ON IMPROVING DISCRIMINATION CAPABILITY OF AN RNN BASED RECOGNIZER

Tan Lee and P.C. Ching
Department of Electronic Engineering
The Chinese University of Hong Kong, N.T., Hong Kong
Tel: (852) 2609 8266 Fax : (852) 2603 5558
e-mail: {tlee1,pcching}@ee.cuhk.edu.hk

ABSTRACT

This paper presents a set of effective and efficient techniques to improve the discrimination capability of a recurrent neural network (RNN) based isolated word recognizer. The recognizer contains a set of individually trained RNN speech models (RSMs). Each of them represents a different word in the vocabulary. Speech recognition is performed by selecting the RSM that best matches the input utterance. For temporal supervised training of the RSMs, a new error function is introduced, in which the contributions of all phonetic components are equalized regardless of their difference in duration. The learning rate for recurrent connections is amplified. This is aimed at strengthening temporal dependency in the RSMs to capture dynamic characteristics of speech signals. Furthermore, a hierarchical training strategy is employed to facilitate more efficient discriminative training among the RSMs. A series of speaker-dependent recognition experiments are performed to evaluate the effectiveness of the proposed techniques.

1 INTRODUCTION

Recurrent neural networks (RNN) refer to a broad class of neural networks that contain delay and feedback connections [1]. With these connections, the present network state is determined not only by the instantaneous external input but also by the past states of the network itself. RNNs have found many successful applications in solving temporal problems, such as time series prediction, grammatical inference, dynamic system identification and control [2]. It is also generally agreed that the temporal properties of RNNs are suitable for the recognition of dynamic speech patterns. However, applications in this area are still very limited [3,4]. New network architectures and training algorithms are being explored widely provide more appropriate representation and learning behavior for complex speech features.

In [5,6], we have proposed a novel method of using RNNs for isolated word recognition. This method adopts the one-class-one-network (OCON) architecture. Each word is represented separately by a fully connected RNN, which is referred as the RNN speech model (RSM). To recognize an unknown utterance, the best matching RSM is obtained through a multi-pass selection-by-elimination process. Discrimination among the words relied mainly on their distinctive static and dynamic features, which are captured by individual training of the respective RSMs. In addition, a discriminative training algorithm has been developed to achieve better recognition performance.

Although simulation experiments have consistently validated the effectiveness of the proposed approach in a number of recognition tasks [6], the overall system discrimination capability still needs further enhancement, especially for highly confused vocabulary. The work presented in this paper is aimed at developing a set of effective and efficient techniques to improve discrimination capability of the RNN based recognizer. Firstly, the temporal error function employed in RSM training is modified to equalize the effect of phonetic components with different duration. Secondly, in RSM training, the learning rate for recurrent connections is augmented pragmatically. As a result, better temporal dependency between the phonetic components is established as a useful discriminating feature. Lastly, a knowledge based grouping strategy is used to facilitate more efficient mutual discriminative training among the RSMs.

2 THE RNN BASED RECOGNIZER

In this section, we shall briefly describe the fundamental design of the RNN based recognizer. As shown in Figure 1, the recognizer consists of a total of C RSMs, each being used to model a specific word. The basic phonetic components of this word are represented by different output neurons in the RSM. When an input utterance is presented to the RSM, the output neurons are activated one after another, following the sequential order of the corresponding phonetic components. Indeed, the RSM performs time alignment automatically for each input

utterance and a phonetic segmentation can be derived directly from the RSM output sequence.

The recognition process is divided into three passes. Pass 1, called the *state sequence screening pass*, removes the RSMs whose output neurons are not activated in the desired sequential order. Pass 2, called the *duration screening pass*, examines the duration of individual phonetic components using a set of pre-determined constraints. If an RSM generates a segmentation violating the constraints, it will be removed in Pass 2. Among the remaining RSMs, Pass 3 selects the nearest one based on a temporal error function defined as follows,

$$E = \frac{1}{T}\sum_{t=1}^{T} e(t) \tag{1}$$

where T denotes the utterance length (in frames) and $e(t)$ denotes the average output error at time t.

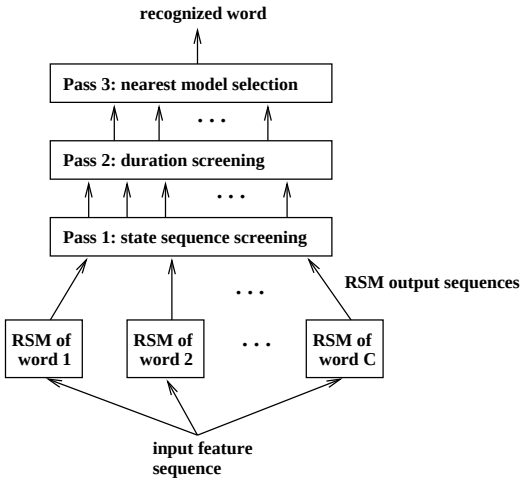

Figure 1: The RNN based isolated word recognizer

The system training is carried out in two stages, which are referred as *individual training* and *mutual discriminative training* respectively. During individual training of an RSM, the error function E is minimized for each training utterance of the designated word. It is supposed to learn the static features of different phonetic segments and, at the same time, to establish temporal dependency between these segments. Whilst discriminative training is applied among all of the C RSMs to minimize the probability of misclassification.

3 A MODIFIED ERROR FUNCTION

In equation (1), E is defined as a simple average of all output errors over time. This equation can be re-written as

$$E = \frac{1}{T}\sum_{k=1}^{K}\sum_{t=\tau_k}^{\tau_{k+1}-1} e(t) \tag{2}$$

where K is the total number of phonetic segments in the word, τ_k denotes the beginning of the k^{th} segment and $\sum_{t=\tau_k}^{\tau_{k+1}-1} e(t)$ is equal to the accumulated output error of the k^{th} segment.

It is obvious that a segment with longer duration tends to have greater contribution in E and consequently dominates the whole word. This contradicts the commonly accepted fact that short phonemes like consonants carry the most useful information for speech recognition [7]. For example, the English alphabets "B" and "D" share the same long vowel /i/ which may occupy over 90% of the utterance duration. To distinguish them relies mostly, if not completely, on the difference of initial consonants /b/ and /d/, which are very short in duration. In such applications, the existing error function is unable to provide sufficient discrimination power to achieve high recognition accuracy.

To alleviate this problem, a modified error function, denoted by E_w, is defined as follows,

$$E_w = \frac{1}{T}\sum_{k=1}^{K} w_k \sum_{t=\tau_k}^{\tau_{k+1}-1} e(t) \tag{3}$$

When $w_k = 1$ for all k, we have $E_w = E$. Therefore E_w can be considered as a generalized form of E. It allows different weighting factors to be applied on the phonetic segments. To reduce the effect of duration difference between the segments, let

$$w_k = \frac{T/K}{\tau_{k+1} - \tau_k} \tag{4}$$

where T/K is equal to the average segment duration and $\tau_{k+1} - \tau_k$ is the duration of the k^{th} segment. Furthermore, we define

$$E(k) = \frac{1}{\tau_{k+1} - \tau_k}\sum_{t=\tau_k}^{\tau_{k+1}-1} e(t) \tag{5}$$

as the average output error specifically for the k^{th} segment. Then E_w can be expressed as

$$E_w = \frac{1}{K}\sum_{k=1}^{K} E(k) \tag{6}$$

In simple words, E_w treats all phonetic segments with equal weighting while the original error function E considers all time frames as the same. From speech recognition point of view, E_w provides a more appropriate basis for comparing and discriminating speech patterns.

4 IMPROVING TEMPORAL STATE DEPENDENCY IN RSM

In the RNN speech model, the total input to each neuron at any time instant consists of two components, namely the external input component and the recurrent component. The external input component is determined by the current input features while the recurrent component provides contextual information of the previous network states. Generally it is required that the two components have about the same magnitude. If the external input component dominates over the recurrent one significantly and persistently, the output response of this neuron would be very sensitive to the instantaneous input. As a result, the network becomes more like a static pattern classifier which takes the successive input vectors as uncorrelated individuals. In this case, the RSM might not be able to capture the sequential order of phonetic segments and the state sequence screening pass in Figure 1 would be less efficient.

To attain the desired temporal dependency, it is necessary to balance the magnitudes of external input component and the recurrent component during the RSM training. Using Williams and Zipser's temporal supervised training algorithm [8], the overall weight change is obtained by accumulating the local adjustment made at each time step. If the training utterance contains T frames which form K stationary phonetic segments, the average training time spent on capturing the static features of each segment will be equal to T/K. Whilst the transition between each pair of adjacent segments is encountered exactly once during the presentation of this utterance. In terms of training time being occupied, the stationary part is T/K times more than the transition part. Under normal speaking rate, the value of T/K is about $10-15$. To make up such a huge difference, the learning rate for recurrent connections should be augmented at each segment transition. Let η be the normal learning rate and η_a be the augmented recurrent learning rate. A simple rule of thumb is,

$$\frac{\eta_a}{\eta} = \frac{T}{K} \tag{7}$$

Hence, the effective training time for each segment transition is increased and better temporal dependency among the segments could be established.

5 HIERARCHICAL DISCRIMINATIVE TRAINING PROCEDURE

In [6], a discriminative training algorithm for the RNN recognizer has been developed based on the method of generalized probabilistic descent (GPD). Although a significant improvement on recognition accuracy can be achieved with this algorithm, the training time required is found to be excessively long. To facilitate more efficient discriminative training, a hierarchical training procedure is described below.

In general, any vocabulary for speech recognition can be divided into a number of confusing groups based on our prior knowledge of phonetic similarity. Each of these confusing groups may be further divided into smaller sub-groups. As a simple example, one possible grouping for the nine E-set alphabets is shown in Figure 2. Discriminative training should start within the groups at the bottom level and deals with the most severe confusion first, e.g. {"B", "D"}, {"C", "Z"}. Then the coverages of training is expanded progressively to include more and more words. Eventually, the entire vocabulary are involved in the global discriminative training.

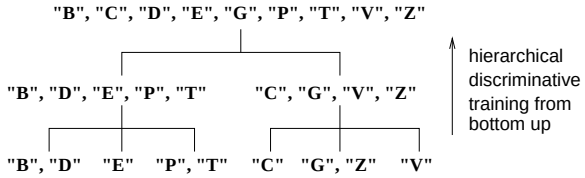

Figure 2: An example of knowledge-based grouping for hierarchical discriminative training

6 EXPERIMENTAL RESULTS

To demonstrate the effectiveness of the proposed performance improvement techniques, recognition experiments are carried out on three small vocabulary tasks: (1) 58 Cantonese CV syllables; (2) English E-set alphabets; and (3) 20 English words. In these experiments, speech features are extracted every 10 ms with a 20 ms Hamming window. Each short-time feature vector consists of 16 components:

$$[energy, \Delta energy, cep_1, \cdots, cep_8, \Delta cep_3, \cdots, \Delta cep_8]$$

where cep_i denotes the i^{th} LPC derived cepstral coefficient.

6.1 58 Cantonese CV Syllables

Cantonese is a commonly used Chinese dialect. It is a typical monosyllabic and tonal language. Cantonese syllables have the general structure of CVC, where C means consonant and V means vowel. In this experiment, we focus on the recognition of CV syllables only. This is regarded as a difficult problem since the vowel part in a CV syllable usually dominates in terms of both

energy and duration. Syllables sharing the same or similar vowel are easily confused. Disregarding the difference in tones, there are totally 58 distinct CV syllables in Cantonese. Therefore the recognizer is composed of 58 RNN syllable models, each with 10 neurons.

Twenty sets of speech data, containing totally 3,420 monosyllabic utterances, are obtained from a male speaker. Half of them are used as training data and the remaining half are used as independent test data. Previous experiments reported an overall recognition accuracy of 86.7% with the original system design [6]. By using the modified error function and applying the augmented recurrent learning rate, a recognition accuracy of 90.4% is attained. Major improvement is observed on those syllables with short initial consonants. In particular, the total number of misclassifications between: 1) /ba/ and /da/; 2) /bo/ and /do/; and 3) /be/ and /de/, are much reduced from 60 to 22. The use of the hierarchical discriminative training procedure has also made a noticeable improvement on training efficiency. The total number of training cycles [1] required is reduced from 30,000 to 8,000.

6.2 English E-set Alphabets

This experiment investigates the recognition problem of the nine English E-set alphabets, {"B", "C", "D", "E", "G", "P", "T", "V", "Z"}. In terms of phonetic composition, these spoken alphabets are very similar to the Cantonese CV syllables described above. They all share the same vowel /i/ and are considered to be very confusing.

Speech data are extracted from the TI46 Isolated Word Corpus and include the voice of two male speakers. 10 data sets are used to train the recognizer and another 16 sets are used for performance evaluation. Without using the improvement techniques, the original RNN recognizer gives an accuracy of only 78%. With the proposed techniques, the recognition performance is improved to 85.6%, which is quite satisfactory for such a confusing vocabulary.

6.3 20 English Words

The vocabulary consists of English digits "0" – "9" and 10 command words ("enter", "erase", "go", "help", "no", "rubout", "repeat", "start", "stop", "yes"). Speech data are obtained from the TI46 Isolated Word Corpus and contain the voice of seven male speakers. 10 data sets are used to train the recognizer and another 16 sets are used as test data. Table 1 compares the recognition performance of the original and the improved RNN recognizers in both single-speaker and multi-speaker cases.

% Accuracy	Recognizer	
	Original	Improved
single-speaker	93.2%	97.8%
multi-speaker	85.3%	91.9%

Table 1: Recognition results for the 20 English words

[1] One training cycle means a complete presentation of all training utterances

7 CONCLUSION

In this paper, we have described a set of new techniques to improve the discrimination capability of an RNN based speech recognizer. These techniques are applied mainly to modify the existing training algorithm for individual RNN speech models, aiming at capturing useful discriminative features for recognition purpose. In addition, a hierarchical procedure is proposed to make the mutual discriminative training more efficient. Experimental results show that the recognition performance can be improved significantly.

ACKNOWLEDGEMENT

This work is partly supported by a research grant from the Croucher Foundation.

REFERENCES

[1] J.L. Elman, "Finding Structure in Time", *Cognitive Science*, Vol.14, pp.179 – 211, 1990.

[2] C.L. Giles, G.M. Kuhn and R.J. Williams, "Dynamic Recurrent Neural Networks: Theory and Applications", *IEEE Trans. Neural Network*, Vol.5, No.2, pp.153 – 156, 1994.

[3] A.J. Robinson, "An Application of Recurrent Nets to Phone Probability Estimation", *IEEE Trans. Neural Network*, Vol.5, No.2, pp.298 – 305, 1994.

[4] W.-Y. Chen, Y.-F. Liao and S.-H. Chen, "Speech Recognition with Hierarchical Recurrent Neural Networks", *Pattern Recognition*, Vol.28, No.6, pp.795 – 805, June 1995.

[5] Tan Lee, P.C. Ching and L.W. Chan, "Recurrent Neural Networks for Speech Modeling and Speech Recognition", *Proc. ICASSP-95*, Vol.5, pp.3319 – 22.

[6] Tan Lee, P.C. Ching and L.W. Chan, "An RNN Based Speech Recognition System with Discriminative Training", *Proc. EUROSPEECH-95*, Vol.3, pp.1667 – 70.

[7] L.R. Rabiner and B.H. Juang, *Fundamentals of Speech Recognition*, Prentice-Hall, Inc., 1993.

[8] R.J. Williams and D. Zipser, "A Learning Algorithm for Continually Running Fully Recurrent Neural Networks", *Neural Computation* 1, pp.270 – 280.