Modelling of Speech Recognition using ANN and UI for Controlling Wheelchairs Indoor

Md. Ibrahim Chowdhury, Amer Khoulani and Oluwafemi Samuel

Abstract— Paralyzed people need to control their electrical wheelchair. One of the most suitable means is by voice. So we aimed to combine Neural Networks, and Utterance Isolating techniques in speech recognition algorithm to make the control of wheelchair by voice command which is reliable, and secure. First the silence is deducted, and the pure word will then be filtered, and processed. So certain features Cepstral Coefficients will be extracted, and sent as input to the neural network which will compare them with the stored ones and give the recognized output.

Index Terms— Speech Recognition, Artificial Neural Network, and, Cepstral Analysis.

——————————   ◆   ——————————

1 INTRODUCTION

Paralyzed people's mobility shall be more comfortable if they can do it relatively independent. One way to achieve this purpose is by using their voice to command wheelchairs. A system like this, however, must be stable at all times, with no exceptions, because the slightest mistake might be fatal.

So we propose a system that will be extremely promising to make handicap people’s life easier.

2 REVIEW OF THE STATE OF ART

A three stage noise reduction system is considered for noise reduction in the Cepstral algorithm, and is implemented to have an excellent feed back of the voice signal. It is a well tested algorithm even in a noisy environment, and with a good response. The output is fed into the neural network classifier for further processing. [1]. The processing time for the speech signals needs to be as small as possible, therefore a great need for the isolation of the exact voice signal minus the noise is a necessity. The main characteristic of the algorithm is the detection of the endpoints of the voice signals for efficient analysis [3]. An improved error rate reduction method, non-conventional Cepstral coefficient extractor was used to produce the Cepstral coefficients, which in turn are used to feed the neural network. The method acts more on the harmonic structure of the frequency domain rather than the conventional classification of the signal spectrum [2].

3 PROBLEM STATEMENT AND MAIN CONTRIBUTION

How will the paralyzed people have more accurate control over their speech control electrical wheelchairs?
Implementing speech recognition using Neural Networks, and Utterances Isolating techniques will increase the accuracy of control over speech controlled electrical wheelchairs. The combination of these techniques will improve the rate of speedy speech recognition in noisy environment. We can state our contribution as the following:

- Modeling of speech recognition using ANN and UI.
- Implementing the whole model on MATLAB.
- Validating the model in MATLAB.

4 PROBLEM SOLUTION

Our vocabulary is limited to five words: forward, backward, right, left, stop. The voice signal will be sampled with rate of

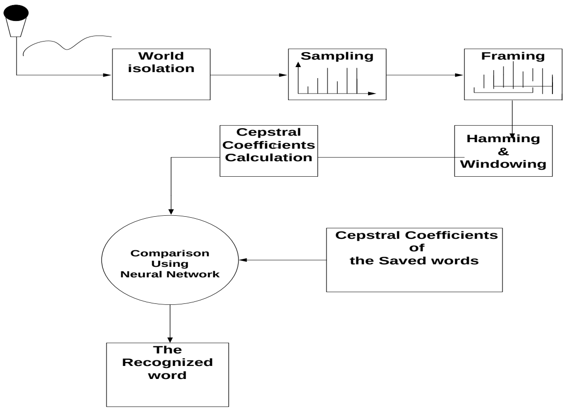
World isolation
Sampling
Framing
Cepstral Coefficients Calculation
Hamming & Windowing
Comparison Using Neural Network
Cepstral Coefficients of the Saved words
The Recognized word

Fig. 1. The model of the whole system with the major steps.

- *Md. Ibrahim Cowdhury is currently working as a Lecturer in Computer Science & Engineering Department at City University, Bangladesh. E-mail: mic@ieee.org*
- *Amer Khoulani has pursued his masters degree in Telecommunications from Blekinge Institute of Technology (BTH). Email: akhoulani@gmail.com*
- *Oluwafemi Samuel has pursued his masters degree in Telecommunications from Blekinge Institute of Technology (BTH). Email: soma2ng@yahoo.com*

10 kHz. The world is isolated in order to get the pure word without the silence. The speech signal is blocked into frames of N samples with adjacent frames being separated by M (M < N). The first frame consists of the first N samples. The second frame begins M samples after the first frame, and overlaps it by N - M samples [3].

During the first 50 ms no speech is recorded, and statistics of the existing silence is measured. We compute the maximum energy in the whole signal *EMAX*. Then we compute the minimum energy in the silence range *EMIN* [3]. Now we get the following constants

$$X = 0.03 * (EMAX - EMIN) \tag{1}$$

$$Y = 5 * EMIN \tag{2}$$

Then we get the lower energy threshold *ITL*, and the upper energy threshold ITU as follows:

$$ITL = Min\ (X, Y) \tag{3}$$

$$ITU = 5 * ITL \tag{4}$$

The zero-reaching threshold *ZRT* equals to the maximum zero-reaching rate of the whole signal multiplied by a constant of 2.5 [3].

$$ZRT = 2.5\ ZCMAX \tag{5}$$

Now we begin by searching frames until the lower threshold *ITL* is exceeded. This point *START*, is tentatively identified as the beginning of the voiced input and is unchanged unless the energy falls below *ITL* afterwards, before it can exceed *ITU*.

Similarly, the algorithm starts searching for the point *END* from the end of energy array until the lower threshold ITL is exceeded [3].

But maybe some of the desired pure signal doesn't lie within the limits *START*, and End but can possibly lay outside them in the case when there is a weak fricative or nasal sound. This happen because these sounds have low energy content but high *ZCR* [3].

So we search the frames until the zero-reaching threshold *ZRT* is exceeded. This point, *ZCSTART*, is identified as the beginning of the unvoiced sound. Similarly, the algorithm starts searching for the point *ZCEND* from the end of the zero-crossing array until the zero-crossing threshold *ZRT* is exceeded. The beginning point of the pure signal *FSTART* will be the minimum of *START*, and *ZCSTART*. The end point of the pure signal *FEND* will be the maximum of *END*, and *ZCEND* [3].

$$FSTART = Min\ (START, ZCSTART) \tag{6}$$

$$FEND = Max\ (END, ZCEND) \tag{7}$$

Then, the Cepstral coefficients are produced. The Cepstral parameters are suitable for speech recognition applications in noisy environments since they are processed from high values using the spectrum domain [2].

$$C(q)=IF(\log|F(s(t))|) \tag{8}$$

Now an input array containing the coefficients of all recorded word will be the input to the feed-forward back-propagation neural networks [1].

Deciding the network parameters is totally experimental with some constraints. The number of the neurons in the input layer should be less the number of the neurons in the input array. We chose it to be 40. We set the output neuron to 3.

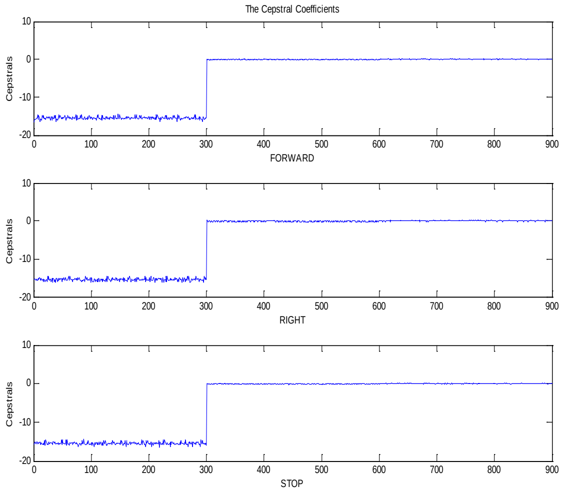

Fig. 2. The 900 Cepstral coefficients for three different words: Forward, Right, & Stop.

The more hidden neurons, the faster training, and performance we get [1]. But we should be aware not to get over-learning situation. We chose it to be 60 neurons. In our network we followed the supervised learning, and we set the number of epochs to be 70000. After the training of the networks the error of zero wasn't reached but the error we get was less than 10^{-10}.

At the black box testing of our model, the results were satisfying. In overall the error rate was about 25% inside, while it rose to 40% in a normally crowded room. A crucial factor in the performance of the neural network is the sensitivity of the microphone. Also the distance between the microphone and the user must be constant. Another important factor is the number of the Cepstral coefficients taken from every word. We took 900 coefficients from every word: 300 from the beginning, 300 from the middle of the word, and 300 from the end. But when we had taken these 900 coefficients, and ignored the others of every word, we lost some features.

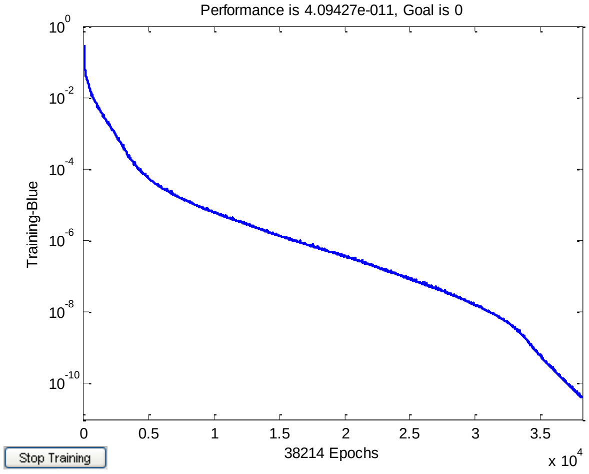

Fig. 3. The net performance after the training of 38214 epochs.

4 Conclusion

We successfully implemented a model of speech recognition using Neural networks, and Utterances isolating techniques which can be applied for any electrical vehicles, and is highly recommended for electrical wheelchair. We observed a huge accuracy of ANN responses by increasing the number of the Cepstral coefficients taken from the processed words. The Linear Predictive coefficients and the Mel-Filter Bank techniques can be used for comparisons, and in further research work.

References

[1] P. D. Polur, Ruobing Zhou, Jun Yang, F. Adnani and R.S. Hobson, ” Isolated speech recognition using artificial neural networks”, presented at the *23rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, Istanbul, Turkey, Oct. 2001.

[2] R. Vergin, D. O'Shaughnessy and A. Farhat, “Generalized mel frequency cepstral coefficients for large-vocabulary speaker-independent continuous-speech recognition,” *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 5, pp 525-32, Sept. 1999.

[3] J. Huang and B.-D. Tseng, “A Walsh transform based endpoint detection of isolated utterances”, presented at the *25th Asilomar Conference on Signals, Systems and Computers*, Pacific Grove, California, Nov. 1991.

Authors’ biography

Md. Ibrahim Chowdhury is working as a Lecturer in Computer Science and Engineering Department at City University Bangladesh since 2011. He has completed his bachelor studies from International Islamic University Chittagong, Bangladesh in 2007 as a CSE (Computer Science and Engineering) graduate. He pursued Masters degree in Electrical Engineering from Blekinge Tekniska Högskola (BTH), Sweden in 2010. His major area of interests are in signal processing, seamless roaming, mobility protocols, simulations, modelling, analysis and performance measurement of communication systems, mobility protocols, wireless networks, digital communications, tele-traffic theories, artificial intelligence and neural networks. He is a member of IEEE and IEEE Communications Society.

Amer Khoulani was born in Damascus, Syria in 1980. He obtained a BSC (Computer & Communications Engineering) from American University of Science & Technology, Lebanon in 2006 and a Bachelor of Law from Damascus University, Syria in 2007. Between March 2007 and October 2007, he worked as technical support engineer at ETISALAT GSM Operator in UAE. He has done his master’s program in Electrical Engineering with emphasis on Telecommunications at Blekinge Institute of Technology, Sweden. His major interests in research are Signal and Speech Processing.

Oluwafemi Samuel Ajayi was born in Ilesa town, Nigeria in 1980; he bagged his B.Tech Electronic and Electrical Engineering from Ladoke Akintola University of Technology, LAUTECH, Nigeria in 2004. He worked as an Operator at PPMC/NNPC from Feb 2005 to Feb 2006. He worked as a Network and IP broadcast Engineer with DNA Communication-Alchemist. Lagos, Nigeria between Aug. 2004 and Nov. 2006. He also worked as the Assistant Head I.T and Engineering dept Soman Global Technologies. Lagos. Nigeria from Nov 2006 to Aug. 2008. He has done his master’s programme in Electrical Engineering with emphasis on Telecommunications at Blekinge Institute of Technology, Karlskrona Sweden in 2011.