be published in the near future.

6. Discussion

The results show that each method could achieve more than 15% error reduction(compared with the baseline system). When all enhancements are combined, approximately 40% error reduction on the whole-utterance test and 25% error reduction on the ten-second test were achieved. The main reason for more error reduction on the

Approach	whole utterance	10-sec. utterance
BS	78.46%	67.04%
BS_NN	83.59%	71.68%
BS_BK	82.56%	70.08%
BS_OP	83.08%	71.36%
BS_EN	87.69%	73.76%
BS_UD	90.76%	77.12%

Table 1: Results for all the experiments on 11-language task

Approach	whole utterance	10-sec. utterance
BS_EN	87.78%	73.97%
BS_UD	91.11%	77.49%

Table 2: Results for the best systems on nine-language task

whole utterance part is that our techniques focus on phone strings. Since whole-utterance segments are four times as long as the ten-second utterance, the improvement attainable on long utterances is larger than that for the short utterances. When phone constraints are not well reflected in short utterances, acoustic information may become a dominant feature. How to incorporate acoustic features into our current LID system is under consideration.

Comparative experiments between approaches to information integration using linear classifiers and neural-networks supports our previous conclusion on a smaller task[2], which showed the necessity of non-linear combination of all the LID scores from different information sources for better system performance.

The performance of the forward and backward bigram based language model proposed shows that this model is a good compromise between detailed modeling and data limitations. In addition, the results of the experiments demonstrate the feasibility of improving LID system performance by optimizing the model numerically.

The preliminary experiment with channel normalization achieved approximately 20% error reduction on our best system, which suggests that a robust front end is essential to LID approaches based on phone recognition. The best results we obtained compare favorably with previously reported results on these 11-language and nine-language tasks.

7. Acknowledgements

The authors wish to thank Prof. R.A. Cole at OGI for his help and valuable suggestions to this work.

8. References

[1] Y.K. Muthusamy, K.Berkling, T.Arai, R.A.Cole and E.Barnard. A comparison of approaches to automatic language identification using telephone speech. In *Proceedings 3rd European Conference on Speech Communication and Technology (Eurospeech '93)*. Vol.2, pages 1307-1310, September 1993.

[2] Y. Yan, E.Barnard. An Approach to Automatic Language Identification Based on Language-dependent Phone Recognition. In *International Conference on Acoustic, Speech and Signal Processing(ICASSP) 1995*. Vol.5, pages 3511-3514, May 1995.

[3] E.Barnard and R.A.Cole. A neural-net training program based on conjugate-gradient optimization. *Technical Report CSE 89-014*, Oregon Graduate Institute, 1989.

[4] Y.K. Muthusamy, R.A. Cole and B.T. Oshika. The OGI multi-language telephone speech corpus. In *Proceedings International Conference on Spoken Language Processing 92*. Vol.2, pages 895-898, October 1992.

[5] M.A. Zissman and E.Singer. Automatic language identification of telephone speech messages using phoneme recognition and N-gram Modeling. In *1994 International Conference on Acoustic, Speech, and Signal Processing Proceedings*. Vol.1, pages 305-308, April 1994

[6] M.A. Zissman. Language Identification Using Phoneme Recognition and Phonotactic Language Modeling. In *1995 International Conference on Acoustic, Speech, and Signal Processing Proceedings*. Vol.5, pages 3503-3506, May 1995.

[7] S.B. Davis and P. Mermelstein, Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences. In *IEEE Transactions on Acoustic, Speech and Signal Processing*, Vol.28, pages 357-366, August, 1980.

[8] B.S. Atal. Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification. In *The Journal of the Acoustical Society of America*, Vol.55, pages 1304-1312, June 1974.