An Investigation into the Voice Energy Level of Pronounced Persian Explosive Consonants by Signal Processing Approach

Hadi JAFARI[1], Peyman JALALI[2], Sina VARAHRAM[2,*],
Reza HASSANNEJAD[2], and Mir Mohammad ETTEFAGH[2]

[1] Faculty of Humanities, Payame Noor University of Qazvin, P.O.Box 34199-14917, Qazvin, Iran
[2] Vibration and Modal Analysis Research Lab, Faculty of Mechanical Engineering, University of Tabriz, P.O.Box 51666-14766, Tabriz, Iran

* Corresponding Author E-mail Address: s.varahram90@ms.tabrizu.ac.ir

Abstract

A voice is a sound which is created by vibration of the vocal cords, which means that, the vocal cords approach to each other by passing the air through the larynx and followed by different sounds. This vibration is under the influence of the statuses of tongue, teeth and lips, as well as the other effective factors. This vibrating air causes some minor changes around the individual speaker, which is called the voice. Some of the most important procedures in the voice signal analyzing is the signal processing methods in frequency and time-frequency domains; the most conventional approach in frequency domain is the Fast Fourier Transform (FFT) method, on the other hand the time-frequency signal processing methods can identify the frequency components of the signal and extract its time-varying characteristics; moreover they are effective tools for extracting the information of non-stationary signals. Time-frequency analysis methods can be classified into linear and nonlinear categories. The Short-Time Fourier Transform and the Wavelet Transform are the most conventional linear methods. This paper investigates the differences in the energy levels of explosive consonants in voiced mode and unvoiced mode using the most conventional frequency analysis and linear time-frequency analysis methods; these approaches are very effective and profitable in phonetic science and biomedical engineering science in order to diagnosing and determining the severity of laryngeal diseases and brain injuries. Results of this study indicate that the energy level of explosive consonants in voiced mode is more than unvoiced mode; also using Fast Fourier and Wavelet Transforms presents the results in better resolution and better image quality than the short-time Fourier Transform.

Keywords: Explosive consonant, Fourier transform, Fast Fourier transform, Short time Fourier transform, Wavelet transform.

1. Introduction

The collision of two or more objects causes the displacement of the air molecules, which create a voice. The speed of this transition is 343 meters per second and it is in the form of sine wave, in which each rise and fall constitutes a cycle. The number of cycles per second is equal to the frequency of the voice, which its unit is Hz. Thus, as the number of cycles per second increases, the voice is soprano and on the other hand, as the number of cycles per second decreases, the voice is bass. Since there are infinite possibilities of collision of the objects, then there are infinite kinds of voices; but just the voices which are in the frequency range of 20-20000 Hz are in the human hearing threshold; in the meantime, the frequency of language voices are between 100-4000 Hz and the number of the sounds which can be produced by human are infinite.

One of the aspects of consonant description in phonology is the position of glottis; if the vocal folds vibrate during the consonant articulation, it will be a voiced consonant; but if the vocal folds don't vibrate, it will be an unvoiced consonant. One of the main principles of acoustics in speech perception is the Voice Onset Time (VOT), which is the cause of distinguish between the explosive consonants in different languages. The VOT can be positive, zero or negative. Lisker and Abramson [1] expressed that if the vibration of vocal folds occurs after the release of consonant, the VOT will be positive, which is called voice lag; on the other hand, if the vibration of vocal folds occurs before the release of consonant, the VOT will be negative, which is called voice lead and the time interval is called prevoicing. If the vocal folds vibrate simultaneously with release of consonant, the VOT will be zero. Klatt [2] presented that the amount of VOT depends on the explosive consonant place of articulation as well as the nature of vowel or the sonorant consonant which comes after the explosive one. Volaitis and Miller [3] investigated the effect of the consonant place of articulation on the initial-syllable occlusive consonants distinguishing based on the VOT. The results indicated that the VOT increases by changing the consonant place of articulation from the lips to the soft plate and this status also depends on the length of syllable. Whiteside and Irving [4] researched into the differences of the VOT of English explosive consonants in the beginning place of the syllable and preceding the vowel, according to the gender of articulation by 5 men and 5 women; the results proved that generally the VOT value by women is more than men. In another research, Whiteside and Irving [5] investigated the gender differences effects on the VOT of English explosive consonants and the stress placement of the beginning of a word and preceding vowel. The results indicated that the average of the VOT of unvoiced explosives consonants by women is more than men and the average of the VOT of voiced explosive consonants by women is less than men. Koenig [6] expressed that the control of voicing in occlusive consonants is accomplished by the VOT and in form of timing between the speech organs. Allen et al. [7] demonstrated that the VOT is a time characteristic and it can firmly determine the voiced occlusive consonant, which varies by an individual speaker to another one. They also expressed that one of the main effective factors on VOT is the consonant place of articulation. However, there are some differences about the place of articulation from one language to another one. Whiteside et al. [8] investigated the effect of gender on VOT of explosive consonants /p,b,t,d,k,g / in vowel textures /i,a/. The results indicated that the VOT by girls is significantly more than boys.

This paper investigates the energy levels of voiced and unvoiced explosive consonants, which can be used as a recognition method to determine the severity of injury in the larynx of patients and to diagnose the disease. The conventional method for analyzing the signals is the fast Fourier transform method. A signal can be decomposed to the sinusoidal elements, which means the frequency constituents of the signal [9]; but this transformation has some limitations and disadvantages. In process of Fourier transformation, some information (space-time) is lost. The condition for using the Fourier transform is that the signal should be stationary, while the voice signals is not usually stationary. The Fourier transform specifies if a particular frequency exist in a signal or not, but it does not specify the place of that frequency [10]. In order to avoid these problems, the time-frequency transforms can be used. Time-frequency transforms can be classified into linear and nonlinear categories. One of the linear transformation methods is the short time Fourier transform, which divides the signal into the smaller segments, so that the signal can be assumed to be stationary in the small segments, then the Fourier transform of each segment (window) is taken and the results are adjoined to each other. The main problem of the short time Fourier transform is that it cannot determine what spectral component exist in the particular moment and it is only possible to determine the existing frequency band in each interval. As much as the width of the window is narrower, the time resolution will be better and the stationary assumption of the signal will be taken in better way, but the frequency resolution will be worse; on the other hand, a wide window presents a good frequency resolution and a weak time resolution, moreover, the wide window may be in conflict with the stationary assumption [10]. In the wavelet transform, both the time and frequency resolutions vary in time-frequency diagram, without the

Heisenberg's uncertainty principle is violated. The difference of the wavelet transform from the short time Fourier transform is that the width of the window varies for each spectral component. This method presents a good time resolution and a weak frequency resolution in high frequencies, but presents a good frequency resolution and a weak time resolution in the low frequencies [11]. In this paper the frequency transformation (Fast Fourier Transform) and the linear time-frequency transformations including the short time Fourier transform and the wavelet Transform are utilized to investigate the distribution of energy level of explosive consonants in Persian language.

2. Phonology of Persian explosive consonants

Persian language has 23 consonants and 6 vowels; 8 of the 23 consonants are explosive, including bilabial explosive: /p,b/, dental explosive: /t,d/, palatal explosive: /k,g/, velar explosive: /G/ and glottal explosive: /?/. The main feature of all the explosive consonants is their two-stage of articulation: closure and release of the speech organs [12]. Since the closure phase suddenly turns into the release phase and the airflow is pushed to outward with pressure, so the produced consonant is called explosive consonant. The mean of closure is full connection between the active and passive speech organs during the explosive consonant articulation, so that prevents the passage of airflow. The active organs are the movable parts of the vocal tract such as lips, tongue and lower jaw; and the passive organs are unmovable parts, such as palate and upper teeth [12]. Oral explosive consonants are the group of occlusive consonants which are articulated by the closure of oral cavity. Explosive sounds can be articulated by vibration of vocal folds (voiced) or without the vibration of vocal folds (unvoiced). Since the aim of this study is only the consideration of oral consonants, the glottal explosive is not investigated. Moreover, usually the closure and release phases do not occur during the articulation of velar explosive /G/, therefore the data associated with these consonants were eliminated from the study. In table 1, the consonants are presented in three states of explosion.

Table 1: Voiced and unvoiced explosive consonants

	Bilabial explosive	Dental explosive	Palatal explosive
voiced	b	d	g
unvoiced	p	t	k

Data acquisition of explosive consonants which are pronounced by linguistics specialists are performed with MATLAB. Time histories of these signals are depicted in figure 1.

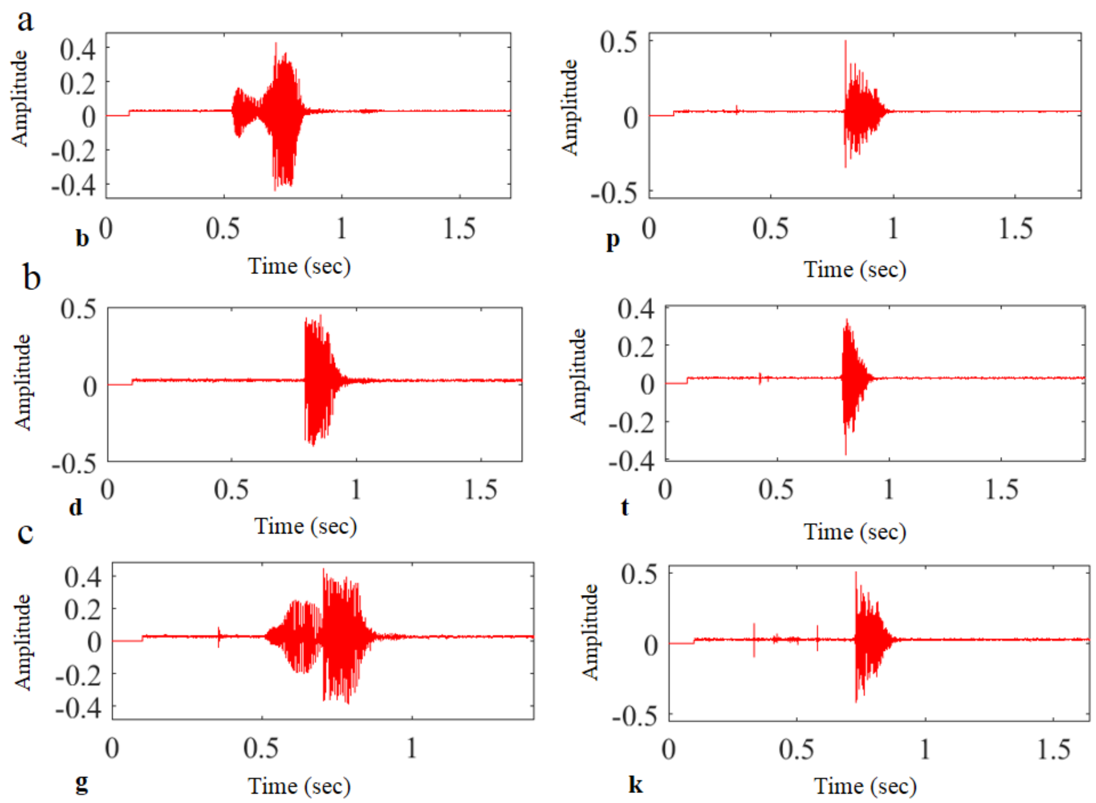

Fig. 1: Time history diagrams of pronounced explosive consonants
a. Bilabial explosive b. Dental explosive c. Palatal explosive

3. Frequency and time frequency analysis in signal processing

Assigning time or space indexes to the measured data set of a measurable phenomenon generates other data, which can be used for extracting specific information. In the most cases, the investigated data sets are in argument space (time or space), which are not always the best type of displaying. In other words, in the most cases, many useful information of signals, lies in their frequency domain.

In order to extract useful information from the signals, mathematical transformations are applied to them. This information is not recoverable from the primary signals (argument space) easily. In the following sections, the most important transformation functions in frequency and time-frequency domains are investigated and applied on the voice signals.

4. Frequency domain transform

4.1. Fourier Transform

In the nineteenth century, Fourier who was a French mathematician demonstrated that any periodic function can be represented as the sum of numerous complex exponential and periodic functions. In other words, Fourier transform decomposes a signal into the complex exponential functions with periodic frequencies, this procedure is a continuous time process and is called continuous time Fourier transform (CFT) which is carried out by the following equation [13].

$$X(f)=\int_{-\infty}^{\infty} x(t)e^{-j2\pi ft}dt \qquad (1)$$

For discrete approach, Eq. (1) is defined as follows, which is called Discrete Time Fourier Transform (DTFT):

$$X_k = \sum_{n=0}^{N-1} x_n e^{-j2\pi kn/N} \qquad k = 1, 2, ..., N \tag{2}$$

4.2. Fast Fourier Transform

Fast Fourier transform (FFT) decomposes a data set to the components with different frequencies [14]. Sometimes, direct calculation of discrete time Fourier transforms by its main definition is very time-consuming; therefore, FFT is an approach to calculate the same results in a shorter time. Calculation of discrete time Fourier transform for n points, takes n^2 mathematical operations by the main definition, while the FFT can present the same result by $n.log^n$ operations. This difference of speed can be noticeable, especially for big data sets; calculation time can be reduced several levels in some cases, where the number of points (n) may actually be more than thousands or millions, therefore, the majority of researchers implement the FFT instead of the discrete time Fourier transform because of this great improvement, and accordingly the FFT is applicable in wide variety of digital signal processing, which is expressed as the following equation:

$$\begin{aligned} X(k) &= \sum_{n=0}^{N-1} x(n) W_N^{kn} \\ &= \sum_{r=0}^{N/2-1} g(r) W_N^{k(2r)} + \sum_{r=0}^{N/2-1} h(r) W_N^{k(2r+1)} \qquad k = 0,1,...,N-1 \\ &= \sum_{r=0}^{N/2-1} g(r) W_N^{2kr} + W_N^k \sum_{r=0}^{N/2-1} h(r) W_N^{2kr} \end{aligned} \tag{3}$$

where $W_N^{kn} = e^{-j\frac{2\pi}{N}kn}$ and also:

$$\begin{aligned} W_N^{2kr} &= \left(e^{-j2\pi/N}\right)^{2kr} = \left(e^{-j2\pi/(N/2)}\right)^{kr} = W_{\frac{N}{2}}^{kr} \\ \Rightarrow X(k) &= \sum_{r=0}^{N/2-1} g(r) W_{N/2}^{kr} + W_N^k \sum_{r=0}^{N/2-1} h(r) W_{N/2}^{kr} \\ &= G(k) + W_N^k H(k) \end{aligned} \tag{4}$$

and finally, general form of FFT is defined as:

$$\begin{aligned} X(k) &= G(k) + W_N^k H(k) \qquad k = 0,1,...,N-1 \\ G(k) &= \sum_{r=0}^{N/2-1} g(r) W_{N/2}^{kr} = \sum_{r=0}^{N/2-1} x(2r) W_{N/2}^{kr} \\ H(k) &= \sum_{r=0}^{N/2-1} h(r) W_{N/2}^{kr} = \sum_{r=0}^{N/2-1} x(2r+1) W_{N/2}^{kr} \end{aligned} \tag{5}$$

The fast fourier transfor of voiced and unvoiced consonants are depicted in figure 2.

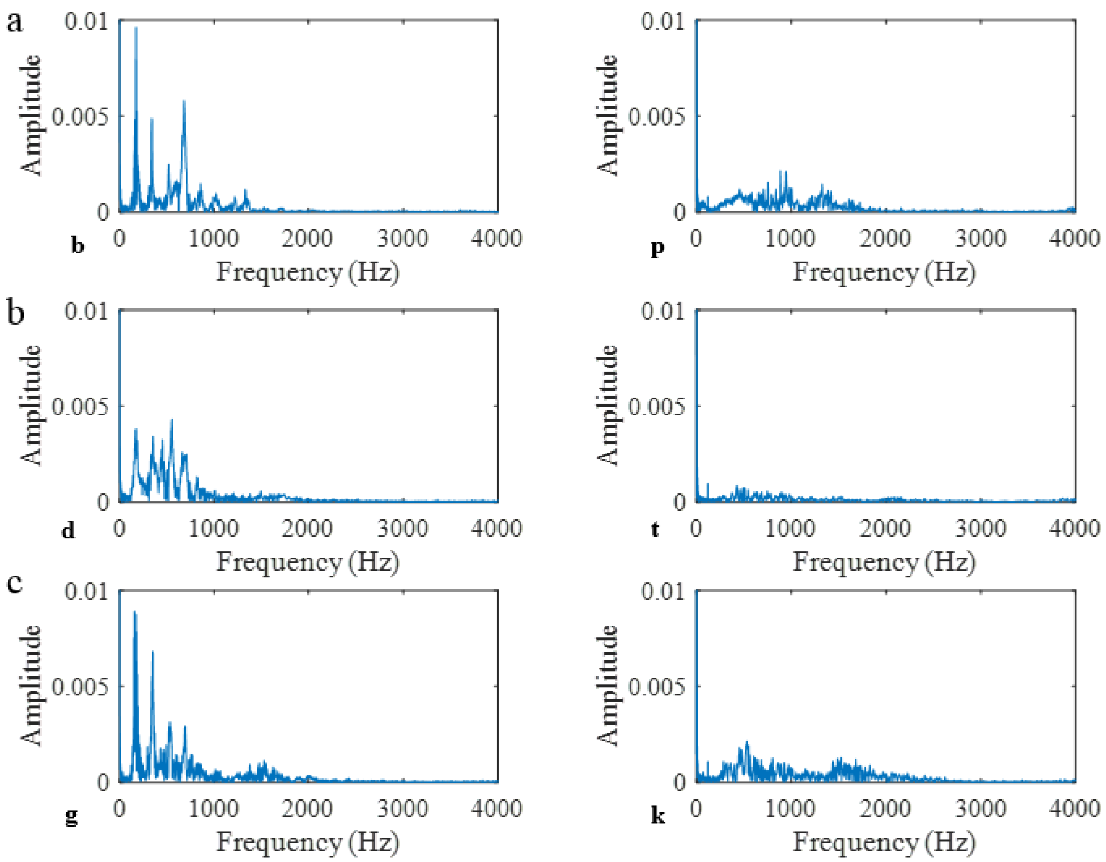

Fig. 2: Fast Fourier transform diagrams of explosive consonants
a. Bilabial explosive b. Dental explosive c. Palatal explosive

As seen in figure 2, energy level of voiced consonants is more than unvoiced consonants, which is in lower frequencies.

5. Time-frequency domain transform

5.1. Short Time Fourier Transform

In the process of short time Fourier transform (STFT) calculation, the non-stationary data set or the signal is divided into the smaller segments so that the signal in each segment is stationary; for this purpose, the window function $\omega(t-\tau)$ is multiplied by the function which is supposed to be transformed; the range of effects of mentioned window function corresponds to the length of the segments in which the signal is assumed to be stationary. By sliding the window function along the time axis, the Fourier transform (which is a one-dimensional function) is taken from the signal, which results in a two-dimensional representation of the primary signal [15,16] and can be expressed as:

$$STFT\{x(t)\}(\tau,\omega) \equiv X(\tau,\omega) = \int_{-\infty}^{\infty} x(t)w(t-\tau)e^{-j\omega t}dt \tag{6}$$

where $w(t)$ is the window function, $x(t)$ is the measured signal and $x(t)w(t-\tau)$ is actually the complex function, which specifies the phase and magnitude of the signal over time and frequency; if the magnitude of the STFT raises to the power of 2, it will present the spectrogram of the function which is represented by the following equation:

$$spectrogram\{x(t)\}(\tau,\omega) \equiv |X(\tau,\omega)|^2 \tag{7}$$

In this paper, the Hanning window is used for spectral estimation of the time-frequency functions, which is defined as follows:

$$Hann(\omega)=\frac{1}{2}+\frac{1}{2}\cos(\frac{\omega}{\tau}) \quad (8)$$

STFT not only presents the frequencies of the signal, but also specifies their occurrence time.

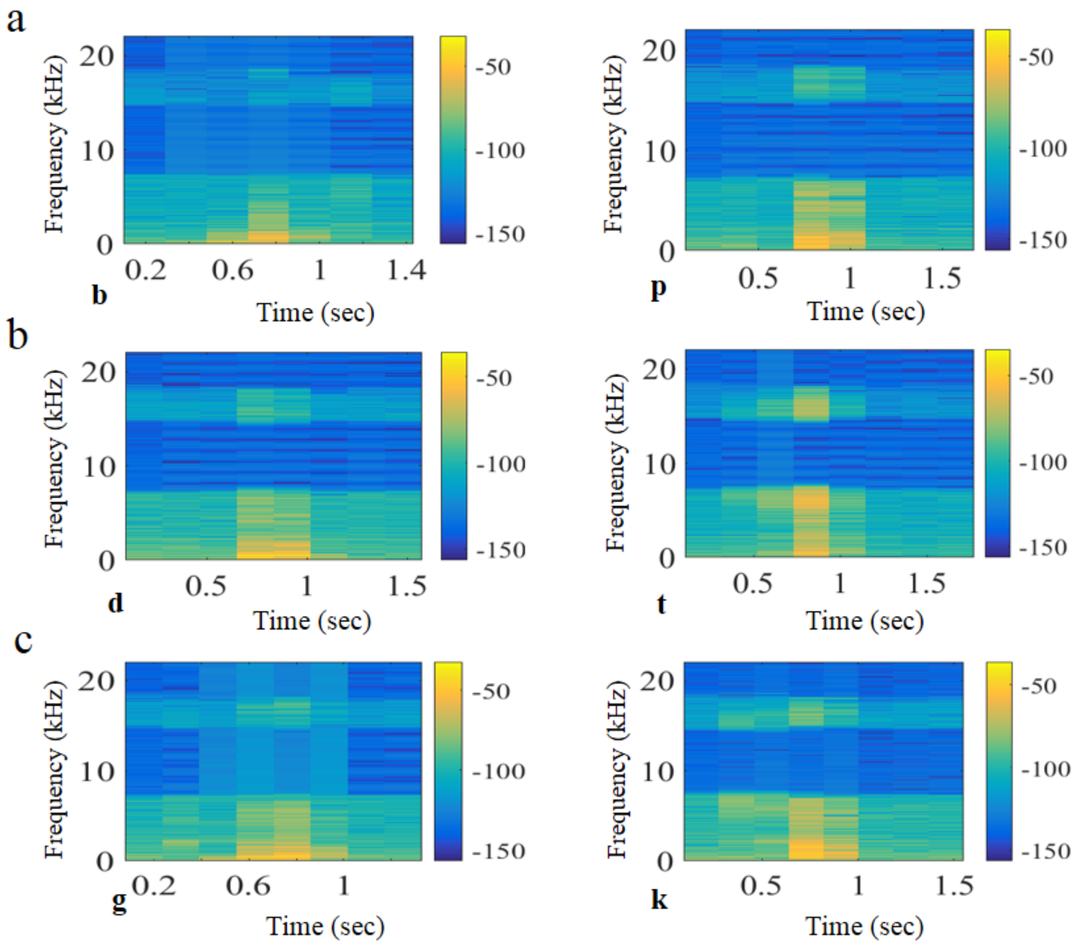

Fig. 3: Short time Fourier transform diagrams of explosive consonants
a. Bilabial explosive b. Dental explosive c. Palatal explosive

As seen in figure 3, the dominant frequency distribution of all three voiced explosive consonants occurs in lower frequencies and greater time intervals than the unvoiced explosive consonants; in addition, determination of exact occurrence time of frequency is difficult and only the frequency bands in a time interval is detectable.

5.2. Wavelet Transform

The wavelet theorem is presented in order to overcome the confronted problems in Fourier transformation. In this method, the procedure of signal dividing into the different segments is carried out by scaling and transmitting of a function; this function is transmitted along the data set and the spectrum of each section is calculated. This procedure is repeated by other functions with different scales, which finally results in argument-frequency data sets. The main feature of wavelet transform compared to the STFT is that the entire base functions are obtained by transmitting and scaling of a parent function (parent wavelet).

Wavelet transform is defined by the inner product of signal *x(t)* and the base functions. The family of wavelets include the functions which are obtained from the parent functions *ψ(t)*: [17, 18]

$$\psi_{a,b}(t) = \frac{1}{\sqrt{a}}\psi\left(\frac{t-b}{a}\right), \qquad a > 0, \qquad b \in R \tag{9}$$

where *a* and *b* denote the transmitting and scaling parameters; transmitting of a wavelet means to delaying it, which causes the wavelet to transmit along the signal in the right direction to the end of it.

Unlike the Fourier transform, the frequency parameter does not exist directly in the wavelet transform, but instead the scaling parameter exists, which is correlated inversely with frequency. The scaling parameter, as its meaning, operates as a mathematical processor to expand and contract the signal; just like the concept of map scaling, large scales correspond to the general overview of the signal, regardless of any details (correspond to the low frequencies) and the small scales correspond to details of the signal (correspond to the high frequencies).

Actually the wavelet transform is evaluation of the similarity between the signal and the base functions (wavelets), which means the similarity between the frequency information. The result of wavelet transformation is a matrix that the columns denote the displacement per time and the rows denote the investigated scale. In other words, the coefficients of wavelet transform denote the proximity of signal to the wavelet in desired scale.

The differences between the various wavelet functions, which are obtained by a family, are originated from the variance of coefficients *a* and *b*. If the coefficient *a* is small, high frequency components of the signal will be processed and if the coefficient *a* is large, low frequency components of the signal will be processed [17, 18]. The continuous wavelet transform is expressed by Eq. (10):

$$CWT(t,a) = \frac{1}{\sqrt{a}}\int_{-\infty}^{\infty} x(u)\psi\left(\frac{u-b}{a}\right)du \tag{10}$$

where *ψ(t)* is the wavelet parent function, and *x(t)* is the considered signal. In this research, the Morlet parent function is used due to the harmonic nature of the signal of pronounced voice. The Morlet function is represented by Eq. (11):

$$\psi(t) = \cos(5t)e^{-t^2/2} \tag{11}$$

The energy spectrum of wavelet transform is called scalogram, which is defined as Eq. (12):

$$sca\log ram(t,a) = \frac{1}{a}\left(\int_{-\infty}^{\infty} x(t)\psi\left(\frac{u-b}{a}\right)du\right)^2 \tag{12}$$

The continuous wavelet transform of voiced and unvoiced explosive consonants are presented in Fig. 4.

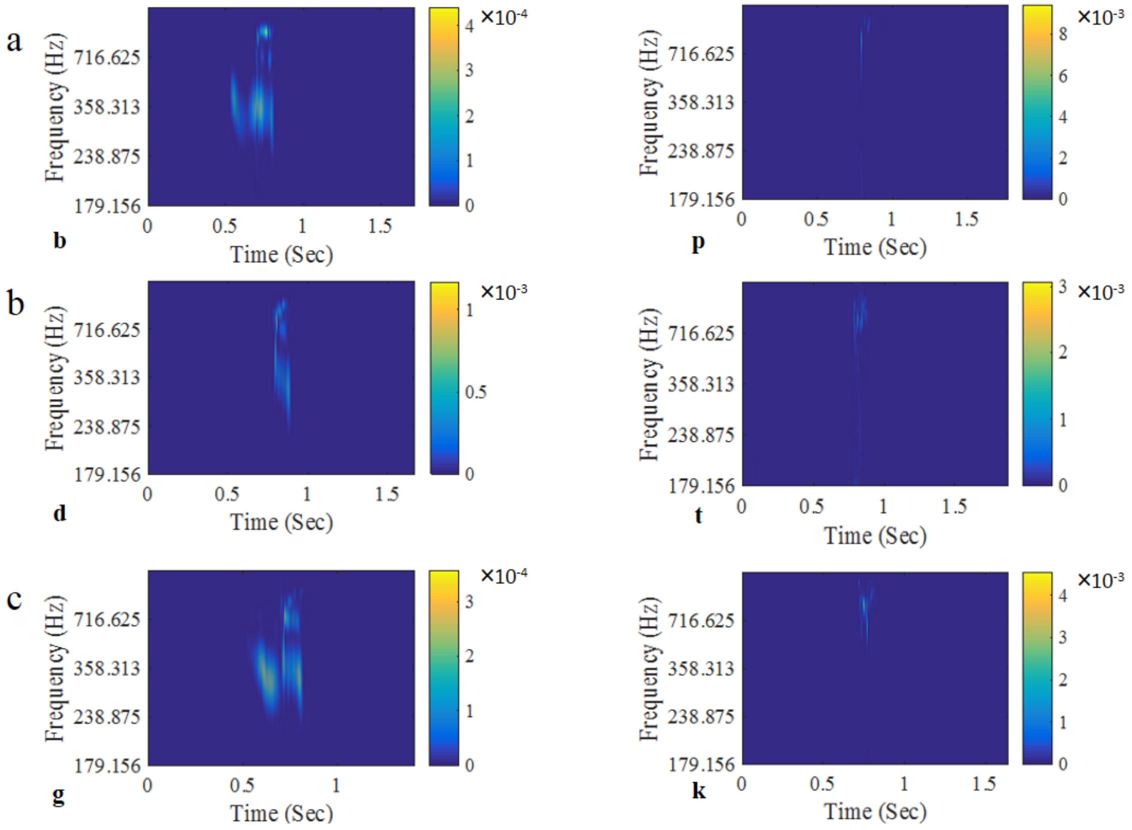

Fig. 4: Wavelet transform diagrams of explosive consonants
a. Bilabial explosive b. Dental explosive c. Palatal explosive

As seen in figure 4, unlike the STFT, which the simultaneous detection of frequency and time in signal was difficult and only the determination of existing frequency bands in each interval was possible, the mentioned problems are not observed in wavelet transform.

6. Conclusion

In this paper the conventional time-frequency signal processing methods were used to investigate the energy level of voice signals of pronounced Persian explosive consonants. In the STFT method, the resolutions of time and frequency are in contrast with each other, while the wavelet transform presents the better resolution than the STFT method. The results indicated that the energy level of pronounced voiced consonants is more than unvoiced consonants which is obvious in the frequency domain by FFT; also the pronunciation of letter /g/, among the voiced consonants and the pronunciation of letter /p/, among the unvoiced consonants have the most energy level and the labial consonants generally have the least energy level. According to results, utilizing the wavelet transform can be useful for analyzing the energy of laryngeal voice signals and also can be effective and advantageous in biomedical engineering science and phonetic science.

References

[1]. Lisker, Leigh, and Arthur S. Abramson. "A cross-language study of voicing in initial stops: Acoustical measurements." *Word* 20.3 (1964): 384-422.
[2]. Klatt, Dennis H. "Voice onset time, frication, and aspiration in word-initial consonant clusters." *Journal of Speech, Language, and Hearing Research* 18.4 (1975): 686-706.
[3]. Volaitis, Lydia E., and Joanne L. Miller. "Phonetic prototypes: Influence of place of articulation and speaking rate on the internal structure of voicing categories." *The Journal of the Acoustical Society of America* 92.2 (1992): 723-735.

[4]. Whiteside, Sandra P., and Caroline J. Irving. "Speakers' sex differences in voice onset time: some preliminary findings." *Perceptual and motor skills* 85.2 (1997): 459-463.

[5]. Whiteside, Sandra P., and Caroline J. Irving. "Speakers' sex differences in voice onset time: a study of isolated word production." *Perceptual and motor skills* 86.2 (1998): 651-654.

[6]. Koenig, Laura L. "Laryngeal factors in voiceless consonant production in men, women, and 5-year-olds." *Journal of Speech, Language, and Hearing Research* 43.5 (2000): 1211-1228.

[7]. Allen, J. Sean, Joanne L. Miller, and David DeSteno. "Individual talker differences in voice-onset-time." *The Journal of the Acoustical Society of America* 113.1 (2003): 544-552.

[8]. Whiteside, Sandra P., Luisa Henry, and Rachel Dobbin. "Sex differences in voice onset time: A developmental study of phonetic context effects in British English." *The Journal of the Acoustical Society of America* 116.2 (2004): 1179-1183.

[9]. Howell, K. B. "The Transforms and Applications Handbook: Ed. Alexander D. Poularikas Boca Raton: CRC Press LLC, 2000." (2000).

[10]. Polikar. The wavelet tutorial. Available http://users.rowan.edu/~polikar/WAVELETS/WTtutorial.html, Acessed October 21,(2009).

[11]. Cohen, Albert, and Jelena Kovacevic. "Wavelets: The mathematical background." *Proc. IEEE*. 1996.

[12]. Hyman, Larry M. *Phonology: theory and analysis*. Harcourt College Pub, 1975.

[13]. Bracewell, Ron. "The fourier transform and iis applications." *New York* 5 (1965).

[14]. Katoh, Kazutaka, et al. "MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform." *Nucleic acids research* 30.14 (2002): 3059-3066.

[15]. Hlawatsch, Franz, and G. Faye Boudreaux-Bartels. "Linear and quadratic time-frequency signal representations." *IEEE signal processing magazine* 9.2 (1992): 21-67.

[16]. Auger, François, et al. "Time-frequency toolbox." *CNRS France-Rice University* (1996): 46.

[17]. Boashash, Boualem. *Time-frequency signal analysis and processing: a comprehensive reference*. Academic Press, 2015.

[18]. Rioul, Olivier, and Martin Vetterli. "Wavelets and signal processing." *IEEE signal processing magazine* 8.LCAV-ARTICLE-1991-005 (1991): 14-38.