NN with DTW-FF Coefficients and Pitch Feature for Speaker Recognition

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

[1]Faculty of Electrical Engineering
[2]Faculty of Science
Universiti Teknologi Malaysia, 81310 Skudai, Johor
Tel.: 607-553 5738, Fax:607-553 5681, rubita@fke.utm.my

Abstract

This paper proposes a new method to extract speech features in a warping path using dynamic programming (DP). The new method presented in this paper described how the LPC 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 is needed for every input class especially in the application of multiple inputs. The new feature processing used the modified version of traditional DTW called as DTW-FF algorithm to fix the input size so that the source and template frames have equal number of frames. Then the DTW-FF coefficients are retained and later being used as inputs into the MLP neural network training and testing. 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 for speaker recognition which can be done by reducing the number of inputs into the network by using warping process, so the local distance scores of the warping path will be utilized instead of the global distance scores. The speaker recognition is performed using the back-propagation neural network (BPNN) algorithm to enhance the recognition performance. The results compare DTW using LPC coefficients to BPNN with DTW-FF coefficients; BPNN with DTW-FF coefficients shows a higher recognition rate than DTW with LPC coefficients. The last task is to introduce another input feature into the neural network, namely pitch. 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.

Keywords:

dynamic time warping, normalization, linear predictive coding, pitch feature, back-propagation neural network

1. Introduction

Speech recognition describes a group of special technologies that allow callers to speak words, phrases, or utterances that are used to control some particular applications. In the case of voice processing, speech recognition is used to replace touch-tone input, make for more intuitive menu structures, and add a level of simplicity and security to some systems. Speech recognition, on the other hand, is a technology that uses the spoken word as input that has an effect on the logic flow and execution of the program in query.

The recognition technique like the NN has been widely used as a recognition engine in speech recognition and other pattern recognition applications. There are also various form of input can be recognized by the network depending on the network setting whether it can accept and process single input or multiple inputs at a time. The mechanism of the network itself plays an important role to determine the suitable parameters and inputs for a particular application. In speech recognition application, a back-propagation NN can be used as the recognition engine and modified according to the norm of the problem.

From the literature reviews, past and most current research are using the global distance scores [2][3][4][9], or LPC coefficient as an input to the neural network one sample at a time. In that respect, a new method called Dynamic Time Warping Frame Fixing (DTW-FF) which is based on DP [1] is proposed to extract another form of feature which has a smaller number of input size so that it can reduce the amount of computation and network complexities in the back-propagation neural network.

Neural Network is chosen as the back-end recognition engine due to its past good and reliable performances in speech recognition. As mentioned in the earlier paragraphs, NN is considered as one of the popular method used especially when dealing with isolated word speech recognition. This study considers mainly on isolated words, NN is chosen as an engine to perform the recognition task. Since the main task of the study is to find an alternative way of reducing the number of inputs into the NN, this should light up a new form of input representation into the NN, which is simpler and smaller when compared to LPC feature.

2. Objective

The main objective of the study is to propose a method of time normalization to speech signals so that it can be used concurrently with other samples at a time using the back-propagation NN. Result from the time normalization which is composed from traditional DTW which then we called as the DTW-FF is used with another feature, namely pitch to compare the recognition performance using DTW-FF itself and both features.

3. Methods of Feature Extraction

The features used in this study are extracted separately before they are combined for the input into the NN. They are extracted as follows:

Step 1: LPC Feature Extraction

The LPC feature is extracted from the raw signal and it goes through some procedures which involve pre-emphasis, frame blocking, windowing, autocorrelation, and LP coefficient computation itself. The details can be found in [4][5][8].

Step 2: DTW-FF Feature Extraction

Every frame in a template and test speech pattern must be used in the matching path. Considering DTW path type 1 (Fig. 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*). If a reference template with feature vector *R* and an input 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.

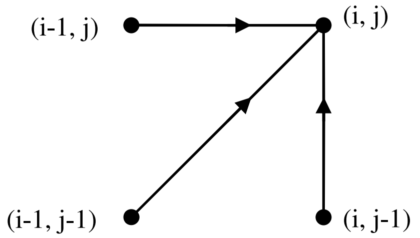

Fig. 1 DTW path type I

The global distance, *D* of a warping path is defined as

$$D(i,j) = min[D(i-1,j-1), D(i-1,j), D(j-1,i)] + d(i,j) \quad (1)$$

The DTW-FF feature is extracted using the DTW-FF algorithm which is designed based on the traditional DTW. The mechanism in the DTW algorithm is meant to calculate and find the shortest warping path between an input and the reference template. Similar occurs in the DTW-FF algorithm except that the algorithm is altered in such a way that it can fix the input frames number to the same amount as the frame number of the reference template. This is what we called as DTW frame fixing or DTW-FF in short. This frame alignment is also known as the expansion and compression method [6].

In this algorithm, frame fixing mechanism also includes the compression and expansion of frame numbers of the input, the reference template frame number is fixed, thus the input is adjusted according to the reference frame numbers. If the input has less frame number, it means that the frame number has to be expanded otherwise if the input frame number is more than the reference frame number, then the input frame number has to be compressed. The expansion and compression has to follow some rules according to the slopes of the warping path:

i- *Slope is ~0 (horizontal line)*

When the warping path moves horizontally, the frames of the speech signal are compressed. The compression is done by taking the minimum calculated local distance amongst the distance set, i.e. compare *w(i)* with *w(i-1), w(i+1)* and so on, and choose the frame with minimum local distance. The search is represented as

$$F^{-} = F(min\{d_{(i,j)\ldots(I,J)}\}) \quad (2)$$

ii- *Slope is ~∞ (vertical line)*

When the warping path moves vertically, 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. In other words, the reference frame duplicates the local distance of that particular vertical warping frame. The expansion can be represented as

$$F^{+} = F(r(i)) \quad (3)$$

iii- *Slope is ~1 (diagonal)*

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

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

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

Step 3: Pitch Feature Extraction and Optimization

Pitch is another feature considered in this study to ace the recognition rate of using the DTW-FF coefficients only. An application that strictly requires pitch into the system is the cochlear implant; the implant device is a custom design device which only suits a particular patient because each patient has different amounts of pitch and periodicity information (which determines the *F0* of a speech).

In this study the pitch feature is extracted using a scaled harmonic filter algorithm. The flow diagram of the algorithm is shown in Fig. 2. According to the diagram, firstly speech in *.wav* is used to obtain the initial values of fundamental frequencies, 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 and yield to an optimized pitch, F_o^{opt} [12]. This F_o^{opt} is used as an added input feature to the DTW-FF feature described in Step 2.

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 [12]. The algorithm find the optimum pitch value for a particular time by minimizing the difference between the calculated and the measured smearing of the spectrum due to the window length.

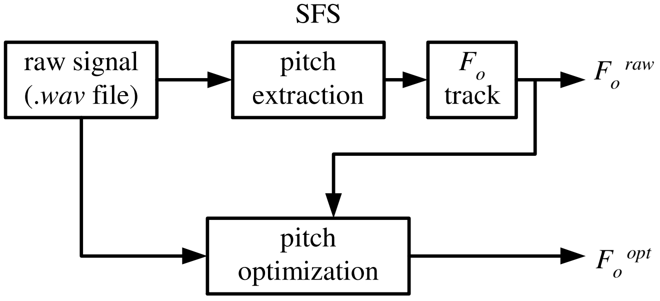

Fig. 2 Pitch feature extraction flow diagram

4. Experiment

The experiments consist of 2 phases. They are experiments of BPNN speech recognition comparing between using DTW-FF feature only and using DTW-FF feature combined with the pitch feature. Earlier experiment has showed that the DTW-FF feature does not alter the information contains in the speech since it showed no difference in its recognition rate when compared with the usage of LPC coefficients using the traditional DTW recognition engine. The result of Phase 1 and Phase 2 experiments are presented and discussed in the next section.

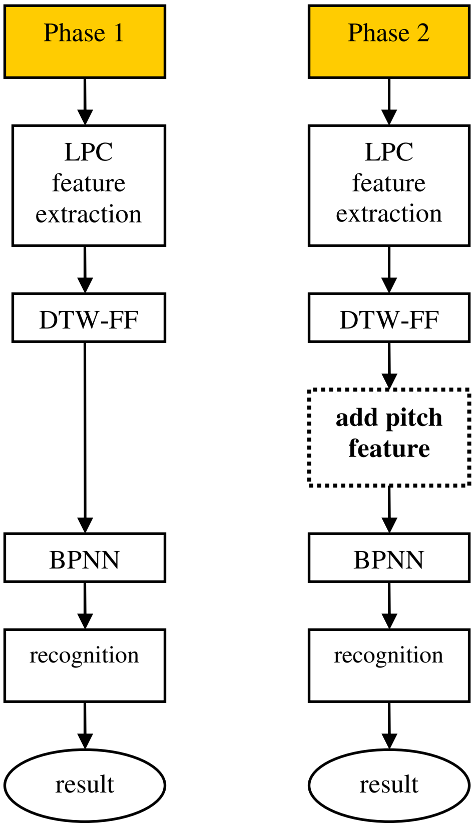

Fig. 3 Phase 1 and Phase 2 of the experiments

5. Results and Discussion

Phase 1: BPNN with DTW-FF Feature

TABLE 1: Comparison of using DTW-FF coefficients in DTW vs. BPNN

Subjects	1	2	3	4	5	6
DTW (%)	92	92	90	94	90	96
BPNN (%)	94	100	96	100	96	100

This experiment is to show that there is an improvement when using BPNN compared to the traditional DTW recognition engine. On average, the improvement from using DTW-FF in typical DTW to using DTW-FF in back-propagation neural networks is increased by 5.34 % for this particular set of experiment using 50 utterances by each subject.

DTW-FF features are obtained from the matching process in the DTW-FF algorithm. The scores have been scaled down from LPC coefficient which is a 10-order feature vector, into a coefficient (which is called as DTW-FF coefficient) derived from each frame. 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. Calculation to the input size reduction for example for 50 samples of 49 frames with LPC order-10 is as follows: For input using the LPC coefficients,

$$\begin{aligned} Input_{LPC} &= \#\ utterance\ x\ \#\ \ frame/utterance\ x\ \#\ coeff/frame \\ &= 50 utterance \text{ x } 49 frame/utterance \text{ x } 10\ coeff/frame \\ &= 24500\ coefficients \end{aligned}$$

For input using the local distance score,

$$\begin{aligned} Input_{LD} &= \#\ utterance\ x\ \#\ frames/utterance\ x\ \#\ coeff/frame \\ &= 50\ utterances \text{ x } 49\ frames/utterance \text{ x } 1\ coeff/frame \\ &= 2450\ coefficients \end{aligned}$$

Therefore, the percentage of number coefficients reduced is

$$\#\ coefficien t\ reduced\ = \frac{Input_{LPC} - Input_{LD}}{Input_{LPC}} x100\% = 90\%$$

However, this percentage is slightly higher if higher LPC order was used. For example, if LPC of order 12 is used, the percent reduction is 91.7. These means a lot of network complexities and amount of connection weights computations during the forward pass and backward pass can be reduced. Thus a faster convergence is achieved and this also allows more parallel computing of the speech patterns can be done at a time (more patterns can be fed into the neural networks at the same time).

Phase 2: BPNN with DTW-FF and Pitch Feature

TABLE 2: Recognition percentage before and after pitch feature is added to DTW-FF feature

Subjects	1	2	3	4	5	6
Before (%)	84	100	86	72	80	100
After (%)	100	100	95.9	98	90	100

The statistical test, called as T-Test has been conducted to the data in Table 6-3. This test assesses weather the means of two groups are statistically different from each other. The formula of the T-Test, *t* is:

$$t = \frac{\overline{X}_T - \overline{X}_C}{\sqrt{\frac{\sigma_T}{n_T} + \frac{\sigma_C}{n_C}}} \qquad (5)$$

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 was found that the value of *t* for DTW-FF in traditional DTW and in BPNN is -2.571 and -3.1247 respectively with a level of significance of α=0.05, this implied that $\mu_{before<}\mu_{after}$. Thus the results rejected the null hypothesis of $\mu_{before}=\mu_{after}$. Since H_1 is true where $\mu_{before<}\mu_{after}$, then it can be concluded that the percent improvement of using DTW-FF coefficients from typical DTW to using BPNN is significant. This also implies that the back-propagation neural network is a better choice for speech recognition for this particular set of data.

On the other hand, bear in mind that a lot of network complexity and amount of connection weights computations during forward and back pass have been reduced, thus faster convergence is achieved.

6. Summary

The frame alignment based on DTW method for pre-processing of linear predictive coefficients into a new form of compressed data called DTW-FF coefficients as input into back-propagation neural networks are described in this paper. The back-propagation neural network is used as the back-end speech pattern recognition engine. Having DTW-FF algorithm, frame matching is performed to fix the different frame numbers into a suitable desired frame number. The output of the frame fixing process, which is the local distance scores are then retained because these scores later are used for recognition using the back-propagation neural networks.

From the experiments, it has been proven that DTW can be modified to suit the needs of a particular situation or application as a front-end processing of speech recognition for back-propagation neural networks, 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 dynamic programming method. The introduction of DTW-FF coefficients in to the back-propagation neural networks also would be a good sign of fast parallel processing. This is important especially when there are a lot of data sets need to be processed at the same time so that the recognition can be obtained simultaneously.

The DTW-FF coefficients were compared to the LPC coefficients using typical DTW algorithm to identify whether or not any loss of information has occurred. From the experiments, it has been proved that there were no changes in the recognition rate, so we conclude that there is no loss of information during the frame fixing.

Pitch contains spectral information of a particular speech and this is the feature that is being used to determine the fundamental frequency, *F0*. The result of the experiment showed improvement on the recognition rate compared to only using the DTW-FF coefficients, this gives good sign how important pitch is when combined with other feature like the DTW-FF feature beside its least significant when it being used alone in speech recognition.

7. Conclusion

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.

References

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

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

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

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

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

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

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

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

[9] Tsai, H. L. and Lee, S. J. (1997 October). A Neural Network Model for Spoken Word Recognition. *IEEE International Conference on Systems, Man, and Cybernetics*. 5: 4029-4034.

[10] Sudirman, R., S. H. Salleh, and S. Salleh (2006). Local DTW Coefficients and Pitch Feature for Back-Propagation NN Digits Recognition**,** *Proceeding of IASTED International Conference on Networks and Comunications*, Thailand. 201-206.

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

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