NN Speech Recognition Utilizing Aligned DTW Local Distance Scores

Rubita Sudirman, Sh-Hussain Salleh, Ting Chee Ming
Center of Biomedical Engineering
Fakulti Kejuruteraan Elektrik
Universiti Teknologi Malaysia
81310 Skudai, Johor
MALAYSIA
email: rubita@fke.utm.my

Abstract

This paper presents the neural network (NN) speech recognition using processed LPC input features. But NN has a limitation that the network must have a fixed amount of input nodes. The input feature processing method will use frame matching based on Dynamic Time Warping (DTW) algorithm to fix the input size to a fix amount of input vectors. The LPC features are aligned between the input frames (test set) to the reference (training set) using our 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 or also known as time normalization, the test set and the training set frames are adjusted so that both sets will have the same number of frames according to the reference set. The neural network with back-propagation algorithm is used as the recognition engine at the back-end processing to enhance the recognition performance. The results compare DTW with LPC coefficients to back-propagation NN with LPC coefficients adjusted using DTW.

1. Introduction

A hybrid approach combines conventional time normalization methods with a highly competitive back-propagation neural network [5]. NN has been known in speech recognition since late 1980s (NN itself was first introduced in the 1950s). Other methods are HMM (surfaced in mid 1980s) and DTW being the most popular due to its ability to search the best path between two time-series signals [3, 6], furthermore it is a cost minimization matching technique, in which a test speech signal is expanded or compressed according to the reference feature vectors[2].

Most of all, research in NN has change its focus to producing a more precise recognition with less network complexity but with fast processing. In order to achieve a high precision, a back-propagation NN algorithm is used with varies hidden nodes and respective training times. Using neural network as a recognition tool requires feeding of training and testing data of the same length into the network.

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, if required [4, 7]. Since then, it has been the basic method for compression and expansion of speech pattern vector.

The pre-processing method also applies trace segmentation method, in which the idea of trace segmentation is to reduce the number of stored feature vectors for the stationary portion during the speech [9]. Trace segmentation was not fully used as the normalization technique because of its bad past performance in speech recognition, cannot even provide the same performance as DTW even though they share a common compression technique, but not the expansion technique. Furthermore, the distance segmentation is inappropriate and as well as the spatial sampling rate along the trace [9], in addition to that it can only perform frame reduction during the stationary speech portion. DTW is a non-linear time normalization technique that can perform both frame expansion and reduction, and still can preserve important features during the process [11]. Thus, DTW time normalization is used to obtain uniform speech length, later NN is used to perform the recognition.

Other works in speech recognition have employ combination of MLP/HMM, DTW/MLP, HMM/MLP [12]. In DTW/MLP, the DTW is used just to time-aligned the input pattern and use MLP as the recognizer. [13] use combination DTW and sequential NN. In this particular research work, combination of DTW/NN with back-propagation algorithm utilized DTW to normalize all input patterns with respect to the template pattern and then perform frame matching between them. In addition to that, their local distance scores are used as input into the back-propagation neural network.

In this paper, a glance of feature extraction method is presented, frame normalization technique are described next, and then followed by NN recognition experiment using a new set of data derived from the new algorithm during frame normalization. Some results to compare typical DTW and the new derived features, its discussion is also included. The conclusion section concludes the finding of this particular research and near future plan to improve current work.

2. Feature Extraction and Frame Matching

LPC coefficients will be used as the feature vectors in the DTW frame matching stage. LPC feature extraction method is described in many earlier works, also found in [4], so only a flow diagram in Fig. 1 is shown here to flash back the process of obtaining the LPC coefficients.

After LPC feature extraction, then the coefficients went through frame normalization process by using DTW-FF algorithm [2]. The algorithm is designed to match the unknown samples with the reference sample. The reference sample is selected by their average frame numbers over the samples. Matching or we also called it as frame fixing, comprised of frames compression and expansion techniques. Other advantage that we are keen of this technique is the reduction of input numbers; this eventually will reduce the amount of network complexity and weight computations, but increase the convergence speed.

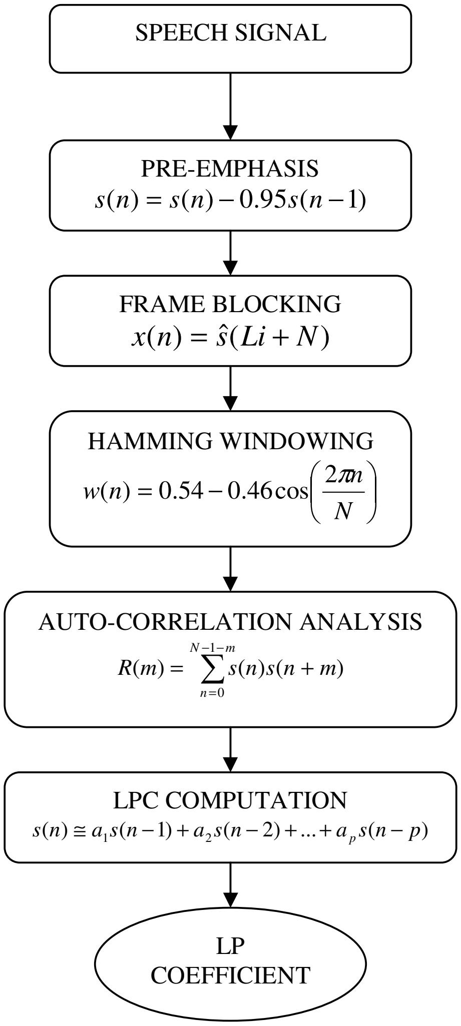
SPEECH SIGNAL
PRE-EMPHASIS
$s(n)=s(n)-0.95s(n-1)$
FRAME BLOCKING
$x(n)=\hat{s}(Li+N)$
HAMMING WINDOWING
$w(n)=0.54-0.46\cos\left(\frac{2\pi n}{N}\right)$
AUTO-CORRELATION ANALYSIS
$R(m)=\sum_{n=0}^{N-1-m}s(n)s(n+m)$
LPC COMPUTATION
$s(n)\cong a_1s(n-1)+a_2s(n-2)+...+a_ps(n-p)$
LP
COEFFICIENT

Fig. 1 Flow diagram of LPC processing

The rules of the DTW-FF algorithm are based on DTW path type 1. The heuristic DTW path type 1 is shown in Fig. 2, and the slope conditions are:

(i) **Slope is 0 (horizontal movement)**

The speech signal frames are compressed: This is done by taking the minimum local distance amongst the feature vectors. In this experiment, there are 10 feature vectors in each set. If w(i) represent the current frame; the frame is chosen by comparing the local distances among the feature vectors in the set, i.e., compare w(i) with w(i-1) and choose the frame with minimum local distance.

The distance for this slope condition is calculated as

$$D(i,j)=D(i-1,j)+d(i,j) \qquad \text{Eqn. 1}$$

(ii) **Slope is ∞ (vertical movement)**

The frame of speech signal is expanded, i.e., movement from (i,j-1) to (i,j); the reference frame gets the identical frame as w(i) of unknown input, which is at (i,j-1) and (i,j) and so on. In other words, reference will have multiple identical frames of unknown input of a particular frame.

The distance is for vertical slope movement is

$$D(i,j)=D(i,j-1)+d(i,j) \qquad \text{Eqn. 2}$$

(iii) **Slope is 1 (diagonal movement)**

The reference frame has the same frame as the frame of unknown utterance. So, there is nothing done to this path because diagonal movement always has the least distance to move from (i-1,j-1) to (i,j), compared to the other two movements.

The distance for a diagonal movement is

$$D(i,j)=D(i-1,j-1)+d(i,j) \qquad \text{Eqn. 3}$$

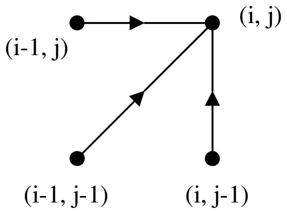

Fig. 2 DTW path type 1

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, the distance is calculated as

$$d(x,y)=\sqrt{\sum_{j=1}^{p}(x_i-y_j)^2} \qquad \text{Eqn. 4}$$

These compression (denoted as $\boldsymbol{F}^{-}$) and expansion (denoted as $\boldsymbol{F}^{+}$) are done by using new 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. Frame compression involves searching minimum local distance out of distances in a frame set within a threshold value, it is represented as

$$F^{-} = \mathrm{F}(\min\{d_{(i,j)\ldots(I,J)}\}) \qquad \text{Eqn. 5}$$

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

$$F^{+} = \mathrm{F}(\mathrm{w(i)}) \qquad \text{Eqn. 6}$$

If the uppermost warping path coordinate of a reference pattern was (M,N), then the fixed frame number is equal to M. It should be understood at this point that DTW is not used as the recognition engine during the DTW-FF algorithm. It is used for frame normalization purpose and then the retained output from the algorithm is used as input into the NN.

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 reference template.

3. NN with Back-Propagation

Back Propagation Neural Networks (BPNN) are one of the most common neural network structures, as they are simple and effective, and have been used widely in assortment of machine learning applications. The term back-propagation refers to the manner in which the gradient is computed for multilayer networks. Properly trained back-propagation networks likely to give reasonable answers when presented with inputs that they never seen before.

The NN use in this research work utilizes back-propagation algorithm so that minimum error between the training and test set can be achieved. In this research, back propagation algorithm is used due to its known ability to minimize errors in their connection weights, especially during the back-pass in the algorithm.

Log sigmoid activation function,

$$f(\theta) = \frac{1}{1+e^{-\theta}} \qquad \text{Eqn. 7}$$

is applied to activate the connection weights and readjust those weights during the iterations. Mean square error method is used to compute the weights adjustments. Method of steepest descent is employed in the direction search for fast convergence of the algorithm.

To summarize, Fig. 3 shows the flow process of the experiment for back-end recognition that has being carried out upon obtaining the LPC coefficients. The data used are 10-order LPC coefficients, using 10ms frames of Hamming windows. An average of 49 frames is selected for the reference and this number is used against the unknown input during the frame fixing process.

After the frame fixing process which happens in DTW-FF algorithm, local distances of the fixed frames are collected and the data now are ready to be used for NN recognition stage. The normalized data/sample has being tested and compared to the typical DTW algorithm and results in Table 1 showed the same global distance score. This enhanced recognition is a new approach used in this particular research to improve the recognition after a slight low percentage for typical DTW method alone (described in Table 1 column 2). Further results and findings are described in Section 5.

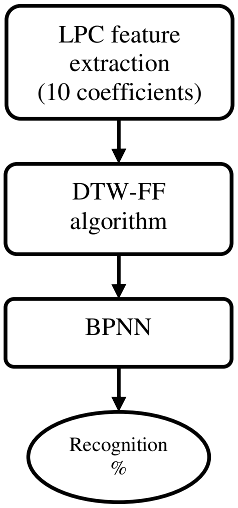

Fig. 3 Flow diagram of DTW-FF with BPNN

4. Experimental Setup

The experiment was conducted using learning rate, η=0.1, momentum rate, α=0.9 with 10 hidden layers and 49 input nodes. There are 5 subjects tested which each uttered digit 0-9, 5 times each digit, in Malay. Since there is no fix formula to determine the learning rate, momentum rate, and hidden layer, experiments were conducted and after several trials optimum numbers for the back-propagation NN algorithm in our particular samples are obtained. Experiments agreed to 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, η=0.1 and momentum rate, α=0.9 for their best recognition, see Fig. 4.

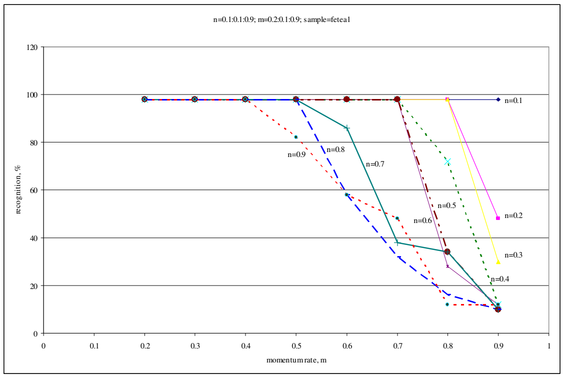

Fig. 4 Recognition vs. momentum rate for learning rate between 0.1-0.9

5. Results and Discussion

An earlier experiment compared the recognition between typical DTW and DTW-FF algorithm using LPC coefficients, the results in Table 1 showed the same recognition percentage. This might due to same feature vectors pattern matching between frames template in both algorithms. So, either the LPC feature vectors are fixed or not, do not affect the recognition rate. This makes a strong argument that the recognition before and after DTW-FF is identical and no loss of information during the DTW-FF algorithm occurred.

Table 1: Recognition percentage using typical DTW with LPC coefficients and DTW-FF coefficients input

	Recognition Rate (%)	
Subject	**LPC**	**DTW-FF**
1	92	92
2	92	92
3	90	90
4	84	84
5	84	84

For the second experiment, the local distance scores are preserved from DTW-FF algorithm. These scores are fed into the NN back-propagation algorithm and the results are shown in Table 2. Improvement can clearly be observed across the subjects using combination of methods as an extended recognition tool, recognition as high as 99% is successfully identified. This would signify that over a small size Malay digits vocabulary of 0-9 that have being uttered 5 times by each subject, the DTW-FF is able to produce relevant form of input data in much smaller number for the BPNN, compared to the amount of using LPC coefficients itself. Remember also, the number of inputs to the BPNN has been reduced about 90% by using the local distance scores instead of LPC coefficients. These means a lot of network complexity and amount of connection weights computations during forward and back pass have been reduced, thus faster convergence is achieved. The following calculations showed the number of input to the BPNN that have been reduced.

Input using local distance score,

Input_{LD} = 50 utterances x 49 frames
= 2450 inputs

Input using LPC coefficients,

Input_{LPC} = 50 utterances x 49 frames/utterance
x 10 coefficient/frame
= 24500 input coefficients

$$\%\ \text{reduced} = \frac{24{,}500 - 2450}{24{,}500} x100\% = 90\ \%$$

Table 2: Recognition between typical DTW and back-propagation NN fed with local distance scores

	Recognition Rate (%)	
Subject	**DTW**	**BPNN**
1	92	98
2	92	98
3	90	99
4	84	94
5	84	92
average	88.4	96.2

Graphically, the improvement can be seen in Fig. 5, from the figure it is clearly shown that NN has improved the recognition performance to a very high percentage with an amount of improvement is greater than 5% and the improvement is almost uniform throughout the samples with Subject 3 being the most remarkable with an increase of 9%. Achievement in high recognition result indicates that the fixed local distance scores can replace LPC coefficients as the input to NN.

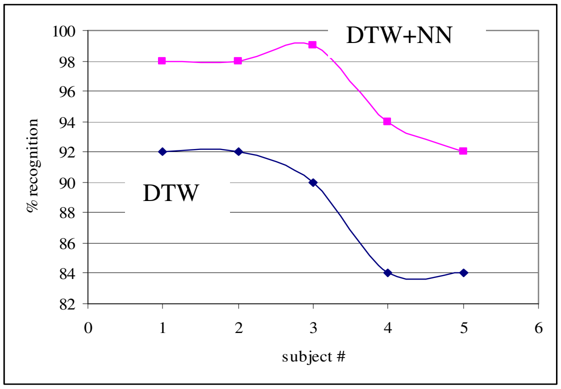

Fig. 5 Graphical comparison between DTW+NN and DTW itself.

6. Conclusions

The frame alignment based on DTW method for pre-processing LP coefficients and a new form of compressed data feeding into BPNN are described in this paper. Besides that, the back-propagation NN is used as the back-end speech pattern recognition engine. 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 or other subsequent pattern matching using the well known DP method.

Combination of methods presented in this work is an alternative to represent feature data and showed high recognition percentage. The frame alignment adopted DTW to normalize the spoken word length, in which these normalized templates then being used as the input to BPNN for the recognition part is proven could improve recognition performance on the samples tested to a higher percentage.

Warping path can also show the characteristics of speaker or words spoken. This information can be used together with acoustic feature, like pitch to study speaker recognition or word recognition. This is definitely a news to improve further the performance of speech recognition.

In conclusion, the 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. This was done by the reduction of inputs from using LPC coefficients to using local distance scores of a particular sample. A high recognition rate is achieved.

Further improvement will investigate the use of pitch information if it could identify the unidentified utterance using our current 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] R. Sudirman, S. H. Salleh, and T. C. Ming. Pre-Processing of Input Features using LPC and Warping Process. *International Conference on Computers, Communications, and Signal Processing*, 16-17 November 2005.

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

[4] 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.

[5] 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.

[6] J. Tebelskis, A. Waibel, B. Petek, and O. Schmidbauer. Continuous Speech Recognition using Linked Predictive Neural Networks. *International Conference on Acoustics, Speech, and Signal Processing.* 1: 61-64, April 1991

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

[8] S. Sae-Tang and C. Tanprasert. Feature Windowing for Thai Text-Dependent Speaker Identification using MLP with Back-Propagation Algorithm. *IEEE International Symposium on Circuits and Systems*, Geneva. 3: 579-582. May 2000.

[9] E. F. Cabral Jr. and G. D. Tattersall. Trace-Segmentation of Isolated Utterances for Speech Recognition. *International Conference on ASSP* 1:365-368, May 1995.

[10] S. H. Salleh. *A Comparative Study of the Traditional Classifier and the Connectionist Model for Speaker Dependent Speech Recognition System*. Universiti Teknologi Malaysia: Master Thesis, 1993.

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

[12] W.Y. Chen, S.H. Chen, and C.J. Lin. A Speech Recognition Method Based on the Sequential Multi-Layer Perceptrons. *Neural Networks*, Vol. 9(4): 655-669, 1996.

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