ISCIT 2004

International Symposium on Communications
and Information Technologies 2004 (ISCIT 2004)
Sapporo, Japan, October 26- 29, 2004

Recurrent Neural Network with Backpropagation through Time for Speech Recognition

Abdul Manan Ahmad*, Saliza Ismail+, Den Fairol Samaon^{L}
* Assoc. Prof of Universiti Teknologi Malaysia
manan@fsksm.utm.my*, chukiwa@hotmail.com+, den_fai@lycos.com^{L}
Department of Software Engineering,
Faculty of Computer Science and Information System,
Universiti Teknologi Malaysia, 81310 Skudai, Johor Darul Takzim, Malaysia
Tel: 607-5532201* Fax: 607-5565044*

Abstract- **The study on speech recognition and understanding has been done for many years. In this paper, we propose a fully-connected hidden layer between the input and state nodes and the output. Besides that, we also investigate and show that this hidden layer makes the learning of complex classification tasks more efficient. We also investigate difference between LPCC and MFCC in feature extraction process. The aim of the study was to observe the difference of Arabic's alphabet like "alif" until "ya". The purpose of this research is to upgrade the people's knowledge and understanding on Arabic's alphabet or word by using Fully-Connected Recurrent Neural Network (FCRNN) and Backpropagation through Time (BPTT) learning algorithm. 6 speakers (a mixture of male and female) are trained in quiet environment.**

Neural Network is well-known as a technique that has the ability to classified nonlinear problem. Today, lots of researches have been done in applying Neural Network towards the solution of speech recognition [1] such as Arabic. The Arabic language offers a number of challenges for speech recognition [2]. Even though positive results have been obtained from the continuous study, research on minimizing the error rate is still gaining lots of attention. This research utilizes Recurrent Neural Network, one of Neural Network technique to observe the difference of alphabet "alif" until "ya"".

1. INTRODUCTION

Speech is a human's most efficient communication modality. Beyond efficiency, human are comfortable and familiar with speech. Other modalities require more concentration, restrict movement and cause body strain due to unnatural positions. Research work on Arabic speech recognition, although lagging that other language, is becoming more intensive than before and several researches have been published in the last few years [3].

The conventional neural networks of Multi-Layer Perceptron (MLP) type have been increasingly in use for speech recognition and also for other speech processing applications. Those networks work very well as an effective classifier for vowel sounds with stationary spectra, while their phoneme discriminating power deteriorates considerably for consonants which are characterized by variations of their short-time spectra.

This may be attributable to a fact that feedforward multi-layer neural networks are inherently unable to deal with time-varying information like time-varying spectra of speech sounds. To cope to this problem, we incorporate feedback structure in the networks to provide them with an ability to memorize incoming time-varying information. This incorporation of feedback structure in feedforward networks results called Recurrent Neural Networks (RNNs) which have feedback connections between units of different layers or connections of self-loop type [4].

The motivation for applying recurrent neural nets to this domain is to take advantage of their ability to process short-term spectral features but yet respond to long-term temporal events. Previous research has confirmed that speaker recognition performance improves as the duration of utterance is increased [5]. In addition, it has been shown that in identification problems RNNs may confer a better performance and learn in a shorter time than conventional feedforward networks [6].

Now, students not interested in lessons regarding Arabic language such as Jawi writing even the lessons have been teaching at primary school. The purpose of the lessons is to teach the students how to pronoun and to write the alphabet. Therefore, students can read Holy-Quran properly.

But the students only can understand that pronoun and writing while they in Standard 6. So after that, they will forget all the lessons [7].

2. ARCHITECTURE

Recurrent neural networks (RNNs) have feedback connections. They address the temporal relationship of inputs by maintaining internal states that have memory. RNN are networks with one or more feedback connection. A feedback connection is used to pass output of a neuron in a certain layer to the previous layer(s) [8]. The different between MLP and RNN is RNN have feedforward connection for all neurons (fully connection). Therefore, the connections allow the network show the dynamic behavior. RNN seems to be more natural for speech recognition than MLP because it allows variability in input length [9].

Recently a simple recurrent neural, which has feedback connections of self-loop type around hidden layer units, has been proposed as an attractive tool for recognizing speech sounds including voiced plosive sounds [10]. We introduce a fully-connected hidden layer between the input and state and the output is shown in Fig. 1. Fully-connected means all nodes have been connected with another node and itself.

The Backpropagation through Time (BPTT) algorithm is based on converting the network from a feedback system to purely feedforward system by folding the network over time. Thus, if the network is to process a signal that is time steps long, then copies of the network are created and the feedback connections are modified so that they are feedforward connections from one network to the subsequent network. The network can then be trained as if it is one large feedforward network with the modified weights being treated as shared weights [11]. The recursive back propagation or Real-Time Recurrent Learning (RTRL) algorithm is based on recursively updating the derivatives of the output and error.

These updates are computed using a sequence of calculations for iteration. The weights are updated either after for iteration or after the final iteration of the epoch. The major disadvantage of this algorithm is that it requires an extensive amount of computation for iteration [12]. Additionally, this algorithm is very slow because the RTRL has many weights to compute and therefore, the training process will be more slowly [8].

3. SPEECH RECOGNITION SYSTEM

Speech recognition is the process of converting an acoustic signal, captured by a microphone or a telephone, to a set of words. The recognized words can be the final results, as for applications such as commands and control, data entry, and document preparation. They can also serve as the input to further linguistic processing in order to achieve speech

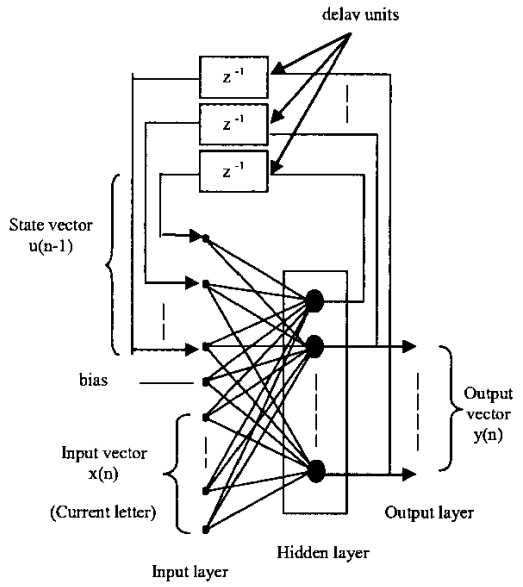

Fig. 1. Fully-Connected Recurrent Neural Network Architecture

understanding, a subject covered in section [13]. As we know, speech recognition performs their task similar with human brain. Start from phoneme, syllable, word and then the sentence which is an input for speech recognition system [14]. Many researches that have been prove to decrease the error and also any disruption while doing the recognition.

Generally, speech recognition process contains three main stages for processing the speech which is acoustic processing, feature extraction and recognition, as shown in Fig. 2. First, we digitize the speech that we want to recognize. In this paper, we digitize the Arabic's alphabet from speaker and also digital filtering that emphasizing important frequency component in signal. Then we analyze the start-end point depends the signal of the phonemes. To filter and conversion the analog to digital, we used GoldWave. From that, we can analyze the start-end point that contains the important information of speeches.

The acoustic processing obtains the sequence of input vector that will be used in next stages, feature extraction. For comparison purposes, Linear Prediction Coding (LPCC) and Mel frequency cepstral coefficients (MFCC) are performs.

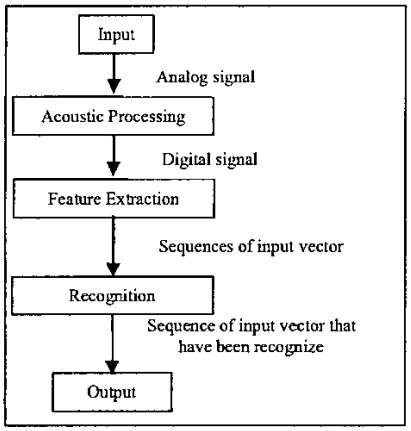

Fig 2. Process of Speech Recognition

The most popular feature set has been the vector of Mel frequency cepstral coefficients (MFCC) traditionally used also in speech recognition. MFCC's are cepstral coefficients computed on a warped frequency scale based on known human auditory perception. In a typical MFCC processing [15], the first step is windowing the speech signal to divide the speech into frames. Since high frequency formants have smaller amplitude than low frequency formants, high frequencies may be emphasized to obtain similar amplitude for all formants. After windowing, FFT is used to find the power spectrum of each frame. Then perform filter bank processing to the power spectrum, which uses mel-scale. Discrete cosine transformation is applied after converting the power spectrum to log domain in order to compute MFCC coefficients.

Linear Predictive Coding (LPCC) is used to extract the LPCC coefficients from the speech tokens [16,17]. The LPCC coefficients are the converted to cepstral coefficients. The cepstral coefficients are normalized in between +1 and -1. The cepstral coefficients are served as the input to the neural networks. The speech is blocked into overlapping frames of 20ms every 10ms using Hamming window. LPCC was implemented using the autocorrelation method. A drawback of LPCC estimates is their high sensitivity to quantization noise. Converts LPCC coefficients into cepstral coefficients where the cepstral order is the LPCC order and to decrease the sensitivity of high and low-order cepstral coefficients to noise, the obtained cepstral coefficients are then weighted [18].

Finally, we classify and recognize the speech with learning algorithm Backpropagation through Time in Recurrent Neural Network. In this paper, the domain of Arabic's alphabet is being experimented. A sample of the Arabic's alphabet "alif" before and after start-end point detection process is shown in Fig. 3. As we can see from the Fig. 3, the start point is 7900 and the end point is 21300 for "alif".

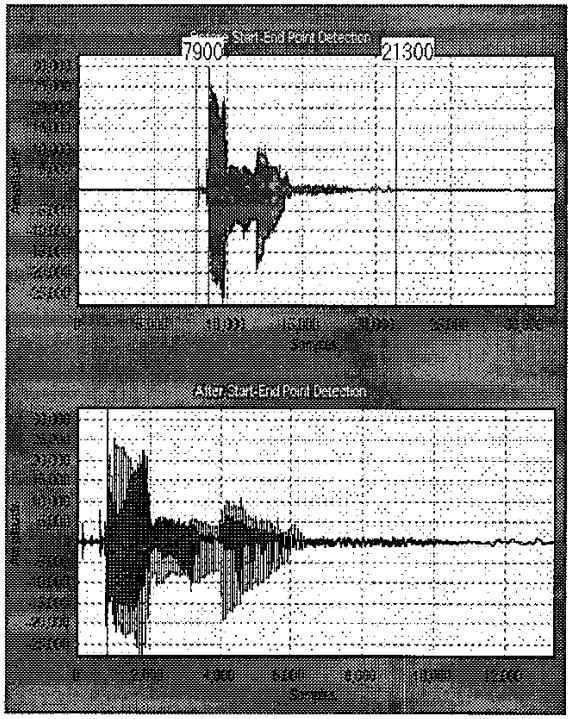

Fig 3. Speech waveform for Arabic's alphabet "alif" (before and after start-end point detection)

The Arabic alphabet has 29 letters and it is written from right to left. Letters change shape depending on which other letters are before or after them, much like American or Continental handwriting.

The Arabic's alphabet in this research contains 29 letters such as "alif", "ba", "ta", "tha", "jim", "ha", "kha", "dal", "zal", "ra", "zai", "sin", "syin", "sad", "dhad", "to", "za", "ain", "ghain", "fa", "qaf", "kaf", "lam", "mim", "nun", "wau", "hah", "hamzah" and "ya", and pronoun by 4 speakers.

4. EXPERIMENTS

The testing of this system has been pronounced by 4 Malay speakers (2 men and 2 women). Every speaker must repeat the Arabic's alphabet about 10 times sequentially for 29 alphabets. So, total of the pronoun for this training and testing that includes 4 speakers x 29 alphabets x 10 times for every alphabet (4x29x10), are 1160 speeches. So, the inputs for the training are 1160 nodes.

For the evaluation of the proposed RNN, both implementation of spectral analysis is tested to determine the best method and the performances are shown in Table 1 and Table 2 with fixed frames=60 and learning rate=0.25.

Table 1 : Implementation of Linear Prediction Coding (LPCC)

Speaker	Hidden Unit	% Acceptance
F1	40	94.5
	50	99.3
	60	98.6
F2	40	95.1
	50	98.6
	60	98.6
M1	40	95.1
	50	99.3
	60	98.6
M2	40	97.2
	50	98.6
	60	97.9

Table 2 : Implementation of Mel-Frequency Cepstral Coefficients (MFCC)

Speaker	Hidden Unit	% Acceptance
F1	40	96.6
	50	97.9
	60	97.9
F2	40	95.9
	50	97.2
	60	97.2
M1	40	97.2
	50	98.6
	60	97.9
M2	40	96.6
	50	97.9
	60	97.2

Table 1 and Table 2 compare LPCCC and MFCC performances with different hidden node (40, 50 and 60) for 4 speakers. Overall, performances of LPCC surpass the performance of MFCC about 0.7%.

5. CONCLUSION

Currently development of speech recognition is widely used in industrial software market. The main contribution of proposed speech recognition system is encouraged to recognize the Arabic's alphabet properly. Investigations of difference between LPCC and MFCC in feature extraction process for Arabic recognition. The aim of the study was to observe the difference of Arabic's alphabet like "alif" until "ya". The purpose of this research is to upgrade the people's knowledge and understanding on Arabic's alphabet or word by using Fully-Connected Recurrent Neural Network (FCRNN) and Backpropagation through Time (BPTT) learning algorithm. Findings from results of the training and testing can be summarized as follows:

1. The best performance of both cepstral analyses is LPCC with 50 hidden units (99.3%).
2. In this paper, rate of recognition more efficient with 50 hidden unit either in LPCC or MFCC.

Hopefully, this system will help us to recognize and differentiate the Arabic's alphabet from "alif" until "ya".

References

[1]Turban E., (1992). "Expert Systems and Applied Artificial Intelligence." Republic of Singapore: MacMillan Publishing Company, 623-640.

[2] Mayfield T. L., Black A. and Lenzo K., (2003). "Arabic in my Hand: Small-footprint Synthesis of Egyptian Arabic." Euro Speech 2003, Geneva, Switzerland.

[3]Jihene El Malik, (1998). "Kohonen Clustering Networks For Use In Arabic Word Recognition System." Sciences Faculty of Monastir, Route de Kairouan, 14-16 December.

[4]Medser L. R. and Jain L. C., (2001). "Recurrent Neural Network: Design and Applications." London, New York: CRC Press LLC.

[5]He J. and Liu L., (1999). "Speaker Verification Performance and the Length of Test Sentence." *Proceedings ICASSP 1999*, vol. 1, pp. 305-308.

[6]Gingras F. and Bengio Y., (1998). "Handling Asynchronous or Missing Data with Recurrent Networks." *International Journal of Computational Intelligence and Organizations*, Vol. 1, no. 3, pp. 154-163.

[7]Siddiq Fadzil, (2001). "Martabat umat Islam Melayu menurut Hamka." Utusan Melayu, 30 April.

[8]Ruxin Chen and Jamieson L. H., (1996). "Experiments on the Implementation of Recurrent Neural Networks for Speech Phone Recognition." *Proceedings of the Thirtieth Annual Asilomar Conference on Signals, Systems and Computers*, Pacific Grove, California, November, pp. 7790782.

[9]Lee S. J., Kim K. C., Yoon H. and Cho J. W., (1991). "Application of Fully Neural Networks for Speech Recognition." Korea Advanced Institute of Science and Technology, Korea, Page(s): 77-80.

[10]Koizumi T., Mori M., Taniguchi S. and Maruya M., (1996). "Recurrent Neural Networks for Phoneme Recognition." Department of Information Science, Fukui University, Fukui, Japan, Spoken Language, *ICSLP 96, Proceedings, Fourth International Conference*, Vol. 1, 3-6 October, Page(s): 326 -329.

[11]Werbos P., (1990). "Backpropagation through Time: What It Does and How To Do It." *Proceedings of the IEEE*, 78, 1550.

[12]Sato M., (1990). "A Real Time Running Algorithm for Recurrent Neural Networks." Biological Cybernetics, 62, 237.
[13]Lippman R.P., (1989). "Review of Neural Network for Speech Recognition." *Neural Computation 1*.1-38.

[14]Joe Tebelskis, (1995). "Speech Recognition using Neural Network." Carnegie Mellon University: Thesis PhD.

[15]Becchetti, C. and L.P. Ricotti, Speech Recognition, John Wiley & Sons Ltd., 1999.

[16]Ting H. N., Jasmy Yunus and Sheikh Hussain Salleh, (2002). "Speaker-Independent Phonation Recognition For Malay Plosives Using Neural Networks." Faculty of Electrical Engineering, Universiti Teknologi Malaysia, Malaysia.

[17]Greenberg S., Carvey H., Hitchcock L. and Chang S., (2001). "Beyond the Phoneme: a Juncture-Accent Model of Spoken Language." *Proceedings of the Human Language Technology Conference (HLT - 2002)*, San Diego, California, March 24-27.

[18]Rabiner L. R. and Juang B. H., (1993). "Fundamentals of Speech Recognition." 1st. Ed. New Jersey, United States of America: Prentice Hall. 58-148.