

TIME NORMALIZATION OF LPC FEATURE USING WARPING METHOD

Rubita Sudirman, Sh. Hussain Salleh, Puspa Inayat Khalid, Abd Hamid Ahmad

Biomedical Engineering Research Group

Fakulti Kejuruteraan Elektrik,

Universiti Teknologi Malaysia

81310 Skudai, Johor, Malaysia

email:rubita@fke.utm.my

ABSTRACT

This paper presents pre-processing of input features to artificial neural network (NN). This is for preparation of reliable reference templates for the set of words to be recognized. 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 (PSHF) algorithm [12]. Another task is to align the input frames (test set) to the reference template (training set) using DTW fixing frame (DTW-FF) algorithm. This proper time normalization is needed since NN is designed to compare data of the same length whilst same speech can varies in their length. By doing frame fixing (time normalization), the test set and the training set is adjusted to the same number of frames. Having both pitch and LPC features fixed frames, speech recognition using neural network can be performed.

1. INTRODUCTION

DTW has been one of the prime speech recognition methods since its birth more than 30 years ago, first introduced by [1]. It works by matching the unknown speech input template to a pre-define reference template, and this method is an easiest speech recognition method compared to others like HMM (surfaced in mid 1980s) or NN (in late 1980s; NN itself was first introduced in the 1950s). DTW has being more prominent with its ability to search the best path between two time-series signals [3], furthermore it is a cost minimization matching technique, in which a test speech signal is expanded or compressed according to a reference template [2].

Using neural network as a recognition tool requires the same length of training and testing data to be fed into the network. Time normalization is a typical method to interpolate input signal into a fixed size of input vector. Thus, a pre-processing method of the data frame based on DTW time normalization is proposed in this paper. The proposed pre-processing method also applies the trace segmentation method, in which the initial idea of trace segmentation is to reduce the number of stored feature vectors for the stationary portion [13].

In this paper, firstly the feature extraction methods are described. Secondly is the DTW alignment algorithm and then followed by an experimental results and some discussion. The winding up of the experiment is presented in the conclusion section.

2. SIGNAL PITCH OPTIMIZATION

Pitch information is one of speech features that rarely taken into consideration while doing speech recognition. In this research, pitch is optimized and will be used as a feature into NN along with LPC feature that will be discussed in the next section. Pitch contains spectral information of a particular speech, in which it is the feature that being use to determine the fundamental frequency, F_0 .

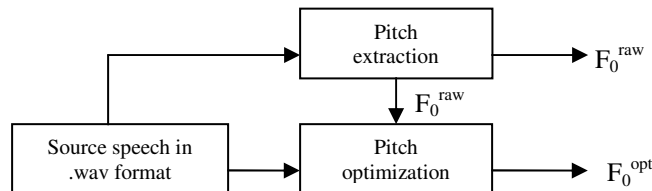


Figure 1: Process flow of pitch optimization

Figure 1 shows a flow diagram of the pitch optimization process. In short, firstly pitch extraction is done to sample speech in .wav format to obtain the initial values of fundamental frequencies, or referred as F_0^{raw} ; F_0^{raw} can be obtained by pitch-tracking manually or by using available speech-related applications. Then this F_0^{raw} is fed into the pitch optimization algorithm (described in the following paragraphs) and yield to an optimized pitch, F_0^{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 described in detail in [10] and they referred their work to previously done by [11]. The optimization is done by taking 4 pitch periods of the signal in each window, accounting for the first 8 harmonics of the pitch optimization and the window offset is 4ms; using the pitch-estimation algorithm. The pitch tracking algorithm is to estimate the pitch period τ by sharpening the spectrum at the first H harmonics, $h \in \{1, 2, 3, \dots, H\}$.

The lower and higher spectral spreads, S_h^+ and S_h^- described the sharpness of the spectrum. Their spectral equations are:

$$S_h^+(m, p) = |S_w(4h+1)|^2 - \frac{|S_w(4h)|^2}{|W(h\Delta f_0)|^2} \left| W\left(h\Delta f_0 - \frac{1}{M}\right) \right|^2 \quad (1)$$

$$S_h^-(m, p) = |S_w(4h-1)|^2 - \frac{|S_w(4h)|^2}{|W(h\Delta f_0)|^2} \left| W\left(h\Delta f_0 + \frac{1}{M}\right) \right|^2 \quad (2)$$

where $\Delta f_0 = \frac{1}{\Delta \tau} = \frac{4f_s}{\Delta M}$, M is the window length ($M(p) = 4\tau(p)$), and f_s is the sampling frequency.

The Hanning window used:

$$w(k) = \frac{M}{2} \left(\text{sinc } \pi k M + \frac{\text{sinc } \pi(kM-1) + \text{sinc } \pi(kM+1)}{2} \right) e^{-j\pi \Delta_0 M} \quad (3)$$

The algorithm find the optimum pitch value for a particular time by minimizing the difference between the calculated and the measured smearing¹ of the spectrum due to the window. The difference is calculated by the minimum mean-squared error, according to the cost function for window length, M

$$J(M, p) = \sum_{h=1}^H [S_h^+(M, p)^2 + S_h^-(M, p)^2] \quad (4)$$

This cost function is used to match the pitch of the decomposed signals and optimization is done throughout the signal by repeating the process with an increment time p . The optimized pitch is compared to other available method such as Speech Filing System (SFS) to ensure its reliability before they are ready to be fed into NN. The sampling frequency used in this processing is 16 kHz. The result of pitch optimization shows a very good estimation when they differ only by ± 1 Hz (refer to Figure 2).

3. LPC FEATURE EXTRACTION

There are many feature extraction methods for speech signals like MFCC, LPC, and 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. Features are represented in vectors form of a chosen dimension, which is called as the LP order.

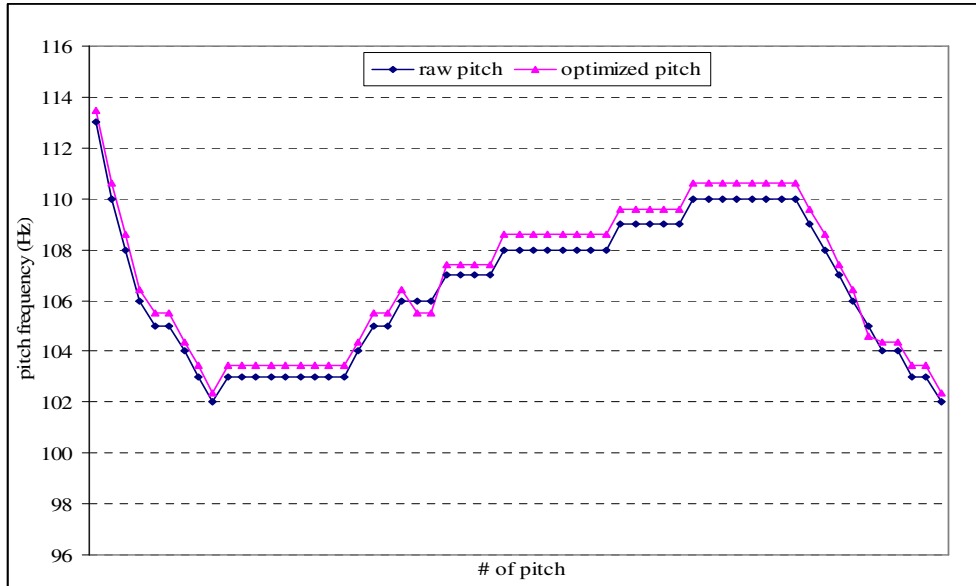


Figure 2: Plot to show original (raw) and optimized pitch of a zoomed-in part of utterance [aɔa]; very small pitch differences are spotted between the extracted pitches.

¹ Smearing and leakage occurs when the frequency of the n th fourier coefficient not exactly aligned with one of the discrete frequency bins, as a consequence errors in bias form are produced.

Pre-emphasis

The front end process the speech signal using Linear Predictive Coding to obtain the coefficients, which represent its feature. The first step to the process is to pre-emphasize the signal so that the signal is spectrally flatten,

$$s(n) = s(n) - 0.95s(n-1) \quad (5)$$

Frame Blocking

Having the signal from pre-emphasis process, then frame blocking is applied to that signal; it is blocked into N equal frames. The start of each frame is offset from the start of the previous frame by L samples, meanwhile the start of the second frame begins at L. The third frame blocking would begin at 2L and so on. But, if $L \leq N$, then adjoining frames will overlap and the linear predictive (LP) spectral estimates will show a high correlation.

$$x_i(n) = \hat{s}(Li + N) \quad (6)$$

where $n = 0, 1, 2, \dots, N-1$ and $i = 0, 1, 2, \dots, I-1$.

Windowing

After pre-emphasis and frame blocking, the signal is windowed using Hamming window function, where N is the window length.

$$w(n) = 0.54 - 0.46 \cos\left(\frac{2\pi n}{N}\right), \text{ for } 0 \leq n \leq N-1 \quad (7)$$

Autocorrelation

The windowed signal then go through autocorrelation process, represented in Equation (8) and p is the order of LPC analysis.

$$R(m) = \sum_{n=0}^{N-1-m} x(n)x(n+m), \text{ for } m = 0, 1, 2, \dots, p \quad (8)$$

LP Coefficients Computation

The common LPC analysis is using Durbin's recursive algorithm, which is based on Equations (9)-(13).

$$E(0) = R(0) \quad (9)$$

$$k_i = \frac{R(i) - \sum_{j=1}^{i-1} a_j^{(i-1)} R(i-j)}{E_{i-1}}, \text{ for } 1 \leq i \leq p \quad (10)$$

$$a_i^i = k_i \quad (11)$$

$$a_j^i = a_j^{i-1} + k_i a_{i-j}^{i-1}, \text{ for } 1 \leq j \leq i-1 \quad (12)$$

$$E_i = (1 - k_i^2) E_{i-1} \quad (13)$$

These equations are solve recursively for $i = 0, 1, \dots, p$, where p is the order of the LPC analysis. Then, the final solution is when $i = p$, which is $a_j^p = a_j^p$, for $1 \leq j \leq p$.

4. DTW FRAME ALIGNMENT

Template matching is an alternative to perform speech recognition; the template matching encountered problems due to 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.

After feature extraction process, speech pattern can be represented by a feature vector sequence. For the sake of example, let consider two feature vectors T and R, let T be the unknown/test speech pattern and R the speech reference template pattern.

$$\begin{aligned} T &= t_1, t_2, t_3, \dots, t_i, \dots, t_I \\ R &= r_1, r_2, r_3, \dots, r_j, \dots, r_J \end{aligned} \quad (14)$$

Translating sequence T and R into Figure 3, the warping function at each point is calculated. Calculation is done based on Euclidean distance measure as a mean of recognition rate, means the lowest distance between the test utterance and the reference templates will has the best match. For each point, the distance called as local distance, d is calculated by taking the difference between two feature-vectors t_i and r_j :

$$d(i, j) = \|r_j - t_i\| \quad (14)$$

Every frame in a template and test speech pattern must be used in the matching path. Considering DTW type 1, if a point (i,j) is taken, in which i refers to the test pattern axis (x-axis), while j refers to the template pattern axis (y-axis), a new path must continue from previous point with a lowest distance path, which is from point (i-1, j-1), (i-1, j), or (i, j-1).

Given a reference template (training) with feature vector R and an input (test) pattern with feature vector T , each has of N_T and N_R frames, the DTW is able to find a function $j=w(i)$, which maps the time axis i of T with the time axis j of R . The search is done frame by frame through T to find the best frame in R , by making comparison of their distances. After the warping function is applied to T , distance $d(i,j)$ becomes

$$d(i, j(i)) = \|t_j' - r_i\| \quad (15)$$

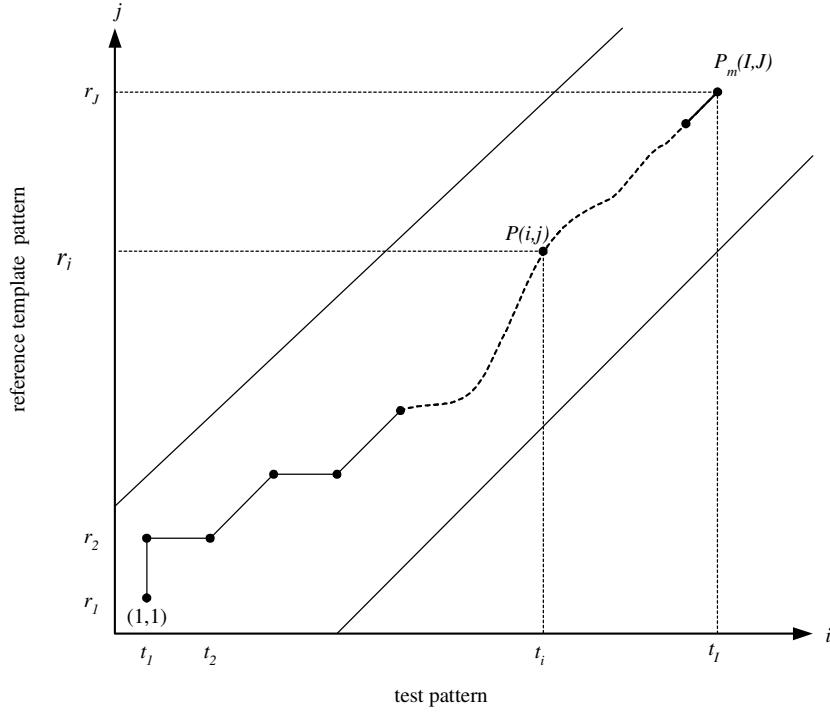


Figure 3: Fundamental of warping function

Then, distances of the vectors are summed on the warping function. The weighted summation, E is:

$$E(F) = \sum_{i=1}^I d(i, j(i)) * w(i) \quad (16)$$

where $w(i)$ is a nonnegative weighting coefficient. The minimum value of E will be reached when the warping function optimally align the two pattern vectors. However, the minimum residual distance between T and R is the distance that still remains after the time difference between them is minimized. Thus, the time-normalized difference is defined as:

$$D(A, B) = \underset{F}{\text{Min}} \left[\frac{\sum_{i=1}^I d(i, j(i)) * w(i)}{\sum_{i=1}^I w(i)} \right] \quad (17)$$

A few restrictions have to be applied to the warping function to ensure close approximation of properties of actual time axis variations. This is to preserve essential features of the speech pattern. [4] outlined the warping properties, also found in [6], [8], and [9]. The warping function slope is more rigidly restricted by increasing slope, M , but if it is too severe then time normalization is not effective, so a denominator to time normalized distance, N is introduced:

$$N = \sum_{i=1}^I w(i) \tag{18}$$

However N independent of the warping function. So, the time normalized distance becomes

$$D(A,B) = \frac{1}{N} \underset{F}{Min} \left[\frac{\sum_{i=1}^I d(i, j(i)) * w(i)}{\sum_{i=1}^I w(i)} \right] \tag{19}$$

Having this time normalized distance, minimization can be achieved by DP principles.

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. This frame alignment is also known as the expansion and compression method [2], this done following the slope conditions as described follows.

There are three slope conditions that have to be dealt with in this research work, based on the DTW type 1:

- i- Slope is ~ 0 (horizontal line) - The frames of the speech signal are compressed.: This is done by taking the minimum local distance amongst the distance set, i.e.: compare $w(i)$ with $w(i-1)$ and choose the frame with minimum local distance
- ii- Slope is $\sim \infty$ (vertical line) - The frame of the speech signal is expanded, i.e.: the reference frame gets the identical frame as $w(i)$ of the unknown input.
- iii- Slope is ~ 1 (diagonal) - The frame is left as it is because it has the least local distance compared to other movements.

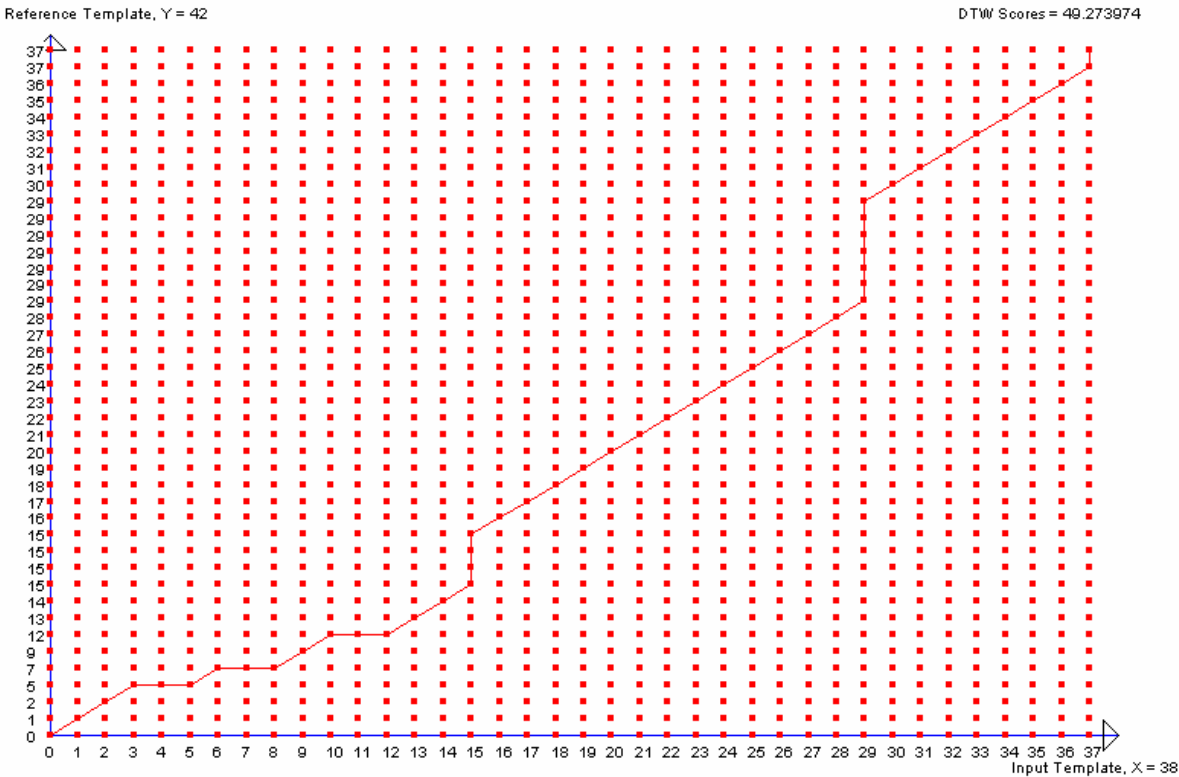


Figure 4: The DTW frame alignment between an input and a reference template; the frame number 0-38 on reference template axis actually contains 42 frames of reference template itself.

This frame adjustment is done by using DTW frame fixing algorithm (DTW-FF), after this procedure the data are ready to be used for neural network recognition. The normalized data/sample has being tested and compared to the typical DTW algorithm and results showed a same global distance score. Further findings are discussed at the results and discussion section.

5. RESULTS AND DISCUSSION

Figure 4 shows an input frame has been matched to a reference template of same utterance. In this example, initially the input template has 38 frames and reference template 42 frames. Then by using the DTW-FF algorithm the input frames have been expanded to 42 , i.e. equals to the number of frames for reference template.

According to slope condition (i), the local distances of unknown input frame $w(3), \dots, w(5)$ are compared and $w(5)$ appears to have the minimum local distance among the three frames, so those 3 frames is compressed to one occupies only $r(4)$. Same goes to $w(6), \dots, w(8)$ which $w(7)$ has the least local distance with respect to the reference, so it is compressed an occupies only $r(5)$. The distance is calculated using Equation (17). On the other hand, slope condition (ii) shows an expansion, for example while $w(15)$ of input are expanded to 4 frames, in which these 4 consecutive frames of the reference template are identical; 4 frames of reference template at $r(10), \dots, r(13)$ have the same feature vectors as frame $w(15)$ of the input vectors, so $w(15)$ occupies $r(10), \dots, r(13)$. These means that frame $w(15)$ of the input has matched 4 feature vectors in a row from 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.

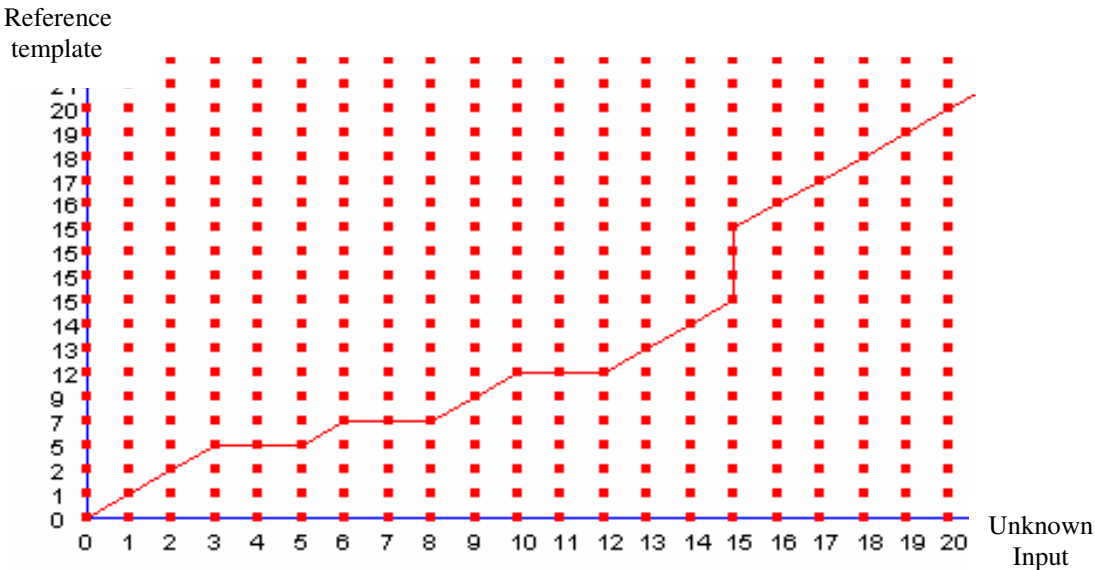


Figure 5: A zoomed-in of Figure (4) to show compression and expansion of template frames; input template frame $w(3), \dots, w(5)$ being compared and select $w(5)$ of reference template due to its minimum local distance score, while $w(15)$ of input are expanded to 4 frames and occupies $r(10), \dots, r(13)$ of reference template.

Having done the expansion and compression along the matching path, the unknown input frame is matched to the reference template frames. Thus, frame fixing/matching is a mean of solution to speech frame variations, however it still preserved the global distance score; the DTW fixing frame (DTW-FF) algorithm only make adjustment on the feature vectors of the horizontal and vertical local distance movements, leaving the diagonal movements as it is with their respective reference vectors. The frame fixing is done throughout the samples, also taking considerations the sample which has the same number of frames as the averaged frames for the reference template.

As for the recognition comparison between the typical DTW and DTW-FF algorithm, the results (refer to Table 1) shows the same recognition rate due to the same pattern matching between the template frames. Moreover, the global distance score is preserved, this makes a stronger argument that the recognition before and after DTW-FF is identical. The tokens used are digits 0-9 recorded in Malay language for 6 sessions and the tokens are spoken five times each session. The recognition accuracy can be increased by increasing number of subjects; increase size of database.

Table 1: Percent recognition between typical DTW and DTW-FF algorithm

Subject	DTW (%)	DTW-FF (%)
1	92	92
2	92	92
3	90	90
4	84	84
5	84	84

6. CONCLUSION

The time alignment based on DTW method for pre-processing the pitch and LP coefficients is described in this paper. The result from experiments conducted shown that the DTW-FF algorithm can perform frame matching between an input and a reference speech frame as good as typical DTW algorithm. From this obtained result, further use of the fixed frame speech along with pitch feature can be applied to neural network speech recognition.

In conclusion, the DTW-FF algorithm can be used as a front-end processing of speech recognition for NN, 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 or other subsequent pattern matching using the well known DP method.

REFERENCES

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

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

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

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

[5] H. K. Sze. *The Design and Development of An Educational Software on Automatic Speech Recognition*. Masters Thesis, Universiti Teknologi Malaysia, 2004.

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

[7] M. Ahmadi, N.J. Bailey, and B.S. Hoyle. Phoneme Recognition using Speech Image (Spectrogram). *3rd International Conference on Signal Processing*. Vol 1: 675-677. 14-18 Oct 1996.

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

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

[10] P. J. B Jackson and C. H. Shadle. Pitch-Scaled Estimation of Simultaneous Voiced and Turbulence Noise Components in Speech. *IEEE Transactions on Speech and Audio Processing*. 9(7): 713-726, 2001.

[11] H. Muta, T. Baer, K. Wagatsuma, T. Muraoka, and H. Fukada. A Pitch Synchronous Analysis of Hoarseness in Running Speech. *Journal of Acoustical Society of America*. 84(4): 1292-1301, 1988.

[12] P. J. B. Jackson and D. Mareno. PSHF Beta Version 3. 10, CVSSP – University of Surrey, Guilford, UK. <http://www.ee.surrey.ac.uk/Personal/P.Jackson>, 2003.

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

[14] Sudirman, R., Salleh, S.H., and Ming, T. C. (2005). Pre-Processing of Input Features using LPC and Warping Process. *International Conference on Computers, Communications, and Signal Processing*, 14-16 Nov., Kuala Lumpur.

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