Jurnal Teknologi

Cascade-forward Neural Networks for Arabic Phonemes Based on k-Fold Cross Validation

Nurul Ashikin Abdul Kadir[a], Rubita Sudirman[a*], Nasrul Humaimi Mahmood[a], Abdul Hamid Ahmad[a]

[a]Faculty of Electrical Engineering, Universiti Teknologi Malaysia, 81310 UTM Johor Bahru, Johor, Malaysia

*Corresponding author: rubita@fke.utm.my

Graphical abstract

Place of Articulation	Symbol, Phoneme	Manner of Articulation
1 - Bilabial	[م], /m/	Nasal
2 - Dental	[ن], /n/	Nasal
3- Alveolar	[ر], /r/	Trill
4-Post-alveolar	[ل], /l/	Lateral

Abstract

The study of Malaysian Arabic phoneme is rarely found which make the references to the work is difficult. Specific guideline on Malaysian subject is not found even though a lot of acoustic and phonetics research has been done on other languages such as English, French and Chinese. In this paper, we monitored and analyzed the performance of cascade-forward (CF) networks on our phoneme recognition system of Standard Arabic (SA). This study focused on Malaysian children as test subjects. It is focused on four chosen phonemes from SA, which composed of nasal, lateral and trill behaviors, i.e. tabulated at four different articulation places. Cascade neural networks are chosen as it provide less time for samples processing. The method, *k*-fold cross validation to evaluate each network architecture in *k* times to improve the reliability of the choice of the optimal architecture. Based on this method, namely 10-fold cross validation, the most suitable cascade-layer network architecture in first hidden layer and second hidden layer is 40 and 10 nodes respectively with MSE 0.0402. The training and testing recognition rates achieved were 94% and 93% respectively.

Keywords: Cascade-forward network; nasal; lateral; trill; *k*-fold cross validation

Abstrak

Kajian mengenai penggunaan fonem bahasa Arab dikalangan rakyat Malaysia adalah sangat terhad, ini menyukarkan penyelesaian kajian. Garis panduan yang khusus tidak didapati sedangkan kajian berkaitan akustik dan fonetik sangat pesat dikalangan bahasa asing seperti Bahasa Inggeris, Bahasa Perancis dan Bahasa Mandarin. Bagi kajian ini, kami mengawasi dan menganalisis tahap prestasi jaringan lata ke-depan untuk sistem pengenalan fonem Standard Arab. Kajian ini memfokuskan kanak-kanak Malaysia sebagai bahan kajian. Berfokus pada empat fonem pilihan dari bahasa Arab standad, yang terdiri daripada penghasilan bunyi sengauan, sisian dan getaran, yang bertabur pada empat tempat artikulasi. Kaedah *k*-lipatan pengesahan bersilang untuk menilai setiap struktur jaringan sebanyak *k* kali bagi meningkatkan kebolehpercayaan pilihan jaringan yang optimum. Berdasarkan kaedah ini, iaitu 10-lipatan pengesahan bersilang, struktur rekabentuk jaringan lapisan lata bagi lapisan pertama yang tesembunyi dan lapisan kedua tersembunyi masing-masing adalah 40 dan 10 nod di mana MSE 0.0402. Kadar pengecaman latihan dan pengujian masing-masing adalah 94% dan 93%.

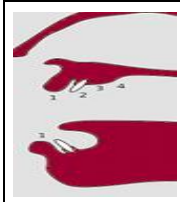*Kata kunci*: jaringan lata kedepan; sengauan; sisian; getaran; *k*-lipatanpengesahanbersilang

∎1.0 INTRODUCTION

Neural network (NN) technology is widely spread as the commercialize technology in various applications. Nowadays the NN technologies are becoming essential in electronics, medical, telecommunications, financial, speech and other industries as a method to perform complex functions. Specifically in the speech applications, neural network is introduced as a function of pattern recognition (speech recognition) and text-to-speech synthesis. The use of NN in speech applications has been proven through several studies.[1-9]

The networks have a few architectural properties which include the number of layers, the number of neurons and the chosen input and output processing functions. There are several inputs under consideration of this study, which represents the features of each speech samples in Linear Predictive Coding (LPC) form. In LPC model as (1), speech signals are compressed, which beneficial as the inputs for neural network and undergo training process to achieve corresponding target.[10]

$$s(n) = a_1s(n-1) + a_2s(n-2) + \ldots + a_ps(n-p) \qquad (1)$$

Figure 1 Standard Arabic and corresponding place of articulation

Place of Articulation	Symbol, Phoneme	Manner of Articulation
1 - Bilabial	[م], /m/	Nasal
2 - Dental	[ن], /n/	Nasal
3 - Alveolar	[ر], /r/	Trill
4 - Post-alveolar	[ل], /l/	Lateral

where, $s(n)$ is speech sample at time n. a_1, a_2 and a_p are assumed constant over the speech analysis frame while minimizing the mean-square error over the entire speech sample and p is the most current samples or the order of LPC.

The numbers of hidden neurons, h in hidden layers were chosen based on (2).[11]

$$h \geq (T-1)/(i+2) \qquad (2)$$

where T is the number of training examples and i is the number of network inputs.

This study concerns the neural network performance that is cascade-forward (CF) networks which was developed in Matlab.

k-fold cross validation (k-fold CV) is the best NN architecture to be relied on.[12] As training session of NN tend to learn the most gross behavior of the training data and ignore subtleties.[13] By dividing the training data into k fold, the average of all k accuracies is known as the k-fold CV accuracy. k-fold CV performs to estimate the performance of the predictive NN model. The estimated performance is the mean of these errors.[14]

According to International Phonetics Alphabet (IPA) system standard, Standard Arabic (SA) composed of six pronouncing behaviors (fricative, plosive, nasal, lateral, trill and approximant). In order to create a small vocabulary speech recognition system, only nasal, lateral and trill were under considerations. The articulation places are originated from frontal part of the mouth as shown in Figure 1. Nasal is produced with a lowered velum in the mouth, allowing air to flow out through the nose. Lateral is produced by raising the tip of the tongue against the roof of the mouth so that the airstream flows past one or both sides of the tongue. While trill, is produced by tongue vibration against alveolar.[15-17]

1.1 Previous Research

It is suggested in the literature that great efficiency improvements can be made in the development of prosody models for languages using cascade architecture.[7] The model was used to predict three prosodic variables which are phrase-boundary strength, word prominence and phoneme duration. There are six languages have been investigated, namely Dutch, English, French, German, Italian and Spanish with recognition rate of 94.9 %, 95.5 %, 91.0%, 96.3 %, 97.0 % and 97.3 % are achieved respectively.

In 2007, a research was done to identify Cipher System from cipher texts. In this research, the accuracy of 90.9 % in cascade network is higher when compared to multi-layer back-propagation network with accuracy of 73.8 %.

[13]The mean of the MSE, accuracy and precision between k-fold CV and neural networks are compared to determine the optimal architecture that suits their applications. MSE for each architecture were recorded for every training time to find a neural network architecture that shows the lowest difference between k-fold CV MSE. By using $k = 10$, overall accuracy was 99 %.

[18]A 10-fold CV was applied in the study of Arabic stop words elimination text classification algorithms. The classifier was studied along with Support Vector Machine and Naïve Bayesian. For Standard Arabic dataset used, accuracy was 91.37 % and error rate 8.62. Results after eliminating the stop words were 90.9 % and error rate 9.1 respectively.

Table 1 summarized previous research findings on cascade networks and k-fold cross validation.

Table 1 Previous research findings

Study	Findings
7	Recognition rate using cascade architecture were: Dutch – 94.9 % English – 95.5 % French – 91.0 % German – 96.3 % Italian – 97.0 % Spanish – 97.3 %
8	Cipher text recognition rate of 90.9 % using cascade network
13	$k = 10$, MSE = 99 %
18	$k = 10$, accuracy = 91.37 %

■2.0 EXPERIMENTAL SETTINGS

A recording session involved primary school children age eight to eleven years was conducted in a quiet room. There were 75 children involved, which were 45 girls and 30 boys. These children are native Malaysian who in their early age had been taught the basic Arabic words by learning Quran.

The requirements to perform this study include:
1. Recording software: Easy Hi-Q Recorder with sampling rate of 16 kHz.
2. Analysis software: Goldwave and Speech Filing System (SFSWin) version 1.7 2008.
3. Recording type and format: *.wav, 16-bit mono.
4. Recording device: External Mic.
5. Recording equipment: A notebook with built-in microphone.

During recording, the children were taught to utter the letters appropriately in one tape. Therefore, a total of 75 sets of Arabic phonemes were collected. By using the analysis software, the speech was cut and grouped into its utterances. For this study, 300 (4 phonemes × 75 subjects) samples were collected to be trained. The resulting 75 sets Arabic phonemes were divided into training (70 %) and testing (30 %) set for neural network system. The results of eliminating mispronounced samples which heard manually by *Maahad Tahfiz* school's teacher was required.

2.1 Data Processing

By applying the digital speech processing technique to all of those samples, a set of pre-processing samples were obtained. The pre-processing stage is needed to ensure the signals are less susceptible to noise. Therefore, a visual image of a speech signal can be seen through a spectrogram after applying FFT technique.

Formant frequencies can be seen through a spectrogram. These valuable methods are proven to be effectively and fastest way to obtain the formants. This process was done to make sure the selected dataset for training purpose of NN is reliable to be the baseline for this study.

From the spectrogram, the formants (F_1, F_2, F_3 and F_4) were studied and samples which fall in the average formants values were extracted.[19-20] These frequencies were obtained as in (3).

$$F_N = (2N-1)c / 4L \qquad (3)$$

This conventional equation is used to calculate the Nth formant frequencies value where N is the formant; c is the speed of sound in warm and moist air (approximately 35000 cm/sec); and L is the length of the vocal tract in cm.

Besides, *k*-fold cross validation are used to evaluate the performance of cascade networks.[13-18] *k* is set to 10, namely 10-fold cross validation. The predicted MSE are calculated.

2.2 Neural Networks Training Process

Only selected samples (4 phonemes of < 75 subjects) from training set were used and converted to LPC before being trained in the networks. By referring to equation (2), the number of hidden neurons must be at least 15, if all training datasets are used since (75 subjects, 4 phonemes + 1) / (19 LPC + 2) ≈ 15. Neural networks with different numbers of hidden-neuron have been trained separately and the performance was evaluated. The following are the architectures of the neural networks:

   i. No. of phonemes: 4
   ii. Analysis Software: Matlab
   iii. Network type: Cascade-forward network.
   iv. Performance function: Mean-square error (MSE)
   v. No. of hidden neurons: 10, 20, 30, 40, 50, 60, 70.
   vi. No. of iterations (Epochs): 1000
   vii. Transfer function for hidden layers: Log-sigmoid
   viii. No. of hidden layers: 2.
   ix. Network training function: Scaled conjugate gradient method.

The training process was repeated up to 50 times. The highest training recognition rates for all neurodes combinations were chosen and tested, to know their testing recognition rates. MSEs for all networks were calculated. The MSE produced during networks training sessions are compared with MSE of 10-fold cross validation. The correspond MSE of k-fold cross validation and cascade networks architecture are chosen as the optimal NN architecture that can be relied on for further application of the recognition system.

■3.0 RESULTS AND DISCUSSION

The characteristics of speech samples are observed through spectrogram. Formants are collected and summarized in Figure 2 to Figure 5.

Only 40 subjects formants frequency for each phoneme are plotted to identify their characteristics through spectrograms. The formants average values are summarized according to its place of articulation as in Table 2 to Table 4. Range of formant frequencies also included for all consonants involved.

Figure 2 shows four formants plotted of selected recorded samples for phoneme /m/, [م]. In Figure 2, only F_1s for bilabial /m/, [م], seems to be scattered in almost a linear line below 1000 Hz. Other three formants scattered in higher range. Example, F_2s are ranging from 700 Hz to 4231 Hz, F_3s 2000 Hz to 5446 Hz and F_4s are between 3078 Hz to 6000 Hz. Nevertheless, the average formants values are considered as 543 Hz, 2519 Hz, 4231 Hz and 5446 Hz for F_1 to F_4 as in Table 2.

As seen in Figure 3, only F_1s of dental /n/, [ن], seems to be scattered in almost a linear line below 1000 Hz. Other three formants scattered at higher range. Such as, F_2s are ranging from 1000 Hz to 5439 Hz, F_3s 2369 Hz to 6000 Hz and F_4s in between 3455 Hz to 6466 Hz. Nevertheless, the average formants values are considered as 411 Hz, 2561 Hz, 4296 Hz and 5306 Hz for F_1, F_2, F_3 and F_4 respectively.

The average value of each phoneme is summarized in Table 2. By neglecting the changes of F_1 and F_4, it can be seen that F_2 and F_3 are increasing from bilabial to dental place of articulation in Table 2. It is just a light increment of bilabial-nasal's F_2 which is 2519 Hz that a bit higher than dental-nasal's F_2 which is 2561 Hz. Furthermore, F_3 for pronouncing phoneme originated at bilabial to dental increased from 4231 Hz to 4296 Hz.

The difference between /m/, [م], and /n/, [ن], nasal consonants pronunciation is that the lip rounding when the phoneme pronounced.

From the spectrogram, the average value for /l/, [ل], F_1 is 458 Hz, F_2 is 2247 Hz, F_3 is 3945 Hz and F_4 is 5437 Hz as shown in Figure 4. The distribution of F_1s and F_2s are along y-axis of 500 Hz and 2071 Hz respectively. While F_3s and F_4s distribution are ranging between 3000 Hz to 5509 Hz and 3500 Hz to 6323 Hz appropriately. The average values of phoneme /l/, [ل], are summarized in Table 2.

The average value for /r/, [ر], F_1 is 514 Hz, F_2 is 1590 Hz, F_3 is 2560 Hz and F_4 is 5147 Hz as shown in Figure 5. The distribution of F_1s, F_2s and F_3s are along y-axis of 514 Hz, 1590 Hz and 2560 Hz respectively. While F_4s distributions are ranging between 3455 Hz to 6266 Hz appropriately. The formants average for phoneme /r/, [ر] is summarized in Table 2.

Table 2 The formants averages for nasal phonemes

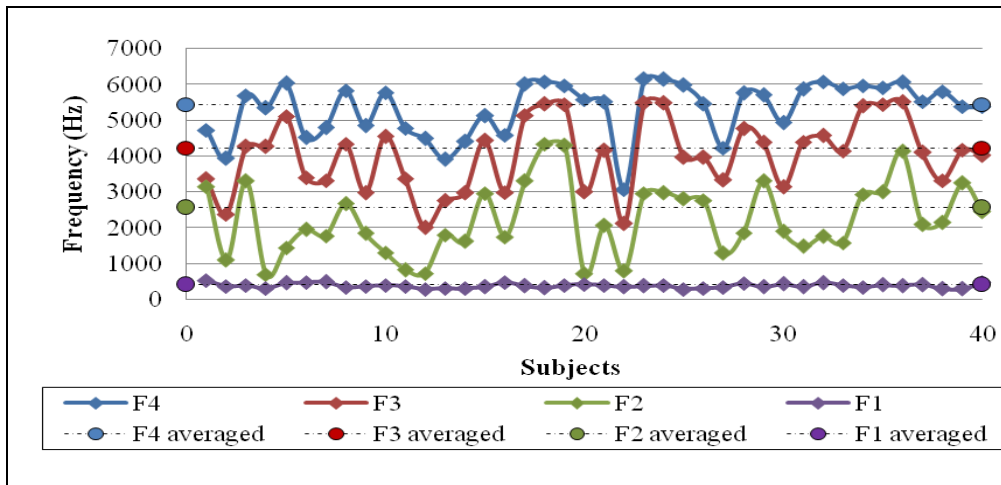| Phonemes Behaviour	Place of Articulation	Phoneme, Symbol	Formants for Voiced Sound (Hz)			
			F_1	F_2	F_3	F_4
Nasal	Bilabial	/m/, [م]	543	2519	4231	5446
	Dental	/n/, [ن]	411	2561	4296	5306
Lateral	Alveolar	/r/, [ر]	514	1590	2560	5147
Trill	Alveolar	/r/, [ر]	514	1590	2560	5147

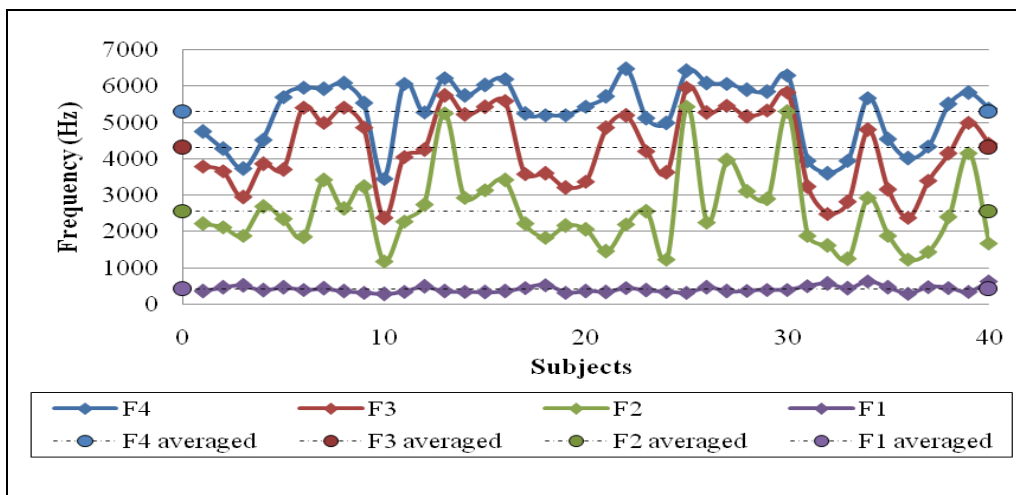**Figure 2** Formants distribution of bilabial /m/, [م]

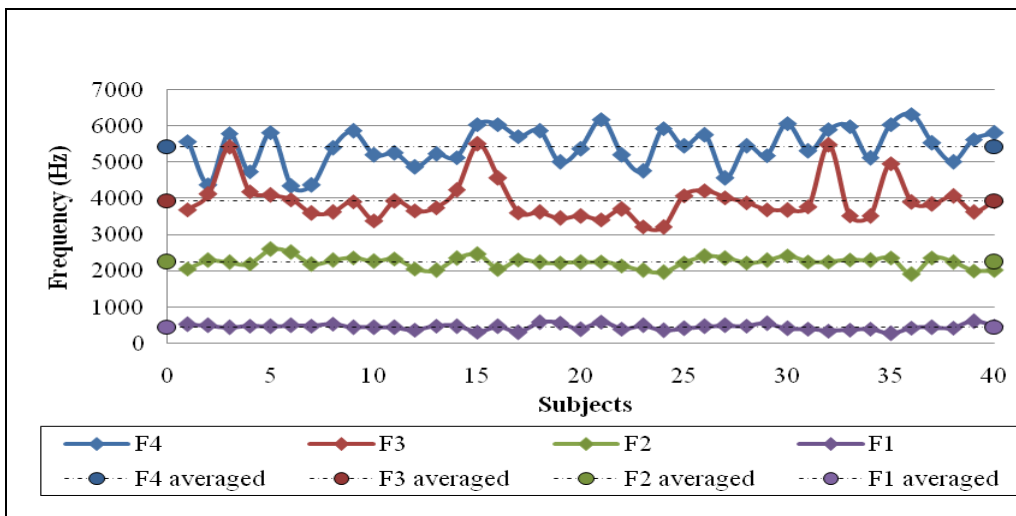**Figure 3** Formants distribution of dental /n/, [ن]

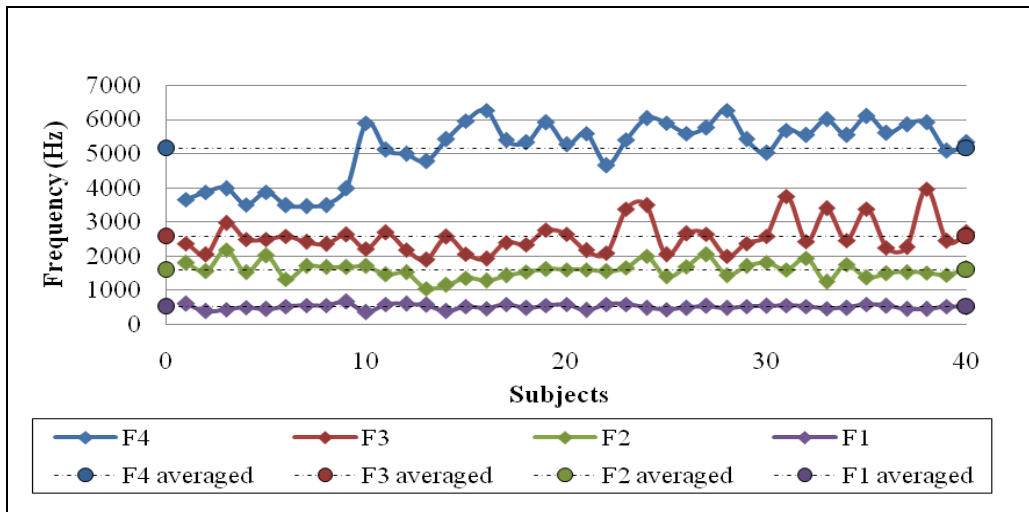**Figure 4** Formants distribution of /l/, [ل]

Figure 5 Formants distribution of alveolar /r/, [ɹ]

Table 3 shows MSE obtained for every fold and average MSE for the samples. The average MSE calculated using *k*-fold cross validation method is 0.0366. According to Table 4, the highest reachable training and testing recognition rates among those 28 hidden neurons combination are 95 % and 93 % respectively.

Table 3 MSE yields from 10-fold cross validation

Fold	Mean-Square Error (MSE)
1	0
2	0.0277
3	0.0256
4	0.0312
5	0.0291
6	0.0400
7	0.0372
8	0.0517
9	0.0591
10	0.0640
Average	0.0366

Total of 4 hidden neurons pairs obtained 95% during training phase, with pairs of 40-20, 50-10, 70-40 and 70-50 hidden neurons in first and second hidden layer. In total of 60, 60, 110 and 120 number of hidden neurons respectively. Also 4 hidden neurons pairs obtained 93 % during testing phase, with pairs of 30-30, 40-10, 50-20 and 60-50 hidden neurons in first and second hidden layer. In total of 60, 50, 70 and 110 number of hidden neurons respectively.

The least testing recognition rate is 82 % that resulted from 20-10, 30-20, 60-30 and 70-40 hidden neurons pairs. While the leasttraining recognition rate is 92 % which resulted from 10-10, 30-20, 40-30, 50-30, 60-50, 70-60 and 70-70 hidden neurons pairs. Hidden neurons pairs that produced recognition rates higher than 90 % are 30-30, 40-10, 50-20 and 60-50. Total numbers of hidden neurodes used are 60, 50, 70 and 110 respectively, while the MSE are 0.0319, 0.0402, 0.0265 and 0.0413 respectively.

Table 4 MSE and recognition rate for cascade networks

Neurodes in First Hidden Layer	Neurodes in Second Hidden Layer	MSE	Training Recognition Rate (%)	Testing Recognition Rate (%)
10	10	0.0459	92	86
20	10	0.0125	90	82
20	20	0.0403	93	89
30	10	0.0270	93	89
30	20	0.0293	92	82
30	30	0.0319	94	93
40	10	0.0402	94	93
40	20	0.0202	95	86
40	30	0.0428	92	89
40	40	0.0289	94	86
50	10	0.0468	95	89
50	20	0.0265	94	93
50	30	0.0578	92	89
50	40	0.0209	90	89
50	50	0.0271	94	89
60	10	0.0269	93	89
60	20	0.0225	94	86
60	30	0.0178	90	82
60	40	0.0204	94	89
60	50	0.0413	92	93
60	60	0.0192	93	86
70	10	0.0291	94	86
70	20	0.0362	93	89
70	30	0.0269	93	89
70	40	0.0195	95	82
70	50	0.0157	95	86
70	60	0.0325	92	89
70	70	0.0676	92	86

In order to choose the best NN architecture, based on literature [12], was based on MSE yielded by *k*-fold cross validation method, which for this study was 0.0366. Therefore, the least difference of MSE obtained from *k*-fold CV method and NN training is chosen as the best network architecture. The criteria suits hidden neurons pair of 40-10, which the MSE only differ by 0.0036 and less hidden neurons needed.

■4.0 CONCLUSION

In conclusion, the characteristics of every Standard Arabic (SA) consonants were identified by implementing Fast-Fourier Transform (FFT) and finding the formant frequencies from the signals representation of spectrograms. 10-fold cross validation was used to build a reliable training method and the estimated MSE for the developed system was 0.0366. A system for recognizing Arabic phonemes sound pronunciation using neural networks for pattern recognition and classification was successfully developed. The chosen NN architecture was 40-10 hidden neurons with 0.0402 MSE and training accuracy of 94 % and testing accuracy of 93 % for network combination of nasal, lateral and trill consonants.

Acknowledgement

References

[1] Z. J. Xiang, and G. Bi. 1990. A neural network model for Chinese speech synthesis. *IEEE Int. Symp. on Circuits And Systems*. 1859–1862.

[2] P. G. J. Lisboa. 1992. *Neural Networks Current Application*. Chapman & Hall.

[3] B. A. St. George, E. C. Wooten, and L. Sellami. 1999. Speech Coding and Phoneme Classification using Matlab and Neural Networks. *Information Sciences Elsevier*. 90: 109–119.

[4] J. Frankel, K. Richmond, S. King, and P. Taylor. 2000. An Automatic Speech Recognition System using Neural Networks and Linear Dynamic Models to Recover and Model Articulatory Traces. *Proceeding ICSLP*.

[5] H. N. Ting. 2002. *Speech Analysis and Classification Using Neural Networks For Computer-based Malay Speech Therapy*. Malaysia: UTM Malaysia.

[6] C. L. Tan and A. Jantan. 2004. Digit Recognition Using Neural Networks. *Malaysian Journal of Computer Science*. 17(2): 40–54.

[7] J. W. A.Fackrell, , H. Vereecken, J-P Martens, and B. V. Coile. 1999. Multilingual Prosody Modeling using Cascades Regression Trees and Neural Networks. *Sixth European Conference on Speech Communication and Technology*. 1835–1838.

[8] B. Chandra and P. P. Varghese. 2007. Applications of Cascade Correlation Neural Networks for Cipher System Identification. *World Academy of Science, Engineering and Technology 26*. 312–314.

[9] J. Macias-Guarasa, J. M. Montero, J. Ferreiros, R. Cordoba, R. San-Segundo, J. Gutierrez-Arriola, L. F. D'haro, F. Fernandez, R. Barra and J. M. Pardo. 2009. Novel Applications Of Neural Networks In Speech Technology Systems: Search Space Reduction And Prosodic Modeling. *Intelligent Automation And Soft Computing*. 15(4): 631–646.

[10] L. R. Rabiner, and R. W. Schafer. 2007. *Introduction to Digital Speech Processing, Foundations and Trends in Signal Processing Vol. 1 Issue 1-2*. USA. The Essence of Knowledge.

[11] N. K. Kasabov. 1996. *Foundations of Neural Network, Fuzzy Systems, and Knowledge Engineering*. 1996. London. The MIT Press Cambridge.

[12] T. Andersen and T. Martinez. 1999. Cross Validation and MLP Architecture Selection. IEEE International Joint Conference on Neural Networks IJCNN'99. 3: 1614–1619 vol. 3.

[13] J. L. Wright. 2010. Neural Network Architecture Selection Analysis with Application to Cryptography Location. *IEEE World Congress on Computational Intelligence*.

[14] R. O. Duda, P. E. Hart and D. G. Stork. 2001. *Pattern Classification*. John-Wiley.

[15] A. Hassan. 1984. *Linguistik Am Untuk Guru Bahasa*. Malaysia. Fajar Bakti.

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

[17] K. Johnson. 2003. *Acoustic and Auditory Phonetics 3rd Edition*. 2003. United Kingdom. Blackwell Publishing.

[18] W. B. Al-Shargabi, Al-Romimah and F. Olayah. 2011. A Comparative study for Arabic text classification algorithms based on stop words elimination. *Proceedings of International Conference on Intelligent Semantic Web-Services and Applications*.

[19] N. A. Abdul-Kadir and R. Sudirman. 2011. Difficulties of Standard Arabic Phonemes Spoken by Non-arab Primary School Children based on Formant Frequencies. *J. Computer Science*. 7: 1003–1010.

[20] N. A. Abdul-Kadir, R. Sudirman, and N. M. Safri. 2010. Modeling of the Plosive Consonants Characteristics based on Spectrogram. *Proceeding of the 4th Asia International Conference on Mathematical/Analytical Modeling and Computer Simulation*. 282–285.