TOWARDS A PARAMETRIC REPRESNTATION of ARABIC SPEECH SIGNALS

Abdel Rahim Elobeid and Iman A. Maaly *

ABSTRACT

The aim of this work is to obtain parameters for a Linear Prediction model of Arabic speech production by focusing on the articulatory analysis of Arabic speech. These parameters would provide the basis for many speech-related applications such as speech synthesis and speech recognition systems. The articulation of the Arabic language distinctive sounds (/ɣ/ع/ʕ/,غ ,/ð/ظ,/ ḍ/ض), as well as that of the Arabic vowels, is discussed and the various parameters are calculated. A new parameter called emphatic/non-emphatic parameter is presented.

1. INTRODUCTION

Many algorithms for vocal tract modeling have been developed for a variety of languages, but in the Arabic language domain, a careful study of the phonetic properties of modern standard Arabic is still needed. The aim of this work is to obtain parameters for a model of Arabic speech production. The parametric model for Arabic speech production provides a basis for many speech production applications, such as artificial intelligent machines and recognizers by reducing the utterance of a word or subword to a string of features (model parameters) extracted from the acoustic speech waveform [1]. Therefore, we have focused on the articulatory analysis of Arabic speech. For example, the area function which is the most important parameter can be helpful in the pathological studies of Arabic speech and as an aid for speech training for deaf persons.

1.1 The All-pole Model

In this work, the presented model for Arabic speech production is the discrete-time model shown in Fig.(1). In this model, the composite spectrum effects of radiation, vocal tract, and glottal excitation are represented by a time-varying digital filter whose steady-state system function is of the form:

$$H(Z) = \frac{S(Z)}{U(Z)} = \frac{G}{1 - \sum_{k=1}^{p} a_k Z^{-k}} \qquad (1)$$

where G is the Gain factor,

p is the predictor number, and

$\{a_k\}$ are the Prediction coefficients.

As shown in Fig.(1), the model used is excited by an impulse train for voiced speech or random noise for unvoiced speech. Thus, the parameters of this model are : voiced/unvoiced classification, pitch period for voiced speech, gain factor, and prediction coefficients $\{a_k\}$of the digital filter. Beside these basic parameters we have obtained the area function, the spectrum, the formants and the emphatic/non-emphatic parameters.

This simplified all-pole model is a natural representation for non-nasal voiced sounds, but for nasal sounds, the detailed acoustic theory calls for both poles and zeros in the vocal tract tranfer function.

* *Department of Electrical Engineering, University of Khartoum SUDAN*

However, if the order p is high enough, the all-pole model provides a good representation for almost all the sounds of speech [2]. The major advantage of this model is that the gain parameter G, and the filter coefficients $\{a_k\}$ can be estimated in a very straight forward and computationally efficient manner by the method of linear predictive analysis technique (LP). For this model, the speech samples S(n) are related to the excitation U(n) by the simplified difference equation

$$S(n) = \sum_{k=1}^{p} a_k S(n-k) + GU(n) \qquad (2)$$

1.2 The LP Model Parameters

A linear predictor with predictor coefficients $\{a_k\}$ is defined as a system whose output is

$$\bar{S}(n) = \sum_{k=1}^{p} a_k S(n-k) \qquad (3)$$

The system function of a p-th order linear predictor is the polynomial

$$P(Z) = \sum_{k=1}^{p} a_k Z^{-k} \qquad (4)$$

The predictor error e(n) is defined as

$$e(n) = S(n) - \bar{S}(n)$$
$$= S(n) - \sum_{k=1}^{p} a_k S(n-k) \qquad (5)$$

It can be seen that the predictor error sequence is the output of a system whose transfer function is

$$A(Z) = 1 - \sum_{k=1}^{p} a_k Z^{-k} \qquad (6)$$

The predictor error filter is the inverse filter for the system function

$$H(Z) = \frac{G}{A(Z)} \qquad (7)$$

The basic problem of linear prediction analysis is to determine a set of predictor coefficients $\{a_k\}$ directly from the speech signal in such a manner as to obtain a good estimate of the spectral properties of the speech signal through the use of Eqn.(6). The basic approach is to find a set of predictor coefficients that will minimize the mean-squared error over a short segment of the speech waveform. The resulting parameters are then assumed to be the parameters of the system function H(Z), in the model of speech production.

In this work, the covariance Lattice method for linear prediction is utilized to investigate the properties of Arabic speech production [3]. This method guarantees the stability of the all-pole filter. It also reduces the number of computations to values comparable to those of the other traditional methods.

In this method the reflection coefficients $\{k_i\}$ are uniquely related to the prediction coefficients $\{a_i\}$ by the following recursive relations for $1 \geq i \geq p$

$$a_i^{(i)} = k_i$$
$$a_j^{(i)} = a_j^{(i-1)} - k_j a_{i-1}^{(i-1)} \quad \text{for } 1 \leq j \leq i-1$$
$$a_j = a_j^{(p)} \quad \text{for } 1 \leq j \leq p \qquad (8)$$

The reflection coefficients $\{k_i\}$ can be computed by minimizing some form of the forward and backward residuals.

The gain factor G is equal to the square root of the minimum total squared error derived in the covariance lattice

method [2].

The pitch period is determined by detecting positions of the samples of the error signal e(n) which are large and defining the period as the difference between pairs of samples of e(n) which exceed a reasonable threshold [4],[5].

The calculation of the voiced/unvoiced parameter is based on the ratio of the mean squared value of the speech samples to the mean squared value of the predictor error samples. This ratio is smaller than 10 for voiced sounds [2],[4].

1.3 The Spectrum Envelope

The spectrum envelope is obtained from the following Equation [5]

$$G(f) = \frac{1}{\left|1 - \sum_{k=1}^{p} a_k \exp\frac{-j2\pi k f}{f_s}\right|^2} \qquad (9)$$

where fs is the sampling frequency. The formants are extracted by detecting the local spectrum maxima by magnitude comparison.

1.4 The Area Function

The filtering process of the inverse filter model and an acoustic tube model of speech are analyzed and they are shown to be identical, with the reflection coefficients in the acoustic tube model as a common factor. The discrete area function can be easily obtained from the set of reflection coefficients [6]. By modelling a non-uniform acoustic tube, the tube is divided into an arbitrary number of sections of equal length. Assuming a lossless tube model, the discrete area

function Si is obtained as follows :

$$S_i = \frac{1 + k_i}{1 - k_i} S_{i+1} \qquad (10)$$

where S_i=the area function at section i.

and k_i = reflection coefficient at section i. Since the area function of each section is obtained in a relative manner, the final area function is obtained by specifying the area at section i+1 as $S_{i+1} = 1$.

1.5 The emphatic/non - emphatic parameter

Arabic sounds are classified into two classes as emphatic sounds and non-emphatic sounds [7]. For emphatic sounds the root of the tongue is elevated towards the palate so that the tongue bulge at this area makes a constriction. The voiced emphatic sounds are (/q/ ق , /ð/ ظ , /t/ ط , /d/ ض, /ɣ/ غ)

The unvoiced emphatic sounds are (/x/ خ , /s/ ص.)

In contrast, for non-emphatic sounds, the root of the tongue is lowered leaving an open area between the root of the tongue and the back of the palate. According to X-ray tracing for a vocal tract length of 17 cm [8], we could approximately calculate the length of the tongue root. As shown in Fig.(2), for a lossless tube model of 10 sections, the root of the tongue starts at point Q1 (k=2) and ends at point Q2 (k=6).

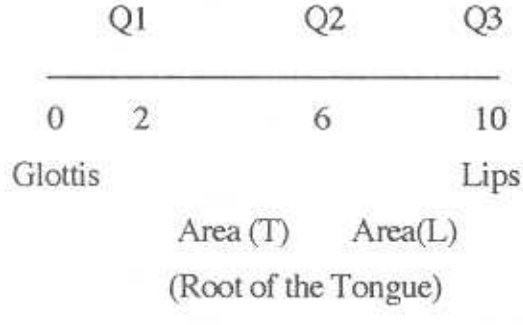


Fig.(2) Vocal Tract Distances for Computing the Emphatic/Non-emphatic Parameter

The emphatic/non-emphatic parameter can be obtained using the following formula

$$R = \frac{area(T)}{area(L)} \qquad (11)$$

where area (T) is the average of areas between points Q1 and Q2, and area (L) is the average of areas between points Q2 and Q3 . We have found that this ratio is less than one for emphatic sounds and greater than one for non-emphatic sounds. Table (1) shows values of R for some Arabic phonemes for 4 different speakers.

2 RESULTS

Our analysis elements are the phonemes. The analysis included all the Arabic non-nasal voiced phonemes. But, only some of them are presented in this paper.

A list of monosyllabic (/CV/ and /CVC/) classical Arabic words are recorded using a sensitive microphone. The microphone signal is fed directly to an IBM computer. The recording is made in a double wall sound treated room in Sudan National Broadcasting studios.

The speakers are five male and two female Sudanese Arabic speakers, plus one American speaker. Three American vowels uttered by a male American speaker were first analyzed as a step to studying Arabic sounds. The purpose of this step is to check the validity of the analysis algorithm by making a real comparison between our results and other's results for similar phonemes of the American English language. We concluded that our results are reasonably accurate

The speech waveform is digitized with the VOICE MASTER Digitizer after selecting the sampling frequency at 10 kHz.

The digitized data is the input to our analysis program. In the program, the frame length (N) of the waveform and the predictor number (P) are determined by the program user. The choice of the frame length is a very important consideration in the implementation of most LPC analysis systems. Therefore, analysis durations from N=50 to N=500 samples at a 10 kHz rate have been tested to find an optimum value. An optimum value of N=215 has been determined.

The choice of the predictor number (P) depends primarily on the sampling rate and is essentially independent on the LP method used. The order of P of the LP analysis can effectively control the degree of smoothness of the resulting spectrum. It is found that the value of P which gives the best results for a sampling rate fs = 10 kHz is P=13. However, in calculating the area function a total of 10 poles is used to represent the vocal tract. The resulting filter stability is checked using the equation

$$-1 < K_i < 1$$

2.1 The Area Function :

The horizontal-axis of the plotted area function is the central line in the vocal tract from the glottis to the lips. The organs of speech can be approximately distributed along the k-axis according to the data derived from the x-ray [8]. Fig.(3a) shows the area function for some American vowels by Wakita [6] where the sampling frequency is 7 kHz and p = 7. Fig.(3b) shows the area function for the same American vowels as a result of our analysis where the sampling frequency is 10 kHz and p = 10. It is obvious that the area function of these vowels has the same features in the two figures. The slight difference may be due to the difference in the predictor number p. Therefore, the analysis scheme is accurate and is applied

to the analysis of the Arabic phonemes.

2.2 Arabic Vowels :

The model parameters of the Arabic are /a/ ا, /i/ ي, /u/ و, long vowels shown in Fig(4), Fig(5) & Fig.(6) respectively. We can notice the quasi-periodic waveforms of these vowels as shown in section (a) of each figure. The point of articulation of each Arabic vowel is near to that of the corresponding American vowel. Fig.(4g) shows the spectrum of the Arabic vowel /i/ which has a low first and high second formants. In Fig.(5g) the spectrum of the Arabic vowel /a/ has a very high first formant and a rather low second formant. First and second formants are therefore close together. This is the typical effect of pharangealization not only in vowels but also in consonants as well. The vowel /u/ is described as a labial sound. It's point of articulation is at the lips as shown in Fig.(6f). It has the lowest first and second formant.

2.3 Arabic Emphatic Sounds :

The model parameters of the voiced emphatic pharyngeal sound / γ / غ is shown in Fig.(7). This sound is proved to be voiced and emphatic. Fig.(7f) shows that it's point of articulation is at the front of the pharynx (the constriction is at section 5) as pronounced by a well trained announcer. This is exactly as described by Sibawaihi [7]. This phoneme is often articulated at the hard palate (the constriction at (k = 6) as pronounced by most of the Arab people nowadays. This wrong articulation which is nearer to that of the phoneme /k/ ك should be considered in Arabic speech recognition studies.

Fig. (8) shows the model parameters of the phoneme / ð /ظ . The point of articulation of this phoneme is at k = 10 as shown in section f. of Fig.(8). This proves that this phoneme is dental

The phoneme /d/ ض is mentioned in [7] to be articulated between the lateral sides of the tongue and the moral teeth at the opposite position. But our results for different speakers gave different area functions. This condition should be considered in the recognition of this sound.

2.4 The Arabic Pharyngeal Sound / ع / ʕ

This sound has a waveform and spectrum nearer in shape to that of the vowel /a / but there is a difference in the area function. Since this sound is an epiglottal sound it's point of articulation is at section 3 as shown in Fig.(9). The results obtained show that this sound is voiced.

3. CONCLUSION

The parameters obtained for the model of Arabic speech production show good results for all the Arabic non-nasal voiced sounds. The voiced/unvoiced decision and the articulation of each Arabic phoneme is compared to the phonetic rules by Sibawaihi and has given good results. The new parameter (emphatic/non-emphatic) is applied to all the Arabic voiced sounds and has given the same classification of emphatics as described by Sibawaihi.

REFERENCES

1. John R. Deller, John G. Prookis, and John H.L. Hansen, Discrete Time Processing of Speech Signals, McMillan Inc., NewYork, 1993.
2. L.R. Rabiner and R.W. Schafer, Digital Processing of Speech Signals, Prentice Hall Inc., New Jersy, 1978.

3. John Makhoul, "Stable and Efficient Lattice Methods for Linear Prediction", IEEE Transaction on ASSP 25, No. 5, Oct. 1977.

4. J. Makoul, "Linear Prediction a Tutorial Review", Proc. IEEE, Vol. 63, pp.561-580, April 1975.

5. B.S. Atal and S.L. Hannver, "Speech Analysis and Synthesis by Linear Prediction of the Speech Wave," J. Acoustics, Soc. of America, Vol. 50, pp. 637-656, Aug. 1971.

6. Hisashi Wakita, "Direct Estimation of the Vocal Tract Shape by Inverse Filtering of Acoustic Speech Waveform," IEEE Transaction on Audio and Electroacoustics, Vol. AU-21, No. 5, Oct. 1973.

7. G. Mirghani, Arabic Speech Sounds, ALESCO Press Khartoum 1985. "In Arabic".

8. Gunnar Fant, Acoustic Theory of Speech Production, Mouton, The Hague, Paris, 1970.

Emphatic Phonemes				
phoneme	R1	R2	R3	R4
/ ɣ / غ	0.65	0.65	0.64	0.64
/ ð / ظ	0.33	0.27	0.51	0.45
/ d / ض	0.42	0.30	0.42	0.34

Non-emphatic phonemes				
Phoneme	R1	R2	R3	R4
/ b / ب	1.29	1.33	1.86	1.62
/ ð / ذ	2.43	2.16	2.43	1.93
/ d / د	1.71	2.66	2.04	1.98

Table (1) Emphatic/Non-emphatic Parameter (R) For Some Arabic Phonemes by 4 Speakers

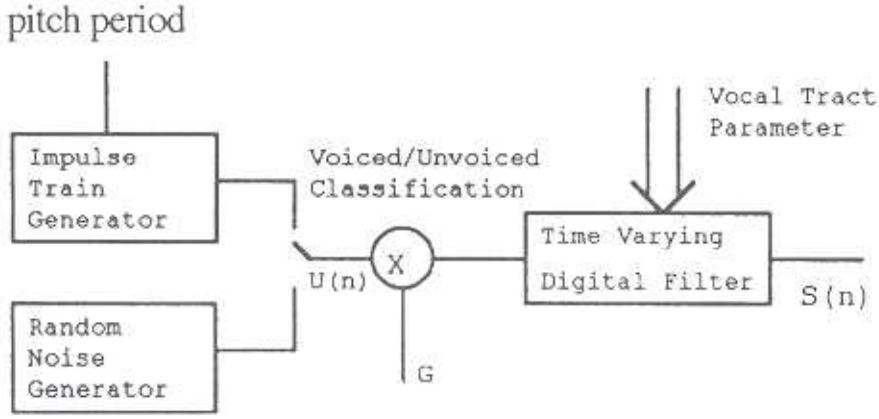
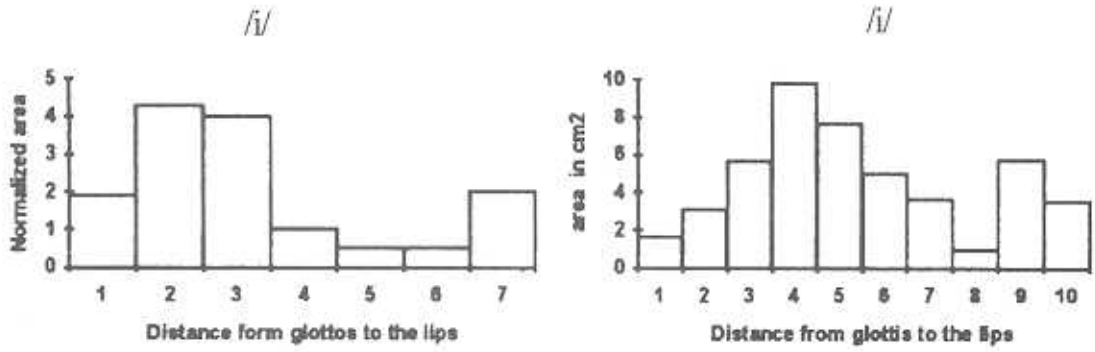
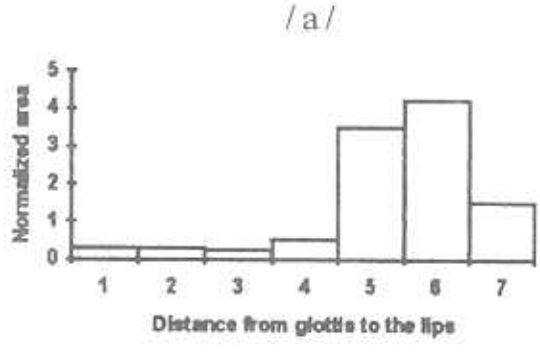


Fig. (1) : Block Diagram of the Model For Speech Production





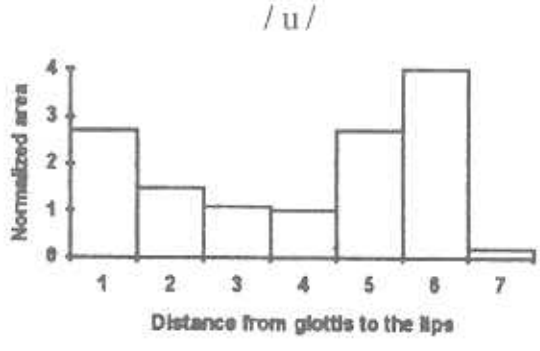
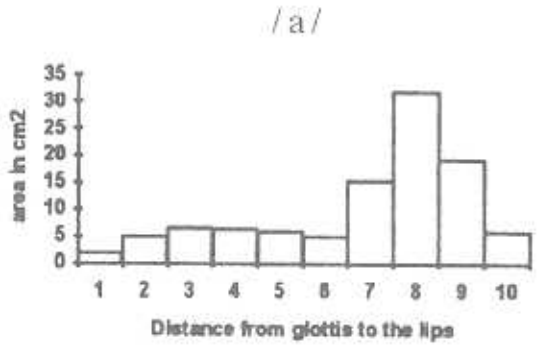


Fig.(3a) The area Function of Some American Vowels by Wakita [6]



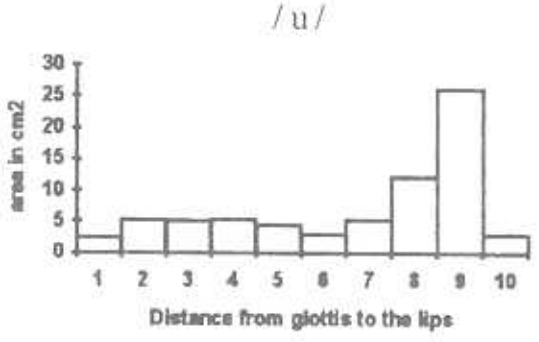


Fig.(3b) The area Function of Some American Vowels (our results)

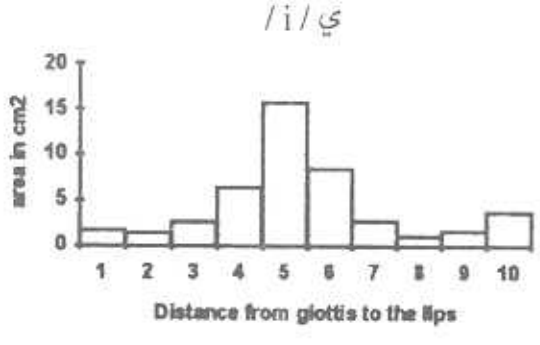


Fig.(4) The area function of /i/ي

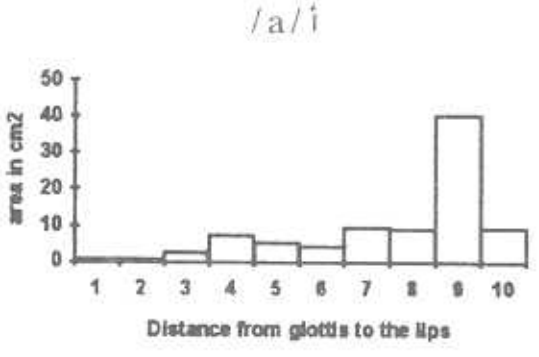


Fig.(5) The area function of /a/i

/ u / و

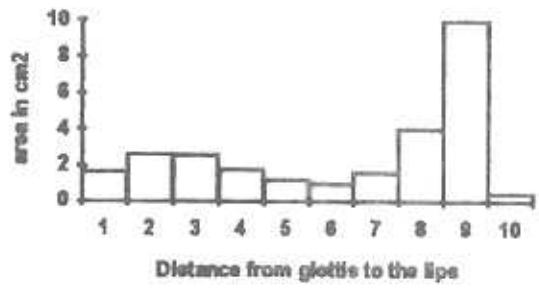


Fig.(6) The area function of / u / و

/ ɣ / غ

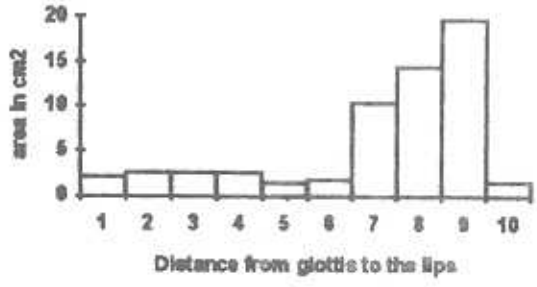


Fig.(7a) Correct articulation of / ɣ / غ

/ ɣ / غ

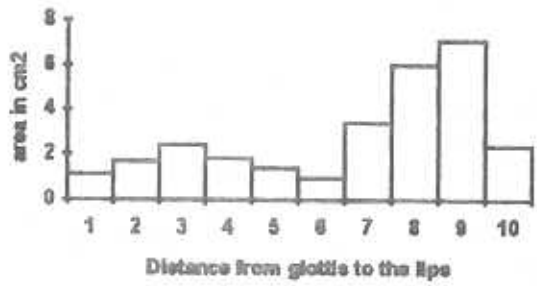


Fig.(7b) Wrong articulation of / ɣ / غ

/ ð / ظ

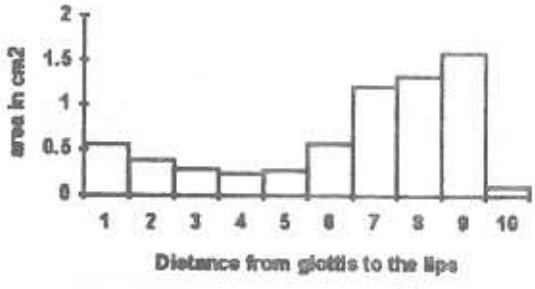


Fig.(8) The area function of / ð / ظ

/ ʕ / ع

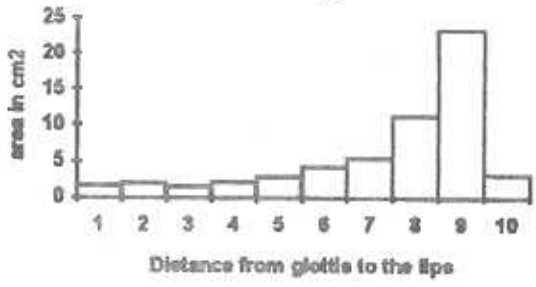


Fig.(9) The area function of / ʕ / ع