Jurnal Teknologi Maklumat & Multimedia 1(2004): 1-14

Malay Speaker Dependent Digits Recognition with Improved Backpropagation

UMMU SALAMAH MOHAMAD, RAMLAN MAHMOD &
SITI MARIYAM SHAMSUDDIN

ABSTRACT

This paper presents a study of a Malay speaker dependent recognition using improved Neural Network (NN). The performances are evaluated for recognition of the isolated Malay digits of "0" through "9". *The Error Backpropagation (BP) and an improved error signal of the BP are used in this study. Experiments are carried out by comparing the recognition rates and convergence time of the standard BP and improved BP, as well as the effects of normalisation techniques on Malay speaker dependent data. The utterances are represented using the Linear Prediction Coding (LPC) method. The results show that the improved BP outperforms the standard BP in terms of its convergence with better recognition rates for unnormalised data. For the effects of normalisation data, the unit simple method gives better result compared to unit range and unit variance with improved BP gives faster convergence and higher recognition rates.*

ABSTRAK

Makalah ini membincangkan kajian pengecaman jurucakap Melayu tidak bebas dengan menggunakan rangkaian neural pembaikan. Prestasi dinilai untuk pengecaman digit terpisah Melayu '0' hingga '9'. *Rambatan balik dan isyarat ralat pembaikan telah digunakan dalam kajian ini. Ujikaji telah dijalankan dengan membandingkan kadar pengecaman dan masa penumpuan rambatan balik piawai, rambatan balik pembaikan dan kesan teknik-teknik penormalan pada datajurucakap tidak bebas Melayu. Ujaran-ujaran diwakili dengan pengkodan ramalan lelurus. Keputusan menunjukkan bahawa, rambatan balik pembaikan mengatasi rambatan balik piawai dari sudut penumpuan dengan kadar pengecaman yang lebih baik untuk data yang tidak dinormalkan. Bagi kesan penormalan data, kaedah mudah memberikan keputusan baik berbanding dengan unit julat dan unit varians dengan rambatan balik pembaikan memberikan penumpuan lebih cepat dan kadar pengecaman yang lebih tinggi.*

INTRODUCTION

An automatic speech recognition system is a goal in speech research for more than six decades (Rabiner 1995). Much effort has been put on this field to produce new idea either in the form of research or commercial system (Markowitz 1996). To achieve this goal, many methods have been introduced such as Hidden Markov Model (HMM) and Dynamic Time Warping (DTW), which are commonly used within the classical method (Torkkola 1994). However, NN method has become more popular recently since its design and performance mimics the biological system of human being, which have attracted many researchers (Nakagawa 1995). The architecture of NN includes the parallel structure of input and output that make it faster in processing data compared to HMM and DTW. In addition NN does not require extra memory for retraining, but it is an excellent classification system. It can classify noisy data, pattern data, variable data streams, multiple data, overlapping, interacting and incomplete cues. Speech recognition is a classification task that has all these characteristics (Kevans 1997).

This paper proposes the performance of digits in Malay Language (ML) speech using standard BP algorithm, an improved error of BP and the normalisation techniques that are applied to the related data. We carried out an extensive comparison on the recognitions of the digits using standard BP and improved BP, also on the normalisation techniques applied to these data accordingly.

FEATURE ACQUISITION AND THE METHODOLOGY

An overview of Malay Digit Speaker Dependent is shown in Figure I and it consists of three phases. These include data acquisition, features extraction and recognition.

Speaker voice is recorded accordingly and uttered number *kosong* to *sembilan* for at least 100 times. Then we divide these data into half in which 50 data for training and the remaining are used for testing. According to Nyquist Theorem (Baert 1995), the original sound can be better replicated when the sampling rate is at least twice the frequency of the original sound. Therefore, the sampling rate is set to WkH with 8 bits format. Each recording consist "silence" at the beginning and ending speech signal (more precisely, non-speech segment). It is important that the non-speech segment be removed from the speech segment so that subsequent processing can concentrate more on the speech segment itself. Average magnitude is utilised as a useful basis for the beginning and endpoint detection.

The feature extraction is done by using LPC analysis method. The LPC analysis could provide a good model of speech signal. It is possible to estimate the important value of acoustic parameter from an incoming sample by using the parameter values from previous sample. The speech signals

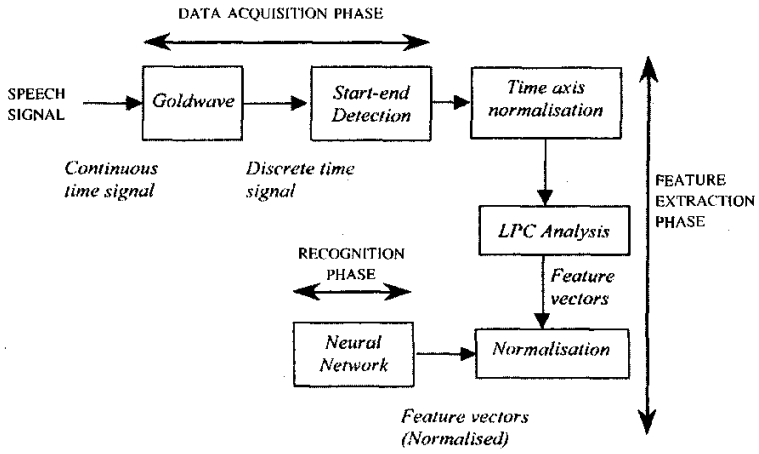

FIGURE 1. Overview of Malay digit speaker dependent

which are encoded using the 8th order LPC analysis (feature extraction) undergo pre-emphasis, frame blocking, windowing and LPC analysis. The function of pre-emphasis is to pre-emphasised the speech data whereas the frame blocking blocks the pre-emphasized data into frames. The produced frames are varied due to the intonation aspects and speaking rates. However the NN structure requires a fixed number of input neurons. Therefore the time-scale normalisation is done out the sampled waveform.

The time-axis normalisation is done after the starting and endpoints have been detected. In this stage, the average speech length is calculated in order to obtain a modification factor. This leads to the compression or enlargement of the sampled waveform, which depends upon the modification factor such that the whole speech signals with the same number of frames are obtained (Figure 2). The average speech length is about 300 millisecond (ms).

Since the sampling rate for the speech signal is 10kHz, 300 ms speech consists of 3000 samples. As a result, 30 ms sliding window frame is overlapping at 10ms and produces 300 samples/frame and 100 samples in the overlap region. Based on the average speech length, 14 frames are produced to represent each Malay digit. Thus, fixed size frames are obtained from each utterance.

The next step is windowing function which is implemented to the individual frame to minimise the signal discontinues at the beginning and end of each frame. Finally LPC analysis is performed to convert each frame into LPC coefficients. As a result, a set of coefficients are obtained from each utterance. Each frame consists of 8 poles. Therefore, the total number of input vectors that are used in these experiments are 104 (8 x 13 frames).

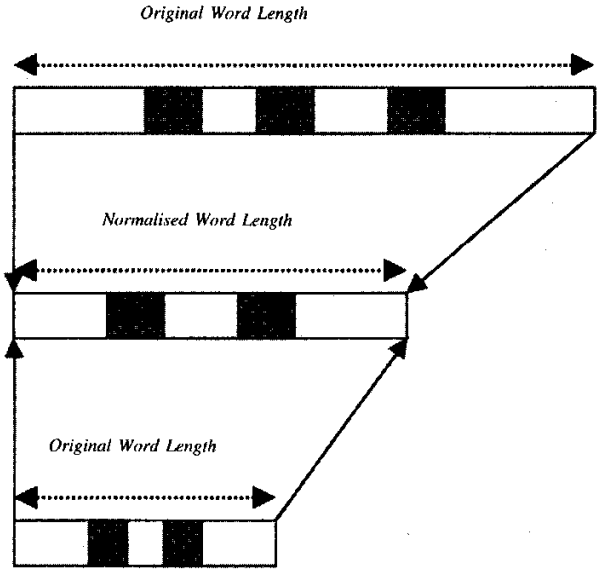

FIGURE 2. Time-Axis scale normalisation

Finally, data normalisation is performed to rescale the data to the range [0,1] or [-1,1] to avoid computational problems. There are 3 methods for input normalisation as summarised (Azoff 1994) along channel normalisation, across channel normalisation, mixed channel normalisation and external normalisation. The external normalisation using the along channel normalisation is utilised in these experiments and these are scaling to unit range technique, the linear scaling to unit variance technique and the simple normalisation method. The formula for each technique is described below.

$$x' = \frac{x-1}{u-1}, \tag{1}$$

$$x' = \frac{x-\mu}{\frac{3\sigma+1}{2}}, \tag{2}$$

$$x' = \frac{x}{\mu}. \tag{3}$$

where x', x, σ 1, u and μ represent the normalised feature, feature vector, sample variance, lower bound of the feature vector, upper bound of the feature vector and sample mean.

BACKPROPAGATION MODEL AND ITS IMPROVED ERROR FUNCTION

Figure 3 shows the training and recognition process are carried out in the final stage. BP is used as a classifier to solve these tasks. The architecture of BP is a feed-forward multi layers, which consists of three layers; input layer, hidden layer and output layer (Figure 4). Each neuron is represented by a circle and interconnections (weight) between neurons are represented by the arrows. The neurons that are labelled as b are the bias neuron. In standard architecture, each neuron of a layer is only connected to the neurons of the next layer (Fausett 1993). As a result, the input signal xi' (i = l , ... n) propagates through the network in the feedforward direction. The net input to a hidden neuron zj.,(j = 1 ... l) is the summation of all the inputs coming to the neuron related which is multiplied by the weight accordingly (Equation 4).

$$net\_input_j = b_j + \sum_{i=1}^{n} x_i v_{ij}..., \tag{4}$$

and,

$x_1..x_n$	input layer neuron
$z_1..z_l$	hidden layer neuron
$y_1..y_m$	output layer neuron
$v_{11}..v_{nm}$	weights between input layer and hidden layer
$w_{11}..w_{lm}$	weights between hidden layer and output layer
$\delta_1..\delta_l$	error signal
$e_1..e_m$	error for the neuron at the output layer.

The output or the activation function of a hidden neuron is computed as,

$$z_j = f(net\_input_j). \tag{5}$$

An activation function from each hidden neuron is sent to the output neuron y_k, $(k = 1,...m)$. This function is used to compute the activation function of the output neuron.

$$y_k = f(z_j). \tag{6}$$

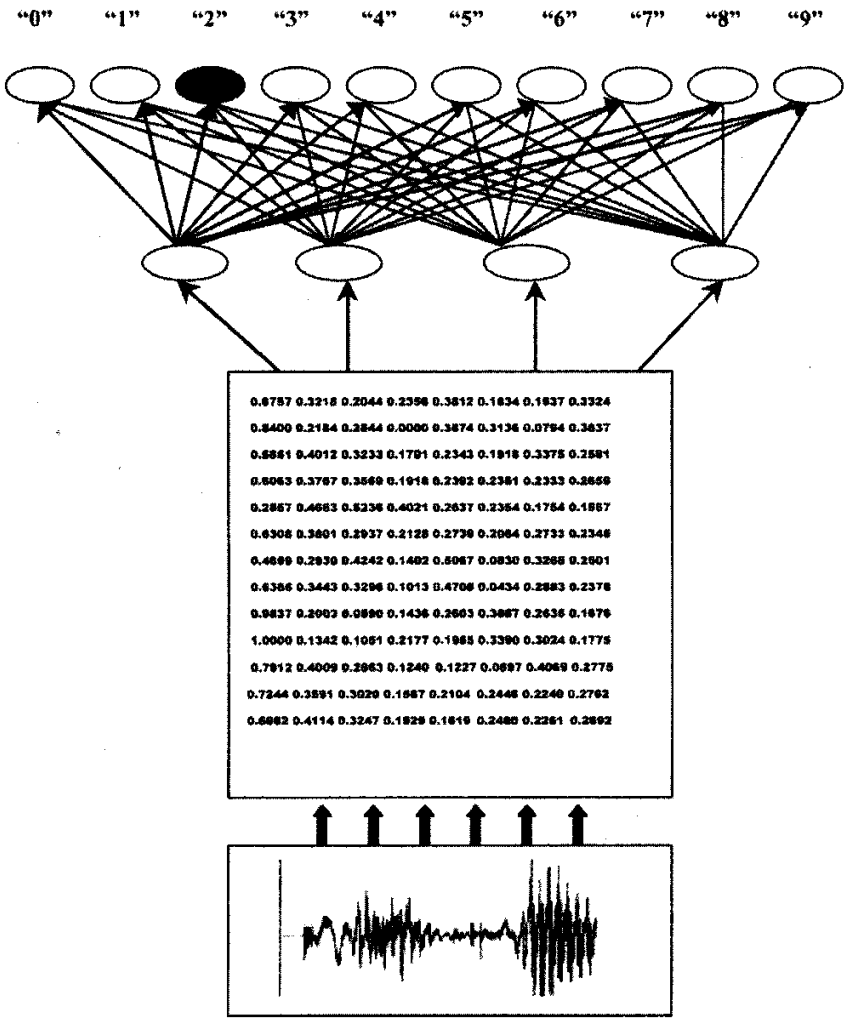

FIGURE 3. The process of recognition for utterance "2"

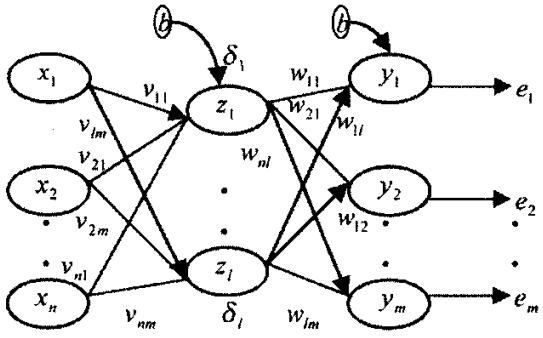

FIGURE 4. The BP architecture

The activation function of hidden and output neuron is computed by using sigmoid (S-shaped) function,

$$f_x = \frac{1}{1+e^{-x}}. \tag{7}$$

The most commonly employed sigmoid function is the logistic function. The advantage of this function is that its derivative is easily found (Shamsuddin et at. 2001),

$$f'(x) = f(x)(1 - f(x)). \tag{8}$$

In the training stage the actual output y_k is compared with the target output t_k. Any different is considered error and will be back propagated to hidden units. The actual output is the result of feedforward calculations. The error term for standard BP is (Figure 5):

$$e_k = \sum_{k=1}^{m} (t_k - y_k)^2. \tag{9}$$

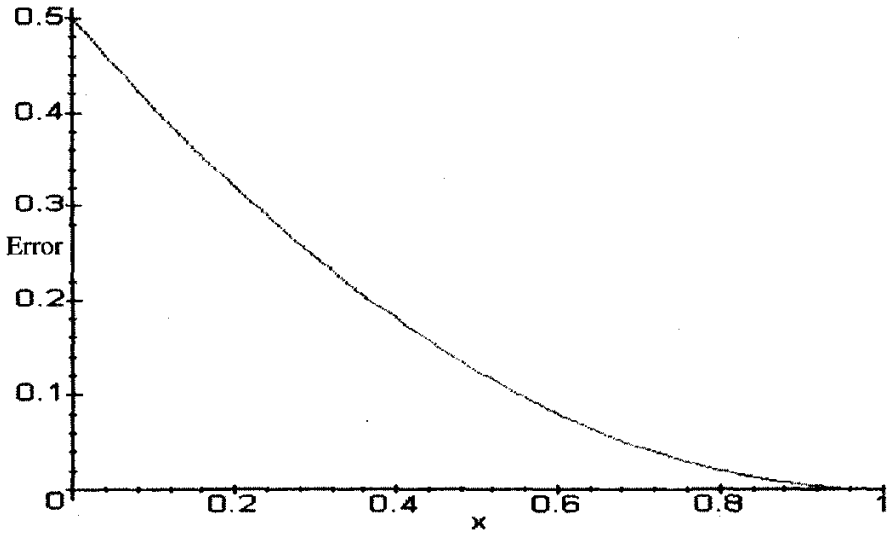

FIGURE 5. Mean square error of backpropagation model

The error is then used to compute the error signal for an output neuron.
The error signal for an output neuron can be represented as,

$$\delta_k = y_k(1 - y_k)(t_k - y_k). \tag{10}$$

The obtained error is than propagate backward for adjustment of weights (between hidden and output neuron) by the following equation,

$$w_{jk}(new) = w_{jk}(old) + \alpha\delta_k y_k + \beta(\Delta w_{jk}\ (old)), \tag{11}$$

Where α is the learning rate, β is the momentum term and *wjk(o[d)* is the previous weight changed.

The error signal for hidden neurons is computed for adjustment of weights (between hidden and input):

$$\delta_j = \sum_{j=0}^{n} v_{ij} zj\left(1 - z_j\right). \tag{12}$$

The weights are then updated using the formula below,

$$v_{ij}(new) = v_{ij}(old) + \alpha\delta_j x_i + \beta(\Delta v_{ij}(old)).$$

And this process continues until the network converges.

The common problems with BP algorithm are to find a local minimum and to minimise the error. Furthermore, the output neuron of standard BP can be zero not only when $t_k = y_k$ but also *whenfinet).* This leads to $\delta k = 0$ for internal units as well (Kalman et al. 1991). Therefore all the derivatives are zero, and the network loses its learning ability. To overcome these problems, a modified BP is utilised and this is based on the enhancement from standard BP and Kalman's BP. Sigmoid activation is used with the gain factor *x.* The introduction of gain factor k, into the BP formulations is to increase the convergence rates of the network (Figure 6).

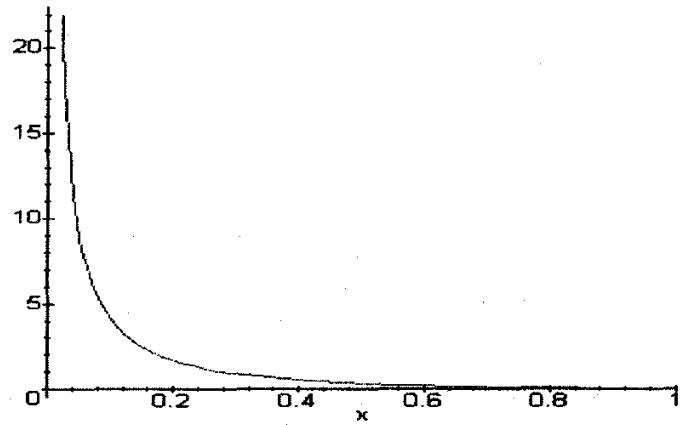

FIGURE 6. Proposed error function of backpropagation model with activation function of $\frac{1}{1+e^{-2x}}$

The proposed error function for standard backpropagation (mm) is defined implicitly as (Shamsuddin et al. 2001):

$$mm = \sum_{K} \rho_K ,$$

with

$$\rho_k = \frac{E_k^2}{2a_k\left(1-a_k^2\right)}, \tag{14}$$

where

$$E_k = t_k - a_k.$$

and

E_k error at output unit k,
t_k target value of output unit k,
a_k an activation of unit k.

An error signal for modified backpropagation for the output layer is,

$$\frac{\partial \rho_k}{\partial Net_k} = \frac{2\left(E+\rho\left(1-3a_k^2\right)\right)}{1+a_k} = \delta_k, \tag{15}$$

and for the hidden layer is the same as standard backpropagation as shown below,

$$\delta_j = \Sigma \delta_k \, w_j f'(a_j). \tag{16}$$

EXPERIMENTAL RESULTS

The experiment is carried out with a speaker who provides 50 tokens of test data for each word and 50 data for the trained template. There are 100 sets of speech data used in this experiment. The input nodes are set to 105 Where the normalisation size of each frame is represented by 8 LPC coefficients. The initial weight value for BP is generated using standard BP. Other parameters involved in this experiment are momentum, learning rate, and hidden node

which is set by trial and error. Maximum error rate of 0.002 is set for termination criteria. Each word is introduced 50 times to the net during training and the learned weights are saved and tested with 500 new patterns that have not been introduced to the net before.

EXPERIMENT ON NORMALISATION TECHNIQUES

The experiment for normalisation techniques is carried out to identify the best data normalisation for Malay speaker dependent among the 3 techniques mentioned above. Using Equation (1)-(3), the data of Table I is normalised accordingly as shown in Table 2 - Table 4. The normalisation data is then used as input to the network. Table 5 shows the recognition rates using normalisation techniques. It shows that the simple method gives better convergence rates compared to unit range and unit variance even though the recognition rates are quite closed. Figure 7 shows that the simple normalisation technique gives faster convergence rates compared to the other two. As a result, the simple normalisation technique is used for the following experiments using the standard BP and improved BP for comparison purposes.

TABLE 1. The unnormalized LPC coefficients of utterance "1"

Frames	LPC Coefficients (8 Poles)							
1	0.8037	0.1404	-0.0796	-0.0212	0.2517	-0.1190	-0.1747	0.1603
2	1.1117	-0.0534	0.0703	-0.4627	0.2259	0.1251	-0.3138	0.2564
3	0.6340	0.2893	0.1433	-0.1438	-0.0235	-0.1033	0.1698	0.0211
4	0.6736	0.2434	0.2063	-0.1033	-0.0144	-0.0220	-0.0254	0.0352
5	0.0727	0.4113	0.5186	0.2910	0.0315	-0.0215	-0.1340	-0.1708
6	0.7196	0.2497	0.0877	-0.0638	0.0507	-0.0758	0.0495	-0.0232
7	0.4180	0.0864	0.3324	-0.2000	0.4871	-0.3071	0.1498	0.0061
8	0.7342	0.1826	0.1551	-0.2729	0.4197	-0.3813	0.0777	-0.0169
9	1.3811	-0.0872	-0.3522	-0.1935	0.0252	0.2620	0.0311	-0.1485
10	1.4116	-0.2111	-0.2658	-0.0547	-0.0963	0.1727	0.1040	-0.1300
11	1.0202	0.2888	0.0740	-0.2303	-0.2328	-0.2945	0.3000	0.0574
12	0.8951	0.2103	0.1033	-0.1653	-0.0683	-0.0043	-0.0429	0.0549
13	0.7859	0.3084	0.1459	-0.1011	-0.1593	0.0021	-0.0390	0.0418

The recognition rates for the standard and improved BP is shown in Table 6. The performance of recognition is exceeds 95% for both standard and improved BP. Although the achievement of the recognition rates are similar for both networks, an improved BP gives faster convergence rates compared to the standard BP. The improved method takes only 25 epoch to converge while the standard BP takes 1022 epoch as illustrated in Figure 8. This is due

TABLE 2. The normalised LPC coefficients of utterance "1" using the Unit Range Technique

Frames	LPC Coefficients (8 Poles)							
1	0.6757	0.3218	0.2044	0.2356	0.3812	0.1834	0.1537	0.3324
2	0.8400	0.2184	0.2844	0.0000	0.3674	0.3136	0.0794	0.3837
3	0.5851	0.4012	0.3233	0.1701	0.2343	0.1918	0.3375	0.2581
4	0.6063	0.3767	0.3569	0.1918	0.2392	0.2351	0.2333	0.2656
5	0.2857	0.4663	0.5236	0.4021	0.2637	0.2354	0.1754	0.1557
6	0.6308	0.3801	0.2937	0.2128	0.2739	0.2064	0.2733	0.2345
7	0.4699	0.2930	0.4242	0.1402	0.5067	0.0830	0.3268	0.2501
8	0.6386	0.3443	0.3296	0.1013	0.4708	0.0434	0.2883	0.2378
9	0.9837	0.2003	0.0590	0.1436	0.2603	0.3867	0.2635	0.1676
10	1.0000	0.1342	0.1051	0.2177	0.1955	0.3390	0.3024	0.1775
11	0.7912	0.4009	0.2863	0.1240	0.1227	0.0897	0.4069	0.2775
12	0.7244	0.3591	0.3020	0.1587	0.2104	0.2446	0.2240	0.2762
13	0.6662	0.4114	0.3247	0.1929	0.1619	0.2480	0.2261	0.2692

TABLE 3. The normalised LPC coefficients of utterance "1" using the Unit Variance Technique

Frames	LPC Coefficients (8 Poles)							
1	0.8298	0.5096	0.4035	0.4317	0.5634	0.3845	0.3576	0.5193
2	0.9784	0.4161	0.4758	0.2186	0.5509	0.5023	0.2904	0.5656
3	0.7479	0.5815	0.5110	0.3725	0.4305	0.3920	0.5238	0.4521
4	0.7670	0.5594	0.5415	0.3920	0.4349	0.4313	0.4296	0.4589
5	0.4770	0.6404	0.6922	0.5823	0.4571	0.4315	0.3772	0.3595
6	0.7892	0.5624	0.4842	0.4111	0.4664	0.4053	0.4658	0.4307
7	0.6436	0.4836	0.6023	0.3454	0.6770	0.2937	0.5142	0.4448
8	0.7962	0.5300	0.5167	0.3102	0.6445	0.2579	0.4794	0.4337
9	1.1085	0.3998	0.2719	0.3485	0.4540	0.5683	0.4569	0.3702
10	1.1232	0.3400	0.3136	0.4155	0.3954	0.5252	0.4921	0.3791
11	0.9343	0.5813	0.4776	0.3307	0.3295	0.2997	0.5867	0.4696
12	0.8739	0.5434	0.4917	0.3621	0.4089	0.4398	0.4212	0.4684
13	0.8212	0.5907	0.5123	0.3931	0.3650	0.4429	0.4231	0.4621

to the proposition of implicit error in BP learning paradigm which leads to a better error signal of hidden layer (Shamsuddin et al. 2001).

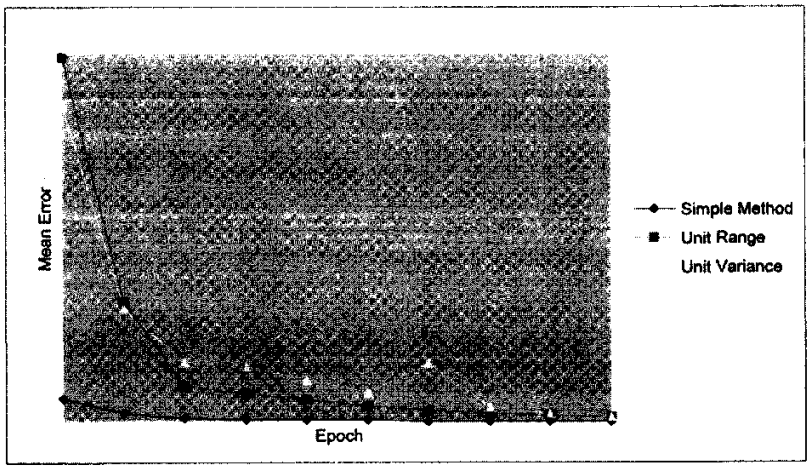

FIGURE 7. Mean error for isolated Malay digits utterance after normalisation

TABLE 4. The normalised LPC coefficients of utterance "1" using the simple technique

Frames	LPC Coefficients (8 Poles)							
1	0.5693	0.0994	-0.0563	-0.0150	0.1783	-0.0843	-0.1237	0.1135
2	0.7875	-0.0378	0.0498	-0.3277	0.1600	0.0886	-0.2223	0.1816
3	0.4491	0.2049	0.1015	-0.1018	-0.0166	-0.0731	0.1202	0.0149
4	0.4771	0.1724	0.1461	-0.0731	-0.0102	-0.0155	-0.0179	0.0249
5	0.0515	0.2913	0.3673	0.2061	0.0223	-0.0152	-0.0949	-0.1209
6	0.5097	0.1768	0.0621	-0.0451	0.0359	-0.0536	0.0350	-0.0164
7	0.2961	0.0612	0.2354	-0.1416	0.3450	-0.2175	0.1061	0.0043
8	0.5201	0.1293	0.1098	-0.1933	0.2973	-0.2701	0.0550	-0.0119
9	0.9783	-0.0617	-0.2495	-0.1370	0.0178	0.1856	0.0220	-0.1052
10	1.0000	-0.1495	-0.1882	-0.0387	-0.0682	0.1223	0.0736	-0.0920
11	0.7227	0.2045	0.0524	-0.1631	-0.1649	-0.2086	0.2125	0.0406
12	0.6341	0.1489	0.0731	-0.1171	-0.0483	-0.0030	-0.0303	0.0388
13	0.5567	0.2184	0.1033	-0.0716	-0.1128	0.0014	-0.0276	0.0296

TABLE 5. Recognition rates for normalisation speech data

Data Types	Unit Range(%)	Unit Variance(%)	Simple(%)
Training	100	100	100
Testing Data	96.83	96.67	97.67

TABLE 6. Recognition rates using standard BP and improved BP

Data Types	Standard BP(%)	Improved BP (%)
Training	100	100
Testing Data	96.67	97.67

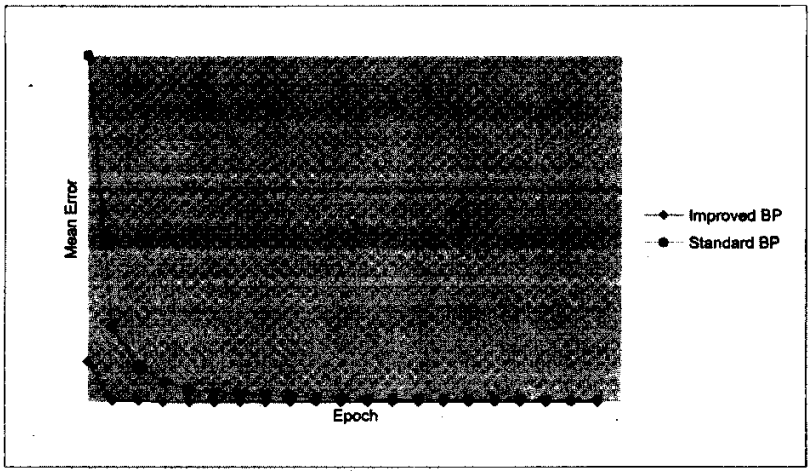

FIGURE 8. Mean error for isolated Malay digits utterances using standard BP and improved BP

CONCLUSION

This study shows that the normalisation of speech data using simple method gives a network with improved BP a better approach to solve Malay digit speech dependent with higher recognitions and faster convergence rates compared to the standard BP. Indirectly, this tell us that using an improved error function would generate less epoch size compared to the mean square error function due to the enhancement of error function at the hidden layer. The graph shows that the improve BP gives less iterations compared to the standard BP on Malay speech data. Therefore, further works will continue and the emphasis shall be on incorporating improved BP with speaker independent domain.

REFERENCES

Azoff, E. M. 1994. *Neural Network Time Series Forecasting of Financial Market.* Chichester: John Wiley & Sons.

Baert, L. 1995. *Digital Audio and Compact Disc Technology,* 3rd Edition. Oxford: Focal Press.

Fausett, L. 1994. *Fundamentals of Neural Networks: Architectures Algorithms and Applications.* London: Prentice Hall.

Kalman, B. and Kwasny, S. C. 1991. A Superior Error Function for Training Neural Network. *International Joint Conference of Neural Network* 2: 42-52.

Kevans, L. and Rodman, R. D. 1997. Voice Recognition. London: Artech House. Markowitz, J. A. 1996. *Using Speech Recognition.* London: Prentice Hall. Morgan, D. P. 1991. *Neural Networks and Speech Processing.* Boston: Kluwer Academic Publishers.

Nakagawa, S. 1995. *Speech, Hearing and Neural Networks Models.* Amsterdam: lOS Press.

Rabiner, L. R. 1993. *Fundamentals of Speech Recognition.* London: Prentice Hall. Rabiner, L. R. 1995. Impact of Voice Processing on Modern Telecommunications. *Speech Communication* 17(3-4): 217-226.

Shamsuddin, S. M., Sulaiman M. N. and Darus, M. 2001. An Improved Error Signal of BP Model for Classification Problems. *International Journal of Computer Mathematics* 76: 297-305.

Torkkola, K. 1994. Stochastic Models and Artificial Neural Networks for Automatic Speech Recognition, In Eric Keller (Eds.) *Fundamentals of Speech Synthesis and By Speech Recognition: Basic Concepts, State of the Art and Future Challenges.* New York JohnWiley.

Ummu Salamah Mohamad, Ramlan Mahmod
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
43400 UPM Serdang
Selangor
umohamad_hussin@hotmail.com
ramlan@fsktm.upm.edu.my

Siti Mariyam Shamsuddin
Faculty of Computer Science and Information System
Universiti Teknologi Malaysia
81310 UTM Skudai
Johor
mariyam@fsksm.utm.my