ARABIC VOWEL DISCRIMINATION USING ADABTIVE THRESHOLD WAVELET ALGORITHM

Amr M. Gody[1]
Cairo University

Mohamed Abdel Fattah[2]
Industrial Education College

Using wavelet algorithm, the signal is represented with different number of parameters in each frequency band [1]. The vowels information has been split into components in a logarithmic frequency bands as the manner in the human auditory system. Then using the proposed system, vowels can be discriminated. The overall performance of 89% is achieved.

1. INTRODUCTION

The superb ability of the human auditory system to process speech in the presence of noise has motivated many researchers to build auditory-based speech processing systems for automatic speech recognition (ASR) applications [2].

Pioneering work of Georg von Bekesy in the 1950s showed that the basilar membrane in the inner ear is responsible for analyzing the input signal into different frequencies. Different frequencies cause maximum vibration amplitude at different points along the basilar membrane [3].

In speaking, vowels are produced by exciting an essentially fixed vocal tract shape with quasi-periodic pulses of air caused by the vibration of the vocal cords. The way in which the cross sectional area varies along the vocal tract determines the resonance frequencies of the tract (the formants) and thereby the sound that is produced. The vowel sound produced is determined primarily by the position of the tongue, but the positions of the jaw, lips, and to a small extent, the velum, also influence the resulting sound.
A convenient and simplified way of classifying vowel articulatory configurations is in terms of the tongue hump position (i.e., front, mid, back), and tongue hump height (high, mid, low), where the tongue hump is the mass of the tongue at its narrowest constriction within the vocal tract. According to this classification the vowels /i/, /I/, /æ/, and /ε/ are front vowels, (with different tongue height) /a/, /Λ/, and /c/ are mid vowels, and /U/, /u/, and /o/ are back vowels.

The front vowels show a relatively high second and third formant frequency (resonance), whereas the mid vowels show well separated and balanced locations of the formants, and the back vowels show almost no energy beyond the low frequency region with low first and second formant frequencies [4]. Table 1 shows the formant frequencies for typical vowels [4].

[1] Department of Electronics and Communication Engineering, Faculty of Engineering , Cairo University , Fayoum Branch, El-fayoum, EGYPT, E-mail: agody@ieee.org.
[2] Department of Electrical Engineering, Industrial Education College, mohafi@hotmail.com

Table 1 the formant frequencies for typical vowels [4].

ARPABET Symbol for vowel	IPA Symbol	Typical Word	F1	F2	F3
IY	/i/	Beet	270	2290	3010
IH	/I/	Bit	390	1990	2550
EH	/ɛ/	Bet	530	1840	2480
AE	/æ/	Bat	660	1720	2410
AH	/Λ/	But	520	1190	2390
AA	/a/	Hot	730	1090	2440
AO	/c/	Bought	570	840	2410
UH	/U/	Foot	440	1020	2240
UW	/u/	Boot	300	870	2240
ER	/ɜ/	bird	490	1350	1690

The wavelet bands correspond to table 1 and it will be our approach. The paper is organized as follows:
Section 2 represents some results in vowel recognition. Section 3 contains our proposed system, which introduce the adaptive threshold technique. Section 4 shows the results. And finally section 5 gives a conclusion.

2. VOWEL RECOGNITION

A lot of research dealt with the problem. Most of them depend on the formant approach of the vowels. Following is presented some of these published researches.

1- A model depends on neural network using formant characteristics succeeded to recognize the five Japanese vowels of trained speakers. However, the model can also recognize some unknown speaker's sounds [5].
2- Artificial neural networks have also been applied successfully in speech recognition applications including multi-layer perceptrons, radial basis functions, and self-organizing maps (SOMs). The most successful trial gets vowels recognition results in the average of 70% [6].
3- Another research reports on a system able to classify different signals containing auditive information based on capture of small signal segments present in specific types of sound. After using a Haar wavelet transform at the preprocessing stage, a neural network known as the O-algorithm compares segments from candidate audio signals against predefined templates stored in the network. There is spread in the results from the word recognition setup due to the difference in pronunciation of the steady state vowels [6].
4- An auditory-based speech processing system, namely the ALSD, is developed to achieve less than 85% accuracy [7].
5- Some results suggest that the formant frequency information without other spectral information is unsatisfactory (mean identification score was only 39.2%, consistent for vowel discrimination) [8].

3. ADAPTIVE THRESHOLD SYSTEM

The proposed system is illustrated in figure 1.

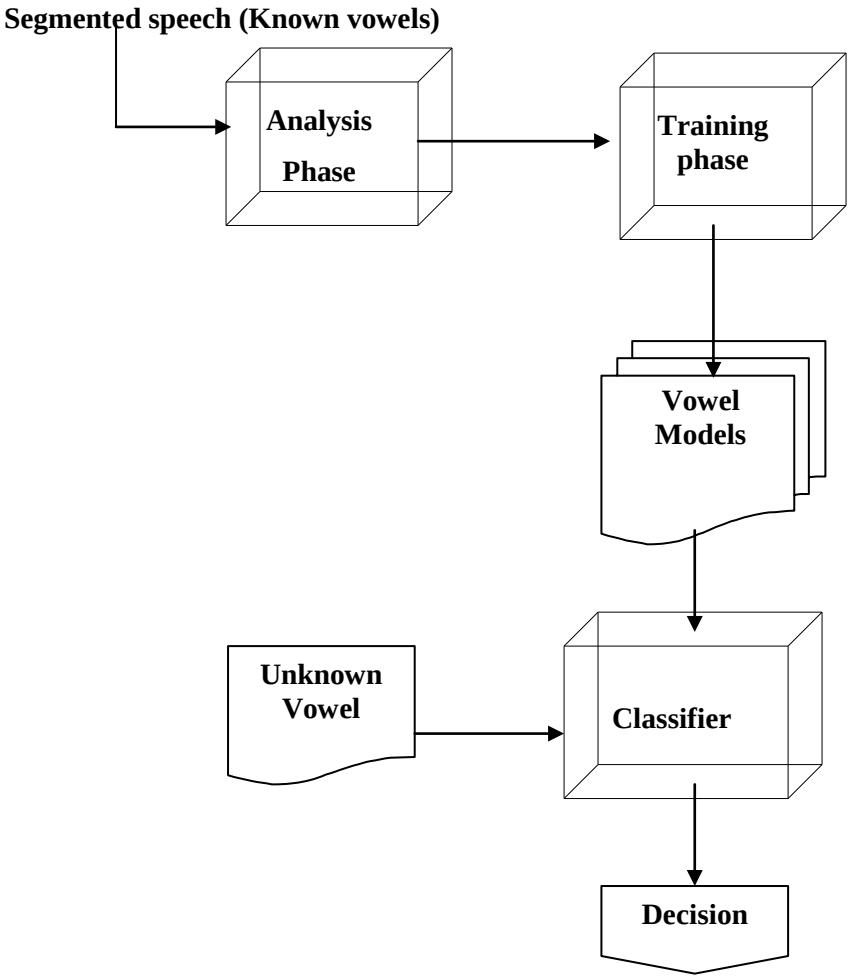

Figure 1 Block diagram of the proposed system

Following subsections will clarify the above system's block diagram.

3.1 SPEECH DATABASE

Speech sampled at rate 11025 Hz and quantized with 16 bit and recorded by two male and two females. Table 2 specifies the details as numbers represents samples count. Half of samples are used for training and the other half are used for test.

Table 2 Database descriptions

	/a/ (فتحة)	*/o/(ضمة)*	*/i/ (كسرة)*	***Total***
Speaker 1	30	30	30	**90**
Speaker 2	30	30	30	**90**
Speaker 3	30	30	30	**90**
Speaker 4	30	30	30	**90**
Total	**120**	**120**	**120**	**360**

3.2 ANALYSIS PHASE

In this phase the speech features are extracted based on wavelet analysis. Daubechies wavelet is selected for its smoothed shapes that make it reasonable to get representing the speech signal's fast-amplitude-variability.

Wavelet filters are 9 levels depth. This level is selected to get more accurate details in the first 3 levels, which are used, for the vowels discrimination process in the following section. This will generates 9 details. Each detail represents certain frequency band as in table 3.

Table 3 Frequency distribution vs. wavelet detail levels

Detail number	*Low frequency limit (Hz)*	*High frequency limit (Hz)*	*Bandwidth (Hz)*
1	2756	5512	2756
2	1378	2756	1378
3	689	1378	689
4	344	689	344
5	172	344	172
6	86	172	86
7	43	86	43
8	21	43	22
9	10	21	11

For mathematical purposes, interpolation is applied to the wavelet parameters. This step grantees that each detail has the same length as the original speech segment. Figure 2 represents the analysis steps to generate the features vector.

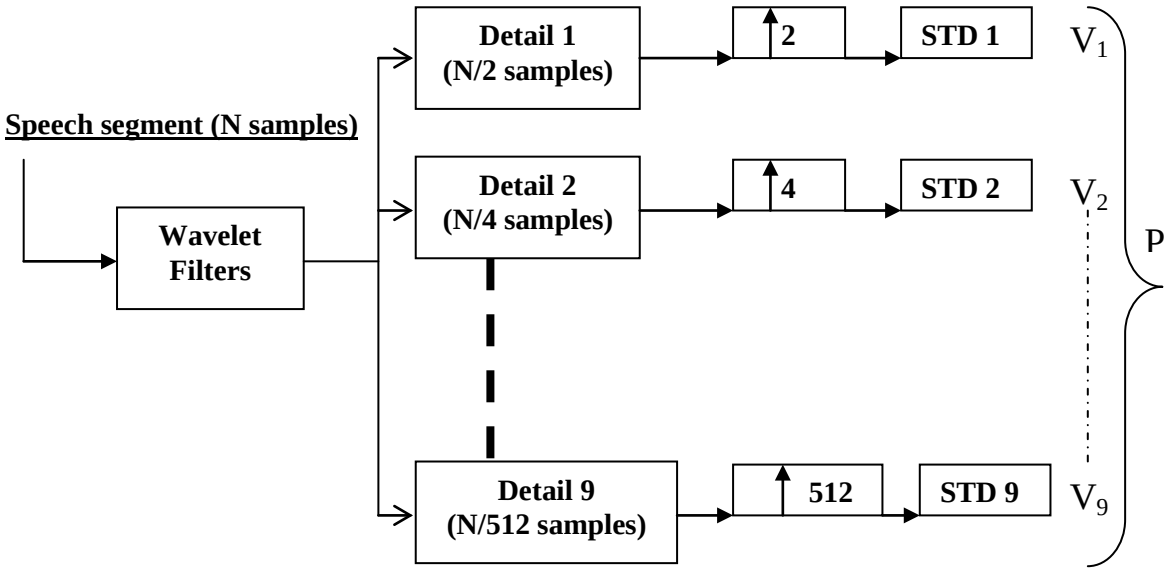

Figure 2. Analysis framework

STD is the abbreviation of standard deviation. The standard deviation is chosen to measure the disturbance. A vector of 9 elements is created. Each element represents detail's STD as shown in figure 2. The vector is normalized to eliminate the effect of

speaker's volume. The generated vector is abbreviated as **P** (**Parametric vector)** for simplicity. P = {v_1, v_2, v_3…v_9}. Where v_i = normalized STD for detail **i**.

3.3 TRANING PHASE

In this phase, the vowel models are generated. The adaptation process is applied to P vectors to get suitable threshold for each vowel, for each speaker. The system is tested as speaker dependent system.
From the P vector, v_1, v_2 and v_3 are selected to get the discrimination between the three Arabic vowels. The selection based on many trials that indicates no big difference is encountered when including the other vector elements.
The training data for certain speaker is used to get the following thresholds as follows:

$$T_1 = \frac{[\overline{v}_1^{/o/} + MIN(\overline{v}_1^{/i/}, \overline{v}_1^{/a/})]}{2} \qquad (1)$$

$$T_2 = \frac{[\overline{v}_2^{/a/} + MAX(\overline{v}_2^{/i/}, \overline{v}_2^{/o/})]}{2} \qquad (2)$$

$$T_3 = \frac{[\overline{v}_3^{/i/} + MIN(\overline{v}_3^{/o/}, \overline{v}_3^{/a/})]}{2} \qquad (3)$$

Where $\overline{v}_n^{/k/}$: the average of v_n for the training data of vowel /k/ from the trained speaker.
Where n = {1,2 or 3} and k = {/a/, /i/, /o/}.
The above threshold equations are adapted for each speaker. They generate adaptive thresholds for each speaker. The equations are formulated after many trials and monitoring of the P vectors.
Figure 3 is a sample that illustrates the first three wavelet filter banks for vowels /a/, /i/ and /o/. It is easy to discriminate the three vowels from figure 3.

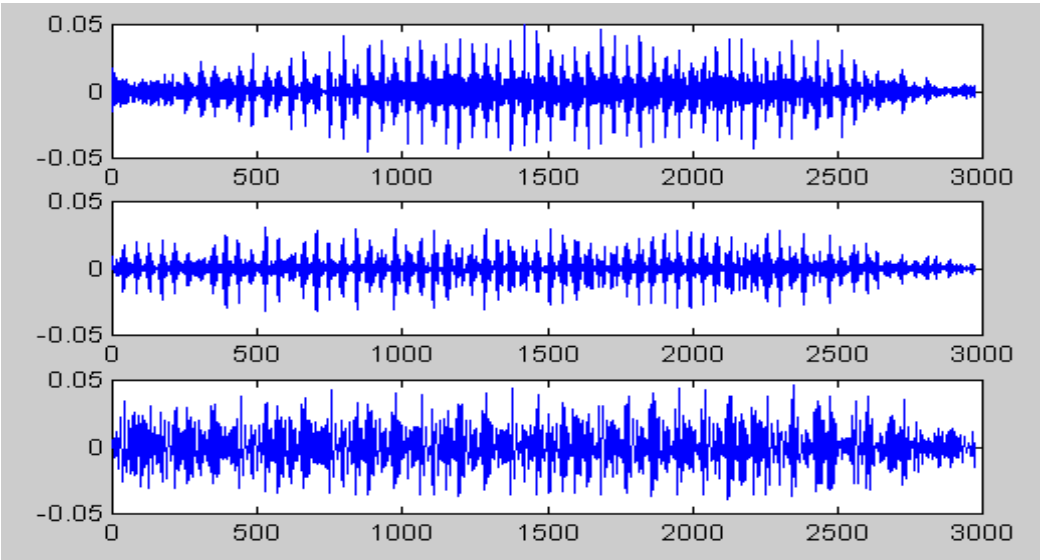

Figure 3-a the first three wavelet Details for vowel /a/. (The top is Detail 1)

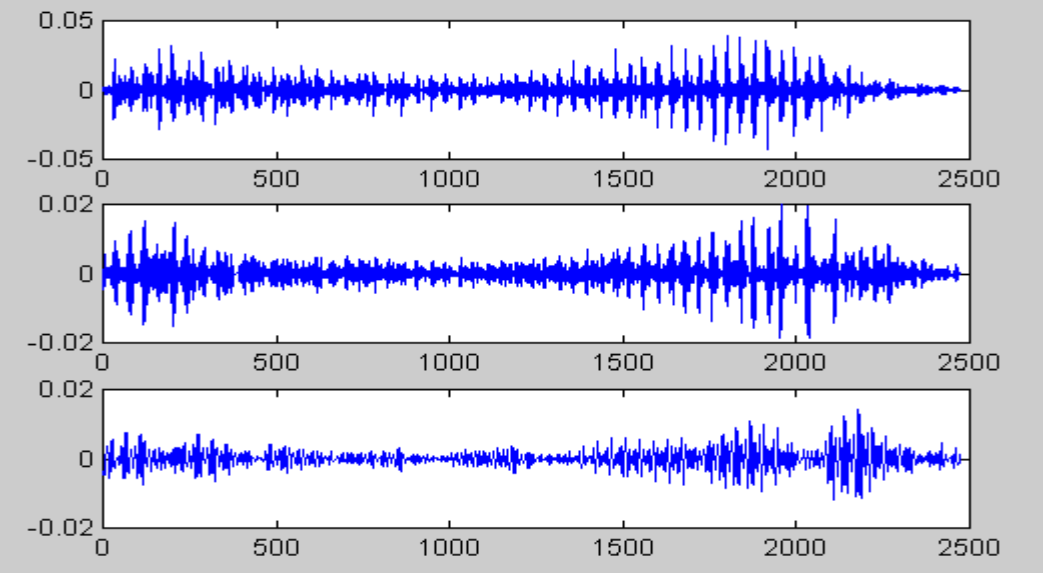

Figure 3-b the first three wavelet Details for vowel /i/. (The top is Detail 1)

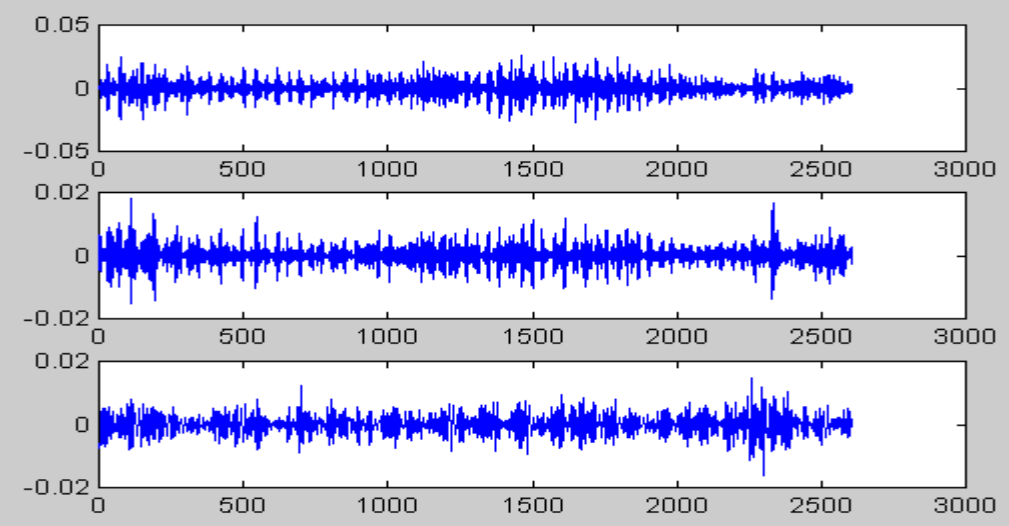

Figure 3-c the first three wavelet Details for vowel /o/. (The top is Detail 1)

Figure 3-a indicates that vowel /a/ contributes high disturbance in the first three details. Figure 3-b shows that vowel /i/ gives high disturbance amplitude in the first detail and very low disturbance amplitude in the other two details. Figure 3-c introduce a little difference of vowel /o/ from vowel /i/. Although vowel /o/ contributes high disturbance amplitude in detail 1 and low disturbance amplitude in the other two details, the disturbance amplitude in detail 2 and detail 3 are relatively higher than those of vowel /i/. Table 4 gives a brief conclusion of the above discussion.

Table 4 Relative disturbance amplitudes of the Arabic vowels for the first 3 wavelet-details

	Detail 1	*Detail 2*	*Detail 3*
Vowel /a/	High	High	High
Vowel /o/	High	Medium	Medium
Vowel /i/	High	Low	Low

3.4 THE CLASSIFICATION PROCESS

After generating the thresholds in the previous section, the system is ready to use them for the recognition process. The P vector of the unknown vowel for the trained speaker is applied to the classifier. It is classified as follows:

$$\begin{cases} (v_1 > T_1)\,\&\,(v_2 \geq T_2)\,\&\,(v_3 > T_3) \rightarrow /a/ \\ (v_1 \leq T_1)\,\&\,(v_2 < T_2)\,\&\,(v_3 > T_3) \rightarrow /o/ \\ (v_1 > T_1)\,\&\,(v_2 < T_2)\,\&\,(v_3 \leq T_3) \rightarrow /i/ \\ ELSE \rightarrow /UNKNOWN/ \end{cases} \qquad (4)$$

The rules are applied from the top to down. If a rule is verified the decision is made and the process stop i.e. the following rules are not applied.

4. RESULTS

The average recognition rate for the four test speakers as a speaker dependent system is tabulated in the following table.

Table 5 Arabic-vowels recognition results.

	/a/(فتحة)	/i/(كسرة)	/o/(ضمة)	Overall
Speaker 1	86.7	93.3	73.3	84.4
Speaker 2	73.3	100	100	91.1
Speaker 3	100	86.7	100	95.5
Speaker 4	93.3	73.3	86.7	84.4
Overall	**88.3**	**88.3**	**90**	**88.9**

5. CONCLUSIONS

Wavelet is used to generate a spectral parametric model that is human like in nature. This similarity to human hearing system gives the model a good result for Arabic vowel discrimination. This system is promising to realize a very speed recognizer with minimum set of database

6. REFERENCES

[1] Thomas F. Quatieri, "Speech Signal Processing", Upper Saddle River: Prentice-Hall inc., 2002, pp. 1-54

[2] Ali, A.M.A.; Van der Spiegel, J.; Mueller, P., "Robust auditory-based speech processing using the average localized synchrony detection", Speech and Audio Processing, IEEE Transactions on, Volume: 10 Issue: 5, July 2002, PP: 279 –292
[3] Amr M. Gody, "Speech Processing Using Wavelet Based Algorithms", PhD thesis, Cairo University, Electronics & communication department, 1999. pp. 1-24 , 126-165.
[4] Lawrence Rabiner, "Fundamentals of speech recognition", Englewood Cliffs, New Jersey: Prentice-Hall inc., 1943, pp. 1-68.
[5] Osamu Hoshinoa, " A Neural Network Model For Encoding and Perception of Vowel Sounds", Neurocomputing, Vols 44-46, Jun 2002, PP. 435-442, www.elsevier.com/locate/neucom.
[6] Najet Arous and Noureddine Ellouze, "Cooperative supervised and unsupervised learning algorithm for phoneme recognition in continuous speech and speaker independent context", Neurocomputing - Special Issue on Neural Pattern Recognition, 23 Jul 2002. (Available online www.elsevier.com/locate/neucom.).
[7] Ali, A.M.A.; Van der Spiegel, J.; Mueller, P., "Robust auditory-based speech processing using the average localized synchrony detection", Speech and Audio Processing, IEEE Transactions on, Volume: 10 Issue: 5, July 2002, PP: 279 –292
[8] Shuichi Sakayori, Toshihiro Kitama, Sohei Chimoto, Ling Qin, Yu Sato, "Critical spectral regions for vowel identification", Neuroscience Research 43 (2002) 155_ 162, www.elsevier.com/locate/neures