USE OF RECURRENT NETWORK FOR 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
E-mail: kan@gavo.t.u-tokyo.ac.jp, hirose@gavo.t.u-tokyo.ac.jp

ABSTRACT

In the past, we attempted to use a multilayer perceptron neural network as a means to prevent those unknown language inputs from being misidentified as one of the target languages in language identification system. However, the use of multilayer perceptron neural network could not utilize the temporal information from the utterances. Results show that with the use of phonemic unigram as input features to a recurrent neural network of Jordan architecture, a 3 target language identification rate of 98.1% can be achieved. By setting the output thresholds to 0.6 to reject 2 more unknown languages, a lower overall rate of 85.9% is obtained.

1. INTRODUCTION

Although the number of languages currently spoken in the world is estimated to be at least about 3,000, most current language identification systems can only handle up to 11 languages[1,2]. 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 language is taken.

In our previous work [3], we attempted to use a multilayer perceptron neural network as a means to prevent those unknown language inputs form being misidentified as one of the target languages in language identification system. A simple multilayer perceptron neural network can have the outputs set to a range between 0 and 1. Assuming 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. However, the use of multilayer perceptron neural network could not utilize the temporal information from the utterance sequences. Therefore, we studied the feasibility of using recurrent neural network to utilize the temporal information more effectively.

From the previous works [1-5], 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 the paper, the performance evaluation of the proposed method, using phonemic unigram features as inputs to a recurrent neural network for target language identification as well as unknown language rejection, are presented.

2. CORPUS

The database used for evaluation was the OGI multi-language corpus of telephone speech. 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. There were two sets of testing data: 1) 45s whole story and 2) 10s NIST.

Usage	Target			Reject	
Language	EN	GE	SP	JA	MA
#Training Data					
45s Whole-story	70	70	70		
#Testing Data					
45s Whole-story	37	31	39	85	90
10s NIST	69	65	58	61	52

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 the training data. During recognition, a testing 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 [5]. 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 experiment here, there are 62 phonemes in the recognizer from the 3 target languages.

[*] Currently with TI Tsukuba R&D Center

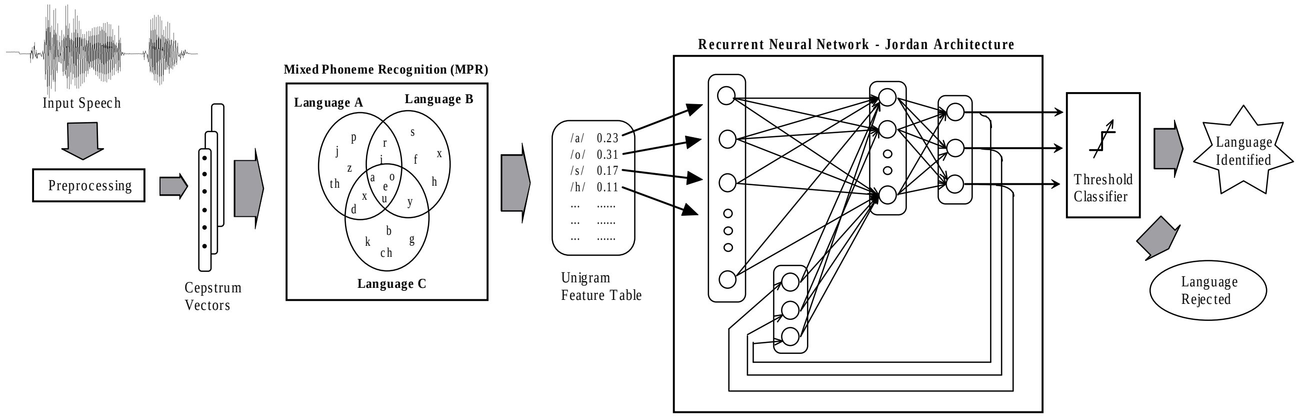

Figure 1. Configuration of Language Identification System

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 [5], the identification decision was based on the maximum score of the unigram model as follows:

$$L = \arg\max_{L_i} \prod_{k=1}^{n} \Pr(p_k \mid L_i) \qquad (1)$$

where L is the target language, L_i is a language set member, p_k is a phoneme and n is the number of phonemes in the language set.

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

$$\Pr(\mathrm{phn}_i) = C(\mathrm{phn}_i) / \sum_{1}^{62} C(\mathrm{phn}_i) \qquad (2)$$

where $C(\mathrm{phn}_i)$ 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 Recurrent Neural Network

A recurrent neural network of Jordan architecture is adopted here as the output normalizer. Comparing to the multilayer perceptron we used before[3], the recurrent property allows the capture of temporal information for classification. In the experiments here, the unigram feature tables obtained from each data are fed into the recurrent neural network as inputs. The context layer is used to provide recurrent links. Three output nodes are set for the 3 target languages. Once the outputs from the recurrent neural network are obtained, final decision can be made by setting an optimal threshold on the classifier. The recurrent network here was simulated by using the free software SNNS ver. 4.0 from the University of Stuttgart.

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 pre-processings were implemented by the commercial software HTK V1.5.

5. EXPERIMENTAL RESULTS

5.1 Comparison to Previous Results using Multilayer Perceptron

When there is no consideration of the unknown language inputs, i.e. only the maximum scores from the network outputs were used to decide the target language, it is found that the recurrent neural network can outperform the multilayer perceptron that we used before (**Figure 2**). The ID rates for whole-story and NIST 10s data are 96.3% and 87.5% respectively.

When the output threshold is set to 0.9 in order to incorporate the rejection capability, the overall performance will be lowered by 20% (**Figure 3**).

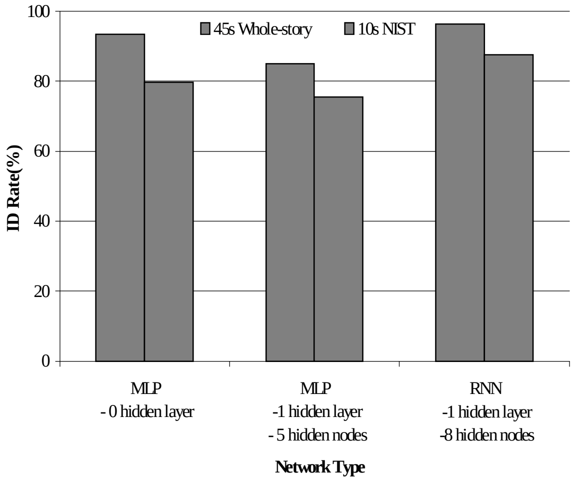

Figure 2. Comparison to Previous MLP Results

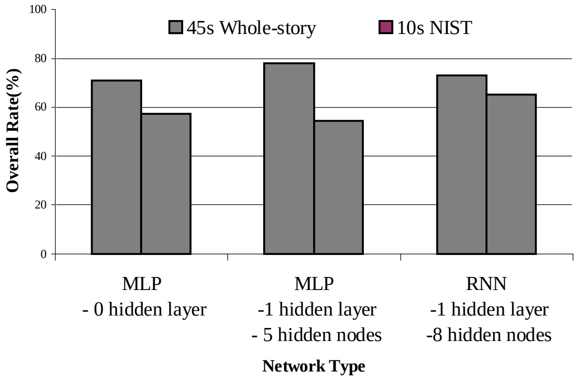

Figure 3. Comparison to Previous Overall Rates

5.2 Effect of Varying Thresholds for Language Rejection

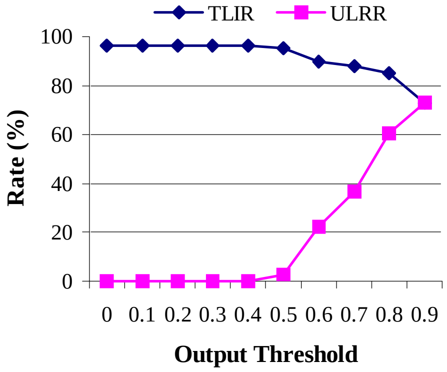

Figure 4. Performances of Varying Thresholds for Whole-story Testing Data

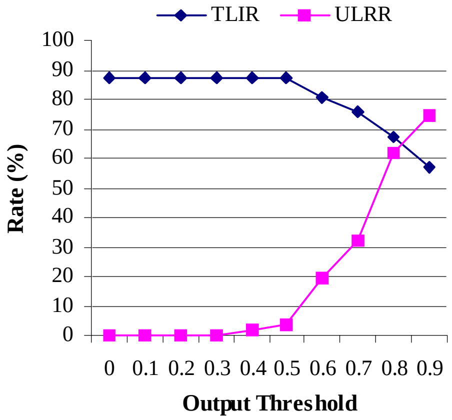

Figure 5. Performances of Varying Thresholds for NIST Testing Data

When the output 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 for rejecting unknown language, an optimal threshold can be found from the trade-off between TLIR and ULRR. From the results (**Figure 4 & 5**), for both whole-story data and NIST data, the threshold has to be set to more than 0.8 or it will fail to reject the unknown language at low threshold value.

5.3 Effect of Varying Time Lengths

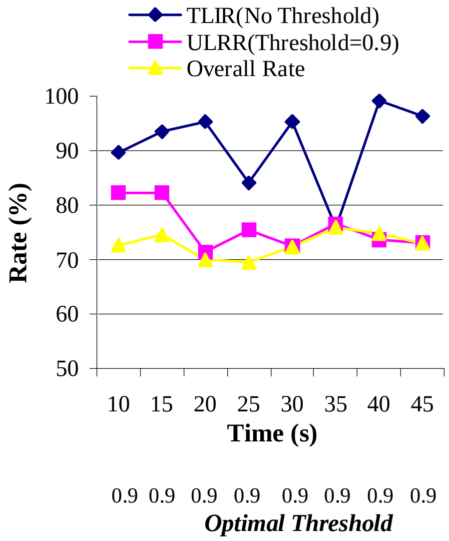

Figure 6. Performance of Varying Time Lengths

From the performance on whole story testing data against time (**Figure 6**), it is found that even though the identification rate quite fluctuates, both the optimal threshold, always 0.9, and the overall correct rates, 70%~75%, are quite consistent with the varying time lengths.

5.4 Modified Decision Scheme

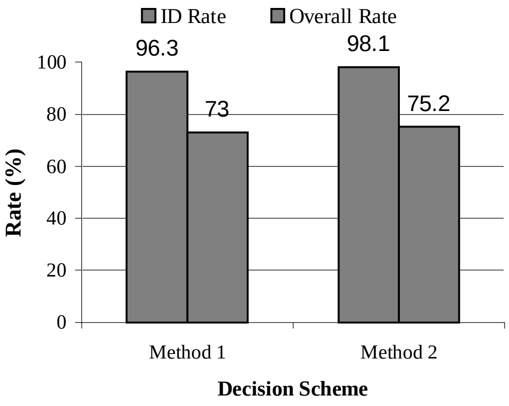

Figure 7. Comparison between Decision Schemes

Instead of classifying the input utterance into a target simply from the final output scores of the recurrent network (Method 1), the candidates with the highest scores from every 5s of the whole-story data are stored up and the target is then selected as the candidate with the most frequent occurrences from all segments in order to reduce any bias due to a small portion (Method 2). Results (**Figure 7**) show that a 2% increase in both the ID rate and the overall correct rate of whole-story testing data can be obtained from such a decision scheme.

5.5 Trained RNN with Unknown Language Data

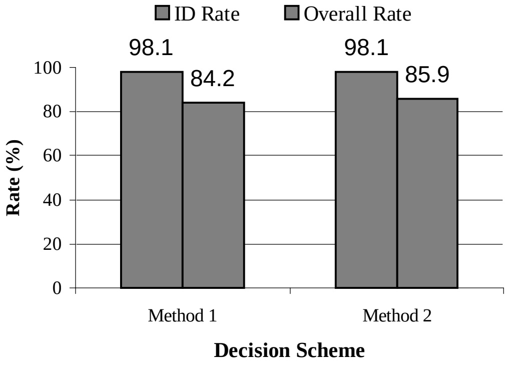

Figure 8. Unknown Language Data Trained Network's Performance

Since no unknown language information were used to train the recurrent neural network, it was naturally more difficult to reject the unknown languages. To deal with this issue, we trained another network with the incorporation of some unknown language data (50 from JA and 55 from MA) to the training set in order to provide information on what to reject. However, due to the assumption that there is no information on what kind of phonemic features the unknown language would have, the number of features remains unchanged. Compared to **Figure 7**, although the identification rate cannot be further enhanced, the overall correct rate is significantly improved by 10% at a lower threshold, 0.6 (**Figure 8**). In addition, this overall rate is quite consistent with the varying time length (**Figure 9**).

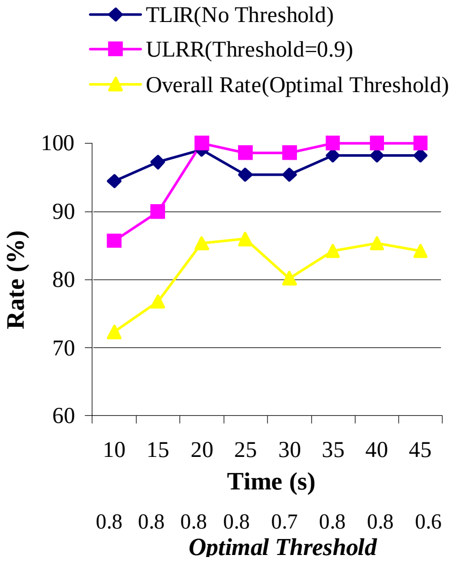

Figure 9. Unknown Data Trained Network Performance of Varying Time Lengths

6. CONCLUSION

The use of recurrent neural network with the proposed decision scheme is found to be competent in yielding a high identification rate of 98.1% for 3 target languages after incorporating some unknown language data into training set. At the expense of a lower overall correct rate of 85.9% obtained from setting the optimal threshold at 0.6, 2 more unknown languages can be rejected effectively.

REFERENCES

[1] Zissman, M.A., "Overview of Current Techniques for Automatic Language Identification of Speech," *Proc. IEEE ASR Workshop*, pp.60-62, 1995.

[2] Yan, Y., and Barnard, E., "Recent Improvements to a Phonotactic Approach to Language Identification," *Proc. Speech Research Symposium XV*, pp.212-219, 1995.

[3] Kwan, H., and Hirose, K., "Unknown Language Rejection in Language Identification System," *Proc. ICSLP96*, pp.1776-1779, Vol.3, 1996.

[4] Kadambe, S., and Hieronymus, J., "Spoken Language Identification with Phonological and Lexical Models," *Proc. Speech Research Symposium XV*, pp.204-211, 1995.

[5] Kwan, H., and Hirose, K., "Recognized Phoneme-based N-gram Modeling in Automatic Language Identification," *Proc. EUROSPEECH '95*, pp.1367-1370, 1995.

[6] *Stuttgart Neural Network Simulator User Manual, Ver. 4.1*, 1995.