AN APPROACH TO AUTOMATIC LANGUAGE IDENTIFICATION BASED ON LANGUAGE-DEPENDENT PHONE RECOGNITION

Yonghong Yan   *Etienne Barnard*

Center for Spoken Language Understanding
Oregon Graduate Institute of Science and Technology
20000 N.W.Walker Road, 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 in this paper. A variety of features and their combinations extracted by language-dependent recognizers were evaluated based on the same database. Two novel information sources for LID were introduced: (1) forward and backward bigram based language models, and (2) context-dependent duration models. An LID system using Hidden Markov Models and neural network was developed. The system was trained and evaluated using the OGI_TS database. For a six-language task, the system performance(correct rate) for 45-second long utterances and 10-second long utterances reached 91.96% and 81.08% respectively. The experiments demonstrated the importance of detailed modeling and the method by which these information sources are combined.

1. Introduction

Recent research[1,2,3,4,5] has shown the importance of acoustic, phonotactic and prosodic information for language identification. In order to get an accurate estimation of these information sources, detailed modeling is found to be necessary[1,2]. In this paper, an approach to language identification based on language-dependent phone recognition is presented. Continuous HMMs are used to build the language-dependent phone recognizers. Acoustic models, language models and duration models are exploited for the LID system. A neural network is trained as the final classifier. Comparisons of the performance of different LID models and their combinations are provided. A forward and backward bigram based language model and context dependent duration model for LID are proposed to enhance the modeling accuracy.

The idea of using the phone set of one language to model the phonotactic constraints of all the languages in the task has been proved to be a powerful technique[1,2]. In this paper, this idea is extended. The rest of this paper is organized as: Section 2 describes the models for LID being used in this paper. Section 3 describes the database. Section 4 describes the LID system implementation and Section 5 presents the experiments and results. Concluding remarks and future work are given in Section 6.

2. Models used

In order to capture the acoustic, phonotactic and duration(prosodic) information, three sets of models for language identification are used in this paper.

2.1. Acoustic model

Acoustic models have been used extensively in language identification[3,4,5] to exploit spectral differences between different languages. In this paper, the acoustic model is used for phone recognition as well as for acoustic likelihood calculation for LID. A three-state left-to-right HMM is used for each phone in each language. Four Gaussian mixtures are used to model the probability density for each state in a model.

By assuming contiguous phones in an utterance are independent of each other, the acoustic score (likelihood) can be calculated as:

$$P_A = \sum_{i=1}^{T} log(P(\vec{f}, b_i, e_i | M(O_i))) \quad (1)$$

where O_i is the *ith* phone in the decoded path by Viterbi search. $M(O_i)$ is its corresponding HMM model. $\vec{f}$ is the set of features in the decoded *ith* segment, initiated at time b_i, ended at time e_i. T is the number of phones in the decoded path.

2.2. Language model

House and Neuburg[6] proposed that sequential constraints on phonemes could be exploited as an efficient approach to language identification. Their work showed that the constraints could be a powerful feature to distinguish the languages even when the speech

0-7803-2431-5/95 $4.00 © 1995 IEEE

events were described by broad-category classes. This idea has been used extensively in recent research[1,2,3], and reflects the phonotactic differences between different languages[7]. Our previous work[2] showed that fine phone categories could lead to better performance than broad categories. Therefore, fine phone categories are used in this work.

One commonly used bigram-based language model is:

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

where O_i is the ith phone in the decoded best path, $P(O_i)$ is the unigram term and $P(O_i|O_{i-1})$ is the bigram term, α and β are the weight coefficients. T is the total number of phones in the decoded utterance. This language model is based on the interpolated n-gram language model proposed in [8]. It exploits left context information; thus only the forward information is captured. Although a trigram-based language model can capture both the right and left context information, a larger database is needed in order to get a well estimated model. For a language with N phonemes, approximately N^2 parameters need to be estimated for a bigram model, while for a trigram model N^3 parameters need to be estimated. With a limited amount of training data, we propose using two bigram models $P(O_i|O_{i-1})$(forward) and $P(O_i|O_{i+1})$(backward) in the language model. The backward bigram based language model can be used to capture the backward phonotactic constraints:

$$P_{LB} = \sum_{i=1}^{T} log(\alpha P(O_i|O_{i+1}) + \beta P(O_i)) \quad (3)$$

One possible way to combine these two bigram models into a language model is as:

$$P_{LFB} = \sum_{i=1}^{T} log(\alpha P(O_i|O_{i-1}) + \beta P(O_i|O_{i+1}) + \gamma P(O_i)) \quad (4)$$

Adding the backward bigram term enables the language model to capture both the right and left context information without adding too many parameters to be estimated.

2.3. Duration model

Duration modeling has been used in [3] to capture a certain class of prosodic information in the different languages. Two representations of duration distribution for each phone were evaluated initially, namely

- Gaussian densities and
- Histograms

Initial experiments showed the performance of these two representations were almost the same. The histogram was selected as our representation of the duration model because of its computational simplicity.

Since the variation in the duration of a phoneme in different contexts can be quite large, context-dependent modeling is desired. In order to decrease the number of parameters in the model, a generalized left context-dependent duration model is proposed. For each phone, six duration models are estimated depending on whether its proceeding phone is a vowel, fricative, stop, nasal, affricative or glide. The duration models used in this paper are given as:

$$P_D = \sum_{i=1}^{T} log((1 - \alpha)P(O_i|O_{i-1} \in S) + \alpha P(O_i)) \quad (5)$$

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 monophone duration model, which is used here as a smoothing factor with weight α. In our experiment, α is set to 0.1.

3. Database

The Oregon Graduate Institute Multi_Language Telephone Speech Corpus (OGI_TS)[9] was used. We limit our attention to the six languages(English, German, Hindi, Japanese, Mandarin and Spanish) which have been phonetically labeled. Two parts of utterances in the database were used: "story-before-the-tone" (story-bt) and "story-after-the-tone"(story-at). The training set had 777 utterances total, which was further divided into following three sets.

- Training set 1. There were 300 utterances in this subset, all were phonetically labeled.
- Training set 2. There were 400 utterances in this subset, most of them were not labeled.
- Development set. There were 77 utterances in this subset, most of them were phonetically labeled.

In each set, the numbers of utterances in each language were approximately balanced. Usage of these data sets will be explained later.

The test set used was the part of the test set used by the National Institute of Standards and Technology(NIST) in their March 1994 evaluation for these six languages. It contained 112 whole utterances and 370 ten second utterances, which were also part of OGI_TS database. None of above sets overlapped.

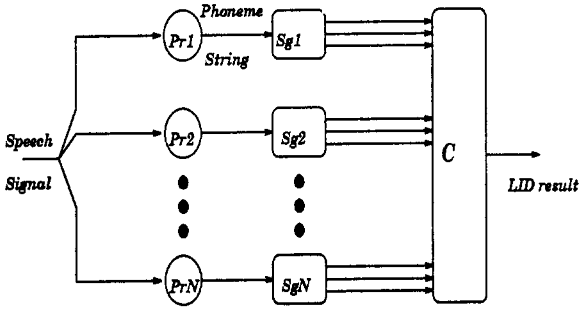

Figure 1: General Structure of the LID system

Language	Number of Phones	Accuracy
English	41	46.79%
German	41	46.57%
Hindi	45	48.13%
Japanese	26	56.25%
Mandarin	41	36.33%
Spanish	30	54.56%

Table 1: Number of phone models and phone recognition accuracies for the six languages

4. Implementation of the LID system

4.1. General architecture of the 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. Here **Pri** and **Sgi** are the language-dependent phone recognizer and score generator for language **i**. **C** is the classifier. The three arrows between each score generator and the classifier represent the scores for acoustic modeling (*1*-dimensional vector), language modeling (*N*-dimensional vector) and duration modeling (*N*-dimensional vector).

4.2. Language-dependent phone recognizer

Six continuous phone recognizers were trained using HMMs, one for each language in the task. The output of the recognizer is the time-aligned phone string with an acoustic probability attached to each phone in the string. The commercial HMM software package HTK V1.5 was used to implement these six language-dependent phone recognizers.

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.

The recognizers were trained using training set 1 and phone recognition accuracies were evaluated by the labeled data in the development set. Table 1 gives the number of phones in each language and the corresponding phone recognition accuracy.

4.3. LID score generator

As shown in Figure 1, the phone recognizer of each language has its own score generator in our system.

After the phone recognizers were implemented, they were used to decode the data in training set 2. The decoded results were used to train the language models and the duration models.

After a recognizer decodes the input test utterance, the score generator for this recognizer will calculate the likelihoods for being each language in the task separately. One advantage of this kind of implementation is: by averaging the likelihoods calculated by all the recognizers for one language, the bias created by different recognizers is alleviated. In this paper, this idea is extended to duration modeling.

The forward language modeling scores, backward language modeling scores and duration modeling scores were calculated as (2),(3) and (5) respectively in our system. Before sending the scores (likelihoods) to the classifier, the acoustic scores are normalized by the number of frames in the test utterance, and the language modeling and duration scores are normalized by the number of phones in the best path decoded by their corresponding phone recognizers.

4.4. Final classifier

When multiple information sources are used, an important issue is how the sources should be combined to make a classification. One approach is to assume that all the information sources are independent probability estimators, so that log probabilities can be added as in [3]. In order to avoid these assumptions, a fully-connected, feed-forward neural network with one hidden layer is used to learn the relations among these scores in our system. The classifier was trained on both the training sets and the development set with conjugate gradient optimization[10].

5. Experiments

In order to test the LID performance of these information sources and their combinations, the nine experiments were carried out. A variety of neural network architectures(number of hidden nodes) were tested; the best results for each experiment are given in Table 2. where **A** denotes acoustic model being used, **F** denotes forward bigram based language model being used, **B** denotes backward bigram based language model being used and **C** denotes the context-dependent duration model being used. To study the effect of the final classi-

Approach	whole utterance	ten second utterance
A	66.96%	61.89%
F	88.39%	77.30%
B	87.50%	75.41%
C	55.36%	41.62%
AF	90.18%	79.19%
AB	90.18%	77.84%
FB	89.29%	78.38%
AFB	91.07%	80.27%
AFBC	91.96%	81.08%

Table 2: Results for all the experiments

Approach	whole utterance	ten second utterance
AF	88.39%	77.03%
AFBC	90.18%	79.19%

Table 3: Results for two experiments with linear classifiers

fier, two of above experiments were repeated with linear classifiers. An optimal linear classifier can be viewed as combining the probabilities with optimal weights. The best results are given in Table 3.

6. Concluding remarks and future work

Within our language-dependent phone recognition based LID system architecture, the best results were achieved when all acoustic, language modeling and duration scores were used. The results of these experiments show that all three information sources are useful for language identification; also a system architecture that can combine these information sources is helpful (the neural-network based systems have 10-20% lower error rate than the comparable systems with linear classifiers). The best system compares favorably with previously-reported results on the six-language task.

7. References

[1] 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

[2] 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.

[3] T.J. Hazen and V.W. Zue. Automatic language identification using a segment based approach. In *Proceedings 3rd European Conference on Speech Communication and Technology (Eurospeech '93).* Vol.2, pages 1303-1306, September 1993.

[4] K.P.Li. Automatic Language Identification Usıng Syllabic Spectral Features. In *1994 International Conference on Acoustic, Speech, and Signal Processing Proceedings.* Vol.1, pages 297-300, April 1994

[5] L.F.Lamel and J.L.Gauvain. Language Identification Using Phone-based Acoustic Likelihoods In *1994 International Conference on Acoustic, Speech, and Signal Processing Proceedings.* Vol.1, pages 293-296, April 1994

[6] A.S.House and E.P. Neuburg, Toward automatic identification of the language of an utterance. I. Preliminary methodological consideration. *Journal of the Acoustical Society of America*, Vol.62, No.3, pages 708-713, September 1977

[7] P.Ladefoged, A Course In Phonetics. *Third Edition*, pages 266-272. Harcourt Brace Jovanovich, Inc. 1993

[8] F.Jelinek. Self-organized language modeling for speech recognition. In A.Waibel and K.F.Lee, editors, *Readings in speech recognition*, pages 450-506. Morgan Kaufmann, Palo Alto, CA, 1990

[9] 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.

[10] 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