Neural Network Based Recognition of Speech Using MFCC Features

[1]Pialy Barua, [1]Kanij Ahmad, [2]Ainul Anam Shahjamal Khan, [3]Muhammad Sanaullah

[1,2]Department of Electrical and Electronic Engineering, Chittagong University of Engineering and Technology,
Chittagong-4349, Bangladesh

[3]School of Computing and Engineering, University of Missouri-Kansas City, MO 64110, USA

Email: pialee_08@yahoo.com, kanij.ahmad08@gmail.com, khanshahjamal@yahoo.com,
mu222@mail.umkc.edu

***Abstract*—Analysis and detection of human voice at workplace such as telecommunications, military scenarios, medical scenarios, and law enforcement is important in assessing the ability of the worker and assigning tasks accordingly. This paper represents the results from a preliminary study to recognize the speech from human voice using mel-frequency cepstrum coefficients (MFCC) features. The 16 mel-scale warped cepstral coefficients were used independently for reorganization of speech from two Bangla commands of our native language. Cepstral coefficients for the utterance of 'BATI JALAO' (i.e., TURN ON LIGHT) and 'PAKHA BONDHO KORO' (i.e., TURN OFF FAN) from a particular speaker under preliminary investigation were used as features in a neural network. Network is trained using the MFCC features of two speakers in such a way that it can recognize only one particular person along with his command and terminate the program for other. Result of matching features in a neural network demonstrates that MFCC features work significantly to recognize speech.**

Keywords—Back-propagation, Artificial Neural Network, Fast Fourier Transform, MFCC, Speech Recognition.

I. INTRODUCTION

As the best ever creature, speech is one of the most important manners of communication for humans, to exchange feelings and information [1]. Since the dawn of civilization, the invention and widespread use of the telephone, audio-phonic storage media, radio, and television has given even further importance to speech communication and speech processing, which led to development of computer based automatic speech recognition system. In computer based speech recognition system, a computer simply attempts to transcribe the speech into the textual representation, rather than understanding the meaning of it**.** It can be used in many applications such as, security devices, household appliances, cellular phones, ATM machines, and computers.

Bangla (which can also be termed as Bengali) is largely spoken by the people all over the world. About 225 million or above people speak in Bangla as their native language. It is ranked seventh based on the number of speakers. It is the native language of our country. Today, most of the computer based resources of automatic recognition system and technical journals are in English. Due to the language barrier, the common masses of our population face big obstacle to enjoy the optimum benefits of modern communication and information technology (ICT) where huge enriched English knowledge databases are there around the globe. Language processing in mother tongue is the only technological way that can be used to remove this barrier. However, a very little number of researches have been performed where many literatures in automatic speech recognition (ASR) systems are available for almost all the major spoken languages in the world. Early researchers have developed Bangla speech recognition system for only phonemes, letters [2], words [3], or small vocabulary continuous speech. But a very few implementation based works have been done.

Speech recognition techniques follow a template matching of features with time normalization by dynamic time warping, or a neural network model. Features used for creating templates are typically derived from spectral or log spectral representation of each frame of speech. Parameters from a linear prediction model and mel-cepstral coefficients have been commonly used for forming templates.

978-1-4799-5180-2/14/$31.00 ©2014 IEEE

This paper aims at developing a speech recognition system that can be further implemented to control any electrical appliance such as lamp, fan, radio, television by using user's voice to help the ill, aged and disabled people of our country, who cannot speak in English and thus are deprived of modern technological facilities.

II. DATABASE

In this research, firstly two Bangla voice commands and four users, from whom only one specific user (referred as recognized user) and his command will be recognized, are selected. Each voice command of each person is recorded for 2 seconds and saved as wave file (.wav) in computer through MATLAB. After several mathematical and computational procedures, the MFCC features are extracted from each recorded voice. Hence acoustic voice signal is converted to a set of numerical values. After that, training data-table is created using MFCC feature data of two persons (recognized user and any one of other users) and target data-table also created as back-propagation neural network was used. Built-in Artificial Neural Network (ANN) is trained with these. Finally, when the test data is given, the ANN compares the test data with train data-table. If speaker is recognized user, it recognizes him and the specific command by finding the maximum possible matches within a given tolerance level. Otherwise, it terminates program.

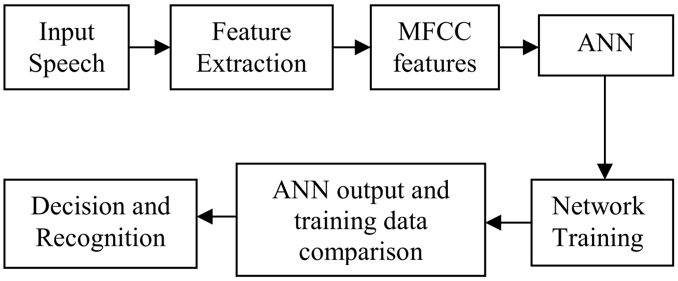


Fig. 1 Block Diagram of Recognition Phase.

III. MFCC FEATURE EXTRACTION TECHNIQUE

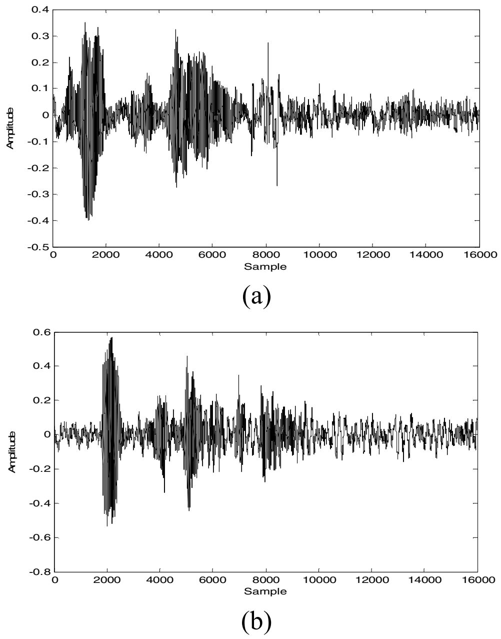
Why MFCC for Feature Extraction

For speech recognition purposes and research, MFCC is widely used for speech parameterization and is accepted as the baseline. This may be attributed because MFCCs models the human auditory perception with regard to frequencies, which then can represent sound better [4]**.** This technique is often used to create the fingerprint of the sound files. The MFCCs are based on the known variation of the human ear's critical bandwidth frequencies with filters spaced linearly at low frequencies and logarithmically at high frequencies used to capture the important characteristics of speech.

Another popular feature extraction technique is Liner Predictive Coding (LPC). This algorithm produces a vector of coefficients that represents a smooth spectral envelope of the DFT magnitude of a temporal input signal. But use of Linear Prediction coefficients alone for speech recognition process is not efficient because all pole assumption of the vocal cord transfer function is not accurate and this method is not efficient enough to separate the convolution of the excitation from the glottis and the pole transfer function [5].

Wave file creation

Firstly the user's speech (which is one of two bangle commands "BATI JALAO" i.e., turn on light and "PAKHA BONDHO KORO" i.e., turn off fan) is recorded for 2 seconds and saved as wave file (.wav) in MATLAB.



(a)

(b)

Fig. 2 Speech Signal of (a) Recognized Speaker's Speech " Bati Jalao"; (b) Recognized Speaker's Speech " Pakha Bondho koro".

Preprocessing

The sampling frequency is 8 kHz. The 16 center frequencies have been set to cover up to actual frequency 4 kHz. To exclude DC, the first critical band is started with the resolution frequency (DF) [6].

978-1-4799-5180-2/14/$31.00 ©2014 IEEE

Table 1 Critical Band filter bank [7]

Band Index	Bandwidth (Hz)	Band Index	Bandwidth (Hz)
1	DF-86	13	1723-1981
2	86-172	14	1981-2326
3	172-258	15	2326-2756
4	258-431	16	2756-3187
5	431-517	17	3187-3876
6	517-689	18	3876-4307
7	689-775	19	4307-4737
8	775-948	20	4737-5254
9	948-1120	21	5254-5857
10	1120-1292	22	5857-6460
11	1292-1464	23	6460-6977
12	1464-1723	24	6977-7585

Triangular Filter-Bank design

The necessary information in a human speech signal contained in such a frequency range, whose band shape looks like a triangle. Therefore, a filter bank of 16 triangular band pass filters is designed.

Frame Blocking

The input speech waveform is cropped to remove silence or acoustical interference that may be present in the beginning or end of the sound file. Here sampling frequency is 8 kHz, it means 8000 samples per second. Taking 20 ms frame length, speech samples are divided into frames, each consisting of 160 samples.

Frame Overlapping

50% overlapping of the frames is done to remove the disadvantage of windowing functions i.e., attenuation of the beginning and end of the signal in the calculation of the spectrum.

Windowing

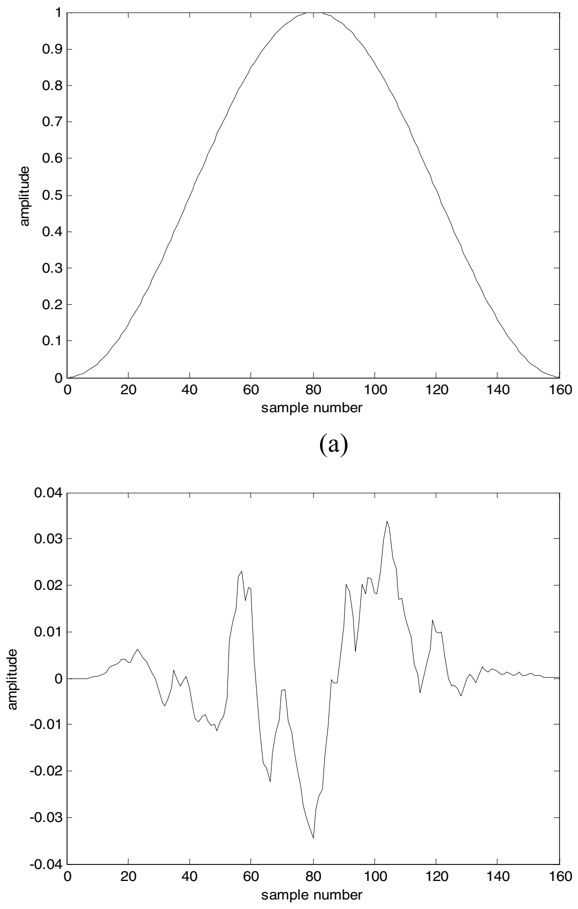
Each frame of speech samples is applied to the windowing block to minimize the discontinuities of the signal by tapering the beginning and end of each frame to zero. A window is shaped so that it is exactly zero at the beginning and end of the data block and has some special shape in between. This function is then multiplied with the time data block forcing the signal to be periodic. In this way, windowing reduces leakage.

Here, Hanning window used due to good frequency resolution property. This cosine window is defined by:

$$w(n) = 0.5\,(1 - \cos\frac{2\pi n}{N-1}) \qquad (1)$$

where, n represents the sample number and N represents the width, in samples, of discrete-time, symmetrical window function *w(n)*.

The ends of the cosine just touch zero, so the side-lobes roll off at about 18 dB per octave.



(a)

(b)

Fig. 3 Waveform of (a) Hanning Window and (b) Windowed Signal.

Fast Fourier Transform (FFT) of windowed frame

One of the most common techniques of studying a speech signal is via the power spectrum. The power spectrum of a speech signal describes the frequency content of the signal over time. Hence the Discrete Fourier Transform (DFT) performed to convert a finite list of equally spaced samples of a function into the list of coefficients of a finite combination of complex sinusoids, ordered by their frequencies, that has those same sample values. Spectral energy is calculated using 512-point DFT. It converts the sampled function from its original domain (often time or position along a line) to the frequency domain. The sequence of N complex numbers $x_1, x_2, \ldots.., x_n$ is transformed into an N-periodic sequence of complex numbers $X_0,X_1,\ldots\ldots\ldots..,X_{N-1}$ according to the DFT formula:

$$X_k = \sum_{n=0}^{N-1} x_n . e^{-i 2\pi k n / N} \qquad (2)$$

where, k = number of filter.

A speech signal contains only real point values, so real-point Fast Fourier Transform (FFT) used for increased efficiency due to its rapid transformation ability.

Mel-frequency wrapping

The spectrum obtained after FFT is wrapped according to the Mel Scale. Human perception of the frequency contents of sound (for speech signal) does not follow a linear scale. Thus for each tone with an actual frequency, F, measured in Hz, a subjective pitch is measured on a scale called the "Mel" scale. The Mel frequency scale is linear frequency spacing below 1000 Hz and a logarithmic spacing above 1000 Hz. As a nonlinear transformation of the frequency scale, the following formula is used:

$$F_{mel} = 2595 \times \log_{10}\left(1+\frac{F_{Hz}}{700}\right) \qquad (3)$$

Mel Cepstrum

Finally, the log Mel spectrum converted back to time domain. The result is called the Mel Frequency Cepstrum Coefficients (MFCC). In cepstrum analysis, previously generated filter bank of k triangular band pass filters is applied to voice signal spectrum. These triangular band pass filters have center frequencies in k equally spaced Mel values. The equally spaced Mel values correspond to different frequency values, i.e.,

$$F_{Hz} = 700\left(10\left(\frac{F_{mel}}{2595}\right)-1\right) \qquad (4)$$

The Mel spectrum coefficients (and so their logarithm) are real numbers, so they were converted to the time domain using the discrete cosine transform (DCT). The discrete cosine transform is done for transforming the Mel coefficients back to time domain.

$$C_n = \sum\nolimits_{k=1}^{k} (\log S_k)\ \cos\left\{n\left(k-\frac{1}{2}\right).\frac{\pi}{k}\right\} \qquad (5)$$

$$n = 1, 2, \ldots\ldots\ldots\ldots, k$$

whereas, $S_{k\,,}\, k = 1, 2, \ldots.., k$ are outputs of last step.

Finally, MFCC features of speech (a 199×16 matrix) signal are obtained.

The complete MFCC feature extraction technique is shown in following block diagram.

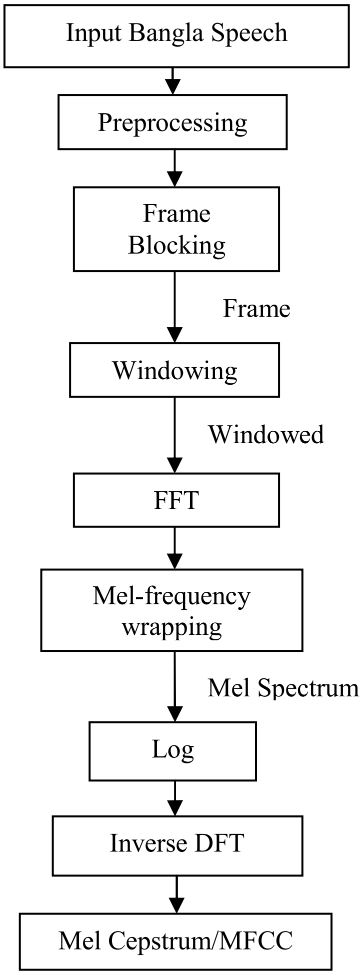


Fig. 4 MFCC Feature Extraction Process.

IV. Recognition of Speaker and Speech through Feed Forward ANN

Feed forward with back propagation algorithm has been used to fit an input-output relationship [8]. The multilayer neural network consists of an input layer, a hidden layer and an output layer.

The hidden layer consists of non-linear sigmoidal activation function [9].

If n is an input and a is an output. Then the input-output relation,

$$a = \frac{1}{(1+e^{n})} \qquad (6)$$

While the output layer has linear activation function [9], then the input-output relation,

$$a = n$$

Each set of Mel-frequency cepstral coefficients (MFCC) for particular speech of speakers is used as a complete input dataset for a neural network, which is used to train the network. In case of MFCC calculation, there are 199 frames for 2 second speech signal each with 16 co-efficient. Each set of MFCC is consist of 199×16=3184 samples per utterance which represents the input layer size. A layer between the input and output layer i.e., hidden layer has 20 neurons. The Levenberg-Marquardt algorithm has been used to train the neural network, as it is fastest back-propagation algorithm. This algorithm updates weight and bias values according to Levenberg-Marquardt optimization. The algorithm convergences to the steepest descent algorithm until the local curvatures are proper to make a quadratic approximation; then it approximately becomes the quasi-Newton algorithm, which can speed up the convergence significantly.

Once the neural network is built, it is trained with input and corresponding targets. Here two neural networks with same configuration have been applied. In first network, the target is 1 for the particular speaker with Bangla speech the network will recognize and 0 for all other speakers. In second network, the target is 1 for the speech “BATI JALAO” (i.e., Turn On Light) and the target is 0 for “PAKHA BONDHO KORO” (i.e., Turn Off Fan). If the first network recognizes the particular speaker, then the second network proceeds and recognizes the speech of particular speaker. If the first network recognizes that the speaker is not the particular one then the system will terminate and the second network will not proceed.

In training phase, it is observed that to reach a low mean square error that means a low average difference between output and target, 4 to 8 epochs are sufficient to train a network with inputs and targets. The value of mean square error zero means no difference between output and target. Here, for neural network, best mean square error is 0.041695 at epochs 6.

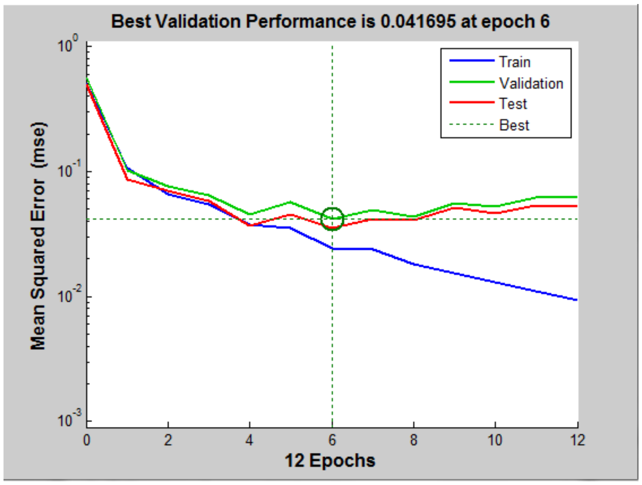


Fig. 5 Performance Plot of neural network (Mean Square Error vs. Epochs).

Regression curve shows the output-target correlation. As much as regression value close to 1, output is more relate with target. Regression values are different in training state, testing state and validation state. Here, regression values are 0.99999, 0.96675 and 0.96743 for training, testing and validation respectively. Overall regression value is 0.98686.

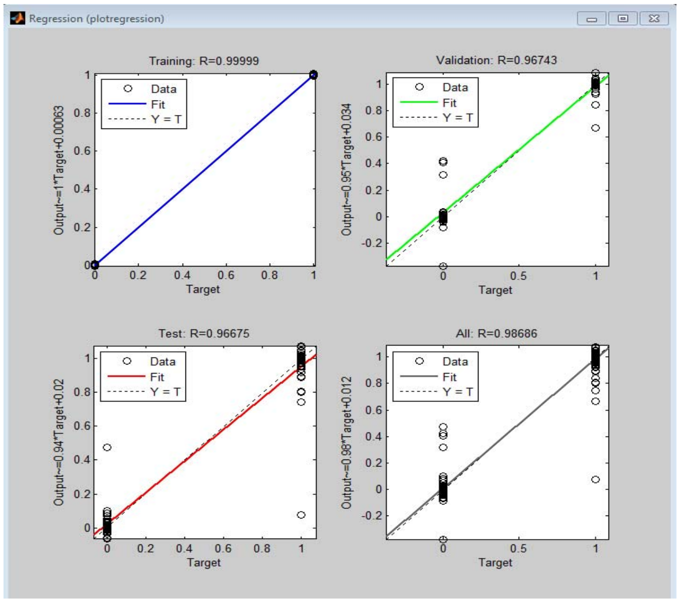


Fig. 6 Regression Plot of Neural Network.

V. Experimental Result

In the experiment, firstly the MFCC features of the speeches of two speakers (Recognized Speaker and one non-Recognized Speaker) were obtained through Feature Extraction process. Databases were created with these MFCC data. Then these MFCC feature databases and the target results of recognized and non-recognized speakers and particular command of recognized speaker were applied to Levenberg-Marquardt neural network for training. The network consists of an input layer, a hidden layer and an output layer.

The system is so designed that, when a particular test command is given, the network firstly compares the test command’s feature data with train data. If they match, the network decides that it is particular recognized speaker and system shows the result. Only after this portion, the network proceeds to determine the particular command of that recognized speaker. But if the test feature data and train data do not match, the network decides that it is non-recognized speaker, hence terminates the program and system shows the result.

The system is tested with four speakers (Speaker 1, Speaker 2, Speaker 3, and Speaker 4) so that it can recognize only Speaker 1 and does not recognize the other speaker’s voices and hence terminate the program. When it recognizes Speaker 1, it continues to the second neural network, which identifies his command. Results are shown in Table 2 and Table 3 respectively. For speaker recognition, the recognition rate is very close to 83% , where for speech recognition exactly 60%.

Table 2 Speaker Recognition Result Table

Speaker	Actual Result	Test Result
Speaker 1	Speaker is Recognized	Speaker is Recognized
Speaker 2	Speaker is non-Recognized	Speaker is non-Recognized
Speaker 3	Speaker is non-Recognized	Speaker is non-Recognized
Speaker 1	Speaker is Recognized	Speaker is non-Recognized
Speaker 4	Speaker is non-Recognized	Speaker is non-Recognized
Speaker 1	Speaker is Recognized	Speaker is non-Recognized

Table 2 Speech Recognition Result Table

Speech	Actual Result	Test Result
BATI JALAO	The command is "BATI JALAO"	The command is "BATI JALAO"
PAKHA BONDHO KORO	The command is "PAKHA BONDHO KORO"	The command is not identified
PAKHA BONDHO KORO	The command is "PAKHA BONDHO KORO"	The command is "PAKHA BONDHO KORO"
BATI JALAO	The command is "BATI JALAO"	The command is not identified
BATI JALAO	The command is "BATI JALAO"	The command is "BATI JALAO"

VI. Conclusion And Future Works

Based on the preliminary analysis of the cepstral features, results indicate several statistical significance to recognize the speech and speaker independently. In pursuit of our primary goal of automatic modeling of speech, we have examined MFCC features-sets for using in artificial machine learning experiments. Further tests using Bangla utterances from more speakers, which are hard to obtain, can confirm the values of the MFCC features as alternative to generally use spectral and cepstral features.

Again, our main target is to help the native people who are disabled or sick so that they can change device states with only voice commands. To accomplish the target, our system needs to be implemented with hardware. So we are working further to combine our system with relay based control circuit through interfacing.

In the system, MFCC feature extraction technique was followed which acts effectively in ideal operating condition. When a mismatch is found between training and testing condition, typically due to background noise, performance of MFCC degrades severely. So, other feature extraction techniques can be applied and compared for betterment of system performance.

Besides, to minimize complexity, no special noise reduction method was included in the system. But practically, the environment surrounding the system may contain a large number of noises. So the inclusion of an effective and strong noise reduction method can make the system perform much better and accurate.

Refrrences

[1] M. M. Rahman, M. F. Khan, M. A. Moni, "Speech recognition front-end for segmenting and clustering continuous bangla speech," Daffodil International University Journal of Science and Technology, vol. 5, issue 1, pp. 67-72, January 2010.

[2] M. A. Hasnat, J. Mowla, M. Khan, "Isolated and continuous bangla speech recognition implementation, performance and application perspective," Seventh International Symposium on Natural Language Processing (SNLP 2007).

[3] M. A. Ali, M. Hossain, M. N. Bhuiyan, "Automatic speech recognition technique for bangla words," International Journal of Advanced Science and Technology, vol. 50, pp. 51-59, January 2013.

[4] H. K. Elminir, M. A. ElSoud, L. M. A. El-Maged, "Evaluation of different feature extraction techniques for continuous speech recognition," International Journal of Information and Communication Technology Research, vol. 2, no. 12, pp. 906-913, December 2012.

[5] W. B. Mikhael, P. Premakanthan, "Speaker verification/recognition and the importance of selective feature extraction: review," 44th IEEE Proceedings on Midwest Symposium on Circuits and Systems, Ohio, vol. 1, pp. 57-61, 2001.

[6] M. Sanaullah, K. Gopalan, "Neural network-based classifier of speech under stress using nonlinear spectral and cepstral features," Illinois/Indiana American Society of Engineering Education Section Conference, Trine University, Angola, issued 6th April 2013.

[7] M. Sanaullah, K. Gopalan , "Distinguish deceptive speech from truthful speech using MFCC features," Proceeding of the WSEAS Conference, Harvard University, Cambridge, USA, pp.167-171, 2013.

[8] C. R. Bharathi, V. Shanthi, "Classification of speech for clinical data using artificial neural network," IJCSI International Journal of Computer Science Issues, vol. 8, issue 6, no. 1, pp. 359-365, November 2011.

[9] M. T. Hagan, H. B. Demuth, M. Beale, Network models and Architecture in Neural Network Design, PWD Publishing Company, China, 1996.