UNKNOWN LANGUAGE REJECTION IN LANGUAGE IDENTIFICATION SYSTEM

Hingkeung Kwan and Keikichi Hirose
Dept. of Information and Communication Engineering, University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo, 113 Japan

ABSTRACT

The number of languages in the world is much larger than the number of target languages that current language identification systems can handle. Therefore, we propose here the use of a multilayer perceptron neural network as a means to prevent these unknown language inputs from being misidentified as one of the target languages. We consider not only the target language identification rate but also the unknown language rejection rate. Results reveal that with the use of phonemic unigram as the input features to the neural network, a target language identification rate of 93.5% can be achieved for 3 languages. By varying the thresholds of the outputs, good unknown language rejection rate can also be obtained at the expense of lower identification rate.

1. INTRODUCTION

Although the number of languages currently spoken in the world is estimated to be at least about 2000, most current language identification systems can only handle up to 11 languages [1, 3]. These systems make use of a hard decision from the maximum likelihood score among the target languages. However, such a method cannot take into account of the possibility that the input utterance belongs to a larger set of languages outside the target set. This is because the decision is made on a relative difference among the scores of the languages but no relative measure for the resemblance of the languages is taken.

In order to reject an unknown language utterance, a criterion has then to be set up to measure the resemblance of the input utterance to the target language, i.e., a probability threshold. To fulfill such a task, a simple multilayer perceptron neural network is considered as an effective means to limit the outputs to a range between 0 and 1. It is assumed that when a proper threshold is determined, the unknown language utterances with score usually lower than the threshold can then be rejected from the target language set.

From the previous works [1]-[4], N-gram modelings of the phonemes are found to be features rich in discriminating information for language identification. Even though bigram of phonemes is supposed to yield more discriminating information than the unigram, unigram of phonemes is used here. This is because of two considerations: 1) the lower dimensions of input features to the neural network and 2) the size of training samples not enough for good estimate of bigram.

Therefore, in this paper, the performance evaluation of the proposed method, using phonemic unigram features as inputs to a neural network for target language identification as well as unknown language rejection, are presented.

2. CORPUS

The database used for evaluation is the OGI multi-language corpus of telephone speech (Table 1). 3 languages: English, German and Spanish were selected as the target languages while 2 languages: Japanese and Mandarin were chosen as the unknown languages to be rejected. The average duration of each whole utterance is about 45s.

<table>
<thead>
<tr>
<th>Usage</th>
<th>Target</th>
<th>Reject</th>
</tr>
</thead>
<tbody>
<tr>
<td>Language</td>
<td>EN</td>
<td>GE</td>
</tr>
<tr>
<td>#Whole-story Training Data</td>
<td>70</td>
<td>70</td>
</tr>
<tr>
<td>#Whole-story Testing Data</td>
<td>70</td>
<td>80</td>
</tr>
<tr>
<td>#NIST 10s Testing Data</td>
<td>69</td>
<td>65</td>
</tr>
</tbody>
</table>

Table 1: OGI Corpus

3. ALGORITHM

3.1. Mixed Phoneme Recognition (MPR)

Stochastic models for all fundamental phonemes of each language are first created from training data. During recognition, test utterance is decoded by a mixed phoneme recognizer in which all the language-dependent and language-independent phonemes of the target languages are covered (Figure 1). The bias due to different language recognizers can then be removed automatically. Since the log likelihood score of optimal phoneme sequence can no longer be utilized from one single phoneme recognizer, the N-gram approach is usually used after this mixed phoneme recognition for language identification. In the experiments here, there are 62 phonemes in the recognizer from the 3 target languages.
3.2. Unigram Feature Table

Stochastic grammar of a language, i.e., N-gram, is shown to be a powerful feature to identify a language in many works [1]-[5]. The two most common language models used are unigram and bigram. In our previous work [4], the identification decision was based on the maximum score of the unigram model as follows:

\[ L = \arg \max_{L_i} \prod_{i=1}^{n} P_r(p_k | L_i) \]  

where \( L \) is the target language, \( L_i \) is a language set member and \( p_k \) is the language specific phoneme.

Instead of using this whole utterance sequence probability approximated by the N-gram modeling, here, a table of the unigram relative frequencies is formed from each recognized utterance sequence, in which the unigram relative frequency is defined as

\[ Pr(phn_i) = \frac{C(phn_i)}{\sum C(phn_j)} \]

where \( C(phn_i) \) is the occurrence frequency of the phoneme \( phn_i \). The model assumes the occurrence of each phoneme is independent from the others as in unigram. Since 62 phoneme models are in the recognizer, there are 62 frequency scores in the feature table.

The use of bigram model may give more language specific combinations of phonemes across languages since many mono-phonemes are not language dependent. However, a lot more training data will be required before the neural network can be properly trained with the much larger dimensions of inputs(62x62 in our case).

3.3. Multilayer Perceptron (MLP)

A multilayer perceptron with error backpropagation training algorithm, is adopted here as the output normalizer. This is because of both its discriminant capability and adaptability. It is also convenient when we do not know how to extract features for complicated classification problem. In the experiments here, the unigram feature tables obtained from each data are fed into the MLP as inputs and then 3 output nodes are set for the 3 target languages while the other parameters of the MLP, such as number of hidden layer, are tunable for optimization (Figure 2). Once the outputs of the MLP are obtained, final decision can be made by setting an optimal threshold on the classifier.

4. SPEECH PREPROCESSING

The speech utterances sampled at 8kHz, were first preemphasized by a filter \( H(z) = 1 - 0.97z^{-1} \). Hamming window of length 25.6ms was then applied at a rate of 10ms. From each frame, feature vector consisting of 12 mel-scale cepstra, 12 delta mel-scale cepstra and 1 delta energy value, was then computed. All the preprocessings were implemented by the commercial software HTK V1.5.

5. EXPERIMENTAL RESULTS

5.1. Comparison to Previous Approach without Language Rejection

![Figure 3: Comparison between Old Method and New Method](image-url)
Comparing the current method using an MLP with 62 input features, no hidden layer and 3 output nodes, with the previous method using only the maximum score for identification decision [4], it is found that the current method can outperform the previous method when whole story data were used to obtain the unigram feature tables (Figure 3). Obviously, the input length affects the effectiveness of the unigram features of the language more seriously than the previous method, which will be discussed in a later section.

5.2. Effect of Varying Thresholds for Language Rejection

![Figure 4: Performances of Varying Thresholds for Whole Story Testing Data](image)

![Figure 5: Performances of Varying Thresholds for NIST 10s Testing Data](image)

In the previous section, there was no consideration of the unknown language inputs, i.e. only the maximum scores from the MLP outputs were used to decide the target language. When a threshold is set to various values, it becomes feasible to reject the unknown language inputs at the expense of a decrease in target language identification rate (TLIR).

Therefore, after defining the unknown language rejection rate (ULRR) as the rate of rejecting unknown languages, an optimal threshold can be found from the trade-off between TLIR and ULRR. From the results (Figure 4), it is observed that for the whole story data, the optimal threshold is found to be 0.9, at which both the TLIR and ULRR are about 70%. On the other hand, for the NIST 10s testing data, a lower optimal threshold, 0.6, was found at which both rejection rate and ID rate are fairly good (Figure 5).

5.3. Effect of Varying MLP Configurations

Hidden layer has been added to the MLP to see whether the MLP would yield more complex decision surfaces for better classification.

![Table 2: Performances of Different Configurations](image)

However, as shown in Table 2, it is found that the addition of a hidden layer did not help to improve the performance. One explanation may be that the number of training data remained the same while the number of weights that has to be trained increased much more. As a result, the more sophisticated MLP was not trained enough for better classification.

5.4. Effect of Varying Time Lengths

![Figure 6: Performances of Varying Time Lengths](image)

As expected, when the input length increases, both the TLIR and the ULRR increase, however, the overall correct rates
do not show such a trend (Figure 6). The overall correct rates do not increase as a saturation point (at about 20s) is reached, on the contrary, it starts to decrease slightly. Furthermore, the optimal threshold is found to vary at different time lengths.

5.5. Weight Distributions of Trained MLP

From the internal weights of the trained MLP without hidden layer, it is hoped that some insights might be obtained to see which features would yield more information for the classification and which would not contribute at all. However, it is found that all features seem to contribute at quite different weights and so no elimination of useless feature input could be done to further simplify the structure of MLP (Figure 7).

Table 3: Performances using Bigram Feature Table of Phonetic Classes

<table>
<thead>
<tr>
<th>Data</th>
<th>TD Rate (Overall Rate at Threshold)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Whole Story</td>
<td>No Hidden Layer: 81.3%</td>
</tr>
<tr>
<td>Testing Data</td>
<td>(62.4% at 0.9)</td>
</tr>
<tr>
<td>NIST 10s</td>
<td>No Hidden Layer: 61.5%</td>
</tr>
<tr>
<td>Testing Data</td>
<td>(52.3% at 0.6)</td>
</tr>
</tbody>
</table>

Figure 7: Weight Distributions of the Trained MLP without Hidden Layer

5.6. Phonetic Class Bigram Features

From our previous results [5], it was found that the bigram of phonetic classes could also achieve comparable performance to the one of phonemes while the complexity could be reduced. Therefore, the bigram of phonetic classes were also evaluated as the input features for the MLP. There are 14 phonetic classes defined and so 14 x 14 = 196 feature inputs to the MLP. Due to the large dimension of features and small number of training data, the estimated bigram was not so reliable. As a result, the use of phonetic class bigram features is worse than the unigram of phonemes (Table 3).

6. CONCLUSION

Results indicate the proposed method is quite effective for unknown language rejection in an automatic language identification system at the expense of a lower identification rate. Currently, the limitation of training data causes it difficult to train more sophisticated MLP for better classification and the time constraint of the approach is not solved. In the future, recurrent neural network will be exploited to tackle these issues. Hopefully, the feedback characteristics of the recurrent neural network can capture the bigram features of the languages for performance enhancement. In addition, this method may be applied together with the previous approaches to improve the performance further.

7. REFERENCES