Hindawi Publishing Corporation
Journal of Computer Networks and Communications
Volume 2014, Article ID 290147, 17 pages
http://dx.doi.org/10.1155/2014/290147

Research Article

Estimation and Statistical Analysis of Human Voice Parameters to Investigate the Influence of Psychological Stress and to Determine the Vocal Tract Transfer Function of an Individual

Puneet Kumar Mongia and R. K. Sharma

School of VLSI Design and Embedded Systems, National Institute of Technology, Kurukshetra, Haryana 136119, India

Correspondence should be addressed to Puneet Kumar Mongia; mongiap@gmail.com

Received 10 June 2014; Revised 24 August 2014; Accepted 11 September 2014; Published 18 November 2014

Academic Editor: Khoa Le

In this study the principal focus is to examine the influence of psychological stress (both positive and negative stress) on the human articulation and to determine the vocal tract transfer function of an individual using inverse filtering technique. Both of these analyses are carried out by estimating various voice parameters. The outcomes of the analysis of psychological stress indicate that all the voice parameters are affected due to the influence of stress on humans. About 35 out of 51 parameters follow a unique course of variation from normal to positive and negative stress in 32% of the total analyzed signals. The upshot of the analysis is to determine the vocal tract transfer function for each vowel for an individual. The analysis indicates that it can be computed by estimating the mean of the pole zero plots of that individual's vocal tract estimated for the whole day. Besides this, an analysis is presented to find the relationship between the LPC coefficients of the vocal tract and the vocal tract cavities. The results of the analysis indicate that all the LPC coefficients of the vocal tract are affected due to change in the position of any cavity.

1. Introduction

1.1. Voice Production Process. The process of voice production involves a sequence of complex biological activities. It originates from the production of airflow in the lungs, which is modulated by the vocal folds (for voice sounds). Spectral shaping of the modulated airflow is done by the vocal tract cavities which transfer the airflow to the lips to radiate the sound in the outside world. This process of voice production is very well discussed in [1–3]. A simplified view of speech production is shown in Figure 1. Here the speech organs are divided into three main parts: lungs, larynx, and vocal tract. Lungs are acting as a power supply which supplies air pressure signals to the larynx stage. The larynx modulates the airflow as is given by the lungs. It consists of two vocal folds or vocal cords. These folds are made up of masses of flesh, ligament, and muscles [2]. The basic functionality of these folds is to stretch between the front and back parts of the larynx. The glottis is a slit like space between the two folds. The vocal folds are open during breathing. But they can either be in open or vibrating condition depending upon the speaking state. In case of voice sources like vowels, the vocal folds are in a vibrating state. This means vocal folds are opening and closing rapidly. For other sources, the vocal folds are not vibrating rapidly [1]. After the larynx stage the signal passes through the vocal tract which consists of three cavities; pharynx cavity, oral cavity, and nasal cavity. These organs are helpful in shaping the modulated airflow spectrally and also in adjusting the quality of speech [2]. The vibration of the vocal folds in case of voice sources can be estimated in the form of a pulse called glottal pulse. A glottal pulse is shown in Figure 2. As we can see, initially the folds are in closed position (air flow is zero above vocal folds); then they are opening slowly (air flow is increasing); then they are fully open (air flow is maximized), and after that they are closing at a faster rate as shown in the figure. From this we can determine the time duration of one glottal cycle, which is known as pitch period and the reciprocal of pitch period is known as fundamental frequency [1]. The value of the fundamental frequency is influenced by many factors

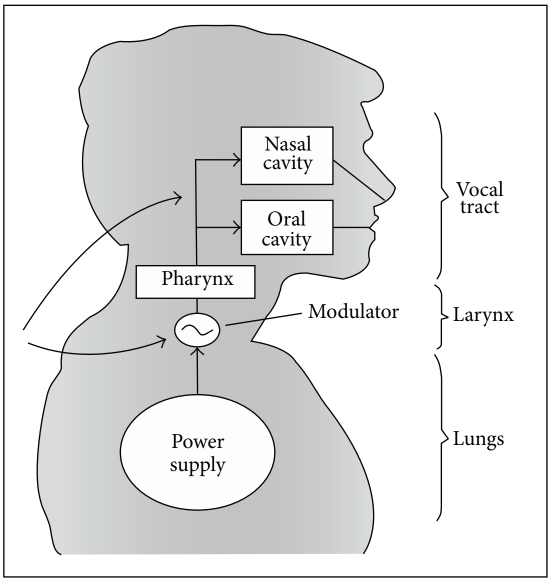


Figure 1: Simplified view of speech production [1].

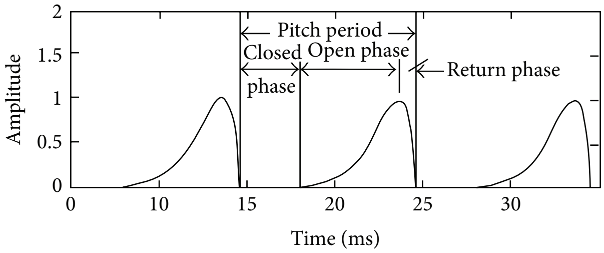


Figure 2: Periodic glottal airflow waveform [1].

like vocal fold muscle tension, vocal fold mass, and the air pressure behind the vocal folds. The average pitch range is roughly 80 Hz to 400 Hz in males and 120 Hz to 800 Hz in females [2].

As the glottal pulse or the excitation signal moves upward on its way through the mouth and nose, it encounters certain obstructions. First the wall of the throat (in the pharyngeal cavity) creates impedance in its path. This impedance causes certain resonance frequencies in the signal. The same effect is caused by the walls of the mouth surrounding the oral cavity and by the walls of the nose surrounding the nasal cavity. The sizes and conformations of these cavities are purely speaker dependent. The resonances of these three cavities (pharyngeal, oral, and nasal) are frequently called formants: the first formant, the second formant, and the third formant, respectively. These frequencies depend upon the shape and dimension of the vocal tract [4]. Because of the motion of organs like tongue, and teeth, higher formant values are likewise possible. As these formant values are immediately linked to the vocal tract cavities so these parameters are also very important and must be measured. After travelling through the vocal tract, the signal is radiated outwards in the form of speech through the lips or nose (in case of nasal voice signals).

The parameters of these organs play a significant role in determining the speaker's characteristics. Getting a true appraisal of these parameters helps us to see the operation of the human speech production mechanism in a more skillful way [5]. These parameters can be beneficial for many speech processing applications such as speaker recognition and speech synthesis [6]. Similarly in biomedical applications or clinical research for the analysis of psychological stress or alcohol intoxication, these parameters play an important role [7, 8]. There is some change in the values of these parameters for normal to diseased or stressed state [9]. So there is a need to effectively estimate these parameters from the voice signal.

1.2. Introduction to Stress. For a number of years the researchers in the field of Speech science and Laryngological studies, are constantly working on the acoustic characteristics of normal and pathological voice. Various methods have been modernized in this subject area for providing the quantitative data [10]. The major reason of growing research in this area is because of the importance of voice signal in determining the effect of clinical disorders like psychological stress. Stress or emphasis is mostly specified as a psychological state that is a reaction to a perceived threat or task demand and is normally accompanied by some specific emotions (e.g., fear, anger, or disgust) [11]. The long term occurrence of stress has serious health consequences [12]. The obvious question that comes to mind is how do we measure stress? The most accurate estimations of a person's stress level can be found by measuring various psychological parameters, such as ECG, EEG or other biological signals, or some biochemical methods [9]. But all these methods require costly and large setup. However, it is very easy to analyze the voice or speech signal; hence this type of analysis is easy and inexpensive. In daily life we often use the term stress to identify negative emotions. However, stress can be classified in two parts, eustress which is a term for positive stress or emotion (like happiness) and distress, which refers to the negative stress or emotions (like anger, fear, or disgust). The positive stress motivates, focuses energy, feels exciting, and improves performance. In contrast, negative stress causes anxiety, feels unpleasant, and decreases performance [9].

1.3. Glottal Pulse Extraction. As discussed in the first section, the glottal airflow is filtered by the vocal tract to provide the air flow at the lip. This airflow is then converted to a pressure waveform at the lips and propagated as a sound signal. So, to get an estimate of the glottal airflow or glottal pulse, one needs to remove the effects of estimated vocal tract filter and lip radiation from the original speech signal. This technique is termed as inverse filtering, since in this process the estimated vocal tract filter and lip radiation effects are inversed to get the glottal flow estimate. MATLAB environment can be used to implement this technique [13–15].

To receive such type of inverse filtering automatically, iterative adaptive inverse filtering (IAIF) algorithm has been used [16–18]. The block diagram of IAIF algorithm used is presented in Figure 3 [7]. Before estimation, the input speech signal is first high pass filtered using a linear-phase

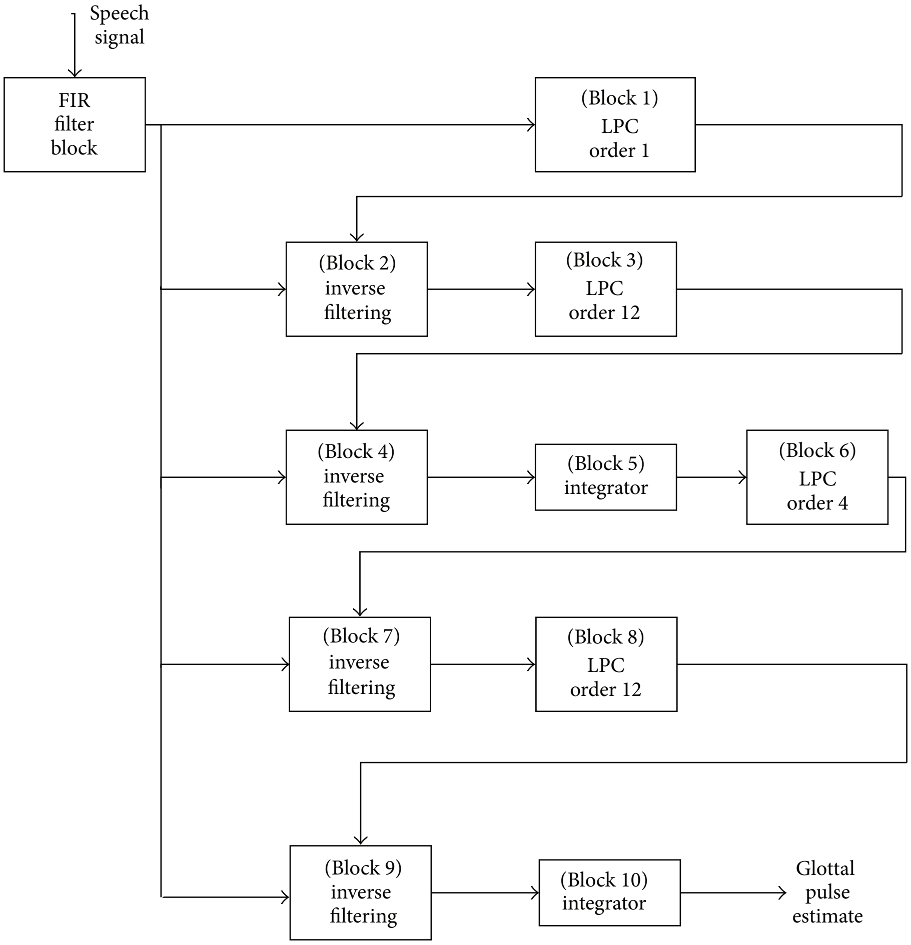


Figure 3: Block diagram of IAIF.

finite impulse response (FIR) filter with a cut-off frequency of 60 Hz to eliminate low frequency fluctuations and DC bias. The high pass filtered signal is used as the input to the next stages. The speech signal is divided into frames before filtering. In block 1, the LPC coefficient fit of order 1 is used to calculate the contribution of the glottal pulse to the speech signal. In the next block 2, this LPC coefficient of order 1 which symbolizes the force of the glottal pulse in the signal is used to design an inverse filter (all zero FIR filters) which is applied to get rid of the glottal effect of the original speech signal. So the input to block 3 represents the speech signal with the glottal flow component filtered out. Next in block 3, LPC fit of order 12 is used to capture the vocal tract filter effect in terms of filter coefficients. Here order 12 is chosen in accordance with the number of formant frequencies which is more than the double number of formants considered for the analysis [19, 20]. So in block 4, the vocal tract filter effect is removed from the original speech signal by inverse filtering. Signal out of this block consists of the effect of glottal flow and lip radiation effect. So to scrub out the radiation issue, a leaky integrator (with coefficient value more than 0.9 and less than 1) is used in block 5, which removes the lip radiation effect from the flow obtained after block 4. The output of block 5 is the first estimate of the glottal pulse. The second repetition runs analogously [7, 15]. The output of block 10 is the glottal pulse estimate of the original speech signal.

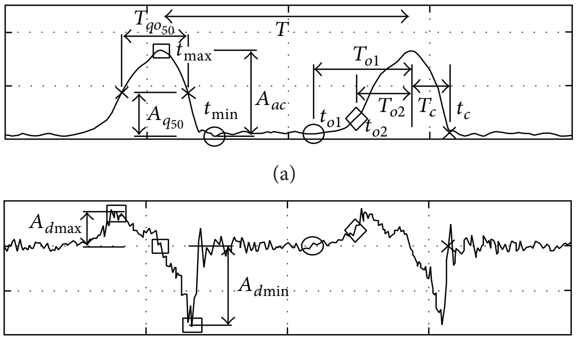
1.4. Glottal Pulse and Its Derivative Parameters.

The parameters of the glottal pulse can provide the quantitative information to examine their importance in the biomedical applications. There are three categories of glottal pulse parameters: time and amplitude domain, frequency domain, and glottal pulse derivative (LF) parameters. The time and amplitude domain parameters involve the extraction of certain time and amplitude instants from the glottal pulse. By counting on these timing instants, several time and amplitude based parameters can be calculated. These time instants can be specified using the glottal pulse and its derivative pulse as shown in Figure 4.

(i) The fundamental time period T is calculated using the fundamental frequency (f_o) of the signal frame.

(ii) $t_{\max}$ is that time instant when the amplitude of the glottal pulse is maximum or when the two vocal folds are completely open. $t_{\min}$ can be defined similarly [21].

(iii) A_{ac} is the peak to peak amplitude level of the glottal pulse which is the difference between the maximum



(a)

(b)

Figure 4: Time and amplitude instants in glottal pulse (a) and its derivative pulse (b) [15].

amplitude to the minimum amplitude of the glottal pulse [21].

(iv) t_c is known as closure time instant which is the time instant when the two vocal folds are just about to close. This time instant is equal to that instant when the glottal pulse derivative pulse crosses to the positive amplitude after $t_{d\,\min}$. Here $t_{d\,\min}$ is the time instant when the glottal pulse derivative pulse is at its minimum value [21].

(v) t_{o1} and t_{o2} are the two opening time instants. To calculate t_{o1} first consider the time sequence which is having 10% amplitude of $t_{\max}$ on the left side of it. Now go left from that time instant up to when the derivative pulse has approached the positive value of its amplitude. This time instant is the first opening time instant. For estimating t_{o2}, first mark the time instant which is 5% more than t_{o1}; then after this time instant look for the maximum positive value of the amplitude of the second derivative pulse of glottal waveform. That time instant is t_{o2}. The importance of considering two opening instants is due to the more gradual opening of the glottal pulse than closing [21].

(vi) t_{qc} and t_{qo} are the time instants where the amplitude of the glottal pulse is 50% of the peak to peak amplitude A_{ac} [21].

(vii) All the time based parameters are calculated with respect to the time instant $t_{\max}$ [21].

From these timing instants, several time and amplitude based parameters can be calculated which are as follows.

(i) OQ (open quotient) measures the relative portion of the open phase compared to cycle duration. Two open quotients can be counted, namely, OQ_1 and OQ_2 [22].

(ii) SQ (speed quotient) measures the ratio of the duration of opening phase to the duration of the closing phase. Possible speed quotients are SQ_1 and SQ_2 [22].

(iii) CIQ (closing quotient) is the ratio of the duration of closing phase to the period length T [23].

(iv) AQ (amplitude quotient) is the ratio of peak to peak amplitude level of glottal pulse and minimum amplitude of glottal pulse derivative [24, 25].

(v) NAQ (normalized AQ) is the normalized value of AQ which is worked out by dividing AQ with the period length T [24, 25].

(vi) QOQ (quasiopen quotient) is same as OQ except that it measures the relative portion of the quasitime instants, that is, t_{qc} and t_{qo}, compared to the cycle duration [26].

(vii) OQ_a is the amplitude counterpart of OQ.

Mathematically, these parameters can be developed as follows:

$$\begin{aligned}
\mathrm{OQ}_1 &= \frac{(t_c - t_{o1})}{T}, \\
\mathrm{OQ}_2 &= \frac{(t_c - t_{o2})}{T}, \\
\mathrm{OQ}_a &= A_{ac}\left(\frac{\Pi}{2A_{d\,\max}} + \frac{1}{A_{d\,\min}}\right) f_o, \\
\mathrm{QOQ} &= \frac{(t_{qc} - t_{qo})}{T}, \\
\mathrm{SQ}_1 &= \frac{(t_{\max} - t_{o1})}{(t_c - t_{\max})}, \\
\mathrm{SQ}_2 &= \frac{(t_{\max} - t_{o2})}{(t_c - t_{\max})}, \\
\mathrm{CIQ} &= \frac{(t_c - t_{\max})}{T}, \\
\mathrm{AQ} &= \frac{A_{ac}}{A_{d\,\min}}, \\
\mathrm{NAQ} &= \frac{\mathrm{AQ}}{T}.
\end{aligned} \tag{1}$$

To estimate frequency domain parameters, the frequency or the power spectrum of the glottal pulse is considered as shown in Figure 5 [15]. There are three main frequency domain parameters of the glottal pulse.

First is $H1$-$H2$ or $dH12$ which is the difference of the first and second harmonics of the glottal frequency spectrum waveform in decibel [27]. Another similar parameter is harmonic richness factor (HRF), which is defined as the ratio between the sums of the amplitudes of harmonics above the fundamental frequency and the magnitude of the fundamental frequency or the first harmonic in decibels [28]. It is shown by the mathematical formula given below:

$$\mathrm{HRF} = \frac{\sum_{r\geq 2} H_r}{H_1}. \tag{2}$$

Here H_r represents the magnitude of the rth harmonic. If $H1$ increases, then $H1$-$H2$ will increase and HRF will

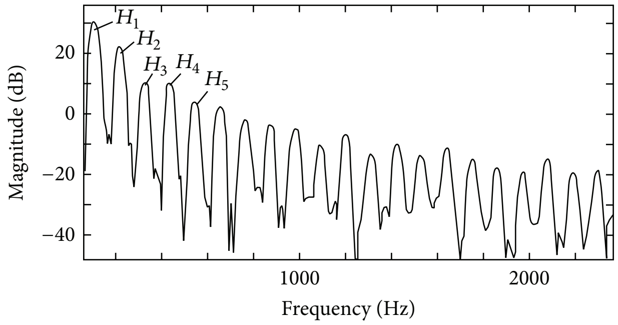


Figure 5: Flow spectrum of a glottal pulse [15].

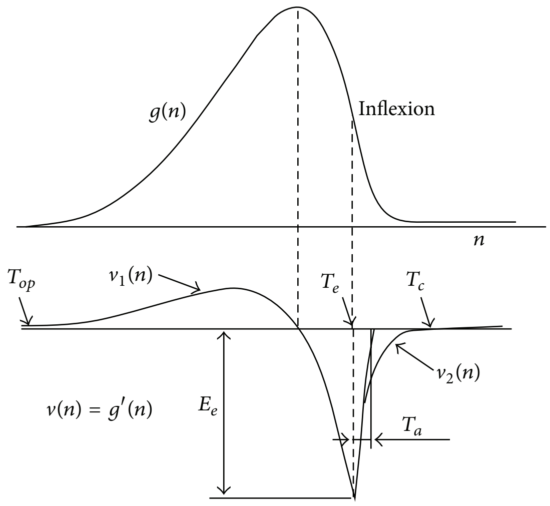


Figure 6: A typical approximation of glottal pulse (upper) and its derivative (lower) [8].

decrease [15]. In [29], the author introduced another similar parameter, parabolic spectral parameter (PSP), which is the second order polynomial to the flow spectrum on a logarithmic scale, computed over a single glottal period [15].

The final type of glottal pulse parameters is the glottal pulse derivative parameters. These parameters are termed as model based parameters because these parameters take on some mathematical expression on the glottal derivative pulse that generates an artificial derivative pulse. With the aid of the artificial pulse the model parameters are estimated. The most used mathematical model is Liljencrants-Fant (LF) model [7, 30]. It is a four parameter mathematical formulation of glottal flow derivative pulse [15]. It accepts applications in both voice analysis and speech synthesis [8, 31–35]. The spectral properties of glottal pulse parameters can also be considered with the aid of this model [36]. The LF approximated glottal derivative pulse is shown in Figure 6 [8].

Following are the timing instants and parameters of LF model.

(i) T_{op} is same as the opening time instant t_{o1} as we have talked about above.

(ii) T_e is that time instant when the derivative pulse is having its minimum amplitude value [37].

(iii) Time instant T_a is the timing instant of the tangent line drawn from the timing instant T_e to the right side of derivative pulse [37].

(iv) Another timing instant T_p, is the instant when the derivative pulse crosses to zero amplitude level for the first time [38].

(v) T_c is same as the glottal pulse closure time instant t_c.

(vi) The parameter E_e is the magnitude of the slope of the negative going glottal pulse [38].

From these timing instants a number of parameters can be obtained:

(i)

$$R_a = T'_a f_o, \tag{3}$$

where T'_a time interval is equal to the difference between T_a and T_e and f_o is the fundamental frequency of the glottal pulse [32].

(ii)

$$R_g = \frac{1}{2T'_p f_o}, \tag{4}$$

where T'_p time interval is equal to the difference between T_p and T_{op} [32].

(iii)

$$R_K = \frac{\left(T'_e - T'_p\right)}{T'_p}, \tag{5}$$

where T'_e time interval is the difference between T_e and T_{op} [32].

(iv)

$$R_d = \frac{(0.5 + 1.2R_K)\left(R_K/4R_g + R_a\right)}{0.11}. \tag{6}$$

OQ (return) is the open quotient for return (closing) phase, which is calculated using the LF model. Consider

$$\mathrm{OQ} = T'_e f_o = \frac{(1 + R_K)}{2R_g}. \tag{7}$$

1.5. Time, Frequency, and Energy Domain Parameters of Voice.

To estimate the glottal parameters one has to apply several steps and algorithms for each frame of data. So if one does not want to look in depth of glottal based parameters, then, he can study the parameters that are directly estimated from the speech signal itself. Here in this section we will discuss time domain, frequency domain, and energy parameters of speech signal.

(i) Autocorrelation function is a time domain parameter of voice. It serves to see the similarity between a speech signal with itself after a little span of time. Let us consider a speech signal $s(n)$ with a frame length of N samples. Let number of frames be m. Then the autocorrelation function of the speech signal for mth frame is defined as

$$r(m) = \frac{1}{2N+1} \sum_{n=-N}^{N} s(n)\, s(n+m). \tag{8}$$

When $m = 0$ then $r(0)$ represents the short term energy of the signal [39]. The value of the autocorrelation function varies between 0 and 1. It yields the value 1 if the speech signal is perfectly coupled with the signal frame just next to it.

(ii) Harmonic to noise ratio (HNR) is the difference between the energies of the speech signal in periodic part and the energies of the signal in the noise in decibels. If HNR = 0 dB, then it implies that the energy in the harmonic part is equal to the energy in the noisy part. A large value of HNR is desirable in speech signals.

(iii) Noise to harmonic ratio (NHR) is the average ratio of the energy of the noise components to the energy of the harmonic components present in the frequency range of speech signal. It evaluates noise present in the speech signal. Variations in amplitude, turbulence noise, subharmonic components, voice breaks, and so forth are considered in NHR. Low value of NHR is desirable in speech signals.

(iv) Short time energy (STE) is defined as the energy of the short segment or frame of speech signal [40]. It can be applied as an effective parameter to differentiate between the voiced and unvoiced segments [41]. The short time energy can be expressed by the following mathematical expression:

$$\mathrm{STE}_n = \sum_{n} [s(n)\, w(n-m)]^2. \tag{9}$$

Here $s(n)$ is the speech signal and $w(n)$ is the window function applied to the speech signal and m varies from 0 to n in a step of the frame size N, which means $m = 0, N, 2N, 3N \cdots n$.

(v) Energy entropy (EE) is a measure of the abrupt changes in energy. This is applied to observe silence and voiced region of speech segments. To calculate EE, first of all each frame is divided into K subframes and energy of each sub frame is computed. Let e_i be the energy of a subframe, then EE of each frame is calculated using the formula [40]:

$$\mathrm{EE} = -\sum_{i=0}^{K-1} e_i^2 \log_2\left(e_i^2\right). \tag{10}$$

(vi) Zero crossing rate (ZCR) is a time domain parameter of speech signal. The number of times per second that the speech signal crosses the zero axis in a frame gives the ZCR in that frame [40]. Overall ZCR of the speech signal is computed by assuming the average value of all the individual ZCRs.

(vii) Spectral centroid (SC) is used to characterize the center of mass of the speech spectrum. It is the weighted mean frequency for a given frame of the speech signal. Weights are the normalized energy of each frequency component in that frame. It can be helpful in detecting frequency peaks in the frame which can either correspond to the location of formants or pitch frequencies [42]. It is given by the formula below:

$$\mathrm{SC} = \frac{\sum_{n=0}^{N-1} f(n)\, x(n)}{\sum_{n=0}^{N-1} x(n)}. \tag{11}$$

Here $x(n)$ represents the weighted frequency value for the frame number n and $f(n)$ represents the center frequency value at that frame [40].

(viii) Spectral flux (SF) is a measure which calculates how quickly the power spectrum of the signal is changing. It is the mean fluctuation of the power spectrum from one frame to the other frame. It is given by the formula below [40]:

$$\mathrm{SF} = \frac{1}{(N-1)(K-1)} \times \sum_{n=1}^{N-1} \sum_{k=1}^{K-1} [\log F(n,k) - \log F(n-1,k)]^2. \tag{12}$$

Here $F(n,k)$ is the FFT of the nth frame of the input speech signal, N is the total number of frames and K is the order of the FFT [40].

(ix) Spectral roll off (SR) is a criterion of the spectral shape of sound like SC. It is that value of frequency for which 85% of the energy of the signal is less than that of frequency [40].

(x) *Jitter* is a measure of period to period fluctuations in the fundamental frequency or pitch of the speech signal [43]. *Jitter* in the signal is mainly affected due to the lack of control in the vibrations of the two vocal folds [44]. *Jitter* can be assessed in many ways given below [43, 44]:

(a) *Jitter* (absolute) is expressed as

$$\mathit{Jitter}\,(\mathrm{abs}) = \frac{1}{N-1} \sum_{k=1}^{N-1} \left|T_k - T_{k+1}\right|. \tag{13}$$

Here N is the number of periods or frames of the signal and T_k is the pitch periods for the frame number k.

(b) *Jitter* (relative) can be expressed equally:

$$\mathit{Jitter}\,(\mathrm{relative}) = \frac{(1/(N-1)) \sum_{k=1}^{N-1} \left|T_k - T_{k+1}\right|}{(1/N) \sum_{k=1}^{N} T_k}. \tag{14}$$

(c) *Jitter* (rap) is the jitter calculated using relative average perturbation:

$$\textit{Jitter}\,(\text{rap}) = \frac{(1/(N-2))\sum_{k=2}^{N-1}\left|T_k - (((T_k + T_{k+1} + T_{k+2}))/3)\right|}{(1/N)\sum_{k=1}^{N} T_k}. \tag{15}$$

(d) *Jitter* (ppq5) is the five point period perturbation quotient jitter. It is computed as the average absolute difference between a period and the average of it and its four closest neighbors divided by the average period.

(xi) *Shimmer* is a measure of period to period variation in the amplitudes of the speech signal [43]. It is affected mainly due to the reduction in the tension of the vocal folds [44]. *Shimmer* can also be assessed in many ways listed below [43, 44]:

(a) *Shimmer* (absolute) is the variation in the peak to peak amplitudes of the speech signal for consecutive periods taken in decibels. It can be expressed as

$$\textit{Shimmer}\,(\text{absolute}) = \frac{1}{N-1}\sum_{k=1}^{N-1}\left|20\log\left(\frac{A_{k+1}}{A_k}\right)\right|. \tag{16}$$

Here A_k is the peak to peak amplitude for the current frame k and N is the number of frames.

(b) *Shimmer* (relative) is the average absolute difference between the amplitudes of consecutive periods, divided by the average amplitude. It can be expressed as

$$\textit{Shimmer}\,(\text{relative}) = \frac{(1/(N-1))\sum_{k=1}^{N-1}\left|A_k - A_{k+1}\right|}{(1/N)\sum_{k=1}^{N} A_k}. \tag{17}$$

(c) *Shimmer* (apq3) is the three point amplitude perturbation quotient which can be computed by considering the mean absolute deviation between the amplitude of a period and average of the amplitudes of its neighbors divided by the mean amplitude of the period. It can be expressed as

$$\textit{Shimmer}\,(\text{apq3}) = \frac{(1/(N-2))\sum_{k=2}^{N-1}\left|A_k - ((A_k + A_{k-1} + A_{k+1})/3)\right|}{(1/N)\sum_{k=1}^{N-1} A_k}. \tag{18}$$

(d) Similarly *Shimmer* (apq5) and *Shimmer* (apq11) can be determined.

It is said that *jitter* (absolute) and *shimmer* (absolute) are useful in speaker recognition [44].

(i) Intensity or vocal intensity of the speech signal refers to the loudness effect of speech signal. Vocal intensity is related to the subglottis pressure of the airflow, which depends on the tension and the vibrations of the vocal folds [44]. A small number of vibrations in the vocal folds make quieter voice as compared to the large number of vibrations of the folds [45]. Mathematically vocal intensity can be expressed as sound intensity level (SIL) or sound pressure level (SPL) [46]. SIL or SPL is measured in dBs. SIL basically tells how much louder a given sound is as compared to the standard (soft) reference vocal intensity, of 10–12 watt/m^2 . This can be determined by [46]

$$\text{SIL} = 10\log\frac{I}{I_0}\ \text{dB}, \tag{19}$$

where I_0 is the standard intensity value and sound intensity can also be expressed in terms of SPL also. Consider

$$\text{SPL} = 10\log\frac{P}{P_0}\ \text{dB}. \tag{20}$$

Here P_0 is the standard pressure level and is having the value of 0.00002 Pascal. SIL and SPL describe the same point of acoustic energy and can be used interchangeably [46].

The formant frequencies can be estimated by taking the frequency response of the vocal tract filter. The peaks of the response are the formant frequencies. The amplitude and bandwidth values at those peaks are also very important parameters and must be considered.

2. Results and Discussion

This section describes the experiments performed and results produced by those experiments. The experimental methodology is first outlined and then followed by the results of the experiment. Let us discuss various experiments performed on the voice parameters.

2.1. Estimation of Glottal Flow. The goal of this experiment was to estimate the glottal flow or glottal pulses from the voice signal of vowels using IAIF algorithm described in the above section by using MATLAB as well as SIMULINK [16–18]. The foremost prerequisite of this algorithm is to obtain the predictor coefficients from the speech signal. For this, lpc function in MATLAB or lpc model of SIMULINK can be used [13, 14]. The speech signal recordings were available in wav format. The speech signals were converted into data samples by taking the sampling frequency of 10 KHz using MATLAB. The workspace block was used to take those samples in SIMULINK. Digital filter design blocks were used

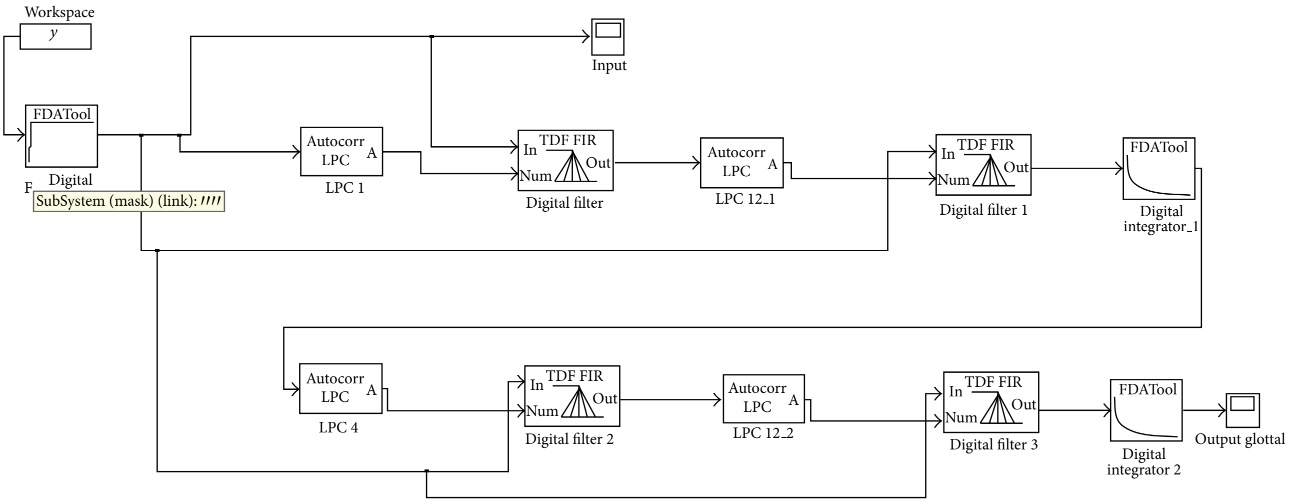
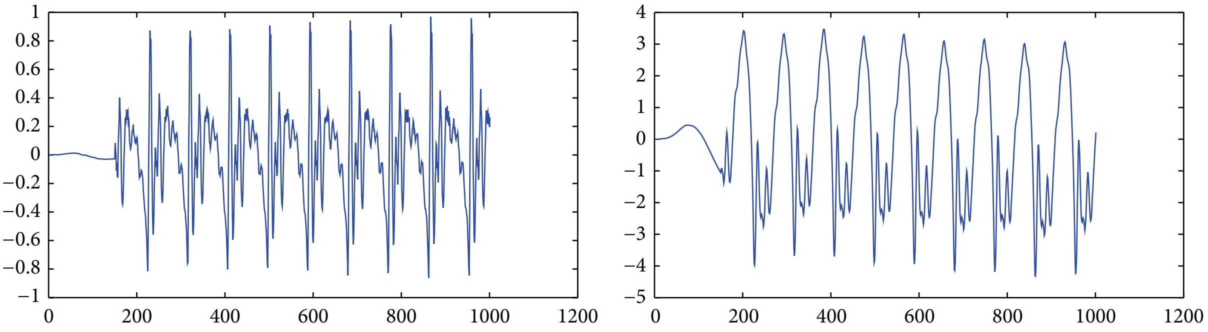


Figure 7: SIMULINK model of IAIF algorithm.



(a)

(b)

Figure 8: Input speech waveform and Output glottal waveform of IAIF algorithm for vowel /a/.

for FIR high pass and inverse filtering. The Autocorrelation LPC blocks were employed to get the predictor coefficients. The digital Integrator block was used for integration. The SIMULINK model of the IAIF algorithm is shown in Figure 7.

The input speech waveform and output glottal waveform for vowel /a/ are shown in Figure 8.

Using the MATLAB code of IAIF algorithm, glottal pulses of five vowels /a/, /e/, /i/, /o/, /u/ obtained are shown in Figure 9.

Table 1: Comparison of computed formant frequencies for male vowel /i/.

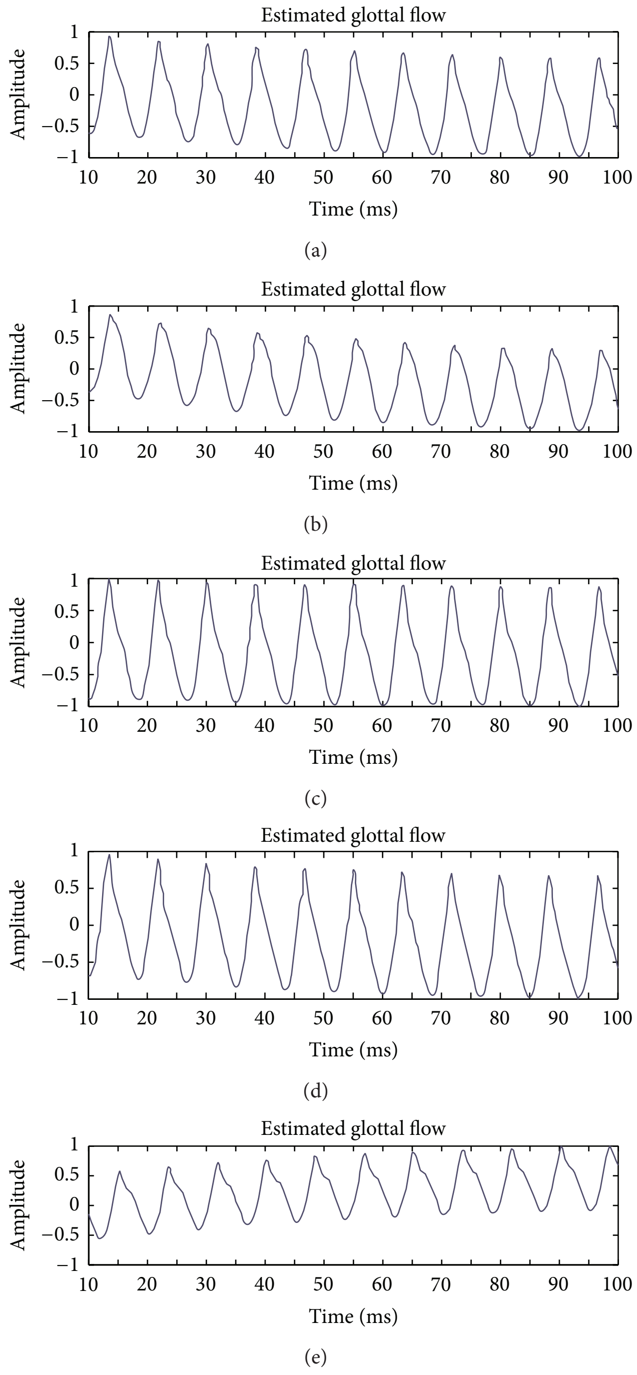
Formant number	By roots	By response	By PRAAT
1	241.3	244.1	233.5
2	2263.6	2270.5	2246.1
3	3194.5	3203.1	3148.6
4	3832.6	3837.9	3828.7

2.2. Comparison of Computed Formant Frequencies. Using the inverse filtering technique the formant parameters can be computed by using two methods. One of them is to find out the peaks of the frequency response of the vocal tract filter and other is to find out the roots of the polynomial equation formed using LPC coefficients of vocal tract filter as explained in [9]. This experimentation was performed to compare the computed formant frequencies by those two methods with the values obtained using phonetic software PRAAT [47].

A total of 15 speech signals were analyzed and four formant frequencies were computed for each case. The speech signals used consist of five vowel segments each for male, female, and child and are available in [48]. In 12 of them (80% of the total), formant values obtained using the two methods above were rather near to the values computed using PRAAT software. In case of LPC polynomial root method, some false formants were also noted. So this idea is not so precise and should be used rarely. By applying these methods, we can also compute the 3 dB bandwidth values and amplitude values for each formant [9].

Tables 1 and 2 are shown for male vowel /i/ and child vowel /a/.

2.3. LPC Coefficients versus Vocal Tract Cavities. As we have discussed in the first section that inverse filtering and LPC coefficients approach can be used to model the human



(a)

(b)

(c)

(d)

(e)

Figure 9: Glottal pulses for five vowels /a/, /e/, /i/, /o/, and /u/, respectively.

Table 2: A comparison of computed formant frequencies for child vowel /a/.

Formant number	By roots	By response	By PRAAT
1	532.5	546.9	549.5
2	1194.1	1196.3	1259.4
3	1807.9	1801.8	1872.6
4	3903.8	3911.1	3893.7

Table 3: Change in the formant parameters when a single coefficient value is changed from 5 to 20%.

Parameters/change	5%	10%	15%	20%
$F1$ (Hz)	551.0	542.0	532.0	512.7
$A1$ (dB)	**31.2**	**29.7**	20.7	16.4
$B1$ (Hz)	37.4	47.6	**138.0**	**225.0**
$F2$ (Hz)	913.0	883.8	849	825.2
$A2$ (dB)	22.1	19.4	16.9	14.6
$B2$ (Hz)	**98.7**	**144.3**	**196.0**	**245.2**
$F3$ (Hz)	1967.0	1958.0	1953.0	1948.2
$A3$ (dB)	3.3	2.8	2.4	1.9
$B3$ (Hz)	**312.0**	**323.9**	**335.0**	**348.1**
$F4$ (Hz)	3291.0	3281.2	3276.0	3271.5
$A4$ (dB)	8.6	7.4	6.4	5.4
$B4$ (Hz)	**228.0**	**253.4**	**277.0**	**310.4**
$F5$ (Hz)	**3842.0**	**3847.7**	**3857.0**	**3867.2**
$A5$ (dB)	11.9	11.0	10.2	9.4
$B5$ (Hz)	**84.6**	**89.0**	**93.5**	**98.2**

vocal tract and is helpful in determining the formant frequencies, so there can be some relationship between the LPC coefficients of the vocal tract and vocal tract cavities. This relationship can be helpful in determining which LPC coefficient of the vocal tract corresponds to which cavity of the vocal tract. It was talked about in the beginning section that each cavity of the vocal tract corresponds to a formant frequency and in the last experiment, we have computed formant frequencies using LPC coefficients of the vocal tract calculated during the final stage of IAIF algorithm. So a relationship can be derived between LPC coefficients and formant frequencies. To derive a relationship 5 speech signals (different persons) were taken. In each signal, each LPC coefficient of the vocal tract was changed (increased and decreased) from 5 to 50%. Corresponding to each change all the formant parameters (frequencies, amplitudes, and bandwidths) were estimated. So for a single signal a total of 24 sets of parameters (both increased and decreased) were tabulated. So for five signals a total of 120 (24 $*$ 5) sets of parameters were tabulated. A single set of the table for the first signal for a change up to 20% is shown in Table 3. This table determines the change in the formant parameters when the 1st LPC coefficient of the vocal tract is increased. Here bold values determine that the corresponding value is more than its original value when no parameter was changed. The original values of the parameters are depicted in Table 4.

After analyzing all the data, the following conclusions were derived.

(i) All the formant parameters were altered due to change in a single coefficient. This signifies that all the portions of the vocal tract are associated to each coefficient.

(ii) Obtained results indicate that these variations follow an individual trend rather than any global trend. So this type of analysis is purely speaker dependent.

Table 4: Original values.

$F1$ (Hz)	556.6
$A1$ (dB)	21.1
$B1$ (Hz)	109.7
$F2$ (Hz)	947.3
$A2$ (dB)	24.7
$B2$ (Hz)	66.3
$F3$ (Hz)	1977.5
$A3$ (dB)	3.7
$B3$ (Hz)	300.2
$F4$ (Hz)	3300.8
$A4$ (dB)	9.9
$B4$ (Hz)	80.4
$F5$ (Hz)	3833.0
$A5$ (dB)	12.8
$B5$ (Hz)	201.9

(iii) Yet a similar trend can be imaged in the change of the value of formant frequencies of all the signals.

(iv) Formant *F1* changes (either increase or decrease) the most, if any individual coefficient is changed.

(v) After that formant *F2* and *F4* come in 2nd and 3rd place in the list.

(vi) In 4 out of 5 signals, *F3* comes after *F4*, and in 1 signal *F5* comes after *F4*.

(vii) No such character of pattern was obtained for amplitudes and bandwidths.

(viii) Nevertheless, in some cases an opposite tendency was seen in bandwidth and amplitude, meaning that if bandwidth was increasing, the amplitude was also decreasing for the whole change.

Figure 10 shows diagrammatically the change in formant values along with bandwidths and amplitudes for a sample.

2.4. Estimation of Vocal Tract Transfer Function for an Individual. According to source-filter theory of speech production, to model the speech production mechanism digitally, we need to consider separate elements of speech production. The speech production system can be modelled with three separate elements: the source, the vocal tract filter, and the radiation effects [17]. The steady state system function of the digital filter is given by the expression:

$$H(z) = \frac{S(z)}{U(z)} = \frac{G}{1 - \sum_{k=1}^{p} a_k z^{-k}}. \tag{21}$$

The primary purpose of this experimentation was to somehow count for a method to forecast or predict the transfer function of vocal tract for an individual. The methodology used was first to calculate the vocal tract predictor coefficients for a signal from the final stage of IAIF algorithm and the gain factor G using lpc function in MATLAB, then by the use of (21) pole zero plot was plotted. As we have discussed before that the LPC order for the vocal tract filter taken is 12 so there will be 12 poles in the transfer function of the vocal tract (Section 1.3).

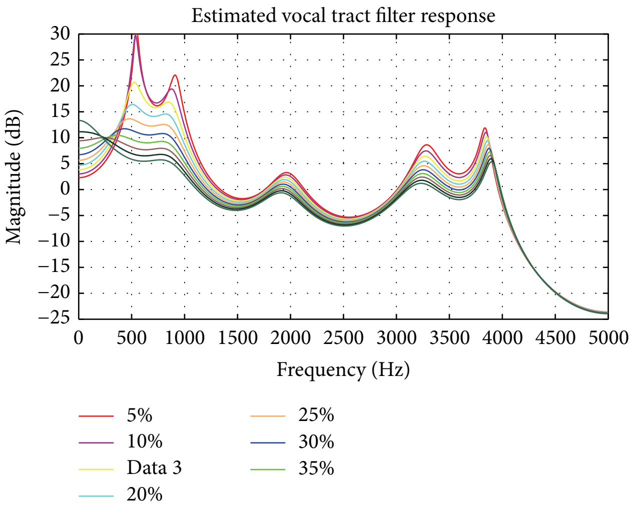


Figure 10: Variations in the formant parameters due to change in LPC coefficients for a signal.

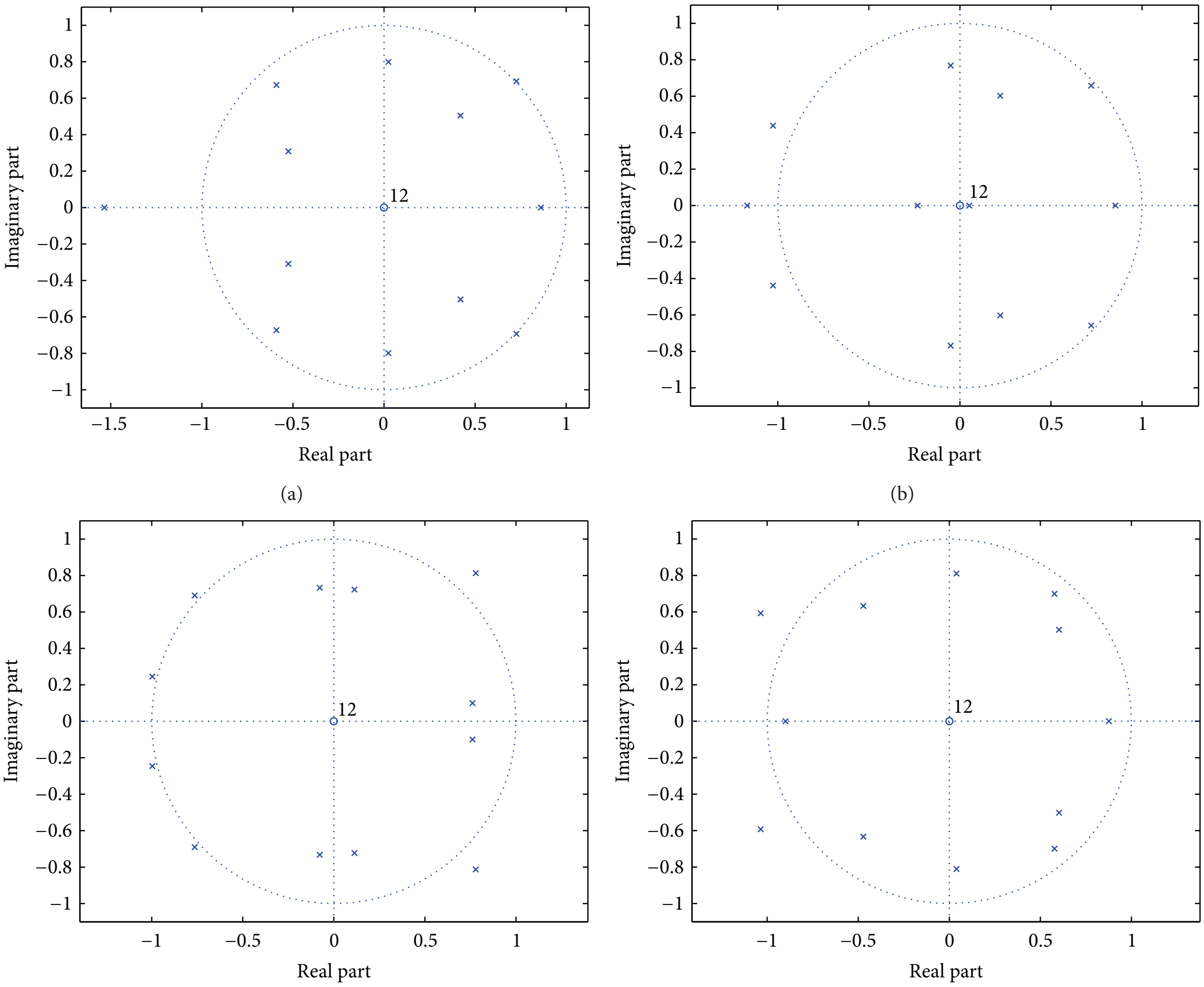
The experimentation was done on two male persons of ages 24 and 26, respectively, by recording their voice samples using Sony IC Recorder (ICD-UX513F) device. Vowels /a/, /e/, and /o/ were taken for the analysis. Each person was asked to pronounce the vowels for at least 3 seconds. Both the persons were asked not to change their day to day activities during the analysis. Total 16 speech samples of each vowel were taken in a single day starting from 7:00 in the morning to 10:00 at night with each sample taken after each hour for each person. So for two persons a total of 96 voice signals of individual vowels were analyzed during two consecutive days. Each vowel signal was pulled out in frames with the help of phonetic software PRAAT [47]. The middle frame was taken for the analysis considering the fact that the speech signal is stationary for a small window of 30–50 msec and has the highest energy at its middle portion [15].

For each signal, parameters like pitch, LPC coefficients of the vocal tract, formant frequencies, pole zero plot, and transfer function were estimated. LPC coefficients were estimated using IAIF algorithm. Formants were estimated using the frequency response method of LPC coefficients of the vocal tract. The pitch was estimated using PRAAT. MATLAB was used for pole zero plot for each signal.

The following are the observations of this experiment.

It was expected that the transfer function for a particular vowel must be unique for a person if calculated at any time of the day. But the experiment showed that the individual shapes of pole zero plots at any time in the day were different from the shapes of pole zero plots calculated at other times. Figure 11 shows pole zero plots for first person at four sampling times.

When the mean value of all the coefficients for each individual vowel for each day was taken and pole-zero plot was plotted for those coefficients, then it was observed that the overall shapes of pole-zero plot for each day were approximately the same. Figure 12 shows overall pole zero



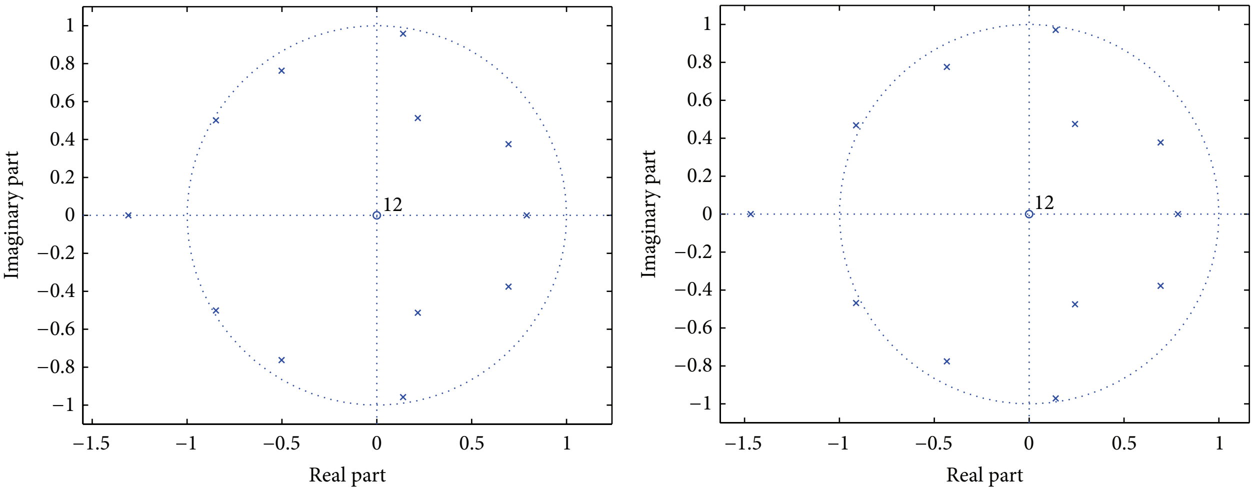
(a)

(b)

(c)

(d)

Figure 11: Pole zero plots of the vocal tract for vowel /a/ at times 7:00 AM day 1 (upper left side) 10:00 PM day 2 (upper right side), 3:00 PM day 1 (lower left side), and 9:00 PM day 2 (upper right side).



(a)

(b)

Figure 12: Mean Pole zero plots for vowel /o/ for person 2 for day 1 (left side) and day 2 (right side).

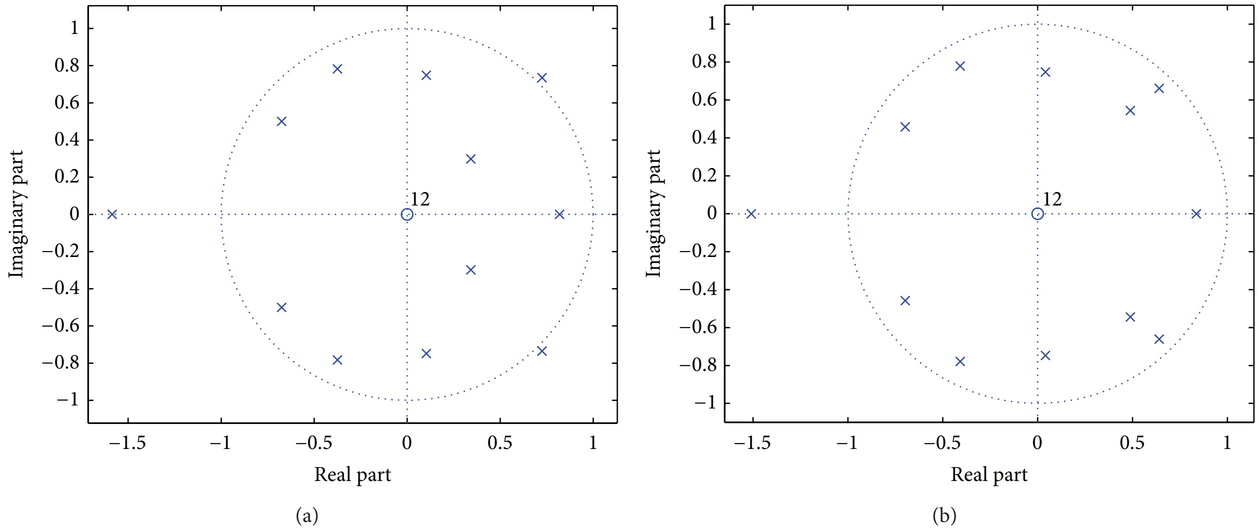


Figure 13: Mean Pole zero plots for vowel /e/ for person 1 for day 1 (left side) and day 2 (right side).

Table 5: Average formant frequencies and pitch for person 1 for both days.

	$F1$ (Hz)	$F2$ (Hz)	$F3$ (Hz)	$F4$ (Hz)	$F5$ (Hz)	Pitch (Hz)
/a/						
Day 1	405.58	1777.6	2413.9	3463.1	4312.0	109.60
Day 2	398.87	1753.2	2427.6	3355.6	4327.0	106.24
/e/						
Day 1	304.56	1982.2	2395.8	3498.2	4101.6	110.12
Day 2	300.60	2062.7	2207.3	3564.1	4207.1	106.18
/o/						
Day 1	389.40	811.16	2430.1	2770.5	4260.8	108.58
Day 2	403.07	862.75	2329.2	3185.8	4207.7	104.85

plot for person 2 for vowel /o/ for both days and Figure 13 shows overall pole zero plot for vowel /a/ for person 1 for both days. So it can be said that the average behaviour of the vocal tract throughout the day is the same which corresponds to its resonance or unique behaviour.

The average pitch value and formant frequencies for person 1 are shown in Table 5.

The following observations can be concluded with this experiment.

(i) This experiment shows that the human vocal tract system tends to change its shape differently in different times of the day.

(ii) This variation in the shape of the vocal tract can be due to day to day activities of that person and can be due to intake of food in the body through the throat or due to lack of energy in the body as the day goes on.

(iii) But in spite of the fluctuations of the vocal tract, the overall shape follows clear uniqueness as we have found out from the pole zero curves.

(iv) The pole-zero plot obtained after taking the mean values corresponds to the vocal tract transfer function for that individual for some specific vowel.

(v) This uniqueness in the pole zero plot can act as a unique signature of that person because the shapes of the pole zero plot were different for same vowels in those two persons.

(vi) So there exists a possibility to find out the biological signature of a person utilizing the vocal system in man.

(vii) This type of analysis can be helpful in studying the vocal tract system behavior in terms of poles.

2.5. Statistical Investigation of Psychological Stress on Human Voice Spectrum. The following work deals with the analysis of speech signal under psychological stress for both positive and negative states of stress. To investigate the influence of stress on speech, acoustic parameters of speech signal were considered. For this type of estimation a suitable database or corpus is required. The most frequently used database among the researchers is the SUSAS (Speech under Simulated and Actual Stress) database of American English which is distributed by Linguistic Data Consortium at the University of Pennsylvania [49]. A German language database called emoDB is also very popular among researchers [50]. A list of existing emotional database is provided in [51, 52]. The database utilized in our analysis was Surrey Audio-Visual Expressed Emotion (SAVEE) database [53, 54]. The database consists of four persons (DC, JE, JK, and KL) of ages 27 to 31 depicting the six basic emotions (anger, disgust, fear, happiness, sadness, and surprise) and the neutral state. The recordings consist of 15 phonetically balanced sentences per emotion (with 15 additional sentences for neutral state) resulting in a corpus of 480 British English utterances. This database is an open source database which can be obtained from the university website on request [55].

The database consists of 15 sentences for each speaker and represents all emotions. Out of these 15, 3 sentences are common and rests are emotion specific. These 3 sentences are considered for the evaluation.

The three sentences were the following.

(i) She had your dark suit in greasy wash water all year.

(ii) Do not ask me to carry an oily rag like that.

(iii) Will you tell me why?

There were three sentences for each speaker and each emotion so a total of 84 signals were considered. 11 vowel segments of 40–60 milliseconds duration were extracted from the individual words of these 3 sentences for each speaker and each emotion using phonetic software PRAAT.

These segments consist of phonemes /aa/ (resemble vowel /a/ sound, e.g., h**a**te), /la/ (resemble long vowel /a/, e.g., h**a**d), /u/ (resemble vowel sound /u/, e.g., b**oo**k), /o/ (resemble vowel sound /o/, e.g., b**oa**t) and /aj/ (resemble vowel /i/ sound, e.g., h**i**de). For each speaker and each emotion, a total of 11 segments were extracted so a total of 308 segments were analyzed.

In the analysis the psychological stress is categorized into three major classes. First is neutral state, the second is positive stress, which was taken as a combination of happiness and surprise emotion, and third is negative stress, which was taken as a combination of anger, disgust, fear, and sadness emotions.

A number of parameters (about 51 parameters) were judged in the depth psychologies which are grouped under the categories as follows.

(i) Group 1 = pitch and intensity (evaluated for all the sentences).

(ii) Group 2 = *Jitter*, *Shimmer*, and *Autocorrelation* (evaluated for all the sentences).

(iii) Group 3 = HNR (harmonic to noise ratio) and NHR (noise to harmonic ratio) (evaluated for all the sentences).

(iv) Group 4 = energy, time, and frequency parameters (energy entropy (EE), short time energy (STE), zero crossing rate (ZCR), spectral roll off (SR), spectral centroid (SC), spectral flux (SF), (evaluated for all the sentences).

(v) Group 5 = formant parameters (frequencies (*F1, F2,* and *F3*), amplitudes (*A1, A2,* and *A3*), and bandwidths (*B1, B2,* and *B3*) (evaluated vowels segment wise).

(vi) Group 6 = glottal pulse timing parameters (NAQ, AQ (milli), CIQ, OQ1, OQ2, Oqa, QOQ, SQ1, and SQ2) (evaluated vowel segment wise).

(vii) Group 7 = glottal pulse frequency parameters (*dH12*, PSP, and HRF) (evaluated vowel segment wise).

(viii) Group 8 = glottal pulse derivative parameters (*Ra, Rg, Rk, Rd,* and *Oq*) (fvaluated vowel segment wise).

(ix) Group 9 = first 12 mfcc feature coefficients (evaluated vowel segments wise).

Groups 1, 2, and 3 parameters were evaluated using PRAAT software. Groups 4, 5, 9, and 10 were assessed by writing their MATLAB codes. Groups 6, 7, and 8 were evaluated using TKK APARAT software [15].

For each signal, all the parameters were evaluated and tabulated emotion wise. After evaluation, they were categorized in terms of positive, negative, and neutral states by combining the appropriate emotion (taking mean values).

The outcomes of the analysis were analyzed by two methods. The foremost objective was to appear for the individual pattern in the decreasing order of values of the parameters in case of all the three states and second aim was to work out the most effective parameters among different groups.

To count on the most effective parameters under each group, DR (discrimination ratio) criteria was used. Consider

$$\mathrm{DR}\,(i) = \frac{\left(m_{N(i)} - m_{S(i)}\right)^2}{d_{N(i)}^2 + d_{S(i)}^2}, \tag{22}$$

where m_N is the mean value of that parameter under neutral state and m_S is the mean value of that parameter under stressed state. d_N and d_S are standard deviations for those parameters.

DR was calculated for positive, negative, and overall stress (by taking averages of DR of both positive and negative). Higher the DR factor more effective is the parameter.

Let us consider the DR calculation for first formant *F1* for vowel /aa/ for speaker DC. By taking the mean values of first formant *F1* for all frames following data was obtained:

$$\begin{gathered} m_N\,(F1) = 656.74\,\text{Hz}, \qquad m_P\,(F1) = 650.64\,\text{Hz}, \\ m_{\text{Neg}}\,(F1) = 639.65\,\text{Hz}, \qquad d_N\,(F1) = 37.979\,\text{Hz}, \\ d_P\,(F1) = 18.989\,\text{Hz}, \qquad d_{\text{Neg}}\,(F1) = 13.81\,\text{Hz}. \end{gathered} \tag{23}$$

Using the above data DR for formant *F1* for positive and negative stressed states can be calculated using (22):

$$\mathrm{DR}\,(F1)\,(\text{Positive}) = \frac{(656.74 - 650.64)^2}{37.979^2 + 18.989^2} = 0.0206. \tag{24}$$

Table 6: DR evaluation table for vowel /aa/ for speaker JE.

Parameter	Mean (N)	Deviation (N)	Mean (P)	Deviation (P)	DR (Pos)
$F1$	615.24	41.43	610.35	34.52	0.01
$F2$	1154.79	58.70	1182.86	44.89	0.14
$F3$	2700.20	75.96	2967.53	84.59	5.53
$A1$	32.26	1.09	22.43	3.05	9.24
$A2$	13.81	1.80	16.67	3.15	0.62
$A3$	10.63	0.70	7.65	0.65	9.65
$B1$	71.70	8.72	173.04	136.03	0.55
$B2$	290.47	10.67	183.69	44.65	5.41
$B3$	105.75	32.67	143.75	30.42	0.72
NAQ	0.09	0.05	0.13	0.03	0.53
AQ (milli)	0.87	0.14	0.56	0.07	4.20
CIQ	0.16	0.10	0.27	0.09	0.73
OQ1	0.44	0.29	0.59	0.11	0.25
OQ2	0.39	0.29	0.49	0.13	0.11

Table 7: DR evaluation table for vowel /la/ for speaker JK.

Parameter	Mean (N)	Deviation (N)	Mean (Neg)	Deviation (Neg)	DR (Neg)
$F1$	755.21	25.06	802.82	54.16	0.64
$F2$	1453.45	28.61	1515.30	76.87	0.57
$F3$	2651.37	119.70	2606.61	173.18	0.05
$A1$	20.15	2.18	19.09	6.11	0.03
$A2$	14.79	4.04	15.59	4.27	0.02
$A3$	15.74	2.52	10.62	2.76	1.88
$B1$	136.33	30.70	208.07	125.56	0.31
$B2$	209.80	107.89	185.96	84.99	0.03
$B3$	141.36	39.45	216.01	85.44	0.63
NAQ	0.08	0.01	0.08	0.04	0.01
AQ (milli)	0.64	0.03	0.52	0.20	0.33
CIQ	0.12	0.02	0.14	0.08	0.06
OQ1	0.55	0.08	0.48	0.11	0.30
OQ2	0.28	0.06	0.35	0.10	0.31

Similarly,

$$\mathrm{DR}\,(F1)\,(\mathrm{Negative}) = \frac{(656.74 - 639.65)^2}{37.979^2 + 13.81^2} = 0.1787. \tag{25}$$

Overall DR can be calculated by taking the mean values of DR (positive) and DR (negative).

Tables 6 and 7 show the DR evaluation table for some parameters of vowels /aa/ for speaker JE for positive stress and for vowel /la/ for speaker JK for negative stress, respectively.

The results from the pattern in the order of stress state of the parameters are as follows.

(i) 8 parameters out of 13 parameters (61.5%), which were evaluated for all the sentences, show a unique rule for all the speakers so they can be helpful in stress detection. Parameters such as pitch, intensity, shimmer, jitter, EE, ZC, SR, and SC show these results. For pitch and intensity, distribution functions were plotted. Figure 14 shows the distribution function of pitch values in case of speaker DC. In 6 out of those 8 parameters, positive stressed signal shows the highest value, followed by negative stress and neutral case.

(ii) 27 out of 38 parameters (71%), which were evaluated for vowel segments, show unique patterns of the values for all the stress states in 3 out of 4 speakers. These 27 parameters were showing results for 37% of the total vowel signals that were analyzed. Out of these parameters, parameter R_a was showing positive results for all the analyzed vowels with positive stressed data having the highest value, followed by negative and neutral data.

(iii) In nut shell, 35 parameters out of 51 parameters are affected due to stress and are showing a singular practice of values in the stressed state for 32% of the examined data.

Results according to the DR criteria were evaluated group wise and are shown in Tables 8 and 9.

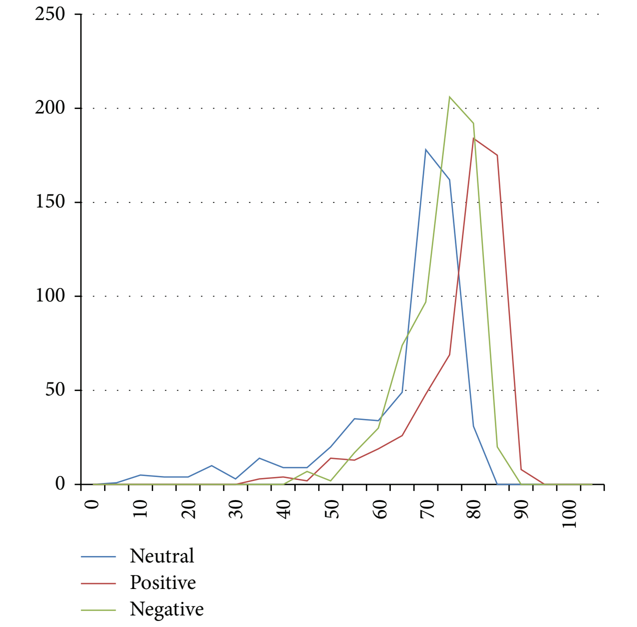


Figure 14: Distribution function for Pitch values for speaker DC.

Table 8: Highest DR values for group numbers 1 to 4.

Group number	Positive effective	Negative effective	Overall
1	Pitch	Pitch	Pitch
2	—	Autocorrelation	—
3	HNR	HNR	—
4	—	—	SC

2.5.1. Final Results

(i) For phoneme /aa/, *F3*, *AQ*, and *Ra* are the most effective parameters for positive stress as well as overall stress detection. *F3* is also the most effective parameter for negative stress detection.

(ii) For phoneme /la/, *A3*, *B3*, and *Ra* are the most effective parameters for positive as well as overall stress detection. *B3* is also the most effective parameter for negative stress detection in this case.

(iii) For phoneme /u/, *A1* and *Ra* are the most effective parameters for positive stress detection; *Ra* is also the most effective parameter for negative stress detection. *F1*, *A1*, and *Ra* are the effective parameters for overall stress detection.

(iv) For phoneme /o/, *dH12* and *Ra* are the most effective parameters for positive, negative and overall stress detection. *F2* is also the effective parameters for positive stress detection.

(v) For vowel independent parameters, pitch and HNR are the most effective parameters for positive stress detection; pitch, autocorrelation, and HNR are helpful in negative stress detection. Pitch and SC are helpful in overall stress detection.

Table 9: Highest DR values for group numbers 5 To 9. (P: positive; N: negative; O: overall).

(a)

Group name	/aa/			/la/		
	P	N	O	P	N	O
Formant freq.	*F3*	*F3*	*F3*	—	—	—
Formant amp.	—	—	—	*A3*	—	*A3*
Formant BWs	—	—	—	*B3*	*B3*	*B3*
Group 6	AQ	—	AQ	—	—	—
Group 7	—	—	—	—	—	—
Group 8	Ra	—	Ra	Ra	—	Ra
Group 9	—	—	—	—	—	—

(b)

Group name	/u/			/o/		
	P	N	O	P	N	O
Formant freq.	—	—	*F1*	*F2*	—	—
Formant amp.	*A1*	—	*A1*	—	—	—
Formant BWs	—	—	—	—	—	—
Group 6	—	—	—	—	—	—
Group 7	—	—	—	dH	dH	dH
Group 8	Ra	Ra	Ra	Ra	Ra	Ra
Group 9	—	—	—	—	—	—

(vi) On the basis of pattern of values of parameters, phoneme /aa/ affects 7 parameters, phoneme /la/ affects 11 parameters, phoneme /u/ affects 5 parameters and phoneme /o/ affects 15 parameters.

(vii) So we can say vowel /o/ should be used for stress detection as it is affecting the most number of parameters.

3. Conclusions

In this paper, we have presented the speech signal analysis using inverse filtering and LPC coefficient approach to estimate some of the important speech parameters like glottal pulse estimation, glottal pulse timing and amplitude parameters, glottal pulse derivative parameters, voice parameters based on time, frequency and energy, MFC coefficients for feature extraction, pitch, intensity, and pole zero plot. The algorithms and methods used for the estimation were studied and discussed in the paper. The formant parameters were compared with the same parameters obtained using phonetic software PRAAT. An analysis was also performed to find out the relationship between the coefficients of the vocal tract and cavities of the vocal tract. Obtained results show that all the coefficients are related to the human vocal tract and no direct correspondence could be held. However, the amount of change in the formant frequencies follow a trend of $F1 > F2 > F4 > F3 > F5$ in most of the cases. Besides this a pole zero evaluation of vocal tract system was discussed to determine the vocal tract transfer function for individuals which shows that the human vocal tract system tends to change its shape in different times of the day for same vowel pronunciations. But the average pole zero plot evaluated follow a unique pattern. This indicates that the ordinary behaviour of human vocal tract system exhibits unique frequency response or resonance. This work can be helpful in simplification of voice related problems in terms of poles and zeros which can be extended further for studying unique voice features in every individual. At last, a speech signal analysis for stress detection was done using SAVEE database. A total of 51 parameters were evaluated and compared for positive stress, negative stress, and neutral state. The features summarized in Tables 8 and 9 have been proven to be the most effective parameters for stress detection among all speakers.

In future, we plan to create our own database, adding other types of stress emotions. We aim to compare the speech features for same emotion for different languages to check whether the emotional content in speech is language dependent or not. Our goal is to detect similar effects with speech with other biological signals like ECG and EEG to identify the correlation among them, which can be helpful in early detection or prevention of many diseases.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

The authors would like to thank Professor Milan Sigmund from Brno University of Technology, for his meaningful discussion on the speech analysis using inverse filtering. The authors express their gratitude to all the colleagues for their immense help throughout the work.

References

[1] T. F. Quatieri, *Discrete-Time Speech Signal Processing, Principles and Practices*, Prentice Hall PTR, 2001.

[2] J. Benesty, M. M. Sondhi, and Y. Huang, *Springer Handbook of Speech Processing*, chapter 2, Springer, Berlin, Germany, 2008.

[3] I. McLoughlin, *Applied Speech and Audio Processing, with MATLAB Examples*, chapter 3, Cambridge University Press, Cambridge, UK, 2009.

[4] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*, Prentice Hall, Englewood Cliffs, NJ, USA, 1978.

[5] P. Alku, "An automatic method to estimate the time-based parameters of the glottal pulseform," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP '92)*, vol. 2, pp. 29–32, San Francisco, Calif, USA.

[6] M. R. Iseli and A. Alwan, "Inter and intra speaker variability of glottal flow derivative using the LF model," in *Proceedings of the International Conference on Spoken Language Processing*, pp. 477–480, Beijing, China, 2000.

[7] M. Sigmund, A. Prokes, and Z. Brabec, "Statistical analysis of glottal pulses in speech under psychological stress," in *Proceedings of the 16th European Signal Processing Conference (EUSIPCO '08)*, Lausanne, Switzerland, August 2008.

[8] M. Sigmund and P. Zelinka, "Analysis of voiced speech excitation due to alcohol intoxication," *Information Technology and Control*, vol. 40, no. 2, pp. 145–150, 2011.

[9] M. Sigmund, "Influence of psychological stress on formant structure of vowels," *Elektronika ir Elektrotechnika*, vol. 18, no. 10, pp. 45–48, 2012.

[10] S. B. Davis, "Acoustic characteristics of normal and pathological voices," *Haskins Laboratories: Status Report on Speech Research*, vol. 54, pp. 133–164, 1978.

[11] J. H. L. Hansen and S. A. Patil, "Speech under stress: analysis, modeling and recognition," in *Speaker Classification I: Fundamentals, Features, and Methods*, C. Müller, Ed., pp. 108–137, Springer, Berlin, Germany, 2007.

[12] P. A. Hancock and J. L. Szalma, *Performance Under Stress*, chapter 1, Ashgate Publishers, 2007.

[13] MathWorks Website, http://www.mathworks.com/.

[14] J. Makhoul, "Linear prediction: a tutorial review," *Proceedings of the IEEE*, vol. 63, no. 4, pp. 561–580, 1975.

[15] M. Airas, "TKK Aparat: an environment for voice inverse filtering and parameterization," *Logopedics Phoniatrics Vocology*, vol. 33, no. 1, pp. 49–64, 2008.

[16] P. Alku, "Glottal wave analysis with Pitch Synchronous Iterative Adaptive Inverse Filtering," *Speech Communication*, vol. 11, no. 2-3, pp. 109–118, 1992.

[17] P. Alku, E. Vilkman, and A. M. Laukkanen, "Estimation of amplitude features of the glottal flow by inverse filtering speech pressure signals," *Speech Communication*, vol. 24, no. 2, pp. 123–132, 1998.

[18] P. Alko, B. Story, and M. Airas, "Evaluation of an inverse filtering technique using physical modelling of voice production," in *Proceedings of the Interspeech 8th National Conference on Spoken Language Processing*, pp. 497–500, 2004.

[19] Formant Analysis Tutorial, http://support.ircam.fr/docs/AudioSculpt/3.0/co/Formant%20Analysis.html.

[20] J. Walker and P. Murphy, "Advanced method for glottal wave extraction," in *Nonlinear Analyses and Algorithms for Speech Processing*, vol. 3817 of *Lecture Notes in Computer Science*, pp. 139–149, Springer, Berlin, Germany, 2005.

[21] H. Pulakka, *Analysis of human voice production using inverse filtering, highspeed imaging, and electroglottography [M.S. thesis]*, Helsinki University of Technology, Espoo, Finland, 2005.

[22] R. Timcke, H. von Leden, and P. Moore, "Laryngeal vibrations: measurements of the glottic wave: part I. The normal vibratory cycle," *AMA Archives of Otolaryngology*, vol. 68, no. 1, pp. 1–19, 1958.

[23] R. B. Monsen and A. M. Engebretson, "Study of variations in the male and female glottal wave," *Journal of the Acoustical Society of America*, vol. 62, no. 4, pp. 981–993, 1977.

[24] P. Alku and E. Vilkman, "Amplitude domain quotient for characterization of the glottal volume velocity waveform estimated by inverse filtering," *Speech Communication*, vol. 18, no. 2, pp. 131–138, 1996.

[25] P. Alku, T. Bäckström, and E. Vilkman, "Normalized amplitude quotient for parametrization of the glottal flow," *Journal of the Acoustical Society of America*, vol. 112, no. 2, pp. 701–710, 2002.

[26] A.-M. Laukkanen, E. Vilkman, P. Alku, and H. Oksanen, "Physical variations related to stress and emotional state: a preliminary study," *Journal of Phonetics*, vol. 24, no. 3, pp. 313–335, 1996.

[27] I. R. Titze and J. Sundberg, "Vocal intensity in speakers and singers," *Journal of the Acoustical Society of America*, vol. 91, no. 5, pp. 2936–2946, 1992.

[28] D. G. Childers and C. K. Lee, "Vocal quality factors: analysis, synthesis, and perception," *Journal of the Acoustical Society of America*, vol. 90, no. 5, pp. 2394–2410, 1991.

[29] P. Alku, H. Strik, and E. Vilkman, “Parabolic spectral parameter—a new method for quantification of the glottal flow,” *Speech Communication*, vol. 22, no. 1, pp. 67–79, 1997.

[30] G. Fant, J. Liljencrants, and Q. G. Lin, “A four-parameter model of glottal flow,” *STL-QPSR*, vol. 26, pp. 1–13, 1985.

[31] C. Gobl, “A preliminary study of acoustic voice quality correlates,” STL-QPSR, 1989.

[32] G. Fant, “The LF-model revisited. Transformations and frequency domain analysis,” *Speech Transmission Laboratory, Quarterly Report*, vol. 2-3, pp. 119–156, 1995.

[33] D. G. Childers and C. Ahn, “Modeling the glottal volume-velocity waveform for three voice types,” *Journal of the Acoustical Society of America*, vol. 97, no. 1, pp. 505–519, 1995.

[34] G. Fant, “The voice source in connected speech,” *Speech Communication*, vol. 22, no. 2-3, pp. 125–139, 1997.

[35] C. Gobl and N. A. Chasaide, “The role of voice quality in communicating emotion, mood and attitude,” *Speech Communication*, vol. 40, no. 1-2, pp. 189–212, 2003.

[36] R. N. J. Veldhuis, “The spectral relevance of glottal pulse parameters,” in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP ’98)*, vol. 2, pp. 873–876, 1998.

[37] J. P. Cabral, S. Renals, K. Richmond, and J. Yamagishi, “Towards an improved modeling of the glottal source in statistical parametric speech synthesis,” in *Proceedings of the of the 6th ISCA Workshop on Speech Synthesis (SSW ’07)*, pp. 113–118, 2007.

[38] H. Li, R. Scaife, and D. O’Brien, “LF model based glottal source parameter estimation by extended Kalman filtering,” in *Proceedings of the Irish Signals and Systems Conference*, 2011.

[39] S. Zhang, Y. Guo, and B. Wang, “Auto-correlation property of speech and its application in voice activity detection,” in *Proceedings of the 1st International Workshop on Education Technology and Computer Science (ETCS ’09)*, pp. 265–268, Wuhan, China, March 2009.

[40] R. Thiruvengatanadhan, P. Dhanalakshmi, and P. S. Kumar, “Speech/music classification using SVM,” *International Journal of Computer Applications*, vol. 65, no. 6, pp. 36–41, 2013.

[41] M. Jalil, F. A. Butt, and A. Malik, “Short-time energy, magnitude, zero crossing rate and autocorrelation measurement for discriminating voiced and unvoiced segments of speech signals,” in *Proceedings of the International Conference on Technological Advances in Electrical, Electronics and Computer Engineering (TAEECE ’13)*, pp. 208–212, Konya, Turkey, May 2013.

[42] D. Hosseinzadeh and S. Krishnan, “Combining vocal source and MFCC features for enhance speaker recognition performance using GMM’s,” in *Proceedings of the IEEE 9th Workshop on Multimedia Signal Processing*, pp. 365–368, Crete, Greece, October 2007.

[43] M. Farrús, J. Hernando, and P. Ejarque, “Jitter and shimmer measurements for speaker recognition,” in *Proceedings of the 8th Annual Conference of the International Speech Communication Association (ISCA '07)*, pp. 778–781, 2007.

[44] M. Farrús and J. Hernando, “Using Jitter and Shimmer in speaker verification,” *IET Signal Processing*, vol. 3, no. 4, pp. 247–257, 2009.

[45] “What is Voice Intensity,” http://science.blurtit.com/2751364/what-is-voice-intensity.

[46] Tutorial on “Control of Vocal Intensity and Efficiency” by “National Center for Voice and Speech”, http://www.ncvs.org/ncvs/tutorials/voiceprod/equation/chapter9/.

[47] PRAAT phonetic software Website, http://www.praat.org/.

[48] Speech samples download, http://www.mattmontag.com/projectsp age/academic/speech.

[49] J. H. Hansen and S. E. Ghazale, “Getting started with SUSAS: a speech under simulated and actual stress database,” in *Proceedings of the European Conference on Speech Communication and Technology (EUROSPEECH ’97)*, pp. 1743–1746, Rhodes, Greece, 1997.

[50] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, “A database of German emotional speech,” in *Proceedings of the 9th European Conference on Speech Communication and Technology (Interspeech-Eurospeech ’05)*, pp. 1517–1520, Lisbon, Portugal, September 2005.

[51] D. Ververidis and C. Kotropoulos, “Emotional speech recognition: resources, features, and methods,” *Speech Communication*, vol. 48, no. 9, pp. 1162–1181, 2006.

[52] “Emotional Speech Database List,” Website, http://download.springer.com/static/pdf/172/bbm%253A978-90-481-3129-7%252F1.pdf?auth66=1415101186_00f43dd61d2c766927757e8949e-1f378&ext=.pdf.

[53] N. Banda and P. Robinson, “Noise analysis in audio-visual emotion recognition,” in *International Conference on Multimodal Interaction*, pp. 1–4, Alicante, Spain, November 2011.

[54] K. V. Krishna Kishore and P. Krishna Satish, “Emotion recognition in a speech using MFCC and wavelet features,” in *Proceedings of the 3rd IEEE International Advance Computing Conference (IACC ’13)*, pp. 842–847, Ghaziabad, India, February 2013.

[55] SAVEE Database Website, http://personal.ee.surrey.ac.uk/Personal/P.Jackson/SAVEE/.