MLFNN ALGORITHM FOR SPEECH CLASSIFICATION USING MMFC FEATURE EXTRACTION IN A SMART WHEELCHAIR

THUẬT TOÁN MLFNN CHO PHÂN LOẠI TIẾNG NÓI SỬ DỤNG TRÍCH ĐẶC TRƯNG MMFC TRONG XE LĂN THÔNG MINH

Nguyen Thanh Hai, Do Duy Tan, Quach Thanh Hai
Ho Chi Minh City University of Technology and Education

TÓM TẮT

Bài báo kiến nghị thuật toán mạng nơron truyền thẳng nhiều lớp (MLFNN) cho phân loại tiếng nói trong xe lăn thông minh, trong đó trích đặc trưng của những lệnh tiếng nói được thực hiện sử dụng những hệ số dựa vào tần số Mel (MMFC). Những lệnh được nhận biết cho điều khiển xe lăn là "Trái", "Phải", "Tới", "Lui" và "Dừng". Hơn nữa, một bộ lọc thông thấp Hamming được áp dụng để giảm nhiễu trước khi trích đặc trưng. Vậy thì những tín hiệu lệnh sau khi được nhận biết sẽ được đưa vào điều khiển xe lăn điện di chuyển. Những kết quả trong nghiên cứu này có thể hỗ trợ những người khuyết tật sử dụng xe lăn một cách dễ dàng và tiện lợi hơn trong cuộc sống và còn minh chứng sự hiệu quả của phương pháp kiến nghị.

ABSTRACT

This paper proposes a Multilayer Feed forward Neural Network (MLFNN) for speech classification in a smart electric wheelchair, in which with extraction of speech commands is performed using a Mel Frequency Cepstral Coefficients (MMFC) method. Speech commands recognized here are Left, Right, Forward, Backward and Stop. In addition, a Hamming low-pass filter is applied to reduce noise before feature extraction. Therefore, the recognized signals will be used to control the electric wheelchair. Results of this study possibly support disabled people using the wheelchair to move easily and more convenient in everyday life and also show to illustrate the effectiveness of the proposed approach.

Keywords: *MMFC feature extraction, Speech classification, Hamming low-pass filter and Multiplayer Neural Networks.*

I. INTRODUCTION

Many severely disabled people, who are increasing in the world, are difficult in community life. Assistive devices always encourage their movement in daily activity. Speech recognition is an active and popular method, used to translate human voice into commands. The model commonly used in identification as: Hidden Markov Model (HMM) [1], Vector Quantization (VQ), MFCC-DTW and neural networks. Identification used in wheelchair control, letter recognition [2], or number count [3]. Moreover other the HMM has high the accuracy, however it is very complex and the training time. The MFFC and DTW model is simple [4], do not take more training time, but lower accuracy than the HMM.

The voice recognition system is used to detect the attendance of speech in a background of noise. The beginning and end point of a word should be detected for processing words. The main difficulty of speech recognition is the same word spoke by different speakers depending on speaking styles, tone, regional, genders and speech patterns. In addition, noise and change of signals over time are problems considered in speech recognition.

Speech recognition plays an important role in an intelligent wheelchair system using

microphone. MFCC and DTW algorithms are applied for feature extraction and identification [5]. Speech commands such as left, right, forward, backward and stop will be recognized for an electrical wheelchair control. Experiments with identified speech commands using the proposed method will be performed by users.

In the word, the number of disable people about 15% of the population. Moreover people with disabilities feel isolated and do not have access to the same opportunities as other within their own communities. Those are reasons why an intelligent wheelchair was. For more convenient in modem life, electric wheelchair is improved day by day and a smart wheelchair is inevitable need. The smart wheelchair is designed to be used for indoor environment and user can easily control it by speech commands. In fact, when the user speech to the wheelchair to move using commands [6], microphone, in which the microphone carries each command to computer through software for wheelchair control.

Neural networks have been applied for identification of signals based on their features. In this research, authors used the neural networks for letter recognition [7] or identification of count numbers [8]. In this paper, the ALFNN will be employed in an electrical wheel chair speech signals corresponding to control commands such as left, right, backward, forward and stop. Experimental results on the wheelchair will be shown to illustrate the effectiveness of the proposed method.

II. MATERIAL AND METHODS

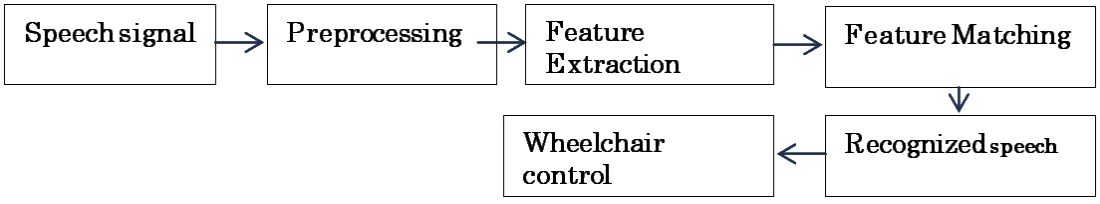

Figure 1. Block Diagram of Methodology.

A user will perform an electronic intelligent wheelchair control using user commands such as: *Left, Right, Backward, Forward* and *Stop*. A speech signal of user is recoded in interval of 2.5 s and all signals are pre-processed with sampling frequency 16 KHz and then feature extracted for identification. The voice signal uses a combination of features based on the MFCC and voice activity detection. The DTW algorithm is used to discriminate the speech into respective classes.

1 Feature extraction

As shown in Fig.2 of Feature extraction using the MFCC that consists of computational processes.

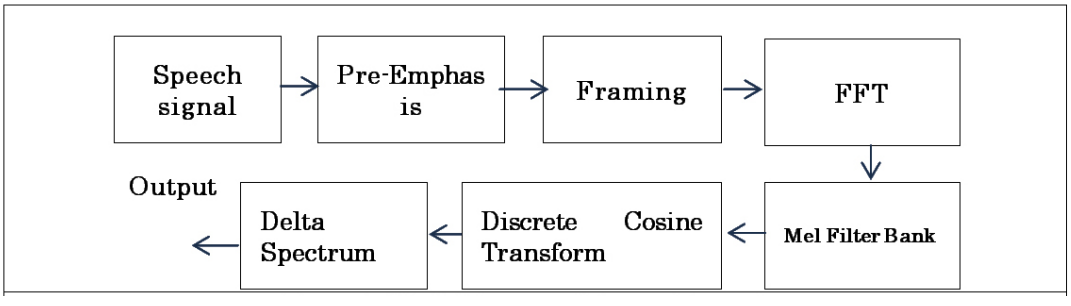

Figure 2. Block Diagram of Feature Extraction.

- ***Pre-Emphasis***

This step processes with purpose of offset high frequency components. In particular speech signal is processed using a filter which emphasizes higher frequency. This process will increase the energy of signal at the higher frequency. The output signal of the Pre-emphasis is computed the following equation:

$$H[n] = u(n) - a.u(n-1) \quad (1)$$

where H[n] is the signal output of the pre-emphasis process, u(n) is the voice signal, typical value of a = 0.95 (>20dB gain for high frequency). The result pre-emphasis process is shown in Fig.3

- ***Framing and Windowing***

The signal after Pre-emphasis is segmented due to the voice signal is continuous with time. The voice reliability can be ensured for a short time. Process frame cannot wait for last sample, which split reduces the signal discontinuities at the beginning and end of each frame, in which the frame length from 10 to 30 msec. The speech signal is divided into frames of N samples. This process is important to retain short term features. Short time analysis is performed by windowing the signal. Normally the Hamming window with (W(n), $0 \le n \ge N-1$) is used and its equation is given as follows:

$$Y(n) = H(n) x W(n) \quad (2)$$

where N is the number of sample in each frame, Y(n) describes the output signal and H(n) is the input signal. In particular, the result Hamming function is shown as:

$$W(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right) \quad (3)$$

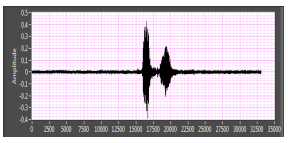

Figure 3. Pre-emphasis of signal

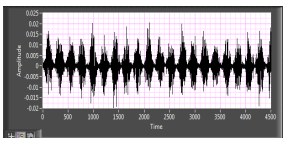

Figure 4. The signal is framed

- ***Mel Filter bank***

Human hearing is not equally sensitive to all frequency bands. It is less sensitive at higher frequency, roughly greater than 1000 Hz, human perception of frequency is non-linear. The Mel spectrum is the total spectrum of the signal spectrum after discrete Fourier transform multiplied by the weight of the Mel filter. Mel filter bank is series of triangular filter of the form at the center frequency and then if decreases linearly to zero at the center frequency of two adjacent filters [10].

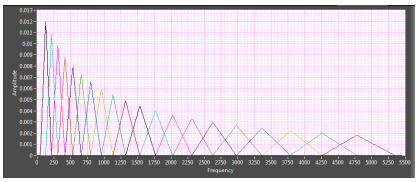

Figure 5. Mel Filter bank with frequency band from 50 to 5400 Hz, the number of filter bank is 20.

Each filter output is the sum of its filtered spectral components. The equation is used to compute the Mel for given frequency f in Hz:

$$f(mel) = 2595 \log_{10}(1 + \frac{f}{100}) \quad (4)$$

- ***Discrete Cosine Transform***

In this final step, the log Mel (mel) spectrum is converted to time. The result is called MFCC. Because the Mel spectrum coefficients are real numbers, we can convert them to the time domain using the DCT. The set of coefficient is called acoustic vectors. Therefore, each input utterance is transformed into a sequence of acoustic vector.

- ***Delta Spectrum and Delta Energy***

The DCT is done on Mel spectral coefficients [9] of each frame, hence obtaining the MFCC. The first 2 coefficients of the obtained MFCC are removed as they varied significantly between different utterances of the same word. Littering is done by replacing all MFCCs except the first 14 by zero. The first coefficient of the MFCC of each frame was replaced by the log energy of that frame. Delta

and acceleration coefficients are found from the MFCC so as to increase the dimension of the feature vector of the frames, thereby increasing the accuracy. The energy in a frame for a signal in a window is stored in an array and the values are used to detect the threshold energy of speech signal and noise removal, the energy is represented using the equation following:

$$E = \sum_{n=0}^{N-1} x^2(n) \tag{5}$$

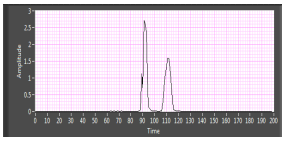

Figure 6a. Energy of signal

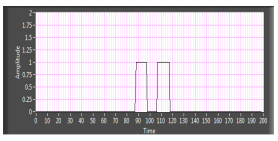

Figure 6b. Smooth energy of signal

2. ALFNN algorithm

For recognition of speech commands, an ALFNN model is applied for training. The most neural networks are trained relationship between the output and input before use for recognition.

The back-propagation network is to minimize the error function in the weight space using the reduced gradient method. Because of this method of calculating the gradient of the error function at each iteration requires that the error function should be continuous and indivisible [11].

The back-propagation algorithm is applied to find the local minima of the error function. Therefore, the gradient of the error function is calculated to change the initial weight values for the network. The weights are the parameters changed to reduce errors.

In this research, the number of hidden layer neurons is calculated dependent on many parameters such as the number of nodes of the input –output, the training sample set. Therefore, the number of the hidden layer neurons is determined using the following equation:

$$H = \frac{Q}{5(I+O)} \tag{6}$$

where I is the number of the input nodes, O describes the output size and Q is the length of the training set.

III. RESULT AND DISCUSSION

1. Signal processing

After running the algorithm, the result obtained. A user speaks words: left, right, backward, forward and stop, and then the system saves them. The input words will be recognized corresponding to the template with the lowest matching score. The ALFNN algorithm is used for distance calculation between the tested speech and the reference word bank. The signal to band-pass filter with cut off frequency interval [80, 1200] Hz. The signal has sample rate 11025 Hz [12] and voice is recorded interval 3 sec per word.

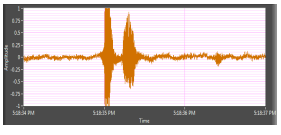

Figure7a. Signal input

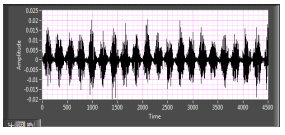

Figure7b. Distributed frame

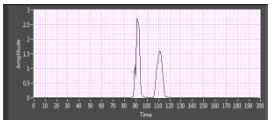

Figure 7c. Energy signal

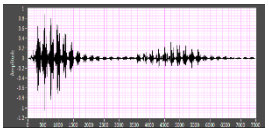

Figure 7d. Identified signal

Figure 7a is the original input signal and figure 7b represents the distributed frame of the signal, in which this frame is to split the discontinuous signal at the beginning and end of each frame having the length from 10 to 30 ms. Fig. 7c shows the energy of signal smoothed using a threshold. In figure7d, the signal is identified. All of the identified signals of user are applied for the electrical wheelchair control to reach the desired target.

2. Signal recognition

After extracting speech features using the MFCC algorithm, one obtained 13 feature coefficients, 13 energy factors and 13 delta coefficients of phonetics, but the first two coefficients of features were removed due to the different pronunciations of the same word. The feature vector inputs to the neural network for recognition.

Type of neural network used here is a multilayer feed forward algorithms trained with back-propagation, consisting of one input layer, one hidden layer and one output layer. The number of input neurons is set to35values of the feature vector, the number of output neurons is equal to five identifiers corresponding to commands such as *left, right, backward, forward* and *stop*.

In this experiment, with five outputs, 5000 training samples for each ward employed. Therefore, the number of the maximum hidden layer neurons is 25 and each neuron is a linear function. Moreover, the performance of the network depends on the quality of the signal related to pre-processing and feature extraction. The result of words recognition using the ALFNN is shown from figures 8a to 8e.

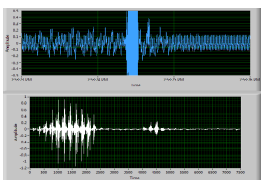

Figure 8a. Classified "left" signal

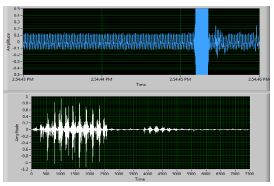

Figure 8b. Classified "right" signal

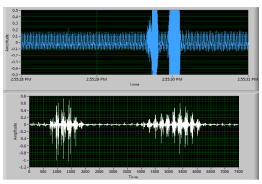

Figure 8c. Classified "backward" signal

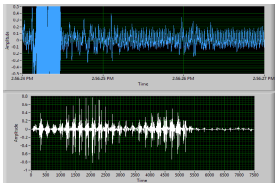

Figure 8d. Classified "forward" signal

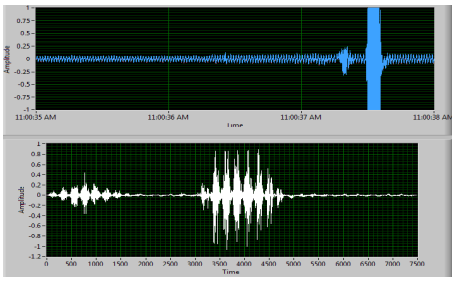

Figure 8e. The signal of "stop" command.

3. Wheelchair control

The signal recognition is the output to the wheelchair control through USB-6008 of NI Company, the output voltage level of USB-6008. The voltage signal output is taken down motors of the wheelchair to control electric wheelchair movement as *left, right, backward, forward* and *stop*. The model of the electric wheelchair is shown in Figure 9, in which user is driving it using speech commands.

Figure 9. The model of an electric wheelchair.

Speech word has been identified using the ALFNN algorithm for controlling an electric wheel chair. The accuracy of words classification for wheelchair control is shown in table 2, in which the test result was performed 40 times.

Recognition results with the neural networks corresponding to changing hidden layers, hidden layer neuron sof1 0 and 15 obtain the different accurate identification results. With 10 hidden layer neurons, the recognition results are lower than the hidden layer neuron

sof15, but the training time is faster. With this ALFNN, the accuracy of the wheelchair control is shown intable1.

Table 1. The accuracy of wheelchair control using the ALFNN

Words	Hidden layer neurons	Accuracy (%)
Left	10	86.25
	15	98.75
Right	10	88.25
	15	94.50
Backward	10	82.00
	15	95.50
Forward	10	89.00
	15	95.25
Stop	10	88.25
	15	94.50

A comparison between two ALFNN and hidden Markov models was made to identify individual words in Arabic [13].In this paper, the NNmethodusing13MFCCand delta coefficients, in which the 256-pointsFFT method was used to find the energy spectrum of the signal, using Mel filter bank with 24 banks. The accuracy is about 89% in a clear environment

IV. CONCLUSION

In this paper, speech signals were filtered by the Hamming low-pass filter and features were extracted for words classification using the Mel Frequency Cepstral Coefficients (MMFC) method. From the MMFC coefficients, the Multilayer Feed forward Neural Network (MLFNN) was utilized to classify speech words. Therefore, the words were used to control the electric wheelchair for disabled people in the indoor environment. The effective results were obtained after classification using the ALFNN.

ACKNOWLEDGEMENTS

The authors would like thank the support of BME department, IU and UTE students, Vietnam.

REFERENCES

[1] Bhupinder Singh, Neha Kapur, Puneet Kaur, "*Speech Recognition with Hidden Markov Model: A Review"*, International Journal of Advanced in Computer Science and Soft Engineering, Vol. 2, pp. 400-403, 2012.

[2] Dipmoy Gupta, Radha Mounima C. Navya Manjunath, Manoj PB, "*Isolated Word Speech Recognition Using Vector Quantization",*International Journal of Advanced in Computer Science and Soft Engineering, Vol. 2, pp. 164-168, 2012.

[3] Talal Bin Amin, Iftekhar Mahmood, "*Speech Recognition Using Dynamic Time Warping",* ICAST, 2008.

[4] A.*Revathi*, R.Ganapathy and Y.Venkataramani, "*Text Independent Speaker Recognition and Speaker Independent Speech Recognition UsingIterative Clustering Approach",* International Journal of Computer science & Information Technology, Vol. 1, No 2, pp. 30-42, 2009.

[5] Lindasalwa Muda, Mumtaj Begam and I. Elamvazuthi, "*Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques*", Journal Of Computing, Vol. 2, Issue 3, pp. 138-143, 2010.

[6] Kohei Arai, Ronny Mardiyanto, *Eyes Based Eletric Wheel Chair Control System,* International Journal of Advanced Computer Science and Applications, 2011.

[7] S. M. Azam, Z.A. Mansoor, M. Shahzad Mughal, S. Mohsin, "*Urdu Spoken Digits Recognition Using Classified MFCC and Backpropgation Neural Network*", Computer Graphics, Imaging and Visualisation IEEE, Vol.1, pp. 7695-2928, 2007.

[8] Chin Luh Tan and Adznan Jantan, "*Digit Recognition Using Neural Networks*", Malaysian Journal of Computer Science, Vol. 17 No. 2,pp. 40-54, 2004.

[9]Chin Luh Tan and Adznan Jantan, "*Digit Recognition Using Neural Networks*", Malaysian Journal of Computer Science, Vol. 17 No. 2, pp2012. 40-54, 2004.

[10] S. M. Azam, Z.A. Mansoor, M. Shahzad Mughal, S. Mohsin, "*Urdu Spoken Digits Recognition Using Classified MFCC and Backpropgation Neural Network*", Computer Graphics, Imaging and Visualisation IEEE, Vol.1, pp. 0-7695-2928, 2007.

[11] R. Rojas, *Neural Networks*. Berlin: Springer-Verlag, 1996.

[12]Fu-Hua Liu; Richard M. Stern; Xuedong Huang; Alejandro Acero, "*Efficient cepstral normalization for robust speech recognition, human language technology*", Proceedings of a Workshop Held at Plainsboro, New Jersey, pp. 21-24, 1993.

[13]Z. Hachkar et al., "*A Comparison of DHMM and DTW for Isolated Digits Recognition System of Arabic Language*", International J. on Computer Science and Enginerring, 2011.