Hybrid Method for Digits Recognition using Fixed-Frame Scores and Derived Pitch

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

[1]Center for Biomedical Engineering, Faculty of Electrical Engineering
[2]Mathematics Department, Faculty of Science
Universiti Teknologi Malaysia, 81310 UTM Skudai, Johore, Malaysia

Abstract— **This paper presents a procedure of frame normalization based on the traditional dynamic time warping (DTW) using the LPC coefficients. The redefined method is called as the DTW frame-fixing method (DTW-FF), it works by normalizing the word frames of the input against the reference frames. The enthusiasm to this study is due to neural network limitation that entails a fix number of input nodes for when processing multiple inputs in parallel. Due to this problem, this research is initiated to reduce the amount of computation and complexity in a neural network by reducing the number of inputs into the network. In this study, dynamic warping process is used, in which local distance scores of the warping path are fixed and collected so that their scores are of equal number of frames. Also studied in this paper is the consideration of pitch as a contributing feature to the speech recognition. Results showed a good performance and improvement when using pitch along with DTW-FF feature. The convergence rate between using the steepest gradient descent is also compared to another method namely conjugate gradient method. Convergence rate is also improved when conjugate gradient method is introduced in the back-propagation algorithm.**

Keywords— **dynamic warping, pitch coefficients, back-propagation neural , conjugate gradient**

I. Introduction

NN has attracted researchers' interest for speech recognition since more than half century ago and the phenomena still growing to improve either ASV or ASR system. Minute avenues are explored because there are possibilities that the system can be refined for more accurate tuning. Areas like dynamic warping has been among the popular methods to accomplish the ASR or ASV system tuning in which they were first introduced by [1].

In this paper, the motivation to the study is initiated after looking at the massiveness of input data presented into the NN and the computation complexities especially when parallel processing is intended. A common avenue that researcher looked into is the speech timing between samples [3][4][5]. Realizing the importance of samples timing relationships, an alignment/normalization method using DTW is investigated and feature vectors manipulations are performed to suit the back-end proposed recognition engine. In our study, back-propagation neural network algorithm is used. The aim is to simplify the input which previously was using the LPC coefficients and produce a faster convergence to the network. By doing this, a new form of input is derived and used into our NN speech recognition system.

Traditionally automatic speech recognition used derived features which represent the vocal tract system characteristics, and leaving the knowledge of voice source characteristics, namely as pitch because pitch is not an ideal source of information for automatic speech recognition [2]. Pitch contains a lot of information such as information about the speaker, it can tell whether the sound is a voiced or unvoiced, as well as it contains prosodic information [6] [7]. In our study, we are considering pitch as another input feature into the NN so that a supra-segmental feature of the vocal tract can be included.

The remainder of this paper is organized according to the flow of the experiments conducted which is arranged as follows: Section II describes the approach and methods used in the study, Section III presents of the results and discussion of the study, while section IV summarizes and concludes the findings of the study.

II. Approach And Methods

A 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. However, it is a typical method to interpolate input signal into a fixed size of input vector which is intended in this study. In this particular study, a hybrid method involves the combination of DTW and NN back-propagation algorithm, utilized DTW to normalize all input patterns with respect to the template pattern of digits utterances for NN recognition.

According to the warping path type 1, three slope conditions are set to perform the compression and expansion to the speech frames [8][9][10]: (i) horizontal slope, (ii) vertical slope, and (iii) diagonal slope. The frame compression and expansion processing used a modified

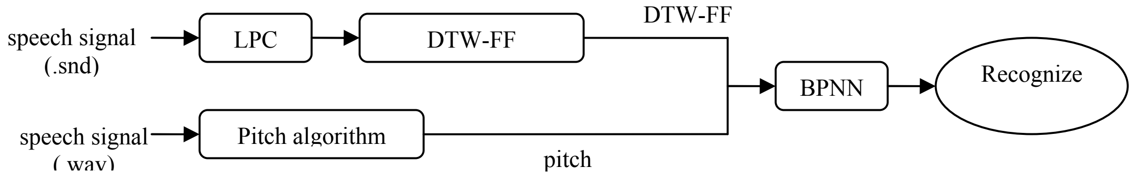

Fig. 1 The experiments process flow

version of DTW matching technique which is renamed as DTW frame fixing (DTW-FF) algorithm. The DTW-FF is utilized to fix the input frames to a fix number of input frames: the source frames are aligned to the template frames. During the frame fixing, the source and template frames are adjusted so that they have the same number of frames. In addition to that, we retained and used the local distance scores (which is called as the DTW-FF coefficient) of the fixed frames as inputs into the MLP neural network instead of using the global distance score which were used by many researchers [3][8][9].

The speech recognition is performed using the back-propagation neural network (BPNN) algorithm to enhance the recognition performance and their results are compared between using the DTW with LPC coefficients to BPNN with DTW-FF coefficients.

The acoustical feature generated caused by the vibration of the vocal fold in the vocal tract, namely pitch is introduced as another input feature into the NN. This is because LPC feature vectors itself sometimes does not give an overall high percent of recognition, pitch feature itself does not give high recognition rate indeed.

The pitch feature is optimized using pitch-scaled harmonic filter algorithm to reduce glitches during the voice activity. The overall approach of the study is illustrated in Fig. 1 which also portrays the flow diagram of the recognition process.

The result for BPNN with DTW-FF plus pitch feature achieved its high recognition rate faster than the combination of BPNN and DTW-FF feature only.

A. The DTW-FF Algorithm - Feature Extraction

The method of time alignment is based mostly on dynamic time warping and part of trace segmentation approach. The method is called the DTW-FF algorithm in which this is a part of feature extraction. In this research, the time normalization is done based on DTW method by warping the input vectors with reference pattern vectors represented by LPC coefficients.

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. 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. The rules are based on the following slopes [10][11]:

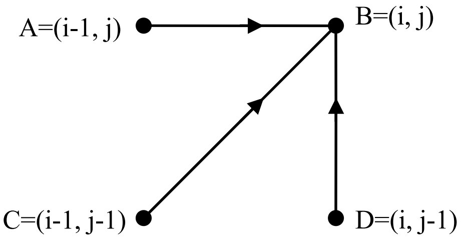

Fig. 2 The warping path type 1

Referring to Fig. 2,

- Slope is 0 (horizontal line)

  If the warping path moves horizontally from one frame to another (example: from point A to B or from point C to D), the frames of the speech signal are compressed. The compression process takes place by taking the minimum calculated local distance amongst the distance set in the frames involved: compare *w(i)* with *w(i-1), w(i+1)* and so on, and choose the frame with minimum local distance.

  Consider the frame vectors of LPC coefficients for input as *i…I*, and reference as *j…J*, while *F* denotes the frame. Frame compression involves searching minimum local distance out of distances in a frame set within a threshold value, it is represented as

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

- Slope is ∞ (vertical line)
  If the warping path moves vertically from one frame to another (example: from point C to A or from point D to B), 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.

  Frame expansion involves duplicating a particular input frame to multiple reference frames of *w(i)*, represented as

$$F^{+} = F(w(i)) \qquad (2)$$

- Slope is 1 (diagonal)
  When the warping path moves diagonally (from point C to B), the frame is left as it is because it already has the least local distance compared to other movements.

The normalized sample has being tested and compared to the traditional DTW algorithm and results showed a same global distance score [10]. Also, from the frame fixing experiment, the fixed frames, N_{ff} is calculated as

$$N_{ff} = N_{if} - N_{cf} + N_{ef} \qquad (3)$$

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

After collection of DTW-FF coefficients, the second feature accounted in this study which is pitch, is extracted [10][11] and introduced into the NN along with the DTW-FF coefficient. This is because pitch feature itself cannot give a good representation of speech signal when used for speech recognition.

B. *Experimental Setup*

In this paper, the experiments are conducted using 11 subjects. Each subject uttered digits 0-9 for five sessions, each digit is uttered fives times in each session giving a total of 50 utterances in each session. The network is tested using different number of hidden nodes with constant momentum rate, α=0.9 and learning rate, η = 0.1 in which these parameters are determined from experiment carried out to the same data. The experiments are described as follows along with their respective results and discussions.

III. Results And Discussion

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

The purpose of this experiment (*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 Fig. 2. 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 [10] and [11], traditional DTW showed same recognition performance when using both features, either LPC or DTW-FF features. One way or the other this proved that no information loss occurred during feature interpolations from LPC to DTW-FF [10][11]. 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.

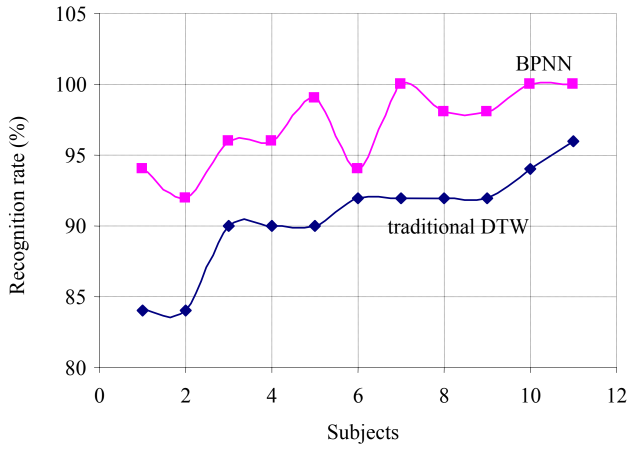

Fig. 3 Comparison of typical DTW and BPNN when using the DTW-FF feature.

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* is using only 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 Fig. 4. 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 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 Fig. 3 in which this test assesses weather the means of two groups are statistically different from each other. The hypothesis is set such that: H_0: μ_{before}=μ_{after} and H_1: μ_{before}<μ_{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 in BPNN. In that case, the results reject the null hypothesis which states that μ_{before}=μ_{after}. Since H_1 is true where μ_{before}<μ_{after}, then it can be concluded that by using DTW-FF coefficients into traditional DTW and BPNN the recognition is significantly improved.

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. The percentage of number coefficients reduced is calculated as follows:

$$\%\ \text{coefficien ts reduced} = \frac{\text{Input}_{\text{LPC}} - \text{Input}_{\text{LD}}}{\text{Input}_{\text{LPC}}} \text{x}100\% \quad (4)$$

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.

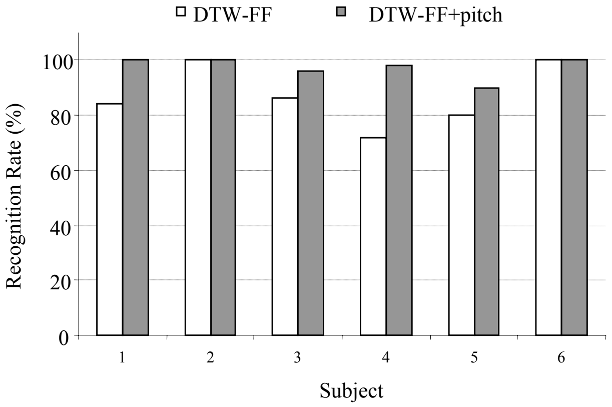

Fig. 4 Before and after pitch addition for 10 hidden nodes

C. Convergence Test

The back-propagation neural network experiments are utilizing the steepest gradient method. The recognition rate achieved is acceptable with their high percentage, sometimes reached to 100%. However, we are looking for a faster convergence time, so the data are tested using other search engine for the back-propagation part, namely the conjugate gradient method. The forward pass mechanism is the same for all architecture except for the backward-pass, so the backward pass is replaced with the conjugate gradient algorithm (CG) [12]. The results using this algorithm are compared to the results using the steepest gradient algorithm obtained in the previous experiments.

In Fig. 5, the curves tell how the search for optimal global minimum behaved for each type of the gradient search. In comparison, the steepest gradient descent (SG) seems to reach the convergence at a faster rate, but not to the optimal value. However, the CG converged at the slower rate but smaller error which determines its optimal global minimum between the methods tested. The result suggested that for a large number of weights like in this experiment, the conjugate gradient is the more efficient compared to other gradient search methods for an optimal global minimum. The oscillation in CG during the early stage shows the search of optimal global minimum in the golden section interval.

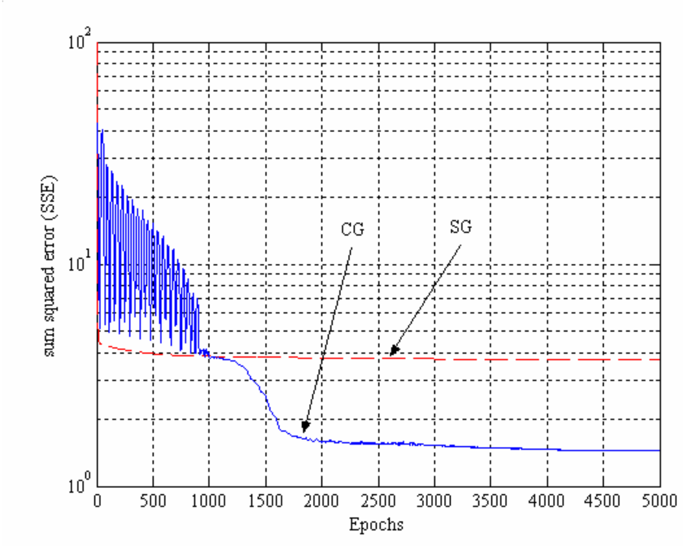

Fig. 5 The convergence comparisons between the steepest gradient descent (SG) and conjugate gradient method (CG).

IV. Summary And Conclusion

In this paper, the frame fixing of speech signal based on DTW method for processing LP coefficients into another form of compressed data called DTW-FF coefficients have been described. These coefficients are used as input into BPNN. 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 coefficients do not affect the recognition accuracy although the coefficients size is reduced by 90% from using an order 10 of LPC to using the DTW-FF coefficients. As a result, the computation and network complexity have been greatly reduced indeed still gain a high recognition rate than the traditional DTW. 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 higher percentage except that network converged to a better optimal global minimum after an extra of 500 epochs for these particular samples. This is due to the 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.

Acknowledgements

An appreciation goes to MOSTI and Universiti Teknologi Malaysia, Skudai for supporting this study.

References

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

[2] M. Magimai-Doss M (2003). Using Pitch Frequency Information in Speech Recognition. Proceedings of 8th European on Speech Communication and Technology. Geneva, Switzerland. 4: 2525-2528.

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

[4] Creany M J (1996). Isolated Word Recognition using Reduced Connectivity Neural Networks with Non-Linear Time Alignment Methods. PhD Thesis, University of New Castle-Upon-Tyne, UK.

[5] Uma S, Sridhar, V, and Krishna G (1992). Time-Normalization Techniques for Speaker-Independent Isolated Word Recognition. Proceedings of Pattern Recognition Conference: Image, Speech and Signal Analysis. 3: 537-540.

[6] Prasanna S R M, Zachariah J M, and Yegnanarayana B (2004). Neural Network Models for Combining Evidence from Spectral and Suprasegmental Features for Text-Dependent Speaker Verification. Proceedings of International Conference on Intelligent, Sensing, and Information Processing. pp 359-363.

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

[8] Botros N M and Premnath S (1992). Speech Recognition using Dynamic Neural Networks. International Joint Conference in Neural Network. 4: 737-742.

[9] Soens P and Verhelst W (2005). Split Time Warping for Improved Automatic Time Synchronization of Speech. Proceeding of SPS DARTS, Antwerp, Belgium.

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

[11] Sudirman R, Salleh S-H and Salleh S (2006). Local DTW Coefficients and Pitch Feature for Back-Propagation NN Digits Recognition. IASTED International Conference on Networks and Communications, Chiang Mai, Thailand. pp 201-206.

[12] Hagan M T, Demuth H B, and Beale M (1996). *Neural Network Design*. Boston: PWS Publishing Company.