The Effectiveness of DTW-FF Coefficients and Pitch Feature in NN Speech Recognition

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

[1] *Center for Biomedical Engineering, Faculty of Electrical Engineering*
Universiti Teknologi Malaysia, 81310 Skudai, Johore, Malaysia
Tel.: 607-5535738, Fax: 607-5535681, email: rubita@fke.utm.my, hussain@fke.utm.my

[2] *Mathematics Department, Faculty of Science*
Universiti Teknologi Malaysia, 81310 Skudai, Johore, Malaysia
Tel.: 607-5537835, email: ss@mel.fs.utm.my

Abstract

This paper presents a method to extract speech features contained in the dynamic time warping path which originally was derived from linear predictive coding (LPC). For the purpose of recognition, the extracted feature will represent the inputs into neural network back-propagation. The new method presented here is how the feature is extracted and those coefficients are normalized against the template pattern according to the selected average number of frames over the samples collected. The idea behind this method is due to neural network (NN) limitation where a fixed amount of input nodes are needed for every input class especially in the application of multiple inputs. Thus, the main objective of this research is to find an alternative method to reduce the amount of computation and complexity in a neural network, in this case is for speech recognition. One way to achieve this is by reducing the number of inputs into the network. This is done through dynamic warping process in which local distance scores of the warping path will be utilized instead of the global distance scores. From the literature review, past and most current research are using the global distance score or LPC coefficients as input to the neural network. LPC certainly presented into the network with a large amount of coefficients in each speech frame.

Keywords:

Dynamic Time Warping, Normalization, Linear Predictive Coding, Pitch Feature, Back-Propagation Neural Network

Introduction

There have been so many intensive research of speech recognition on improving the ASR system. However there are still avenues that can be explored so that a better or a faster method of recognition can be formulated. For this purpose, time alignment/normalization method using DTW is investigated and feature vectors manipulations are performed to suit the back-end proposed recognition engine, which is back-propagation neural network algorithm. The aim is to simplify the input forms 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. In our study, we are considering pitch as another input feature into the NN so that a suprasegmental feature can be included. The pitch is extracted using a method called the pitch scaled harmonic filter.

In short, this paper is to study the effect of using combination of DTW-FF and pitch feature when they are used in conjunction of the back-propagation neural network. The remainder of this paper is organized as follows: Section 2 describes the approach and methods used in the study, Section 3 consists of the experimental setup followed by section 4 of the results and discussion and finally section 5 which concludes the experiments findings.

Approach and Methods

Time normalization is a typical method to interpolate input signal into a fixed size of input vector. 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. In this particular research work, combination of DTW/NN back-propagation algorithm utilized DTW to normalize all input patterns with respect to the template pattern.

Three slope conditions are set to perform the compression and expansion of the speech frames: (i) horizontal, (ii) vertical, and (iii) diagonal slopes. The new feature processing used our modified algorithm of the traditional Dynamic Time Warping (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, whereby the frames of LPC feature vectors are aligned between the source frames to the template frames according to their allocated frames

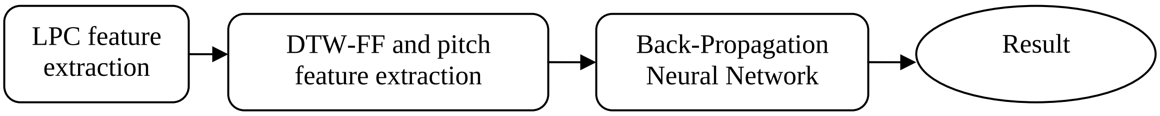

LPC feature extraction
DTW-FF and pitch feature extraction
Back-Propagation Neural Network
Result

Figure 1 - The Process Flow of the Experiments

By doing this 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 of the fixed frames as inputs into the MLP neural network. 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 neural network. 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. This pitch feature firstly is optimized using pitch-scaled harmonic filter algorithm to reduce glitches during the voice activity. 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.

The DTW-FF Algorithm

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 this research, the time normalization 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, ie, 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 [10][11][13].

The frame compression (F^-) and frame expansion (F^+) are done by using our new so called 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:

i) ***Slope is 0***

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

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

ii) ***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.

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

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

iii) ***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 is 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. An example of result of DTW-FF algorithm is shown in Fig. 2, it revealed that the DTW-FF has been able to fix the frame of the input to the number according to the reference template.

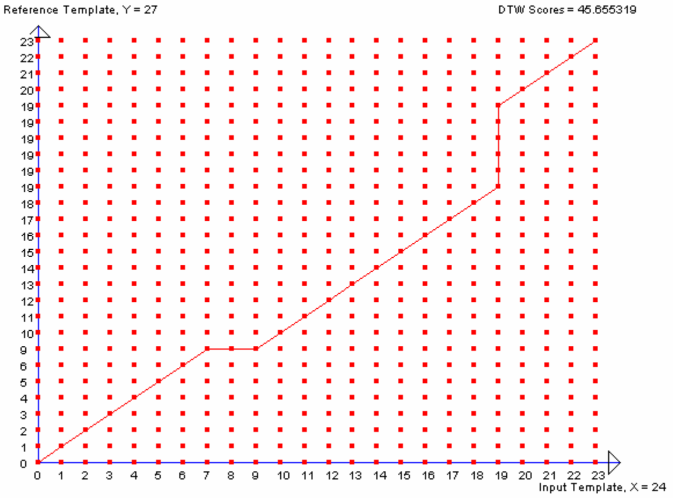


Figure 2 - Result from DTW-FF Algorithm: Input Template Frame Equals Reference Template Frame

Initially the input frame has 24 frames, but the reference template has 27 frames. The DTW-FF algorithm has fixed the input frame to 27 frames through the expansion and compression procedures. From Fig. 2, it seems that there is one compression process among frame 7, 8, and 9 of the input, and also one expansion process of frame 19 of the input into 6 frames at the reference. From the experiment, the number of the fixed frames, N_{ff} can be formulated 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

The Pitch Feature Algorithm

The pitch feature is extracted using a pitch-scaled harmonic filter algorithm (PHF). Alternatively, the pitch can also be extracted using pitch extraction module contained in Speech Filing System (SFS) which is a shareware program developed for research purposes. However, in PHF algorithm, the pitch is optimized so that it will resolve any octave errors if they persist.

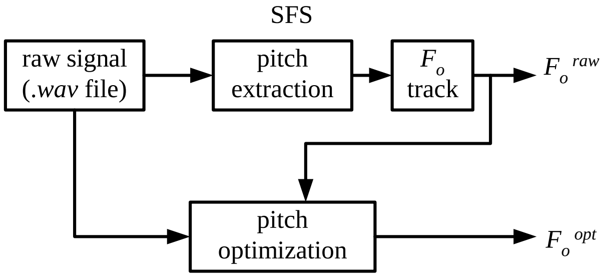


Figure 3 - Flow Diagram of Pitch Feature Extraction Using the PHF Algorithm

According to the PHF diagram in Fig. 3, the speech in *.wav* is used to obtain the initial values of fundamental frequencies, *F0* 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 contain in the PHF and yield to an optimized pitch, F_o^{opt}. This F_o^{opt} is used as another input feature which will be added to the DTW-FF feature extracted earlier.

Experiment Setup

In this paper, the experiments are conducted using digits 0-9 spoken in Malay by 6 subjects, each subject uttered every digit 5 times each session and they make five recording sessions giving a total of 50 utterances in each session for each digit. The network is tested using different number of hidden nodes with constant momentum rate, α=0.9 and learning rate, η = 0.1; these rates are determined experimentally using the same data set. Experiment 1 is to find the recognition rate when DTW-FF is fed into typical DTW and also into the NN.

In the NN experiment of DTW-FF combined with the pitch feature (experiment 2), the same network setting is use so that it will produce a fair result when compared to preceding experiment. The pitch feature is extracted and introduced into the NN along with the DTW-FF feature because pitch feature itself cannot give a good representation of speech signal.

Results and Discussion

Typical DTW versus BPNN with DTW-FF Feature

Result of experiment 1 is illustrated in Fig. 4 - a very clear illustration in Fig. 2 proved that BPNN outperformed typical DTW when both are presented with the DTW-FF coefficients. The results are collected from an average of 20 hidden nodes NN when most of the networks have sufficiently learned.

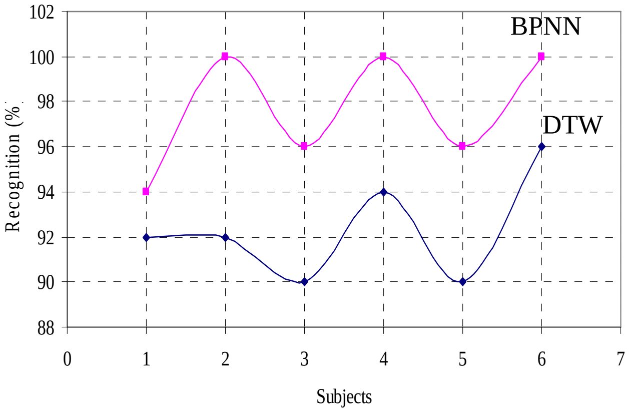
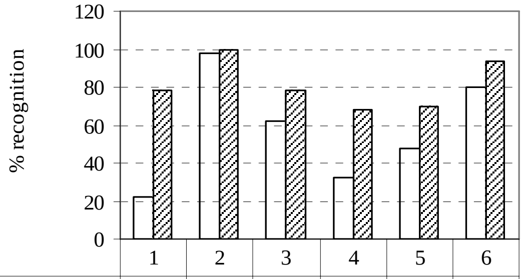


Figure 4- Comparison between Using Typical DTW and Back-Propagation Neural Networks when Both are Fed with DTW-FF Coefficients

DTW-FF and Pitch Feature into BPNN

However in experiment 2, the NN has learned sufficiently and most of the subjects are using only 10 hidden nodes, compared to 20 in experiment 1. This certainly has proven that the pitch feature is an attractive feature to be added to the DTW-FF feature to produce a higher recognition and faster convergence. Even some of the subjects start to show drastic improvement as early when using 5 hidden nodes. Look at the difference that is clearly illustrated in Fig. 3 and Fig. 4. 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; this might due to a speaking variation, intonation and dialect that have been used by the speakers during the recordings of the sentences (uncontrolled situations).



	1	2	3	4	5	6
□ DTW-FF	22	98	62	32	48	80
▨ DTW-FF+pitch	78	100	78	68	70	94

Figure 5 - Before and after addition of pitch feature for 5 hidden nodes

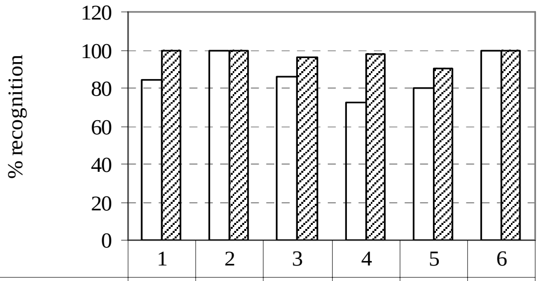
The statistical test, called as T-Test has been conducted to the data collected for Fig. 2. This test assesses weather the means of two groups are statistically different from each other so that it can be decided whether or not any significant difference has been obtained.

The T-Test done by finding the value of *t*:

$$t = \frac{\overline{X}_T - \overline{X}_C}{\sqrt{\frac{\sigma_T}{n_T} + \frac{\sigma_C}{n_C}}} \qquad (4)$$

where subscripts *T* and *C* represent the groups of data and *n* is the number of data in the group.

The hypothesis is set such that: H_0: $\mu_{before}=\mu_{after}$ and H_1: $\mu_{before<}\mu_{after}$. From the test, it is found that the value of *t* for DTW-FF in typical DTW is -2.571 and in BPNN is -3.125 with a significance level of α=0.05. The *t* value implies that $\mu_{before<}\mu_{after}$, therefore the results reject the null hypothesis which states that $\mu_{before}=\mu_{after}$. On the other hand, H_1 is true ($\mu_{before<}\mu_{after}$), thus it can be concluded that by using DTW-FF coefficients into typical DTW and BPNN the recognition is significantly improved. On average, the improvement from using DTW-FF in typical DTW compared to using BPNN is increased by 5.34 % for this particular set of experiment using 50 utterances by each subject.



	1	2	3	4	5	6
□ DTW-FF	84	100	86	72	80	100
▨ DTW-FF+pitch	100	100	95.9	98	90	100

Figure 6 - Before and After Addition of Pitch Feature for 10 Hidden Nodes

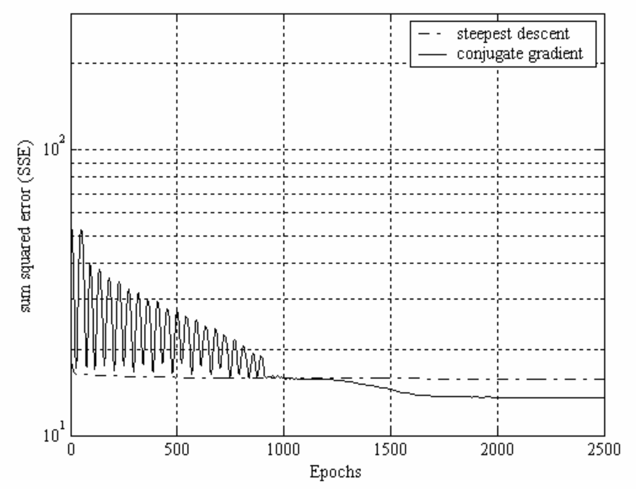


Figure 7 - Convergence comparison between using the steepest gradient descent and the conjugate gradient algorithm

A comparison between using the steepest gradient descent and the conjugate gradient algorithm is also performed. The result is shown in Fig. 7. From the figure, the steepest gradient method converged faster but not at an optimal global minima, however the conjugate gradient algorithm converged at a better global minima compared to the steepest gradient descent.

Summary

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.

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. By using DTW-FF algorithm, frame fixing or also known as frame matching is performed and the outputs are the matched local distance scores.

From the experiments conducted, it was found that DTW-FF algorithm is an alternative method found to pre-process the LPC data before these reprocessed data are fed into neural network algorithm or other subsequent pattern matching method. The traditional DTW algorithm is tested with the LPC coefficients and the DTW-FF coefficients to compare if there was any loss of information occurred. Fortunately, it is proven that there were no changes in the recognition rate, so it can be concluded that there is no information loss during the frame fixing process.

The combination of DTW-FF coefficients and pitch feature have also shown better recognition rate. Therefore, this is a new approach of feature representation and combination that can be used into the back-propagation neural network. In conclusion, the DTW-FF algorithm has successfully been developed and derived new form of feature for the neural network speech recognition. These has also saved the computation cost and network complexity than using the typical DTW. A higher recognition rate is achieved when pitch feature is added to the DTW-FF feature. When lesser coefficients presented into the neural network, shorter computation time is also achieved.

From the convergence comparison between the steepest gradient descent and the conjugate gradient descent, it is suggested that the conjugate gradient algorithm method should being adopted into the NN speech pattern recognition for Malay words database in the future. It is good step towards a better convergence (settling at an optimal global minimum) since there is no work can be found about using Malay words database in NN using the conjugate gradient algorithm.

Acknowledgements

The authors would like to thank the Ministry of Science and Innovation of Malaysia (MOSTI) and the Universiti Teknologi Malaysia, Skudai for funding and supporting this research study.

References

[1] 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.

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

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

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

[5] S. R. M. Prasanna, J. M. Zachariah, and B. Yegnanarayana. 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*. 359-363.

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

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

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

[9] Sudirman, R. Salleh, S.H., 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.