Faculty of Electrical Engineering
Universiti Teknologi Malaysia

VOL. 9, NO. 1, 2007, 9-13

ELEKTRIKA
http://fke.utm.my/elektrika

DTWFF-Pitch Feature and Faster Neural Network Convergence for Speech Recognition

Rubita Sudirman[1*], **Sh-Hussain Salleh**[1] and **Shaharuddin Salleh**[2]

[1]Faculty of Electrical Engineering, Universiti Teknologi Malaysia, 81310 UTM Skudai, Johor, Malaysia.
[2]Mathematics Department, Faculty of Science, Universiti Teknologi Malaysia, 81310 UTM Skudai, Johor, Malaysia

[*]Corresponding author: rubita@fke.utm.my, Tel: 607-5535738, Fax: 607-5566272

Abstract: This paper presents the pre-processing of speech templates for artificial neural network (ANN). The processed features are pitch and Linear Predictive Coefficients (LPC) for input and reference templates, based on Dynamic Time Warping (DTW) algorithm. The first task is to extract pitch features using Pitch Scale Harmonic Filter algorithm. Another task is to align the input frames (test set) to the reference template (training set) using DTW fixing frame (DTW-FF) algorithm. This is a time normalization process in which it is needed for data with unequal length. By doing time normalization, the test set and the training set are adjusted to the same number of frames. Having both pitch and LPC features fixed frames, speech recognition using neural network can be performed. A high recognition rate is obtained using combined features of DTW-FF and pitch for Malay digit words of 0-9, as high as 100% is achieved. Another task included in this paper is to find the optimal global minimum of the NN surface using the conjugate gradient algorithm to replace the steepest gradient descent in the back-propagation algorithm. Results showed that conjugate gradient algorithm is able to find a better optimal global minimum.

Keywords: Artificial neural network, Conjugate gradient method, Dynamic time warping, Frame fixing, Pitch.

1. INTRODUCTION

The emergence of DTW more than 3 decades ago has brought a brighter light for speech recognition problem. The DTW technique works by matching the unknown speech input template to a pre-define reference template. Later in late 1980s NN become popular among speech recognition researchers to utilize the method beside other methods like HMM and DTW itself. NN has been used for many different tasks in several domains and they have proved to be very efficient for learning complex input-output mappings. In the speech processing domain, NN has been used for speech recognition; many experiments were made for isolated word recognition on small vocabularies to continuous speech recognition [1].

Neural network is a method of recognition which requires the same length of training and testing data to be fed into the network especially when the data are fed into the network in multiple numbers for parallel processing. Time normalization is a typical method to interpolate input signal into a fixed size of input vector. A pre-processing method of the data frame based on DTW time normalization is used in this paper which it also applies the trace segmentation method to reduce the number of stored feature vectors for the stationary portion [2][3]. The method proposed in this paper is called the dynamic time warping frame fixing algorithm or DTW-FF, specially designed to fix the frame numbers between the selected reference (average over the samples' population) and the test speech. The idea is also based on [4] and [5] which used the DTW matching between two varies utterances. These great deals of effort are to obtain optimal recognition accuracy particularly for the vocabulary used in this experiment.

It is very delicate to determine which training algorithm will be the fastest to converge for a given problem. It will depend on many factors, including the complexity of the problem, the number of data points in the training set, the number of weights and biases in the network, the error goal, and whether the network is being used for pattern recognition (discriminant analysis) or function approximation (regression). Due to this manner, the performance optimization is required in our system identification so that an optimized performance is achieved. Several methods are available and have been used in the respective channels for the same reason.

Following this section are the description of the rest of the paper: section 2 - the features extraction and pre-processing which includes the details of DTW-FF algorithm and its results, section 3 – some result and discussion of using features extracted, section 4 – application of conjugate gradient algorithm for network optimization, and section 5 – summarize and conclude the finding of the work done.

2. FEATURE EXTRACTION AND PRE-PROCESSING

The features extraction and pre-processing are divided into two: (i) DTW-FF pre-processing and (ii) Pitch feature processing. The flow process of the entire experiment is presented in Figure 1 for clearer illustration.

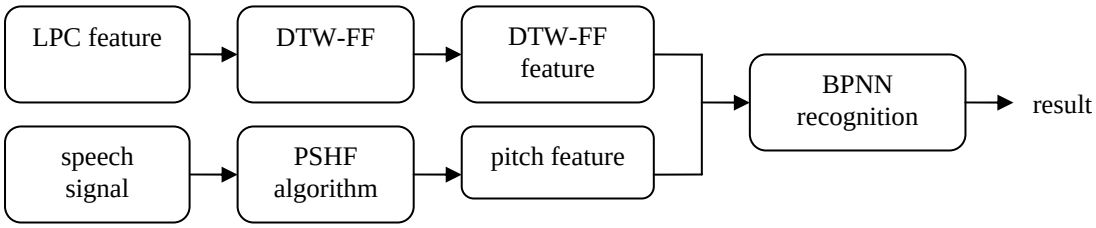

Figure 1. The flow process of the experiment

(i) DTW-FF pre-processing

Template matching is an alternative method to perform speech recognition; the drawback of the template matching method is in speaking rate variability, in which there exist timing differences between the similar utterances. Dynamic Time Warping (DTW) method was first introduced by [1], in which it was used for recognition of isolated words in association with Dynamic Programming (DP). The problem of time differences can be solved through DTW algorithm, which is by warping the reference template against the test utterance based on their features similarities. So, DTW algorithm actually is a procedure, which combines both warping and distance measurement, which is based on their local and global distance.

In this research, the time normalization is done based on DTW method by warping the input vectors with a reference vector which has almost similar local distance, while expanding vectors of an input to reference vectors which shows a vertical movement; shares same feature vectors for a feature vector frame of an unknown input.

Due to its ability to encode speech at low bit rate and its ability to provide the most accurate speech parameters, LPC coefficients are used as the feature to extract the DTW-FF feature coefficients. The method of time alignment is based mostly on dynamic time warping and part of trace segmentation approach. The frame fixing is done based on the traditional DTW method; it is performed by warping the input vectors with reference vector. If an input frame has almost similar feature vectors as the reference within a frame (a frame consists of 10 feature vectors), then they will have almost similar local distances. For this condition, vectors expansion of the input will take place, i.e, reference vectors shows a vertical movement; shares same feature vectors for a feature vector frame of an unknown input.

The frame compression (F^-) and frame expansion (F^+) are done by using the DTW frame fixing algorithm (DTW-FF). Consider the frame vectors of LPC coefficients for input as *i…I*, and reference as *j…J*, while *F* denotes the frame. The rules are based on the following slopes:

a) Slope is 0

If compression vector takes place, the input frames will be compressed and take only a copy the reference feature vector frame, in other words compression is compressing multiple similar input frames into one frame with respect to the reference [6][7]. The Frame compression takes place when the warping path moves horizontally. It is done by taking the minimum calculated local distance among the neighboring frames, i.e. compare *w(i)* with *w(i-1)*, *w(i+1)* and so on, and choose the frame with minimum local distance. In other words, the frame compression involves searching minimum local distance out of distances in a frame set within a threshold value, it is represented as

$$F^- = F(min\{d_{(i,j)\ldots(I,J)}\}) \quad (1)$$

b) Slope is ∞

The frame of the speech signal is expanded when the warping path moves vertically. At this instance, the reference frame is expanded according to frame *w(i)* of the input source. In other words, the reference frame duplicates the local distance of that particular vertical warping frame, it is represented as

$$F^+ = F(w(i)) \quad (2)$$

c) *Slope is 1*

When the warping path moves diagonally, the frame is left as it is because it already has the least local distance compared to other movements.

The DTW-FF algorithm can be summarized as follows:

If slope=0,
Then F- = F(min{$d_{(i,j)\ldots(I,J)}$*})*
Else
if slope =∞,
Then F+= F(w(i))
Else
if slope=1,
Then $x_i = y_j$
End

The normalized data/sample has being tested and compared to the typical DTW algorithm and results showed a same global distance score as reported in [7] and [8].

(ii) Pitch feature processing

Pitch is a feature considered in this experiment although it is one of speech features that is rarely taken into consideration when doing speech recognition. In this research, pitch is optimized and will be used as a feature into NN along with LPC feature. Pitch contains spectral information of a particular speech, in which it is the feature that is used to determine the fundamental frequency, F0. The pitch extraction is done to sample speech in .wav format to obtain the initial values of fundamental frequencies, or referred as F_o^{raw}; F_o^{raw} 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 to yield an optimized pitch, F_o^{opt} [8].

3. RESULTS AND DISCUSSION

Experiment A - Traditional DTW vs. BPNN with DTW-FF Feature

The purpose of *Experiment A* is to find the recognition rate when DTW-FF is fed into traditional DTW and also into the NN. Results of the experiment are illustrated in Figure 2 which used samples digits of 0-9 uttered by 11 subjects for 5 times each word. The momentum rate and learning rate set in this experiment throughout are α is 0.9 and η is 0.1. It is clearly shown that the BPNN using the DTW-FF coefficients outperformed the traditional DTW for all subjects. Traditional DTW gives an average recognition of 90% while BPNN is 97%, however the average improvement is about 6.45% per subject. In earlier experiments reported in [8], traditional DTW showed same recognition performance when using both features, either LPC or DTW-FF features. One way or the other this implies that DTW-FF is a valid feature to be used as an input feature for speech pattern recognition. Indeed the used of DTW-FF feature in BPNN has outperformed the traditional DTW. The results are collected from an average of 20 hidden nodes NN where most of the networks have learned sufficiently.

Experiment B - DTW-FF and Pitch Feature into BPNN

In the NN experiment of DTW-FF combined with the pitch feature (*Experiment B*), the same network setting is use so that it will produce a fair result when compared to *Experiment A*. The same experimental setup as *Experiment A* is used for *Experiment B* except this time, *Experiment B* used only samples by the last 6 subjects. Observation to this experiment found that a faster network convergence is achieved when the network has learned sufficiently using only 10 hidden nodes, compared to 20 hidden nodes in *Experiment A*, refer to Figure 3. This proved that pitch feature is an attractive feature if it is used along with other feature namely the DTW-FF feature to produce a higher recognition and faster convergence. During the experiment, some of the subjects start to show drastic improvement as early as 5 hidden nodes. These have proven that better recognition can be achieved when taking pitch feature into account particularly in isolated digits speech recognition. This method also has been tested on a number of words obtained from TIMIT database. However, the result is not very encouraging: only around 65-70% accuracy, this might be due to a speaking variation, intonation and dialect that have been used by the speakers during the recordings of the words in the sentences.

The statistical test, called as T-Test has been conducted to the data in Figure 3 in which this test assesses whether the means of two groups are statistically different from each other. The hypothesis is set such that: H_0: $\mu_{before}=\mu_{after}$ and H_1: $\mu_{before}<\mu_{after}$. From the test with a level of significance of α=0.05, it is found that the value of t for DTW-FF in traditional DTW is smaller than the value of t in BPNN. In that case, the results reject the null hypothesis which states that $\mu_{before}=\mu_{after}$. Since H_1 is true where $\mu_{before}<\mu_{after}$, then it can be concluded that by using DTW-FF coefficients into traditional DTW and BPNN the recognition is significantly improved.

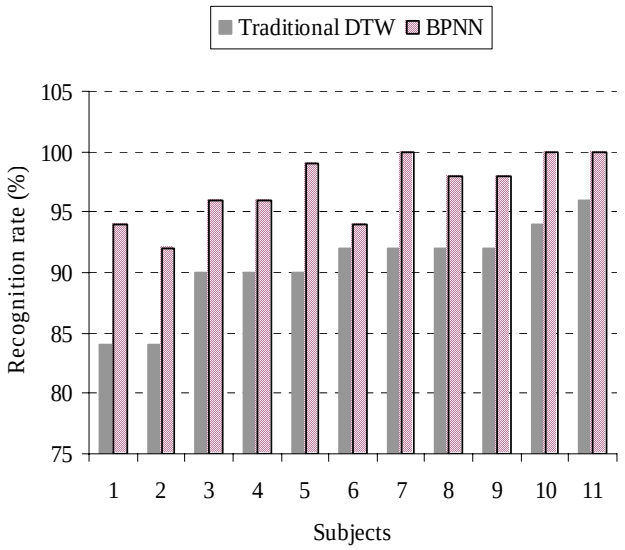

Figure 2. Comparison of the recognition rate between the typical DTW and BPNN when DTW-FF feature coefficients are used.

In addition, a lot of network complexity and amount of connection weights computations during forward and backward pass have been reduced due to replacement of LPC coefficients with DTW-FF coefficients. Besides fixing to equal number of frames between the unknown input and the reference, this activity have also tremendously reduced the amount of inputs presented into the back-propagation neural networks. For example, the input size reduction for 50 samples of 49 frames with LPC order-10 is 90% when using the local distance scores instead of the LPC coefficients. Nevertheless, this percentage will be higher if higher LPC order was used. For 12-order LPC the reduction is about 92%. This means a simpler calculation for connection updates in the NN thus giving faster convergence for the same sample under testing.

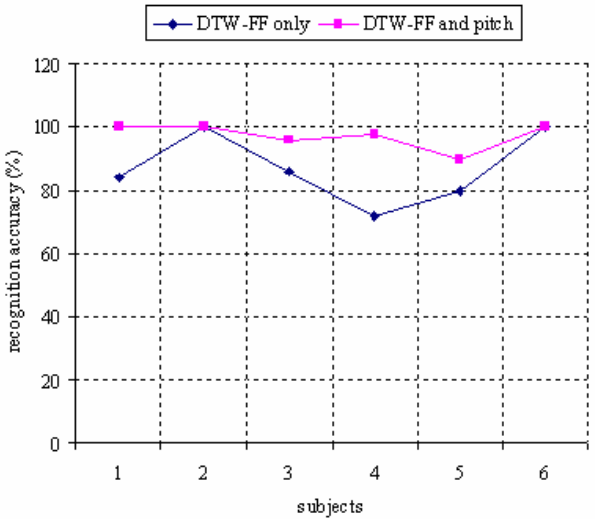

Figure 3. The recognition comparison between using DTW-FF coefficient only and its combination with pitch feature.

4. BACK-PROPAGATION NN OPTIMIZATION

Back-propagation method has shown to be a successful technique in learning the neural networks. However, several deficiencies still need to be resolved, especially the problem of local minima get trapped in the valley, particularly in the non linear problem like speech. When the error is trapped in the local minima, back-propagation may lead to failure in finding the global optimum solution. Another disadvantage of back-propagation is that the convergence rate is still slow although small learning error desired is achieved. The slow convergence of back-propagation is mainly due to the derivative of the activation function, in which it leads to premature saturation of the network output units. This happens when the derivative is approaching the extreme values of the sigmoid function, which is either 1 or 0, the derivative becomes extremely small and the back-propagation error may vanish [9]. In that case, the output can be wrong because it does not produce a large error signal.

The learning process and connections weight adjustment is slow particularly if the process involved a large input size. Back-propagation usually went through thousands of iterations to leave the flat spot. Several methods can be used to improve the convergence rate and make comparison between the methods tested for our particular problem. The back-propagation neural network experiments carried out up to this point was utilizing the steepest gradient method. The recognition rate achieved is acceptable with their high percentage, sometimes reached to 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 conjugate gradient algorithm (CGM).

Even for a large number of weights, conjugate gradient algorithms are usually much faster than the back-propagation, and the memory required by this algorithm is proportional to the number of weights. The back-propagation NN algorithm adjusts the weights in the steepest descent direction of the gradient. The performance function is decreasing rapidly in this direction. However, even if the function decreases most rapidly along the negative gradient, this does not guarantee to produce the fastest convergence. The Conjugate Gradient Method (CGM) by Fletcher and Reeves only requires a slight modification of the discrete-time steepest descent method but often enable to improve the convergence rate dramatically. In CGM, the global search is performed along the conjugate direction, which in turn this algorithm generally produces faster convergence compared to the steepest descent.

The conjugate gradient algorithm starts with the search in the steepest descent direction (negative of the gradient) in the first iteration. The first search direction is denoted as P_o, where

$$P_o = -g_o \tag{3}$$

The simplest form of CGM algorithm is

$$x^{(k+1)} = x^{(k)} + \eta^{(k)} P_k \tag{4}$$

in which it represent the line search to determine the optimal distance to move along the current search direction,

where $\eta_k = arg \min_{\eta \geq 0} E\left(x^{(k)} + \eta P_k\right)$

and P_k is the new search direction. It is combination of steepest descent and the previous search direction.

Thus, $$P_k = -\nabla E\left(x^{(k)}\right) + \beta_k P_{k-1} \tag{5}$$

where $\beta_k = \dfrac{\left\|\nabla E\left(x^{(k)}\right)\right\|_2^2}{\left\|\nabla E\left(x^{(k-1)}\right)\right\|_2^2}$.

The Results

In Figure 4, the curves tell 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 (BPM) seems to reach the convergence at the fastest rate, but not to the optimal point. However, the CGM converged at the slower rate but better smaller error which determines its optimal global minimum between the two methods tested. The result suggested that for a large number of weights like in this experiment, the CGM is more efficient than the BPM search methods for an optimal global minimum.

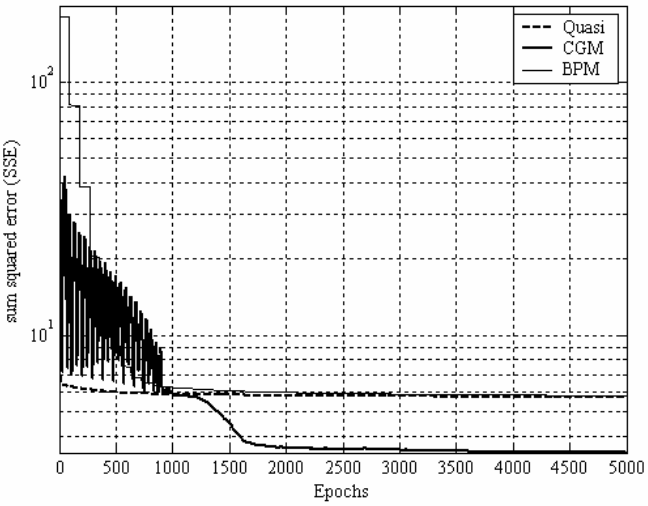

Figure 4. The convergence comparisons between the steepest gradient descent (BPM), Quasi Newton (Quasi), and conjugate gradient method (CGM).

Figure 5 shows the oscillation during the search for optimum global minimum using the CGM. The fluctuations, which is the increase and decrease in sum squared error is due to the search of the 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.

5. CONCLUSION

This paper has described the frame fixing of speech signal based on DTW method for processing LP coefficients into another form of compressed data called DTW-FF coefficients, this coefficients then are used as input into BPNN, in which BPNN is the back-end speech pattern recognition engine.

Initial observation from the experiment conducted leads to a resolution that the DTW-FF algorithm is able to produce a better way of representing input features into the neural networks. These have been proven that the reformulation of the LPC feature into DTW-FF feature coefficients do not affect the recognition performance even though the coefficients size is reduced by 90% for an order 10 of LPC. As a consequence, the computation cost and network complexity have been greatly reduced, but still gain a high recognition rate than the traditional DTW itself. Therefore, this is a new approach of feature representation and combination that can be used into the back-propagation neural networks.

A higher recognition rate is achieved when pitch feature is added to the DTW-FF feature. It can be concluded that even though pitch itself cannot provide a good recognition, eventually it can be an added feature to another very reliable feature to form a very good recognition.

Performance optimization showed that the recognition does not produce any higher percentage except that the convergence rate is faster. This is due to the less iteration in finding the global optimum; this was done by line search technique and followed by the golden section search which only focused on the global point vicinity based on the interval defined from the line search process.

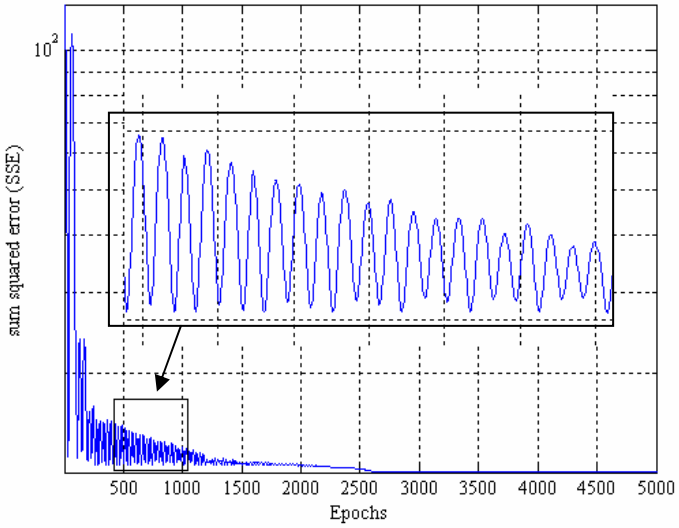

Figure 5. The oscillation in the golden section during the gradient search of optimal global minimum using the conjugate gradient method.

ACKNOWLEDGMENT

The authors would like to thank Universiti Teknologi Malaysia for funding the research study.

REFERENCES

[1] M. Sima, V. Croitoru, and D. Burileanu, "Performance Analysis on Speech Recognition using Neural Networks," *Proceeding of the Device and Applications System Conference*, Romania, pp. 359-266, May 1998.

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

[3] B. R. Wildermoth. *Text-Independent Speaker Recognition using Source Based Features*. Griffith University, Australia: Master of Philosophy Thesis, 2001.

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

[5] M. H. Kuhn, H. Tomaschewski, and H. Ney, "Fast nonlinear Time Alignment for Isolated Word Recognition," *Proceedings of ICASSP.*, pp. 736-740, April 1981.

[6] R. Sudirman and S. H. Salleh, "NN Speech Recognition Utilizing Aligned DTW Local Distance Scores," 9th *International Conference on Mechatronics Technology*, Kuala Lumpur, 5-8 Dec. 2005.

[7] R. Sudirman, S. H. Salleh and S. Salleh, "Local DTW Coefficients and Pitch Feature for Back-Propagation NN Digits Recognition," *IASTED International Conference on Networks and Communications Systems*, Chiang Mai, Thailand, 29-31 March 2006.

[8] R. Sudirman, S. H. Salleh, P. I. Khalid, and A. H. Ahmad, "Time Normalization of LPC Feature Using Warping Method, *Jurnal Elektrika*, Vol 7, no. 2, December 2005.

[9] S. C. Ng, C. C. Cheung, S. H. Leung and A. Luk, "Fast Convergence for Back-Propagation Network with Magnified Gradient Function," *Proceedings of the International Joint Conference on Neural Networks*. pp.1903-1908, 2003.