

## ESTIMATION OF VOCAL SYSTEM INVARIANT POLES USING FREQUENCY DOMAIN ENVELOPE FUNCTION

*A.N.M. Zainuddin, Nasrin Sultana, Md. Ahsanullah and M. Rezwan Khan\**

Department of Electrical & Electronic Engineering, Bangladesh University of Engineering  
Technology \*United International University (UIU)  
Dhaka-1000, Bangladesh  
Email: \*rezwanm@uiu.ac.bd, zainuddin@eee.buet.ac.bd

### ABSTRACT

The paper presents a novel approach to classify certain vocal system properties that are consistent for a particular person irrespective of the pitch or utterances. To reduce the influence of the cavities in the vocal system, envelop of the Fourier coefficients of the recorded speech are taken to estimate the system poles. Assuming an all-pole model for the vocal system, it is observed that a pair of poles remains reasonably consistent irrespective of the pitch or the sound. Such a pair of poles associated with the glottal response of the vocal fold, can be an identifiable characteristic of the person under consideration.

### 1. INTRODUCTION

Though generation of human voice is biologically a well-understood phenomenon [1,2], there remains significant difficulty in modeling speech [3,4] and vocal system mathematically. The main problem of vocal system analysis is the modeling of vocal tract features and movement of vocal folds from recorded speech data [4]. For example, acoustic phonetics deals with inferring the vocal tract characteristic where the models estimate the source waveform by filtering the speech signal through the inverse of the vocal tract transfer characteristic. Usually, the success of this inverse filtering depends on the model accuracy of the biophysical processes accurately. Another aspect of modeling is to approximate the source waveform using rather simple models, such as the mass-spring models [5,6]. These models estimate vocal parameters that determine the vocal tract transfer characteristic. In some recent research, glottal flow waveforms are represented by high order polynomial models [4]. These models can be used for simulating and thus improving the behavior of the speaker verification tools. However, constructing simple predictive models of phonatory acoustics, tissue

mechanics, and glottal aerodynamics is rather difficult. Despite of having varieties of advantages, the models described above fail to represent the vocal fold very accurately. In all conventional models, either the vocal tract or the vocal fold models are over simplified [7]. For instance, in inverse filtering technique, variation in formant bandwidths due to variation in the glottal impedance during the open phase of the glottal cycle is not considered [8]. As a result, from the viewpoint of filter design, it seems almost impossible to estimate the characteristics of the voice source and the vocal tract simultaneously only from the speech [7].

In this paper, attempts have been made to isolate the vocal fold response from the speech through a different approach. Here, we determined the spectral envelope of different voiced sounds for different pitches and then AR model parameters is found out. The system properties so obtained are assumed to characterize the response of the vocal fold.

### 2. FREQUENCY DOMAIN MODEL

The glottal inputs create two biophysical actions. One is the adduction and abduction, and the other, is a natural vibration of the vocal folds. As the size, shape and tension of the vocal folds vary significantly with pitch and phonation, natural vibration of the vocal folds are expected to vary accordingly. Yet, vocal source generation due to opening and closing of the vocal folds during the passage of air from the lungs is believed to vary by a much smaller extent with pitch [7]. As this is fairly a slower process compared to the natural vibration of the vocal cords, the spectral response of any particular sound at a constant pitch taken over several periods is dominated by the multiples of the pitch frequency and the selectivity of the cavities of the vocal system. In most of the situations, the

response corresponding to the opening and closing of the glottis remain difficult to be isolated from the recorded voice. In this section, we have used the spectral response to isolate the low frequency poles due to vocal fold opening and closing.

Let  $r(n)$  represents the sampled voice data with a predetermined sampling rate satisfying the Nyquist criteria for sampling. In frequency domain, we can express the vocal response by the following relation-

$$R(\omega) = G(\omega).H_v(\omega).H_c(\omega) = G(\omega).H(\omega)$$

Where,  $G(\omega)$ ,  $H_v(\omega)$  and  $H_c(\omega)$  are the transfer functions of input, vocal cord and the cavity of the vocal system. As  $H_c(\omega)$  represents the response of the vocal system cavity, it should show sharply varying peaks corresponding to the cavity responses. On the other hand,  $H_v(\omega)$  is the transfer function of the vocal fold that includes the closing & opening of the vocal fold and the vibration of the vocal cord. As the opening and closing of the vocal folds is the slowest of all, the envelope function of the  $R(\omega)$  would be predominantly shaped by the slowest varying function in frequency domain, i.e. the opening and closing of the vocal fold.

In this work, we took the envelope of the frequency spectrum magnitude of a single average pitch period of the speech signal to identify the system poles. This process has two fold advantages. Firstly, the choice of a single pitch period data reduces the effect of input periodicity. Secondly, the envelope function reduces the impact of the transfer function of the cavities. As a result, we got the poles of the opening and closing of the vocal folds

### 3. ESTIMATION OF AR MODEL PARAMETERS FROM FREQUENCY DOMAIN ENVELOPE FUNCTION

**Envelope Detection:** To find the envelope function, the average periodicity  $P$  of a monotonic speech signal is calculated from the recorded data. To avoid the effect of sharp discontinuity at the beginning and at the end, a Hamming window was multiplied to this signal having  $P$  points.

Then a  $2N$  point Fourier transform of the signal has been taken over  $P$  ( $2N > P$ ) data points with  $(2N - P)$  zero in the remaining trailing points. The small peaks (considered to be noise) were removed by setting an amplitude threshold. The peaks having magnitude higher than the threshold value were picked for envelope estimation.

To avoid negative swing of the envelope function, the magnitudes of the Fourier peaks were inverted and their log magnitude values were taken. Then a continuous curve was obtained by fitting through these points with the Spline interpolation technique [9]. Then the desired envelope of the speech spectrum was obtained by taking the antilog of the curve followed by its inversion.

**AR Model:** If the vocal system is modeled as an autoregressive (AR) system having a finite number of poles, the model becomes dynamic in nature and varies with different sounds. A system model having fixed poles is only possible for any particular vowel at a particular pitch.

An  $N$  point Inverse Discrete Cosine Transform (IDCT) of the envelop function is taken by considering first  $N$  points of any envelop. The IDCT function is the inverse of a DCT function  $y(k)$  given by the following relation-

$$x(n) = \sum_{k=1}^N w(k)y(k) \cos \frac{\pi(2n-1)(k-1)}{2N}, \quad n=1,2,\dots,N$$

Where

$$w(k) = \begin{cases} \frac{1}{\sqrt{N}}, & k=1 \\ \sqrt{\frac{2}{N}}, & 2 \leq k \leq N \end{cases}$$

$N$  = length of  $x$ , which is the same as length of  $y$ .

$M$  and higher order AR model is tried on each set of data to find the system parameters using modified Yule-Walker equations.

For a real AR equation:

$$y(n) + a_1 y(n-1) + \dots + a_L y(n-M) + \dots + a_N y(n-N) = x(n)$$

Here  $a_0 = 1$  is assumed. Representing the AR equation in the vector form we can write:  $\mathbf{YA}' = \mathbf{X}$

$$[y(n), y(n-1), \dots, y(n-N)] \begin{bmatrix} 1 \\ a_1 \\ \vdots \\ a_N \end{bmatrix} = x(n)$$

Here  $x(n)$  is an impulse input to the system at  $n = 0$

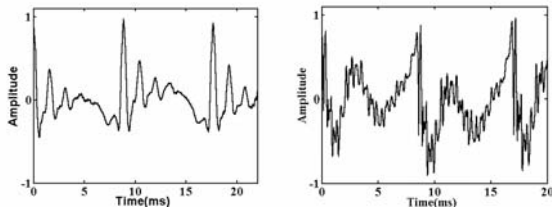
$$\begin{aligned} \text{So } x(n) &= 1, \quad \text{for } n = 0 \\ &= 0, \quad \text{for } n > 0 \end{aligned}$$

To calculate the AR coefficients other than  $a_0$  we solve the AR equations by taking points where input is not present, i.e.  $x(n) = 0$ .

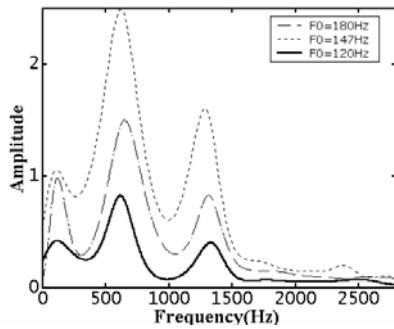
#### 4. DATA ACQUISITION AND OBSERVATION

The experiment was carried out on 4 individuals (2 males & 2 females). Each person performed monotonic speech of utterances in 6 different voiced sounds namely AA, AO, IY, UH, OW, EH, in 7 different pitches within an octave. The sounds were recorded with a microphone that is kept at a certain constant distance from all the speakers' mouth. The voice recording was done at a sampling rate of 44.1k with 16-bit quantization.

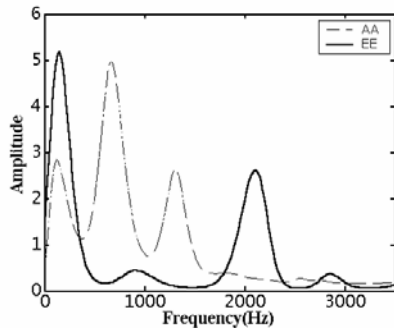
Fig.1 represents time domain speech signals for two different vowel sounds for a male individual under same pitch frequency of 120 Hz. It can be noted that the responses vary markedly for different vowel utterances. Fig.3 shows the frequency domain envelope function for the sound AA for different pitch frequencies.



**Fig. 1:** Speech signals of a Male for AA and EE respectively.



**Fig. 2:** Frequency domain envelope of different pitches for particular vowel sound AA.



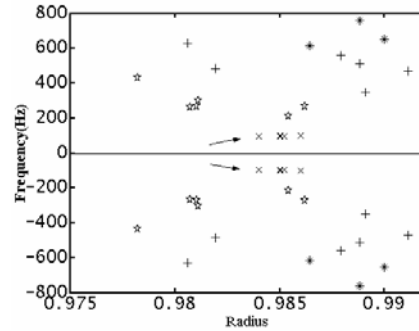
**Fig. 3:** Frequency domain envelope for same pitch ( $f_0=120\text{Hz}$ ) for different vowel sounds AA and EE.

It is interesting to observe that the envelope peaks remain almost unchanged, which clearly indicates that the envelope function is independent of pitch frequency. However, there is some change in the peak of the envelope function when the sound is changed, as evident from Fig.3. Yet, the lowest peak remains unchanged.

Also in the Fourier domain, using 2048 point FFT, we noticed that the envelopes of voice spectrum of different pitches of a particular vowel quality show similar patterns while, the envelopes of different vowel qualities show dissimilar patterns, as shown in fig 2 and 3.

#### 5. RESULTS

It has been observed that, for a particular vowel sound, the poles under different pitch frequencies do not change their positions significantly in the z-plane, while for different vowel sounds under a particular pitch pole positions vary widely as determined by previously stated AR modeling technique.



**Fig. 4:** Distribution of the low frequency poles for a male voice under different pitch and vowel sounds.

However, when AR model was adopted for  $M=16$  or higher order; all the low frequency poles in different vowel sounds vary their positions except the lowest pair. This pole pair is consistently present irrespective of pitches and vowels in all higher order models. Fig 4 shows the average radius vs. average frequency of some of the poles of a male voice extracted from the z-plane diagram for six different vowel sounds. It is quite clear that only one pair of poles remains steady (poles shown by cross marks), where as others generate scattered pattern.

The results are summarized in Table 1 for a male voice considering his vocal system as a 16-order AR model. It can be seen from the table that the selected pole pair remains quite steady with small standard deviations. Table 2 shows the average of

these steady poles for 2 male and 2 female voices. These results show similar degree of invariance as the pitch and the vowel sounds are changed.

**Table 1:** Mean and standard deviation of the frequencies and the radii of the steady poles for a male voice.

Vowels	Frequency(Hz)		Radius	
	$f_{av}$	$\sigma_f$	$r_{av}$	$\sigma_r$
AA	100	3	.9850	.006
AO	98	3	.9830	.004
IY	97	4	.9870	.006
UH	98	5	.9840	.005
OW	100	3	.9850	.005
EH	98	4	.9860	.004

**Table 2:** Mean and standard deviation of the frequencies and the radii of the poles for two male (M1 and M2) and two female voices (F1 and F2).

	M1	M2	F1	F2
$f_{av}$	98.500	146.50	243.30	205.16
$\sigma_f$	4.5400	5.2301	5.2101	5.1232
$r_{av}$	0.9850	0.9865	0.9741	0.9758
$\sigma_r$	0.0035	0.0048	0.0042	0.0052

## 6. CONCLUSION

The pair of poles we have identified from this experimental research represent the vocal property of a person irrespective of the sound uttered. This pair of poles can be utilized to model the vocal chord and also to identify the voice of a particular person. The results show that the distinction between male and female voices can be made quite

easily. However, to distinguish between two male or female voices may require some more parameters as the identified frequency and the radii may be very close to each other.

## REFERENCES

- [1]. Evgeny Karpov, "Real-time Speaker Identification", Department of Computer Science, University of Joensuu, 2003.
- [2]. "Gray's anatomy", page 1651-1652, Thirty-eighth ELBS edition, Churchill Livingstone (Medical Division of Pearson Professional Ltd.), 1995.
- [3]. T Backstrom, P Alku, E. Vilkmann, "Time Domain Parameterization of the Closing Phase of Glottal Airflow Waveform From Voices Over a Large Intensity Range". IEEE Trans. Speech and Audio process., (10) 3, 186-192, 2002.
- [4]. D.G Childers, "Speech processing and synthesis toolbox", John Wiley & sons, Inc., 2000.
- [5]. J.L Flanagan, and K. Ishizaka, "Computer model to characterize the air volume displaced by the vibrating vocal cords", J. Acoust.Soc.Am., 63:1559-1565, 1978.
- [6]. B.H Story and I.R Titze, "Voice simulation with a body-cover model of the vocal folds", J. Acoust. Soc. Am., 97(2):1249-1260. 1995.
- [7]. Yoshinori Shiga and Simon King "Estimation of Voice Source and Vocal Tract Characteristics Based on Multi-frame Analysis", Eurospeech Geneva, Switzerland pp.1749-1752. September 2003.
- [8]. Christer Gobl, "The Voice Source in Speech Communication", Doctoral thesis, Department of Speech Music and Hearing, Stockholm, Sweden, 2003.
- [9]. M Unser "Splines: a perfect fit for signal and image processing", Signal Processing Magazine, IEEE Volume: 16, Issue: 6, Page(s): 22-38, 1999.