A DISTRIBUTED NEURAL-NET TRAINING STRATEGY FOR RECOGNITIO N OF CHINESE NUMBERS USED IN THE VOICE DIALER OF MOBILE HANDSE TS†

Chua-Chin Wang‡, Hsin-Long Wu, and Shau-Guo Huang§

Department of Electrical Engineering
National Sun Yat-Sen University
Kaohsiung, Taiwan 80424

ABSTRACT

An efficient capacity estimator for batteries is the core of the success of the battery management. In this paper, we propose a general capacity measurement module for the current computation in the process of battery charging and discharging. Such a module can be included in a variety of battery management systems (or ICs). The simulation and testing results of the final chip prove the correctness of the design.

1. INTRODUCTION

In many occasions, the users of wireless handset can not dial the phone number at will, e.g., driving a vehicle or typing a paper. Thus, an auxiliary tool to help the dialing task is in demand [3]. The most convenient tool is the human speech [8]. That is why the voice dialer (VD) emerges quickly in the research areas of industry and academy. Despite the background noise problem, the VD is still the most acceptable interactive usage to common users of handsets. The major purpose of the VD, thus, is defined as a speech recognizer for a finite set of certain languages, such as Chinese numbers [5].

The structure of many successful systems for speech recognition typically consists of a feature analysis-extraction procedure (front-end) followed by a statistically pattern classifier (back-end) [8]. Prior studies showed that the signal processing and classification techniques interact with each other to affect phonetic classification. The recent advent of discriminative feature extraction showed that improved recognition results can be obtained by using an integrated optimization of both the preprocessing and the classification stages [2]. These approaches, including filter bank, lifter and dynamic features design, are pretty complicated to be incorporated in a VD for handsets. One of the most powerful speech recognition methods is the model-independent speech features, called mel-frequency ceptral coefficients (MFCC's), use DCT as a linear operator to map mel-warped DFT into a lower dimensional feature space [4]. Despite the empirical superiority of MFCC's over many other types of signal processing techniques, there are no theoretical reasons why the linear transformation associated with DCT, which is fixed a priori and independent of HMM (hidden Markov model) states and speech classes on channel energies.

Instead of removing background noise, the proposed idea is mainly focused on the voice recognition of Chinese numbers 0 to 9 [5] and a finite set of Chinese phrases [1], e.g., dial, shutdown, etc. basing on a neural network strategy. The reason is that Chinese numbers, words, and phrases are easy to discretized by simple algorithms, e.g., zero-crossings. The task to develop an efficient voice dialer can be partitioned into two parts : first, an effective preprocessor to extract the features of given voice, and second, a precise classifier able to tell what the speech indicates. Neural-network-based recognizer [3], hidden Markov model-based classifier (statistical method) [2], and rule-based classifier [1] for the second part have been evaluated to find out the maximum recognition rate. When it comes to the voice recognition in mobile handsets, the demand of low power, small code size, and noise immunity emerge. The neural network approach possesses these features after it is appropriately trained. This paper, thus, presents a novel strategy to train the neural network such that it is quickly converged and able to process the Chinese number recognition.

2. DISTRIBUTED TRAINING STRATEGY

2.1. Speech data representation

Since the goal of this work is aimed at the voice dialer for portable handsets, the small code size to process the speech recognition is one of the basic requirements of systems with limited hardware resources. A proper speech data representation will drastically reduce the size of the codes. We, thus, employ the LPC (linear prediction code) to extract and represent the digitized speech data owing to the facts that LPC possesses several impressive characteristics, e.g., easy to generate

†This research was partially supported by Industrial Technology Research Institute under grant ITRI G4-88007-e, and National Science Council under grant NSC 87-2215-E-110-010 and NSC 88-2219-E-110-001.

‡the contact author

§Mr. Huang is currently the manager of mobile phone section in ITRI.

and fast to calculate. An additional enhancement feature is the cepstral coefficients which can be derived from LPC parameter set [6].

2.2. Integrated training scheme

A simple thought to utilize the back-propagation neural network to recognize a limited set of Chinese vocabulary such as Chinese numbers and other terms used in mobile handsets is as shown in Fig. 1. The neurons at the input layer is assumed to be $I_i, i = 1, \ldots, n$, while the neurons at the output layer is $O_k, k = 1, \ldots, m$. Besides, the hidden layer which is composed of $H_j, j = 1, \ldots, l$, is required to make the back-propagation neural network converge. Entries of the weight matrix between the input layer and the the hidden layer and those of the weight matrix between the hidden layer and the output layer are, respectively, expressed as w_{ij} and w_{jk}. Feature vectors (also called training vectors), t_p, $p = 1, \ldots, M$, generated by the extraction of given Chinese numbers are fed into the input neurons. Thus, the weight updating scheme in the training phase is summarized as follows [7].

Output layer weight updating scheme :

$$\begin{aligned} \delta_{pk} &= (t_p k - O_{pk}) \cdot O_{pk} \cdot (1 - O_{pk}) \\ \triangle_p w_{kj} &= \eta \cdot \delta_{pk} \cdot O_{pk} \end{aligned} \tag{1}$$

Hidden layer weight updating scheme :

$$\begin{aligned} \delta_{pj} &= O_{pj} \cdot (1 - O_{pj}) \cdot \sum_k \delta_{pk} w_{kj} \\ \triangle_p w_{ji} &= \eta \cdot \delta_{pj} \cdot O_{pi} \end{aligned} \tag{2}$$

Notably, the activation function of each neuron is assumed to be a sigmoidal function in the above derivation.

$$\text{sigmoidal} : f(x) = \frac{1}{1 + e^{-x+x_0}}, \tag{3}$$

where x_0 is an offset constant. Besides, the η in the above equation is called the learning rate which indicates how fast an updating step is.

If the neural network defined by Eqns.(1), (2), (3) is trained in order to recognize the phonetic Chinese numbers, the training time for the network to converge will be very long when the number of extracted speech feature vectors are large. By contrast, if the set of training feature vectors is limited, the discriminative rate of the numbers will decrease.

2.3. Distributed training scheme

Although the traditional back-propagation neural networks possess a potentially powerful recognition capability, an inefficient training method might neutralize this feature. We, thus, propose a distributed training scheme for the training of a finite set of Chinese numbers to enhance the recognition capability of the back-propagation neural networks and reduce the training time.

In the proposed scheme, one small network is dedicated for one single Chinese number, as shown in Fig. 2. For instance, the "0" network is trained by the same training vector set as that in the integrated scheme. However, there is a single output neuron at the output layer which is turned "ON" only given a correctly corresponding input vector.

Hence, if there are M terms to be recognized, there are exactly M networks which is corresponding to each term, respectively. The theoretical reasons for such distributed networks to converge faster than the integrated network are as follows.

A. The weights to be adjusted in each distributed network is much less than that of the integrated network. Use the ten Chinese numbers as an illustrative example. In Fig. 1, the number of weights in the integrated training scheme to be trained is as follows.

$$W_{integrated} = n \times l + l \times 10 \tag{4}$$

By contrast, the number of weights in each network of the distributed network, as shown in Fig. 2, is

$$W_{distributed} = n \times l + l \times 1 \tag{5}$$

Hence, those M distributed networks can be trained simultaneously such that the speed of convergence is enhanced.

B. According to the Boltzmann machine learning analysis [7], the learning speed given in a back-propagation-like neural network is defined as

$$\frac{\partial G}{\partial w_{ij}} = -\frac{1}{T}\left[prob_{ij}^{+} - prob_{ij}^{-}\right], \tag{6}$$

where G is an information measure, w_{ij} is the weight between unit i and unit j, T is a constant temperature, $prob_{ij}^{-}$ is the corresponding probability when the network is free running, and $prob_{ij}^{+}$ is the probability averaged over all environmental inputs and measured at equilibrium.

In the integrated training scheme, the weight vectors of the network is randomly "pulled" to the one of states of "1", "2", ..., "0", depending upon the currently fed training vector. This property, thus, decreases the possibility for the network to converge to a final state which is defined by the equilibrium of all ten numbers. By contrast, the individual network in the distributed training scheme is dedicated for the convergence of one single state defined by the equilibrium particular number, say "0". Hence, the induced probability averaged over training vectors in the formal case is less than that of the latter case.

$$prob_{ij\,\text{integrated}}^{+} \leq prob_{ij\,\text{distributed}}^{+} \tag{7}$$

This simple fact indicates the following conclusion.

$$\frac{\partial G}{\partial w_{ij}}_{\text{integrated}} = -\frac{1}{T}\left[prob_{ij\,\text{integrated}}^{+} - prob_{ij}^{-}\right]$$

$$\geq -\frac{1}{T}\left[prob^{+}_{ij\,\text{distributed}} - prob^{-}_{ij}\right] = \frac{\partial G}{\partial w_{ij}\,\text{distributed}} \quad (8)$$

Notably, since the weight updating scheme for back-propagation neural networks is essentially a gradient decent approach, Eqn.(8) shows that the step size for each down-hill updating stage of the distributed training scheme is larger than that of the integrated training scheme. This fact will lead to a fast convergence of the distributed training scheme.

3. SIMULATION AND IMPLEMENTATION

A properly segmented speech input waveform is divided into 6 frames in which a order 10 LPC is used to generate the ceptral coefficients. Hence, 60 LPC ceptral coefficients are selected to be the speech data representation, while 100 hidden neurons are used to precede a series of simulations. A sample speech waveform is shown in Fig. 3, which is the Chinese number "0". Fig. 4, thus, shows the LPC encoded waveform. In the back-propagation neural network training, the learning rate is chosen to be 0.7, and the momentum is 0.6 [7]. The definition of the convergence is defined as $\mid t_{pk} - O_{pk} \mid < 0.01,\ \forall k,\ k = 1, \ldots, M$. In each training iteration, the training vector pair which has the largest difference is selected such that the maximal error could be reduced in each iteration. Thus, Table 1 shows the results of the simulations based upon the distributed training scheme. Table 2 reveals the result of the integrated training scheme.

It is obvious that the distributed training scheme is superior in every category. The prediction of Eqn.(8) is also verified. Although the total number of iterations of "0", .., "9" is larger than that of "0-9", a parallel processing method is adopted such that distributed training is feasible and fast. Besides, the most important of all is that when there is a new phase or word needed to be added to the handsets, there is no need to re-train or re-tune the entire large network. Instead, simple generate a small network dedicated for the new phrase or word.

4. CONCLUSION

The distributed training scheme provides an efficient training alternative for finite Chinese vocabulary used in mobile handsets. When there is a new word or phrase needed to be added into user's speech recognition database, the proposed method indicates that it will be faster to "recognize" the new word by generating a new and small network instead of tuning the original and large network.

5. REFERENCES

[1] M.-I. Chen, "Practical implememtation of PC computer speech recognition, READING: (Chinses version), Flags Publishing Inc., 1994.

[2] R. Chengalvarayan, and L. Deng, "HMM-based speech recognition using state-dependent, discriminatively derived transforms on Mel-Warped DFT features," *IEEE Trans. on Speech and Audio Processing*, vol. 5, no. 3, May 1997.

[3] Interactive Speech Line of Product, "RSC-164," data sheet.

[4] C. R. Jankowski, H. Vo, and R. P. Lippman, "A comparison of signal processing front ends for automatic word recogniti on," *IEEE Trans. Speech Audio Processing*, vol. 3, pp. 286-293, July 1995.

[5] K.-G. Lee, "Speaker-independent speech recognition of Chinese characters 0 to 9," M.S. Thesis, National Cheng-Kung University, 1997.

[6] L. Rabiner, and B.-H. Juang, "Fundamentals of speech recognition," Reading : Prentice Hall, 1993.

[7] D. E. Rumelhart, and J. L. McClelland, "Parallel distributed pocessing," Vol. 1: Foundations, Reading : MIT Press, 1986.

[8] L. R. Rabiner, and R. W. Schafer, "Digital processing of speech signals," Reading : Bell Lab. Inc., 1978.

number	# times	max. iterations	min. iterations	average
"0"	77	1,102,710	9031	103929
"1"	59	858,693	10,374	106,970
"2"	57	1,105,580	33,426	146,789
"3"	88	833,474	8,449	94,390
"4"	101	787,816	6,434	73,252
"5"	51	1,290,080	19,988	145,181
"6"	50	1,017,800	33,007	150,212
"7"	57	1,435,690	15,989	127,303
"8"	59	1,258,040	28,621	138,357
"9"	57	669,882	18,642	125,152

Table 1 : The training speed of the distributed training scheme.

number	# times	max. iterations	min. iterations	average
"0-9"	62	1,603,590	622825	765,324

Table 2 : The training speed of the integrated training scheme.

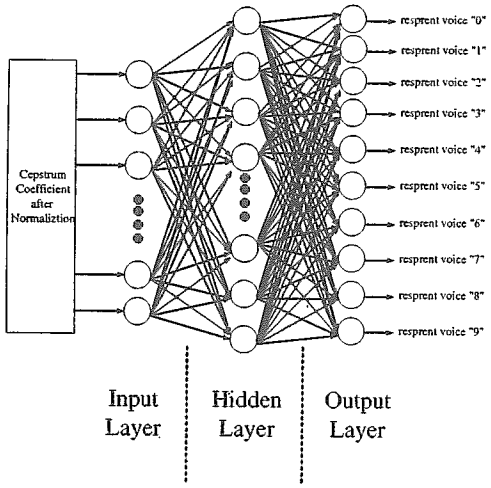

Figure 1: Integrated training scheme

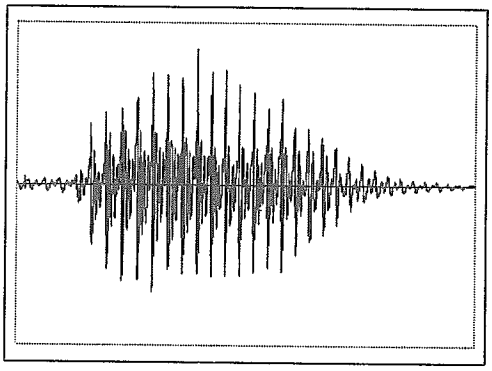

Figure 3: A sample speech waveform Chinese number "0"

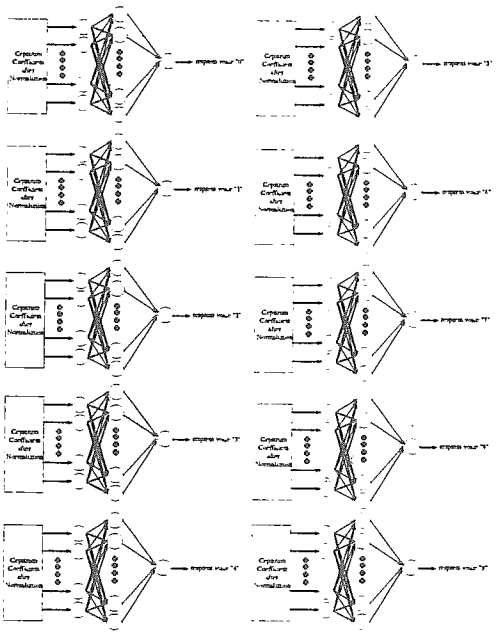

Figure 2: Distributed training scheme

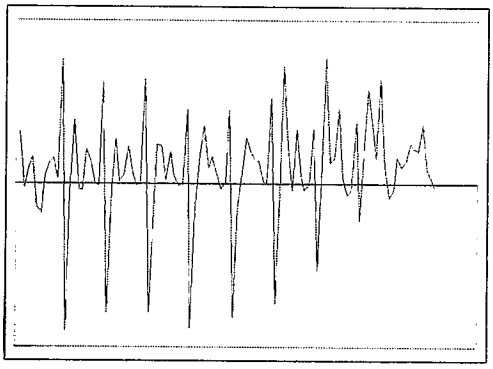

Figure 4: The LPC diagram of a Chinese number "0"