Spoken Arabic Vowel Recognition Using ANN

Fatimah Mohammed Aloqayli
Computer Engineering Department
King Saud University
Riyadh, Saudi Arabia
fatimahaloqayli@gmail.com

Yousef Ajami Alotaibi
Computer Engineering Department
King Saud University
Riyadh, Saudi Arabia
yaalotaibi@ksu.edu.sa

***Abstract* — In general, the phonemes of any language can be classified into two categories: vowels that contain no major air restriction through the vocal tract, and consonants that involve a significant restriction, and are therefore weaker in amplitude, and often "noisier," than vowels. Despite the importance of analyzing vowel phonemes in Arabic language, not a lot of research exist in the published literature yet. Consequently, this study is concerned specifically with the analysis of vowels in the Modern standard Arabic (MSA) dialect. The values of the first three formant frequencies and their derivatives in these vowels are analyzed for the purpose of automatic recognition of the six vowels of MSA through the use of an artificial neural network (ANN) based recognition system. In this paper network was tested with both one and two hidden layers, each with varying numbers of neurons. As an outcome of our experiments with four network architectures, we were able to compare, analyze and discuss the outcomes of these four network architectures.**

Keywords - Classical Arabic; ANN; Formant; MSA; Vowels.

I. INTRODUCTION

Vowels are those fundamental speech units present in every spoken language, produced with a relatively open vocal tract, and an airstream that is not severely impeded. The resulting acoustic signal is therefore relatively loud. In addition, vowels are usually produced with vocal fold vibration. All vowels are phonated, and are normally among the phonemes of largest amplitude [1].

Arabic is a Semitic language, and one of the world's oldest languages. Modern standard Arabic (MSA) consists of 36 phonemes, of which six are vowels, two are diphthongs, and 28 are consonants. The six vowels are /a, i, u, a:, i:, u:/, and are further classified by the duration of their sound as short and long. The short vowels are /a/ (Short fatha), /i/ (Short kasrah) and /u/ (Short dummah), while the long vowels are /a:/ (Long fatha), /i:/ (Long kasrah) and /u:/ (Long dummah). These six vowels can be found in almost all Arabic dialects [2]–[4].

All Arabic syllables must contain at least one vowel, and that means almost 60 to 70% of Arabic speech consists of vowels [5]. As a result, the analysis and investigation of vowels in Arabic is very important when designing reliable and robust speech processing systems. Although the importance of analyzing vowel phonemes in the Arabic language is known, most of the reported studies to-date have been conducted on Arabic language and digital speech processing in general, with only a few focusing specifically on Arabic vowels.

The acoustic properties of phonemes play an important role in phoneme analysis and investigation. One of these acoustic properties is the set of formant frequencies. Formant frequencies are defined as the resonance frequencies of the vocal tract, and are considered to be representative of the underlying phonetic knowledge of speech. The first five formants are denoted as F1, F2, …, F5 [6]. Vowels can be distinguished by the location of their formant frequencies, the first three formants (F1, F2 and F3) generally sufficient for the task. This is because the frequencies of higher formants, such as F4 and F5, seem to be specific to the speaker, and may therefore provide information about the identity of the speaker, rather than the vowel itself [7].

Several analytical techniques are useful the in linguistic field, one of them is artificial neural network (ANN) is a mathematical model, used to perform a particular function, that works much like the human brain. The most beneficial characteristic of ANNs for solving the automatic speech recognition (ASR) problem are their fault tolerance and nonlinear properties [8]. An ANN is made up of many artificial neurons, which are connected together in accordance with an explicit network architecture. The objective of the neural network is to convert its inputs into labeled outputs. One of the most important models of neural networks is the multilayer perceptron (MLP), a feedforward network with one or more layer(s) of nodes hidden between the input and output nodes [9].

The rest of the paper is arranged as follows: A brief literature review of related work, along with the objectives

2473-3539/17 $31.00 © 2017 IEEE
DOI 10.1109/EMS.2017.24

of this study, are presented in Section II. The experimental framework used for this study is described in Section III, and in Section IV, the results are discussed. Finally, our conclusions, and directions of future work, are stated in Section V.

II. RELATED WORK AND OBJECTIVES

There have been a number of studies published related to the spectral analysis of Arabic vowels, with interesting results. For example, a work published in 1970 by Paddock [10] is one of the earliest studies to show interest in analyzing Arabic vowels. The author synthesized vowels out of the formants, and investigated the response of Russian and Egyptian subjects to these vowels.

Another work, by Newman et al. [11], had the goal of conducting a formant based analysis of the six vowels of connected speech in MSA Arabic and Egyptian dialects. Interestingly, their work depended on only one speaker for each part of their work, and their findings did not confirm the existence of a high classical style as an acoustically 'purer' variety of MSA Arabic.

Formant based analysis of the six vowels of the MSA was also carried out for connecting speech during work by Seddiq and Alotaibi [12], for the purpose of vowel identification and characterization. The corpus was built by extracting segments of the desired syllables from recorded Quranic recitation. In this study, the authors compared the values that were captured from the first three formant frequencies with results that were achieved in similar, previously published studies. The comparison was performed from a geometric perspective, using the Euclidean distance. The results of this geometric comparison were found to be consistent with the visual inspection of the vowels.

Another study was published by Alghamdi [13], where he investigated whether the six vowels are at the same phonetic level when spoken by speakers of different Arabic dialects, with the assumption that they are the same at the phonological level. His result found that the phonetic implementation of the MSA vowel system differs according to the spoken dialect. In addition, significant work has been performed by Iqbal et al. [14], whose paper provided an analysis of cues with which to identify Arabic vowels. Their algorithm extracted the formants of pre-segmented recitation audio files and categorized the vowels based on these extracted formants. The scope of their work focused exclusively on short vowels, using a database of THQ recitations that was built in-house for the purpose of that work. The vowel identification system they developed showed up to 90% average accuracy over continuous speech files comprising around 1000 vowel tokens.

Some researchers consider Arabic vowels to number eight in total, counting the two diphthongs as vowels, and this is normally considered to be the case for MSA. Thus, the number of vowels was considered to be eight in the study that was published by Alotaibi and Amir [15]. The goals of this study were to investigate these vowels using both speech recognition and formant based analysis with a spectrographic technique. A recognition system was built to classify and determine similarities and differences among the eight vowels, and the overall accuracy of the speech recognition system was determined to be 91.6%. Their results showed that both methods of investigations found a high degree of similarity between short Arabic vowels and their long counterparts.

There have also been a number of studies published, with interesting results, that target vowels in Arabic dialects other than MSA. One of these studies was done by Yaser Natour et al. [16], and focused on the formant frequency of the six Arabic vowels as an acoustic measure. For this study, Jordanian Arabs (100 females, 100 males, 100 children) were targeted, and the authors showed that male frequencies were significantly different in comparison to the frequencies of females and children. On other hand, while differences between females and children were found in F1, none were observed in either F2 or in F3. In addition that, when the results of this study were compared with other available data, they found formant frequencies for males to be generally lower in F1 and F2, and higher in F3, whereas female formant frequencies were generally higher in F1, and lower in F2 and F3. The formant frequencies of children were generally found to be lower across all formants (F1, F2 and F3).

The objective of this paper is to perform formant based analysis for the purpose of recognition of the six vowels of the MSA. The importance of this work is to expand on the auditory description of the Arabic vowels system in MSA. In addition to this, the results of such study would be a first stage of a bigger project that aims to compare the vowels in MSA to the vowels in other languages.

III. EXPERIMENTAL FRAMEWORK

A. *Corpus*

MSA largely follows the grammatical standard of the classical Arabic dialect and has basically the same vocabulary which excludes words used in any spoken dialect. The Holy Quran (THQ) dialects derived from classical Arabic, accordingly we have decided to extract our

speech corpus from THQ recitations due to the need to build a corpus that, for the purpose of this research, depends on correct pronunciations.

THQ is the holy book of Islam, and its recitation is governed by the rules of a science that is concerned with reaching perfect pronunciation and recitation of the Quran: the science of Tajweed. The Quran consists of 114 "Surahs", typically translated as "Chapters." Each Surah can be read independently, and not necessarily in order, and consists of partitioned verses called "Ayat" [17].

B. Syllable

For the six vowels studied in this paper, the Consonant-Vowel-Consonant (CVC) syllable was considered, where V indicates a vowel (long or short), and C indicates a consonant. The consonant preceding the vowel was chosen to be /t/, while the consonant succeeding the vowel was chosen to be /n/. Therefore, for the six vowels there are six syllables, i.e., /tan/, /tin/, /tun/, /ta:n/, /ti:n/ and /tu:n/.

C. Database

Acoustic analysis was performed on 108 speech files downloaded from official and authenticated websites [18]. These audio files were recorded by six well-known male reciters, shown in Table I. For each one of the six syllables mentioned earlier, three different words that contain the syllable are chosen from THQ. After that, the 18 words are segmented out of each one of the six recorded recitations to give 108 segments constituting the corpus of this research experiment.

The segments were analyzed using PRAAT software [19]. For each vowel phoneme, we obtained the first three formant frequencies F1, F2, and F3, and then used those results to calculate F2-F1 and F3-F2. For this study, we are considering 10 frames, the considered frame readings were from the middle of the vowel (i.e., within the stable interval). The frame duration was set to 30 milliseconds with a step size of 20 milliseconds (i.e., with overlap of 10 milliseconds), therefore the feature vector for each vowel in the carrier word is 10 frames multiplied by 5 formants (F1, F2, F3, F2-F1, F3-F2), and, as a result, we obtained 50 features per phoneme.

TABLE I. RECITERS NAMES AND ASSIGNED CODE

Reciter Name	***Reciter code***
Saud Al Shuraim	01
Ali Al Hodaifi	02
Tawfiq Al Sayegh	03
AbdulRahman Al Sudais	04
Naser Al qetami	05
Adel Alkalbani	06

D. Files coding

In order to organize our research, and lower the cost of its expansion and maintenance in following research efforts, the audio filenames in the system have been coded in a specific format. Each audio file name consists of 13 characters, with six letters that identify the attribute to which some number of alpha-numeric symbols following each letter relate. In general file name format is in the following form SxxxAyyyCxxVzRxxTxx.WAV. Beginning from the left, the three digits following the first (Sxxx) character represent the Surah (chapter) number from the THQ. The three digits following the second character (Ayyy) represent the Ayah (verse) number from the THQ. The two digits following the third character (Cxx) represent the carrier word number; any carrier word with a distinct meaning was given a different number, and in our case there were 12 different carrier words. The symbol following the fourth (Vz) is the identifier of the Vowel number; in our study we numbered the six vowels as follows: /a/=1,/i/=2,/u/=3,/a:/=4, /i:/=5, and /u:/=6. The two digits following the fifth character (Rxx) represents the reciter number, and the two digits following the sixth, and final, character (Txx) represent the trial for each reader of the same verse.

We created another audio files by extracting spoken words only, naming it as WSxxxAyyyCxxVzRxxTxx.wav. This follows the same formula as just described, but with the letter W at the beginning. These files were analyzed with PRAAT software in order to get the values of three first formants of the target vowels. With this file format, we are able to gradually increase our database in the future by increasing the number of reciters, carrier words, and/or trials.

E. ANN system overview

A fully connected feedforward MLP network was used to recognize the six vowels. The network consists of 50 nodes in the input layer (source nodes). Number of nodes in this layer depends on number of formants (F1, F2, F3, F2-F1 and F3-F2) for every frame and number of considered frames in the whole token that is currently applied on the input layer. Number of considered frames is 10. (5 Formants x 10 frames=50). Many transfer functions are included in the Neural Network, the linear transfer function are used in the final layer of multilayer networks that are used as function approximators. And the sigmoid transfer function are used in the hidden layers.

We tested the MLP network with both one and two hidden layers, each with a different number of neurons in the hidden layer(s), Figs. 1 and 2 show the ANN architecture for the MLP with singly and doubly hidden layers. The output layer consists of six neurons, where each neuron in the output layer is expected to be on or off, depending on the vowel applied to the input layer. For the

normal, intended situation, in which the applied features indicate one of the six Arabic vowels, there is expected to be exactly one node in the "on" state, while all others are expected to remain in the "off" state. In the event that the applied features do not indicate one of the six Arabic vowels, all neurons should reflect the "off" state.

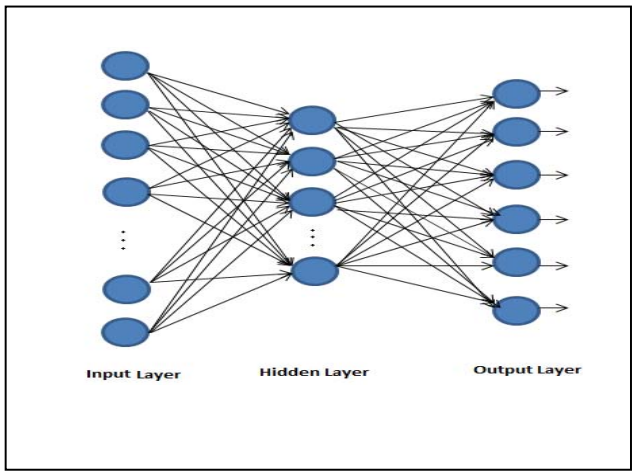

Fig. 1. ANN architecture with single hidden layer.

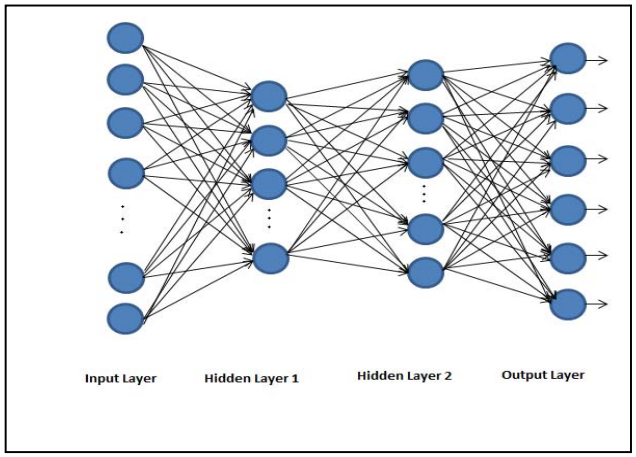

Fig.2. ANN architecture with two hidden layers.

IV. Results

In order to conduct our experiments correctly, the used speech corpus was split into three subsets, a training subset (70% of the total database), a testing subset (25% of the whole database) and a validation subset (5% of the total database). The big problem is the fact, that ANNs cannot explain their prediction, the processes taking place during the training of a network are not well interpretable and this area is still under development. Depending on the task, the model may yield high performance if it matches well with true underlying distributions, but may lead to inferior results if the model is invalid.

We examined the MLP network with both one and two hidden layers, with a different number of neurons in the hidden layer(s). Additional layers have the effect of applying a greater number of non-linear transformations to the data, and provide more opportunities to disentangle the data before it reaches the final classifier, the last layer, which decides what the given instance appears to be.

For the MLP network with two layers (i.e., one hidden layer and the output layer), we conducted two experiments. Table II shows the confusion matrix that was generated by the recognition system in the case of a single hidden layer with 16 neurons. This table represents the overall accuracy as well as the accuracies for each individual vowel. The system must try to recognize 18 tokens for each vowel, where the total number of tokens is 108. The overall system performance was measured at 87.96%. The system failed in recognizing only 13 tokens out of the 108 total tokens. Vowel /a/ got a 100% recognition rate, vowel /a: / got a 94.44% recognition rate, vowels /i: / and /u: / each got an 88.89 % recognition rate, vowel /i/ got an 83.33% recognition rate, and vowel /u/ got a 72.22% recognition rate. We can notice that vowel /a/, vowel /i/, and vowel /i:/ are most frequently selected by the recognition system when it is confused. In addition, we can conclude from the confusion matrix that there are significant similarities between the following vowel pairs (vowel /u/, vowel /a/) and (vowel /i:/, vowel/i/), where the second pair is similar but differ in the length, whereas vowel /i:/ has the long kasrah, but vowel /i/ is the short kasrah.

Table III shows the confusion matrix that was generated by the recognition system in the case of one hidden layer with 30 neurons. This table represents the overall accuracy as well as the accuracies for each individual vowel. The overall system performance was measured at 86.11%. The system failed in recognizing only 15 tokens out of the 108 total tokens. Vowel /a:/ got a 100 % recognition rate, vowel /a / got an 88.89% recognition rate, vowels /i /, /u/ and /i:/ each got an. 83.33% recognition rate, and vowel /u:/ got a 77.78% recognition rate. In this table, vowel /a:/, vowel /i/, and vowel /u/ are most frequently selected by the the recognition system when it is confused. The significant similarities between the following vowel pairs (vowel /i/, vowel/a:/), (vowel/i:/, vowel/i/) and (vowel/u:/, vowel/u/), the second and third pairs are similar but the differ in the

length whereas vowel /i:/ is long kasrah but vowel /i/ is the short kasrah in second pair, and vowel /u:/ is the long dummah but vowel /u/ is the short dummah.

Table IV shows the confusion matrix that was generated by the recognition system in case of two hidden layers with five neurons and seven neurons for the first and second layer, respectively. This table represents the overall accuracy as well as the accuracies for each individual vowel. The overall system performance was measured at 83.33%. The system failed in recognizing only 19 tokens out of the 108 total tokens. Vowels /a:/ and /u:/ got an 88.89% recognition rate, vowels /a /, /u/ and /i:/ each got an 83.33% recognition rate, and vowel /i / got a 72.22% recognition rate. In this table vowel /u/, vowel /a:/, and vowel /i:/ are most frequently selected by the recognition system when it is confused. Also, the significant similarities between the following vowel pairs (vowel/a/,vowel/u/), (vowel/i/,vowel/i:/), (vowel/u/, vowel/a:/), (vowel /a:/, vowel/u/) and (vowel/i:/, vowel/u/) show that the second pair is similar, but they differ in length, whereas vowel /i:/ is the long kasrah but vowel /i/ is the short kasrah.

Finally, Table V shows the confusion matrix that was generated by the recognition system in case of two hidden layers with eight neurons and another eight for first and second hidden layer, respectively. The overall system performance was measured at 87.04%. The system failed in recognizing only 14 tokens out of the 108 total tokens. Vowel /u/ got a 100% recognition rate, vowels /a: /and /i:/ each got a 94.4% recognition rate, vowel /a/ /got an 83.33% recognition rate, vowel /i/ got a 77.78% recognition rate, and vowel /u:/ got a 72.22% recognition rate. In this table vowel /i:/, vowel /i/, and vowel /u/ are most frequently selected by the recognition system when it is confused. The significant similarities between the following vowel pairs (vowel /a/, vowel/i/), (vowel/i/, vowel/i:/) and (vowel/u:/, vowel/u/), show that the second and third pairs are similar but differ in length, whereas vowel /i:/ is the long kasrah in the first pair, while vowel /i/ is the short kasrah in the second pair, and vowel /u:/ is the long dummah but vowel /u/ is the short dummah.

Examining our experiments, the most difficult vowels to classify are those that have high similarity to /u:/ and /i/. Conversely, the vowel that is easy to classify is /a/, because its similarity to the other vowels is minimal.

TABLE II. ANN CONFUSION MATRIX (ONE HIDDEN LAYERS WITH 16 NEURONS)

	/a/	/i/	/u/	/a:/	/i:/	/u:/	Acc. (%)
/a/	18	--	--	--	--	--	100
/i/	--	15	1	1	1	--	83.33
/u/	3	--	13	--	1	1	72.22
/a:/	--	--	--	17	1	--	94.44
/i:/	--	2	--	--	16	--	88.89
/u:/	1	1	--	--	--	16	88.89
Average							87.96

TABLE III. ANN CONFUSION MATRIX (ONE HIDDEN LAYERS WITH 30 NEURONS)

	/a/	/i/	/u/	/a:/	/i:/	/u:/	Acc. (%)
/a/	16	--	1	--	--	1	88.89
/i/	--	15	--	2	1	--	83.33
/u/	1	--	15	1	1	--	83.33
/a:/	--	--	--	18	--	--	100
/i:/	--	2	--	1	15	--	83.33
/u:/	1	1	2	--	--	14	77.78
Average							86.11

TABLE IV. ANN CONFUSION MATRIX (TWO HIDDEN LAYERS WITH [5-7] NEURONS)

	/a/	/i/	/u/	/a:/	/i:/	/u:/	Acc. (%)
/a/	15	--	2	1	--	1	83.33
/i/	--	13	--	1	4	--	72.22
/u/	1	--	15	2	--	--	83.33
/a:/	--	--	2	16	--	--	88.89
/i:/	--	1	2	--	15	--	83.33
/u:/	--	1	1	--	--	16	88.89
Average							83.33

TABLE V. ANN CONFUSION MATRIX (TWO HIDDEN LAYERS WITH [8-8] NEURONS)

	/a/	/i/	/u/	/a:/	/i:/	/u:/	Acc. (%)
/a/	15	2	1	--	--	--	83.33
/i/	--	14	--	--	4	--	77.78
/u/	--	--	18	--	--	--	100
/a:/	1	--	--	17	--	--	94.44
/i:/	--	1	--	--	17	--	94.44
/u:/	1	--	2	1	1	13	72.22
Average							87.04

V. CONCLUSIONS AND FUTURE WORK

In this paper, an ANN-based recognition system was designed to quantify the similarities and dissimilarities between the six MSA vowels, in terms of the degree of success in distinguishing the vowels from each other. The used corpus was built by extracting segments of the desired phonemes from recorded Quranic recitations. The values of formants F1, F2, and F3 frequency values were captured, along with their derivatives, specifically F3-F2 and F2-F1.

The network was tested with both one and two hidden layers, each with varying numbers of neurons. As an outcome of our experiments with four network architectures, we were able to present four different results of overall system performances. In the case of one hidden layer, the system with 16 neurons in the hidden layer had an overall performance of 87.96%, while the system with 30 neurons in the hidden layer had an overall system performance of 86.11%. In the case of two hidden layers, the system with 5 neurons in the hidden layer and 7 neurons in second hidden layer had an overall system performance of 83.33%, while the system with 8 neurons in the first and second hidden layer had an overall system performance of 87.04%.

Adding more vowel features and considering more sophisticated ANN systems is the goal for future phases of this work. More analytical work will be conducted in the field of MSA vowels analysis in order to enrich the literature of this topic and to set a stable ground for the researches and developer who need such results.

REFERENCES

[1] D. J. Fucci and N. J. Lass, “Fundamentals of Speech Science” May 16, 1999.

[2] D. L. Newman, "The phonetic status of Arabic within the world’s languages: the uniqueness of the lu“At Al-d÷AAd", Antwerp Papers in Linguistics, No. 100, pp. 77-86, 2002.

[3] Y. A. Alotaibi and A. Hussain, “Speech Recognition System and Formant Based Analysis of Spoken Arabic Vowels", FGIT 2009, LNCS 5899, pp. 50–60, 2009.

[4] Y. A. Alotaibi and A. Hussain, "Formant Based Analysis of Spoken Arabic Vowels", BioID_MultiComm2009, LNCS 5707, pp. 162–169, 2009.

[5] A. Nabil and M. Hesham, "Formant Distortion After Codecs for Arabic", Proceedings of the 4th International Symposium on Communications, Control and Signal Processing, (ISCCSP), pp TBC, 2010.

[6] Y. A. Alotaibi and A. Hussain, “Speech Recognition System and Formant Based Analysis of Spoken Arabic Vowels", FGIT 2009, LNCS 5899, pp. 50–60, 2009. [

[7] http://isites.harvard.edu/fs/docs/icb.topic482062.files/ReetzJongman%20%20Chapter%2010%20%20Acoustic%20Characteristics%20of%20Speech%20Sounds.pdf.

[8] S. Haykin, Neural Networks: A Comprehensive Foundation, Second Edition, Prentice Hall 1999.

[9] R. Lippmann, Review of Neural Networks for Speech Recognition, Neural Computation, pp.1-38, MIT press, 1989.

[10] H. J. Paddock, "The major pitch features of vocalic quality", Lingua, Volume 25, Pages 142-151, 1970.

[11] D. L. Newman and J. Verhoeven "Frequency analysis of Arabic vowels in connected speech", Antwerp Papers in Linguistics, No. 100, pp. 7786, 2002.

[12] Y. M. Seddiq and Y. A. Alotaibi," Formant-Based Analysis of Vowels in Modern Standard Arabic – Preliminary Results " Iscience,Signal Processing and Their Applications (ISSPA),2012.

[13] M. Alghamdi, “A Spectrographic Analysis of Arabic Vowels: A CrossDialect Study”, Journal of King Saud University, Vol. 10, Arts (1), pp 3–24, 1418 AH (1998).

[14] H. R. Iqbal, M. M. Awais, S. Masud, and S. Shamail, “On Vowels Segmentation and Identification Using Formant Transitions in Continuous Recitation of Quranic Arabic”, New Chall. in Appl. Intel. Tech., SCI 134, pp. 155–162, 2008.

[15] Y. A. Alotaibi and A. Hussain, “Speech Recognition System and Formant Based Analysis of Spoken Arabic Vowels", FGIT 2009, LNCS 5899, pp. 50–60, 2009.

[16] Y. S. Natour, B. S. Marie, M. A. Saleem and Y. K. Tadros “Formant Frequency Characteristics in Normal Arabic-Speaking Jordanians” Accepted for publication October 26, 2010. From the *Department of Hearing and Speech Sciences, University of Jordan, Amman, Jordan; and the yDepartment of Audiology and Speech Pathology, Al-Ahliyya Amman University, Amman, Jordan.

[17] https://www.explorethequran.com/about-the quran/generalinformation.html.

[18] islamweb.net, [Online]. Available: http://audio.islamweb.net/audio/index.php?page=surahlist

[19] praat: doing phonetics by computer, [Online]. Available: http://www.fon.hum.uva.nl/praat/