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

Nurul Ashikin Abdul Kadir, Rubita Sudirman, Nasrul Humaimi Mahmood, Abdul Hamid Ahmad

*Faculty of Electrical Engineering, Universiti Teknologi Malaysia, 81310 UTM Johor Bahru, Johor, Malaysia

*Corresponding author: rubita@fke.utm.my

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

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.

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.

\[
s(n) = a_0s(n-1) + a_2s(n-2) + \ldots + a_ps(n-p) \quad (1)
\]
where, $s(n)$ is speech sample at time $n$. $a_1$, $a_2$ and $a_3$ 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).\(^\text{11}\)

$$h \geq \frac{(T-1)}{(i+2)} \quad (2)$$

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.\(^\text{12}\) As training session of NN tend to learn the most gross behavior of the training data and ignore subtleties.\(^\text{13}\) 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.\(^\text{14}\)

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.\(^\text{15-17}\)

**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.\(^\text{7}\) 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 %.

\(^\text{18}\)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**

<table>
<thead>
<tr>
<th>Study</th>
<th>Findings</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>Cipher text recognition rate of 90.9 % using cascade network</td>
</tr>
<tr>
<td>13</td>
<td>$k = 10$, MSE = 99 %</td>
</tr>
<tr>
<td>18</td>
<td>$k = 10$, accuracy = 91.37 %</td>
</tr>
</tbody>
</table>

**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.
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 (F₁, F₂, F₃ and F₄) were studied and samples which fall in the average formants values were extracted. These frequencies were obtained as in (3).

\[ F_n = \left(2N-1\right)c / 4L \]

(3)

This conventional equation is used to calculate the \(N^{th}\) 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. \(k\) 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) \(\approx\) 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.

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/, [a]. In Figure 2, only F₁s for bilabial /m/, [a], seems to be scattered in almost a linear line below 1000 Hz. Other three formants scattered in higher range. Example, F₂s are ranging from 700 Hz to 4231 Hz, F₃s 2000 Hz to 5446 Hz and F₄s 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 F₁, F₂ as in Table 2.

As seen in Figure 3, only F₁s of dental /n/, [j], seems to be scattered in almost a linear line below 1000 Hz. Other three formants scattered at higher range. Such as, F₂s are ranging from 1000 Hz to 5439 Hz, F₃s 2369 Hz to 6000 Hz and F₄s 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 F₁, F₂, F₃ and F₄ respectively.

The average value of each phoneme is summarized in Table 2. By neglecting the changes of F₁ and F₄, it can be seen that F₂ and F₃ are increasing from bilabial to dental place of articulation in Table 2. It is just a light increment of bilabial-nasal’s F₂ which is 2519 Hz that a bit higher than dental-nasal’s F₂ which is 2561 Hz. Furthermore, F₃ for pronouncing phoneme originated at bilabial to dental increased from 4231 Hz to 4296 Hz.

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

From the spectrogram, the average value for /l/, [l], F₁ is 458 Hz, F₂ is 2247 Hz, F₃ is 3945 Hz and F₄ is 5437 Hz as shown in Figure 4. The distribution of F₁s and F₃s are along y-axis of 500 Hz and 2071 Hz respectively. While F₂s and F₄s distribution are ranging between 3000 Hz to 5509 Hz and 3500 Hz to 6323 Hz respectively. The average values of phoneme /l/, [l], are summarized in Table 2.

The average value for /r/, [r], F₁ is 514 Hz, F₂ is 1590 Hz, F₃ is 2560 Hz and F₄ is 5147 Hz as shown in Figure 5. The distribution of F₁s, F₃s and F₄ are along y-axis of 514 Hz, 1590 Hz and 2560 Hz respectively. While F₂s distributions are ranging between 3455 Hz to 6266 Hz appropriately. The formants average for phoneme /r/, [r] is summarized in Table 2.

<table>
<thead>
<tr>
<th>Phonemes</th>
<th>Place of Articulation</th>
<th>Symbol</th>
<th>Formants for Voiced Sound (Hz)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nasal</td>
<td>Bilabial</td>
<td>/m/, [a]</td>
<td>F₁ 543 F₂ 2519 F₃ 4231 F₄ 5446</td>
</tr>
<tr>
<td></td>
<td>Dental</td>
<td>/n/, [j]</td>
<td>F₁ 411 F₂ 2561 F₃ 4296 F₄ 5306</td>
</tr>
<tr>
<td>Lateral</td>
<td>Alveolar</td>
<td>/r/, [l]</td>
<td>F₁ 514 F₂ 1590 F₃ 2560 F₄ 5147</td>
</tr>
<tr>
<td>Trill</td>
<td>Alveolar</td>
<td>/r/, [j]</td>
<td>F₁ 514 F₂ 1590 F₃ 2560 F₄ 5147</td>
</tr>
</tbody>
</table>
Figure 2  Formants distribution of bilabial /m/, [ŋ]

Figure 3  Formants distribution of dental /n/, [ŋ]

Figure 4  Formants distribution of /l/, [ɻ]
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

<table>
<thead>
<tr>
<th>Fold</th>
<th>Mean-Square Error (MSE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0.0277</td>
</tr>
<tr>
<td>3</td>
<td>0.0256</td>
</tr>
<tr>
<td>4</td>
<td>0.0312</td>
</tr>
<tr>
<td>5</td>
<td>0.0291</td>
</tr>
<tr>
<td>6</td>
<td>0.0400</td>
</tr>
<tr>
<td>7</td>
<td>0.0372</td>
</tr>
<tr>
<td>8</td>
<td>0.0517</td>
</tr>
<tr>
<td>9</td>
<td>0.0591</td>
</tr>
<tr>
<td>10</td>
<td>0.0640</td>
</tr>
<tr>
<td>Average</td>
<td>0.0366</td>
</tr>
</tbody>
</table>

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 least training 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 neurons used are 60, 50, 70 and 110 respectively, while the MSE are 0.0319, 0.0402, 0.0265 and 0.0413 respectively.

In order to choose the best NN architecture, based on literature\textsuperscript{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 pairs 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

This paper is a part of a publication series on Research and Development in Signal, Image and Sensors in Biomedical Engineering Applications. The authors are indebted to MOSTI and Universiti Teknologi Malaysia for supporting and funding this study (Vote 79368) and research university grant (Vote 00J05) respectively. We also would like to express our gratitude to bio-Medical Instrumentation and Electronics Research Group (bMIE) members for their ideas and comments to the study.

References