Jurnal Teknologi, 46(D) Jun 2007: 135–149

AN IMPROVED METHOD IN SPEECH SIGNAL INPUT REPRESENTATION BASED ON DTW TECHNIQUE FOR NN SPEECH RECOGNITION SYSTEM

RUBITA SUDIRMAN[1*], SH. HUSSAIN SALLEH[2] & SHAHARUDDIN SALLEH[3]

Abstract. A pre-processing of linear predictive coefficient (LPC) features for preparation of reliable reference templates for the set of words to be recognized using the artificial neural network is presented in this paper. The paper also proposes the use of pitch feature derived from the recorded speech data as another input feature. The Dynamic Time Warping algorithm (DTW) is the back-bone of the newly developed algorithm called DTW fixing frame algorithm (DTW-FF) which is designed to perform template matching for the input preprocessing. The purpose of the new algorithm is to align the input frames in the test set to the template frames in the reference set. This frame normalization is required since NN is designed to compare data of the same length, however same speech varies in their length most of the time. By doing frame fixing, the input frames and the reference frames are adjusted to the same number of frames according to the reference frames. Another task of the study is to extract pitch features using the Harmonic Filter algorithm. After pitch extraction and linear predictive coefficient (LPC) features fixed to a desired number of frames, speech recognition using neural network can be performed and results showed a very promising solution. Result showed that as high as 98% recognition can be achieved using combination of two features mentioned above. At the end of the paper, a convergence comparison between conjugate gradient descent (CGD), Quasi-Newton, and steepest gradient descent (SGD) search direction is performed and results show that the CGD outperformed the Newton and SGD.

Keywords: Dynamic time warping, time normalization, neural network, speech recognition, conjugate gradient descent

Abstrak. Kertas kerja ini membentangkan pemprosesan semula ciri pertuturan pemalar Pengekodan Ramalan Linear (LPC) bagi menyediakan template rujukan yang boleh diharapkan untuk set perkataan yang hendak dicam menggunakan rangkaian neural buatan. Kertas kerja ini juga mencadangkan penggunaan cirian kenyaringan yang ditakrifkan dari data pertuturan sebagai satu lagi ciri input. Algoritma Warping Masa Dinamik (DTW) menjadi asas kepada algoritma baru yang dibangunkan, ia dipanggil sebagai DTW padanan bingkai (DTW-FF). Algoritma ini direka bentuk untuk melakukan padanan bingkai bagi pemprosesan semula input LPC. Ia bertujuan untuk menyamakan bilangan bingkai input dalam set ujian dengan set rujukan. Pernormalan bingkaian ini adalah diperlukan oleh rangkaian neural yang direka untuk membanding data yang harus mempunyai kepanjangan yang sama, sedangkan perkataan yang sama dituturkan dengan kepanjangan yang berbeza-beza. Dengan melakukan padanan bingkai, bingkai input dan rujukan boleh diubahsuai supaya bilangan bingkaian

[1&2]Center for Biomedical Engineering, Faculty of Electrical Engineering, Universiti Teknologi Malaysia, 81310 UTM Skudai, Johor, Malaysia
[3] Department of Mathematics, Faculty of Science, Universiti Teknologi Malaysia, 81310 UTM Skudai, Johor, Malaysia
* Corresponding author: Email: rubita@fke.utm.my

sama seperti bingkaian rujukan. Satu lagi misi kertas kerja ini ialah mentakrif dan menggunakan cirian kenyaringan menggunakan algoritma penapis harmonik. Selepas kenyaringan ditakrif dan pemalar LPC dinormalkan kepada bilangan bingkaian dikehendaki, pengecaman pertuturan menggunakan rangkaian neural dilakukan. Keputusan yang baik diperoleh sehingga mencapai ketepatan setinggi 98% menggunakan kombinasi cirian DTW-FF dan cirian kenyaringan. Di akhir kertas kerja ini, perbandingan kadar *convergence* antara *Conjugate gradient descent* (CGD), *Quasi-Newton*, dan *Steepest Gradient Descent* (SGD) dilakukan untuk mendapatkan arah carian titik global yang optimal. Keputusan menunjukkan CGD memberikan nilai titik global yang paling optimal dibandingkan dengan *Quasi-Newton* dan SGD.

Kata kunci: Warping masa dinamik, pernormalan masa, rangkaian neural, pengecaman pertuturan, *conjugate gradient descent*

1.0 INTRODUCTION

Since its birth more than 30 years ago, Dynamic Time Warping (DTW) has been one of the prime speech recognition methods. Its mechanism is known as a matching technique of the unknown speech input template to a pre-define reference template, and this method is considered as the simplest speech recognition method compared to others like Hidden Markov Model (HMM) or neural network (NN). DTW is more popular among pattern recognition class methods due to its ability to search for the shortest path between two time-series signals like speech [1]. Through its matching technique, a test speech signal can be expanded or compressed according to a reference template [2, 3].

Now we live in the era equipped with high technology and fast computing devices. Research in speech recognition using back-propagation neural network (BPNN) also focus on developing a more precise recognition device with less network complexity but with fast processing time. This is to accommodate the current living standards and needs. Past research in NN found that when a higher number of hidden neurons is used, higher recognition rate is achieved, but longer processing time is required for the error to converge. To use NN as a recognition tool, one is required to fix the number of frames to the same length for training and testing data. Thus, a method to overcome these problems especially with less amount of input feature representation has to be developed. In that respect, time normalization is required to align the frames to a fix length with respect to the reference that we pick over the samples based on their average value. Time normalization is a typical method to interpolate input signal into a fixed size of input vector. Linear time alignment is the simplest method to overcome time variation, but it is a poor method since it does not account important feature vectors when deleting or duplicating them to shorten or lengthen the pattern vectors, if required [2, 3], nevertheless, it has been the basic method for compression and expansion of speech pattern vector [4, 5].

Many works have employed combinations of NN with multilayer perceptron (MLP) architecture, HMM, and DTW one way or the other [6]. Meanwhile, [7] used DTW and MLP with sequence of dynamic networks. They did not perform time alignment

using DTW, but instead they used DTW to find the global distance score and used that score as the input into their MLP. Other works involved DTW and NN also include [8, 9], however they also used the total distance of the warping path as input into the MLP.

2.0 FEATURE EXTRACTION AND FRAME MATCHING

There are many feature extraction methods for speech signals like mel frequency cepstral coefficient (MFCC), linear predictive coefficients (LPC), and linear predictive cepstral coefficient (LPCC). However, in this work linear predictive (LP) is chosen over other methods due to its ability to encode speech at low bit rate and can provide the most accurate speech parameters, so that least information is lost during feature extraction process. It has been widely used by speech researchers as speech features representation [2, 4, 5]. Features are represented in vectors form of a chosen dimension, which is called as the LP order.

Upon obtaining the LPC coefficients, then they are used as the feature vectors in the DTW frame matching stage. The coefficients are processed for frame normalization using the DTW-FF algorithm [10, 11]. The algorithm is built to match up the unknown word template to the reference template so that the unknown source's frame number is equivalent to the reference template. The reference template is chosen from the samples' population based on the averaged frame numbers. The frame fixing up process involved frames compression and expansion. This means that if the source has less number of frames than the template, then the source frames will be expanded. If the source has more frames than the template, the source frames will be compressed according to the rules outlined in subsection 2.1. Another advantage of this technique is the ability of the algorithm to reduce the number of input coefficients, in which only local distance scores of the warping path is retained for the NN recognition. This surely can reduce the amount of inputs presented to the network. It is proven when comparing between using 10 LPC coefficients per frame (from linear predictive algorithm) and using only a coefficient per frame (from our frame fixing algorithm). It also reduces the amount of network complexity and weight computations, and certainly would increase the convergence speed. Using 10-order LPC coefficients, the input reduction is up to 90% if using the DTW-FF feature, and this percentage will be higher if higher LPC order was used [11, 12].

2.1 DTW Frame Alignment

Dynamic Time Warping method was initially used for recognition of isolated words in association with Dynamic Programming (DP) [1]. It works based on template matching which provides an alternative to perform speech recognition. The template matching encountered problems due to speaking rate variability, where there exist

timing differences between similar utterances. Problem of time differences can be solved using DTW algorithm by warping the reference template against the test utterance based on their features similarities. So, DTW algorithm actually is a procedure that combines both warping and distance measurement, which is based on their local and global distance.

After feature extraction process, speech pattern can be represented by a feature vector sequence. As an example, let's consider two feature vectors t and r. Let t be the unknown/test speech pattern, $t = t_1, t_2, t_3,..., t_i,..., t_I$ and r be the speech reference pattern, $r = r_1, r_2, r_3,..., r_j,..., r_J$ [10]. Based on the DTW type 1 heuristic path in Figure 1, the distance of the pattern similarity matching is determined according to the Euclidean distance (refer to Equation (3)) as a mean of recognition measure.

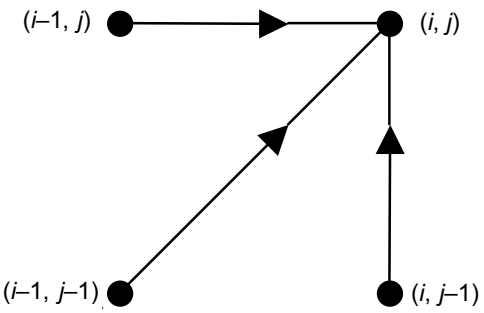

Figure 1 DTW heuristic path type 1

In this research, the frame fixing is done based on DTW method: the input vectors are warped with a reference vector which has almost similar local distance (horizontal warp), while expand the input vectors to reference vectors which shows a vertical warp (they share the same feature vectors for a feature vector frame of an unknown input). This frame alignment is also known as the expansion and compression method according to the slope conditions described as follows. There are three slope conditions that have to be dealt with in this research work, based on the DTW type 1 during the frame compression (denoted as F^-) and the expansion (denoted as F^+).

(i) Slope is ~0 (Horizontal line)
When the warping path moves horizontally, the frames of the speech signal are compressed. The compression takes place by calculating the minimum local distance amongst the distance set, i.e. compare $w(i)$ with $w(i–1)$, $w(i+1)$ and so on, where $w(i)$ represents the speech frame at time i. The search is represented as

$$F = F\left(\min\left\{d_{(i,j)\ldots(I,J)}\right\}\right) \tag{1}$$

(ii) Slope is ~∞ (Vertical line)
When the warping path moves vertically, the frame of the speech signal is expanded. This time the reference frame gets the identical frame as $w(i)$ of the unknown input source. In other words, the reference frame duplicates the local distance of that particular vertical warping frame. The expansion can be represented as

$$F^{+} = F(r(i)) \tag{2}$$

(iii) Slope is ~1 (Diagonal)
When the warping path moves diagonally from one frame to the next, the frame is left as it is because it already has the least local distance compared to other movements.

The distance is calculated using Euclidean distance measure. For a set of LPC coefficients with p feature vectors, which is from j=1, 2, ..., p of (x, y) coordinate, x represents the test set axis while y represents the reference set axis. The distance is calculated as

$$d(x, y) = \sqrt{\sum_{j=1}^{p} (x_i - y_j)^2} \tag{3}$$

The expansion and compression are done throughout the samples along the warping path where the input frames are matched to the reference template frames using the DTW-FF algorithm. After this procedure, the data are ready to be used for neural network recognition. The normalized data/sample has been tested and compared to the typical DTW algorithm and results showed a same global distance score. Further findings are discussed in the results and discussion section.

2.2 Pitch Optimization

Pitch is defined as the property of sound that varies with variation in the frequency of vibration. In speech processing, pitch is defined as the fundamental frequency (oscillation frequency) of the glottal oscillation (vibration of the vocal folds). Pitch information is one of speech acoustical features that is rarely taken into consideration when doing speech recognition. In this work, pitch is taken into consideration, then it is optimized and will be used as another feature into NN along with DTW-FF feature. Pitch contains spectral information of a particular speech, it is the feature that is being used to determine the fundamental frequency, *F0* of a speech at a particular time.

Figure 2 shows a flow diagram of the pitch optimization process. In short, firstly pitch extraction is done to the sample speech to obtain the initial values of fundamental

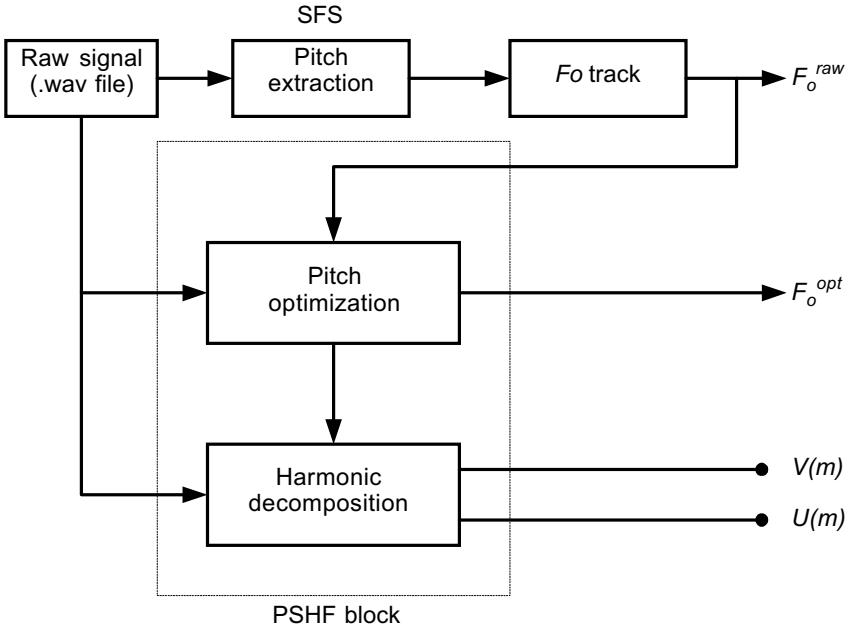

Figure 2 Process flow of pitch optimization

frequencies, or referred as F_o^{raw}; it can be obtained by pitch-tracking manually or by using available speech-related applications. Then this F_o^{raw} is fed into the pitch optimization algorithm and yielded an optimized pitch, F_o^{opt}.

Pitch optimization is performed to resolve glitches in voice activity and pitch discontinuities due to octave errors. The algorithm of the pitch optimization is based on the algorithm initially described in [13] and [14]. The algorithm finds the optimum pitch value for a particular time by minimizing the difference between the calculated and the measured smearing of the spectrum due to the window length.

3.0 BACK-PROPAGATION NN ALGORITHM

Back Propagation Neural Networks (BPNN) is one of the most common neural network structures as they are simple and effective, and has been used widely in assortment of machine learning applications. It has been giving general success in learning the neural networks. The term 'BP' refers to the gradient which is computed in the back-pass manner for multilayer networks. Properly trained BP networks are likely to give reasonable answers when presented with inputs that they have never seen before. Due to its known ability to minimize errors in their connection weights especially during the back-pass in the algorithm, BPNN is used in this research to search for minimum error between the training and test set. Log sigmoid activation function, $f(q) = (1+e^{-q})^{-1}$ is applied to activate the connection weights and readjust those weights

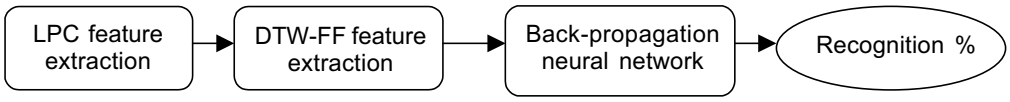

Figure 3 DTW-FF flow diagram with back-propagation NN

during the iterations. Mean square error method is used to compute the weights adjustments. Method of steepest gradient descent is employed in the direction search for fast convergence of the algorithm.

To summarize, Figure 3 shows the flow process of the experiment for back-end recognition that has been carried out upon obtaining the LPC coefficients. The data used for this preliminary study are 10-order LPC coefficients, using 10 ms frames Hamming windows of 6 subjects uttering digits 0-9 in Malay repeated 5 times in 5 different sessions. An average of 47 frames is selected for the reference in the digits recognition and this number is used against the source/unknown input during the frame fixing process.

4.0 EXPERIMENTAL SETUP

The experiment was conducted using learning rate, $\eta = 0.1$, momentum rate, $\alpha = 0.9$ with 10 hidden nodes and 47 input nodes. This initial experiment involves 6 speakers and each of the speakers uttered digit 0-9, 5 times for each digit, in Malay giving a total of 50 utterances for each subject. Since there is no fixed formula to determine the learning rate, momentum rate, and hidden layer, experiments were conducted to find out the optimum. Our experiments agreed with the optimum parameters obtained, in which their values are as suggested by most NN researchers, i.e., in our experiment, all subjects require learning rate, $\eta = 0.1$ and momentum rate, $\alpha = 0.9$ for their best recognition with 10 is the average number for hidden nodes. The recognition process can be summarized in the following order:

The training stage:

(i) LPC feature extraction order-10.
(ii) Use DTW to check the sample number of frames.
(iii) Use DTW-FF to fix the frames; source is fixed to have the same number of frames as the template.
(iv) Retain DTW-FF coefficients of the fixed frames.
(v) Use (d) as input into BPNN.
(vi) Obtain weights and save them for reference in the testing phase.

The testing stage:

(i) LPC feature extraction order-10.
(ii) Use DTW to check the sample number of frames.

(iii) Use DTW-FF to fix the frames; source must have the same number of frames as the template.
(iv) Retain DTW scores of the fixed frames.
(v) Load test samples.
(vi) Set the learning rate and momentum rate, i.e. must be the same as the training setting.
(vii) Recall weights in (f) of training phase.
(viii) Compare and obtain recognition percentage.

After the frame fixing process which uses DTW-FF algorithm, local distances of the fixed frames are collected and the data are ready to be used in the next stage, which is the recognition stage. The normalized data/sample are tested and compared to the LPC coefficients using typical DTW algorithm. The results did not show any changes in the recognition rate which also mean that there is no loss of information even though only fixed frame's local distance scores (DTW-FF feature) that are being used as the input. However, the use of DTW-FF feature can reduce the amount of input to be presented into the NN.

4.1 Results and Discussion

Figure 4 shows the frame being fixed to a fix number according to the reference template. In this particular word example, initially the input template has 24 frames, whereas the reference template has 27 frames. By using the DTW-FF algorithm the input frames have been expanded from 24 to 27, i.e. equal to the number of frames for reference template.

The number of fixed frames, N_{ff} is calculated as:

$$N_{ff} = N_{if} - N_{cf} + N_{ef} \tag{4}$$

where N_{if} = number of input frame
N_{cf} = number of compressed frame
N_{ef} = number of expanded frame

In Figure 4, the local distances of the unknown input frame $x(7)$, ..., $x(9)$ are compared according to the slope condition (i) due to its horizontal warping. Frame $x(9)$ appears to have the minimum local distance among the three frames, so those 3 frames are compressed to one frame and occupies only $y(7)$ (appears as frame 9 on y-axis). However, frame $x(19)$ of input is expanded to 6 frames, in accordance to the slope condition (ii) due to a vertical warping between the utterances. The warping shows an expansion of 6 consecutive frames of the reference template: 6 frames of reference template at $y(18)$, ..., $y(23)$ have the same feature vectors as frame $x(19)$ of the input vectors, so $x(19)$ occupies $y(18)$, ..., $y(23)$. These mean that frame $x(19)$ of the input has

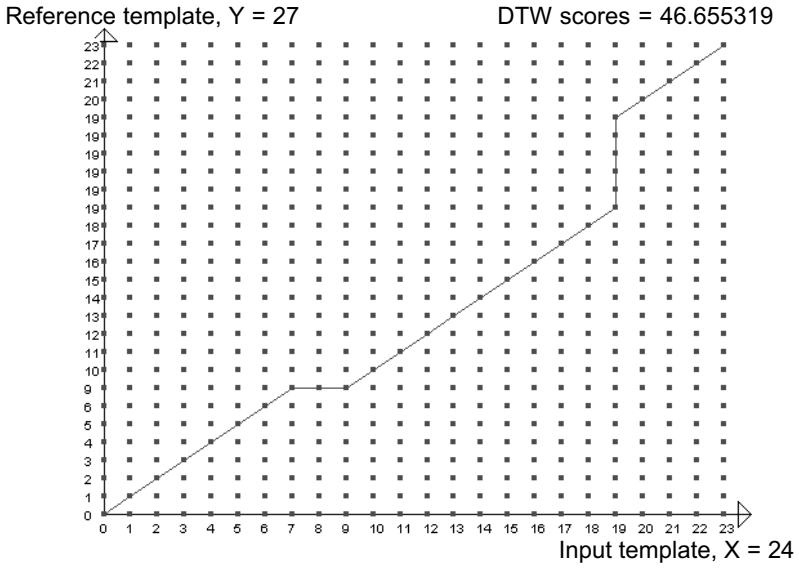

Figure 4 The DTW frame fixing between an input and a reference template by a speaker for digit '1'

matched 6 feature vectors in a row of the reference template set. Since diagonal movements is the fastest track towards achieving the global distance and it gives the least local distance at all time compared to the horizontal or vertical movements, a normal DTW procedure is applied to it.

4.2 DTW with LPC and DTW-FF Feature

An earlier experiment compares recognition of typical DTW using LPC coefficients and DTW-FF coefficients [11]. The results in Table 1 show the same recognition

Table 1 Recognition percentage - Typical DTW with LPC and DTW-FF coefficients, and BPNN with DTW-FF coefficients

	Recognition rate (%)		
	Typical DTW		**BPNN**
Subject	**LPC**	**DTW-FF**	**DTW-FF**
1	92	92	94
2	92	92	100
3	90	90	96
4	94	94	100
5	90	90	96
6	96	96	100

percentage between the two types of coefficients. This might be due to the same feature vectors pattern matching between frames template in both algorithms. Therefore, fixing the frames does not affect the recognition rate. This also supports that the recognition before and after DTW-FF is identical and no loss of information during the DTW-FF algorithm has occurred.

4.3 BPNN with DTW-FF Feature

For the second experiment, the local distance scores are preserved from DTW-FF algorithm. Improvement was observed using combination of methods as an extended recognition tool, which used combination of DTW (for DTW-FF feature) and BPNN. A perfect score is successfully achieved for 3 speakers (refer to Table 1, Column 4) for an average of using 20 hidden nodes. This would signify that during this early experiment the DTW-FF is able to produce relevant form of input data in much smaller size for the BPNN, compared to the amount of using LPC coefficients itself in this early experiment. Remember also, the number of inputs to the BPNN has been reduced about 90% by using the local distance scores instead of LPC coefficients [11].

Table 2 Recognition percentage using BPNN with 5 and 10 hidden nodes before and after pitch feature is added

Subject	**Recognition rate (%)**			
	5 hidden nodes		**10 hidden nodes**	
	Before	After	Before	After
1	22	78	84	100
2	98	100	100	100
3	62	78	86	95.9
4	32	68	72	98
5	48	70	80	90
6	80	94	100	100

The statistical test, called as T-Test has been conducted to the data in Table 1. The results rejected the hypothesis that $m_{before}=m_{after}$. Since $m_{before}<m_{after}$, then it can be concluded that the percent improvement of using DTW-FF coefficients by typical DTW and BPNN is significant. On the other hand, these mean a lot of network complexity and amount of connection weights computations during forward and back pass have been reduced.

4.4 BPNN with DTW-FF and Pitch Feature

Further experiments were conducted to find out the effect of adding pitch feature onto the DTW-FF feature. The pitch feature is extracted using pitch-scaled harmonic filter.

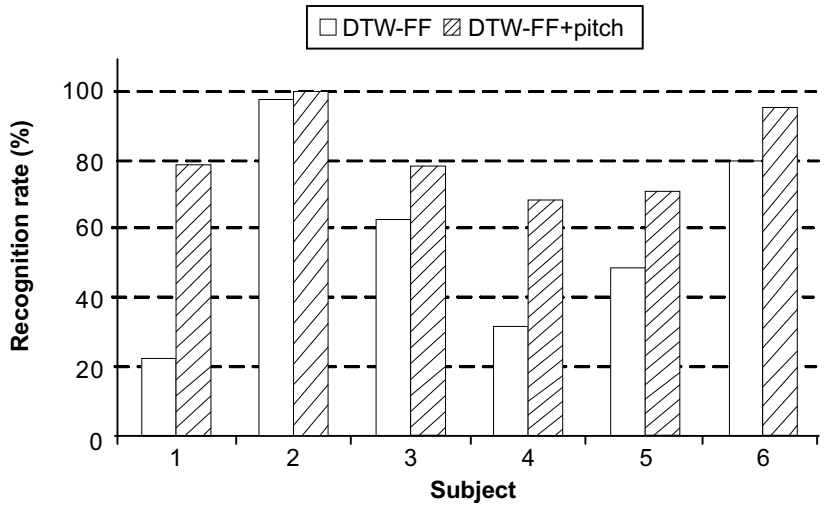

Figure 5 Recognition rate for 5 hidden nodes before and after pitch feature is included

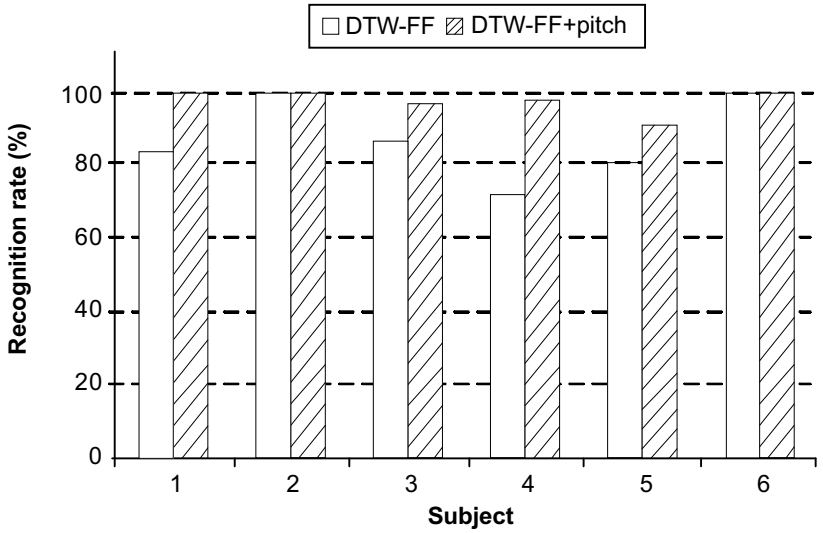

Figure 6 Recognition rate for 10 hidden nodes before and after pitch feature is included

The pitch feature is added after step (iv) in both training and testing phase which is described in Section 4.0. The results are tabulated in Table 2, graphically, the improvement (before and after pitch addition) can be seen in bar chart form in Figures 5 and 6.

Figures 5 and 6 clearly show that pitch feature has improved the recognition performance when added to DTW-FF feature even though pitch itself cannot give a good representation for speech recognition. The networks have learned sufficiently

with 10 hidden nodes rather than 20 hidden nodes to get to a high percentage before pitch feature is added. This means that the network converged faster.

The 5 hidden nodes comparison before and after pitch feature addition is only to show an improvement made even though the network has not learn sufficiently yet. However, the 10 hidden nodes improvement gives a good indication of the importance of pitch when combined with other feature like the DTW-FF feature.

5.0 COMPARISON OF CONVERGENCE RATE

The back-propagation neural network experiments carried out up to this point utilized the steepest gradient method. The recognition rate achieved is acceptable with their high percentage, sometimes reaching 100%. But the convergence time is what matters in this case, therefore the data are tested using other search engine for the back-propagation part. The forward pass mechanism is the same for all architecture except for the backward-pass that differs. The backward pass is replaced with the Newton and conjugate gradient algorithm, one at a time. The results using this algorithm are compared to the results using the steepest gradient algorithm. The comparison is made in terms of their convergence rate and epochs.

Figure 7 shows how the search for optimal global minimum behaved for each type of the gradient search. In comparison of the three curves: the steepest gradient descent seems to reach the convergence at the fastest rate, but not to the optimal point. The Newton method converges at a slower rate and settles at about the same global minimum as the steepest gradient method. However, the conjugate gradient method

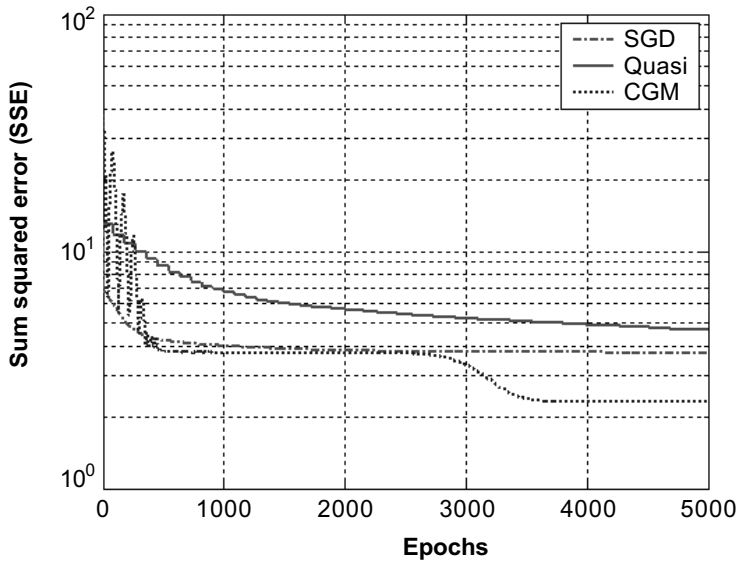

Figure 7 Convergence comparisons between the steepest gradient descent (SGD), Quasi Newton (Quasi), and conjugate gradient methods (CGM)

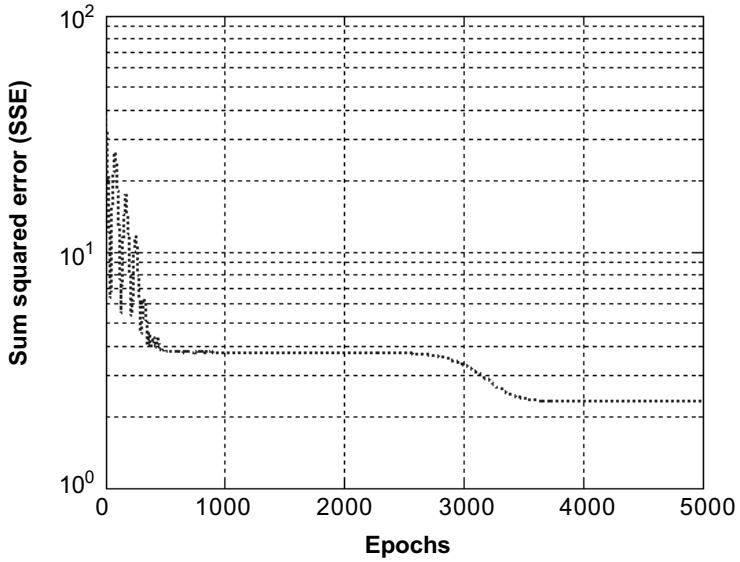

Figure 8 The fluctuations during the search for optimal global minimum using conjugate gradient algorithm

converges at the rate between the steepest gradient and Newton methods, but has the most optimal global minimum among the three methods tested.

Meanwhile, the zoomed-in view of Figure 7 in the interval between epochs 200-600 is shown in the little box in Figure 8. The fluctuations, which is the increase and decrease in sum squared error is due to the search for optimal global minimum that applies the golden section search. In this gradient search, the time interval is subdivided into smaller sections so that the optimal global minimum can be located. That is why in the golden section search, sum of the errors is fluctuating between two points in the selected interval. The sum of the error continues to increase and decrease until the optimal global minimum is obtained, at this time the sum of squared error has the minimum value when the difference of the points in the intervals reached to the set tolerance.

6.0 CONCLUSION

The frame alignment based on DTW method for pre-processing LP coefficients has been used to produce a new form of compressed data called DTW-FF coefficients. These new inputs are used as input to the BPNN as described in this paper. Having DTW-FF algorithm, frame matching is performed and the output, which is the local distance scores are then fed into BPNN. From the experiments, it was proven that DTW-FF algorithm can be used as a front-end processing of speech recognition for BPNN, although DTW itself is a back-end recognition engine. This is an alternative method found to resolve the problem of data feeding into neural network algorithm.

The DTW-FF coefficients were compared to the LPC coefficients using typical DTW algorithm to identify whether or not any loss of information has occurred. From the experiments, it has been proved that there were no changes in the recognition rate, so we conclude that there is no loss of information during the frame fixing. The frame alignment has adopted DTW to normalize the spoken word length. The normalized templates then are being used as the input to BPNN for the recognition part and they are proven been able to improve recognition performance on the samples tested to a higher percentage.

In conclusion, the proposed DTW-FF algorithm is able to produce a better way of representing input features into the NN while saving the computation cost and network complexity with a high recognition rate than the typical DTW itself. A higher recognition rate is achieved when pitch feature is added to the DTW-FF feature.

Network performance also has been tested using other descent method like the Quasi-Newton and Conjugate Gradient which used the second order information of the function that want to be optimized. These methods compared against the traditional back-propagation neural network which used the steepest gradient descent during the back-pass to update the connection weights. The observations found that the conjugate gradient algorithm reached to the optimal global minimum point.

REFERENCES

[1] Sakoe, H. and S. Chiba. 1978. Dynamic Programming Algorithm Optimization for Spoken Word Recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing*. ASSP-26(1): 43-49.

[2] Silverman, H. F. and D. P. Morgan. 1990. The Application of Dynamic Programming to Connected Speech Recognition. IEEE ASSP Magazine. 7-25.

[3] Abdulla, W. H., D. Chow, and G. Sin. 2003. Cross-Words Reference Template for DTW-based Speech Recognition System. IEEE Technology Conference (TENCON). Bangalore, India. 1: 1-4.

[4] Rabiner, L. and B. H. Juang. 1993. *Fundamentals of Speech Recognition*. Englewood Cliffs, New Jersey: Prentice Hall.

[5] Creany, M. J. 1996. Isolated Word Recognition using Reduced Connectivity Neural Networks with Non-Linear Time Alignment Methods. Ph.D. Thesis. University of New Castle-Upon-Tyne.

[6] Kuhn, M. H., H. Tomaschewski, and H. Ney. 1981. Fast Nonlinear Time Alignment for Isolated Word Recognition. Proceedings of ICASSP. 6: 736-740.

[7] Ahmadi, M., N. J. Bailey, and B. S. Hoyle. 1996. Phoneme Recognition using Speech Image (Spectrogram). 3rd International Conference on Signal Processing. 1: 675-677.

[8] Abdul Aziz, M. A. 2004. Speaker Recognition System Based on Cross Match Technique. Master Thesis. Universiti Teknologi Malaysia.

[9] Wildermoth, B. R. 2000. Text-Independent Speaker Recognition Using Source Based Features. Master of Philosophy Thesis Griffith University. Australia.

[10] Sudirman, R. and S. H. Salleh. 2005. NN Speech Recognition Utilizing Aligned DTW Local Distance Scores. Proceeding of 9th International Conference on Mechatronics Technology. Kuala Lumpur.

[11] Sudirman, R., S. H. Salleh, and S. Salleh. 2006. Local DTW Coefficients and Pitch Feature for Back-Propagation NN Digits Recognition, Proceeding of IASTED International Conference on Networks and Comunications. Thailand. 201-206.

[12] Sudirman, R., S. H. Salleh, and T. C. Ming. 2005. Pre-Processing of Input Features using LPC and Warping Process. Proceeding of 1st International Conference on Computers, Communications, and Signal Processing. Kuala Lumpur. 300-303.

[13] Jackson, P. J. B. and C. H. Shadle. 2001. Pitch-Scaled Estimation of Simultaneous Voiced and Turbulence Noise Components in Speech. *IEEE Transactions on Speech and Audio Processing*. 9(7): 713-726.
[14] Muta, H., T. Baer, K. Wagatsuma, T. Muraoka, and H. Fukada. 1988. A Pitch Synchronous Analysis of Hoarseness in Running Speech. *Journal of Acoustical Society of America*. 84(4): 1292-1301.