Incremental Learning Algorithm for Speech Recognition

Hisham Darjazini
h.darjazini@uws.edu.au

Qi Cheng
q.cheng@uws.edu.au

Ranjith Liyana-Pathirana
r.liyanapathirana@uws.edu.au

University Of Western Sydney
School of Engineering
Locked Bag 1797 Penrith South DC NSW 1797
Australia

***Abstract* - This paper presents an implementation of incremental learning neural networks algorithm for speech recognition. The algorithm has been investigated using the TIMIT speech samples and it has been shown to demonstrate high recognition accuracy.**

KEY WORDS
Incremental learning, speech recognition

I. INTRODUCTION

Incremental learning updates a recognizer using the information obtained from unknown input data (ID). It has the advantage of adapt to changing input without requiring time-consuming in training. Incremental learning algorithms were mostly designed and tested for pattern recognition applications (References [1], [8], [9] and [10]).

Speech signals change from speaker to speaker and from time to time even for the same speaker. Incremental learning is a suitable tool for speech recognition. Though incremental learning has been applied to speech enhancement in Ref. [3], very little research has been reported in the literature on its use in speech recognition.

In this paper, we propose an implementation to feed-forward incremental learning algorithm based on method developed by Darjazini and Tibbitts (Ref. [2]).

II. WEIGHT SET ADDITION ALGORITHM

The weight set addition algorithm is based on a method for speech recognition that employs a comb of phone sub-recognizers (Ref. [2]). As shown in Fig. 1, the method employs 55 sub-recognizers for recognition of 54 phones and silent period.

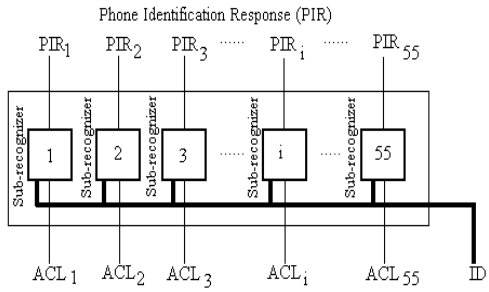

Fig. 1. Comb of Sub-recognizers.

All the sub-recognizers are implemented using an identical feed-forward neural network (FF-NN). Each sub-recognizer has an output referred to as Phone Identification Response (PIR), where the PIR is a continuous variable between 0 and 1. Each sub-recognizer indicates that the input speech contains a specific phone if the value of PIR is close to 1. The tolerance in the network is set to 0.05. Therefore, each PIR with a value of greater than or equal to 0.95 is taken as an indication of a potential match.

The algorithm extracts a new weight set (WS) from a new unknown data set during recognition. In this algorithm, the sub-recognizer contains two phases of back-propagation instead of one. At the initial trial the network behaves as a normal back-propagation network. At the subsequent trial the network performs the incremental learning process firstly by running the network using the available weight set, at this point a measure applies at the output, if the resulted error comes greater than the maximum allowed error, the process

T. Sobh (ed.), Innovations and Advanced Techniques in Computer and Information Sciences and Engineering, 231–235.

will be terminated as a nonrecognized phone. If the error is less

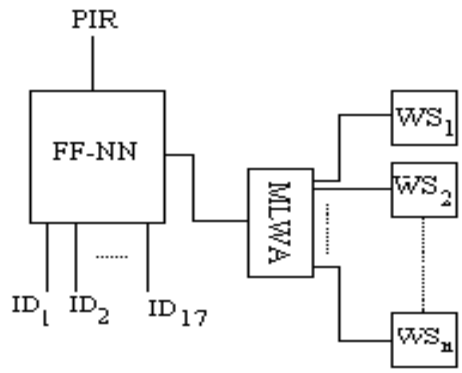

Fig. 2. Choice of Weight Set for Incremental Learning.

than the maximum allowed error and higher than a minimum acceptable output; then the incremental learning phase starts. The goal of the incremental learning phase here is to achieve an acceptable output at the output layer by adjusting the weight set using adaptive learning rate. When this is achieved the new weight set will be saved and sent to the MLWA (see Fig. 2) for later reference.

In the following recognition, the new set, as well as all the existing sets, will be tested as a potential weight set candidate in the FF-NN. The weight set that produces the highest PIR is selected as the most recent updated weight set. This function is performed by the Most Likelihood Weight Activator (MLWA) unit, which is shown in Fig. 2.

In Fig. 2, WS_1 is obtained from training session, obviously, in the early stages of incremental learning. Subsequent WS_S along with WS_1 will have statistical order in the MLWA and the highest probability WS is the one which is used mostly. Other sets will be used more often later on.

Fig. 3 shows the multi-layer structure of a sub-recognizer. The input layer contains 17 processing elements (PE) used to receive 17 input elements, which represents the Mel-scale Frequency Coefficients (MFCC) of the corresponding phone. In this structure, the input layer acts as a buffer to the subsequent hidden layers. There are three hidden layers H_1, H_2, and H_3, each one containing (34 - 51 - 34) PEs respectively. The output layer contains one PE representing a measure of the matching of the input speech (stimulus) to a particular phone.

III. Experimental results and discussion

The input data was extracted from 75 spoken sentences of the TIMIT speech database. The sentences are spoken by 25 speakers (5 female and 20 male). Every speaker posses one of three main dialects from the American English, and the accent was chosen arbitrarily. The data was mixed to produce as much variety as possible to every phone; this is to get the advantage of having the sub-recognizer being exposed and to deal with most varied forms of the same phone. In the primitive representation of the input data, there were 54 distinctive phones appeared in 2440 samples, which were segmented from 637 words. The table in the appendix shows these phones and their number of occurrence.

Experiments were performed firstly by initiating (first run) the sub-recognizers using the back-propagation learning algorithm and applying the Delta rule. The exit condition of this session was the number of iterations, which was set at 500, and the learning rates were all initiated to 0.5. The weights were initiated to random normally distributed values and the learning set contained nonclustered stimulus. Maximum accepted error (tolerance) is 0.01 and the incremental learning width is 0.219, i.e. the range is from 0.989 to 0.77

The initial session provides the first weight set (WS_1) for the MLWA and determines the first cluster of the input data. The number of phone samples for the initial session was in this case 15, and the sub-recognizer converged from the target after 50 epochs. In each epoch, the network manipulated the inner weights of the hidden layers. An error monitor was set to measure the value of mean squared error (MSE) value at each hidden layer and the effects of a particular PE on the overall result of the network. The accuracy of the PIR was within an error value of 0.01, which is below the tolerance value. The overall performance on the initial learning set scored 94.44% accuracy.

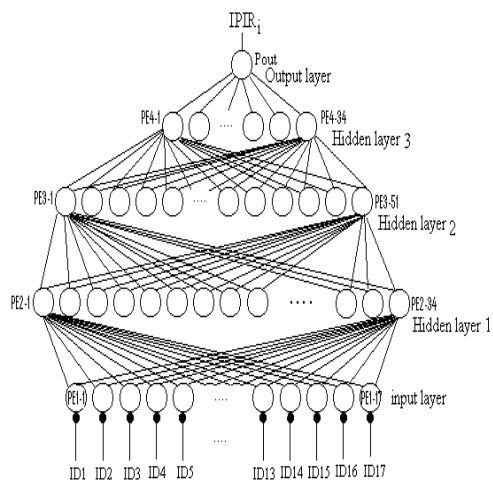

Fig. 3. Structure of Sub-recognizer.