Indian Journal of Science and Technology, *Vol 10(1), DOI: 10.17485/ijst/2017/v10i1/109386, January 2017*
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

Classification of Sex based Speech Differentiation in Healthy Human Beings based on Voiced and Unvoiced Components

Saurabh V Raut[1], Lavanya C Panthangi[1], B. G. Akhil[1], Syed Faisal Ali[2], H. S. Sanjay[3] and S. Bhargavi[4]

[1]Department of Medical Electronics, M S Ramaiah Institute of Technology, Bangalore – 560054, Karnataka, India
[2]MedNXT Innovative Technologies, Bangalore, India
[3]Jain University, Department of Medical Electronics, MSRIT, Bangalore – 560054, Karnataka, India
[4]Department of Telecommunication Engineering, SJC Institute of Technology, Chikkaballapura – 562101, Karnataka, India

Abstract

Background/Objective: The objective of the present study is to classify a given speech signal by using energy as a differentiating parameter into voiced and unvoiced components due to the fact that the voiced components have a higher energy than their unvoiced counterparts. **Method/Statistical Analysis:** This is accomplished by dividing the speech signal into frames and by computing the short time energy of each frame. The recorded speech signal is segmented and then the energy component of these frames are obtained and then classified into voiced and unvoiced components. The current protocol involves 44 subjects, both males and females of no known vocal pathology. Predefined set of words, both in Kannada and English were recorded in a noise proof environment which was then separated into voiced and unvoiced components using MATLAB tool. **Findings:** The results proved a successful discrimination of the speech signal into voiced and unvoiced components based on the statistical parameters calculated for voiced as well as unvoiced components thereby providing a definite cue towards an automated approach to differentiate the speech into voiced and unvoiced components using statistical parameters. **Application/Improvements:** Such an approach can further be useful in various speech processing as well as speech recognition applications.

Keywords: Frame-by-frame Processing, Short Time Energy (STE), Voiced Speech, Unvoiced Speech, Windowing

1. Introduction

One of the most intriguing signals that humans work with is speech. Speech is the expression or the ability to express thoughts and feelings by sounds. The study of speech signals and the processing methods of these signals are known as speech processing.[1] Speech processing is used for understanding the speech as a means of communication, transmission and reproduction, automatic recognition and extraction of information and to discover some physiological characteristics of an individual. One major application of speech processing is speech recognition. Speech recognition or "automatic speech recognition" incorporates knowledge in the fields of linguistics, computer sciences and electronics to develop methodologies and technologies which enables the recognition and translation of spoken language into text using computers and computerized devices which involve smart technologies and robotics.[2]

Each word, when spoken, is generated out of phonetic combination of a restricted set of vowels and consonants. Speech production is spontaneous such as when a person outputs the words for conversation as a reaction to visual stimuli, reading a written word aloud or depiction of a picture, or a vocal imitation such as in speech repetition. The speech production can be broadly classified, based

*Author for correspondence

on the type of input excitation phenomenon, into three components namely the voiced Speech (nearly periodic in nature), unvoiced Speech (random in nature) and the silence region (absence of excitation in the signal)[3].

1.1 Voiced Speech

A speech is said to be voiced when the input excitation is almost periodic impulse sequence and the corresponding speech looks visually periodic. The air exhaling out of lungs through the trachea is interfered periodically by the vibrating vocal folds which takes place during the production of voiced speech. Because of this, the glottal wave which is produced exhilarates the speech production system which results in voiced speech. Voiced signals are associated with periodicity and can be measured by autocorrelation analysis. This period is known as pitch. During the pronunciation of a phoneme, when the vocal chords vibrate, voiced speech is produced. Signals which tend to be louder like the vowels /a/, /e/,/i/, /u/, /o/ are usually voiced signals.[4]

1.2 Unvoiced Speech

If the input excitation is random nose-like, then the following speech is also random noise- like without any periodicity. This is known as unvoiced speech. The air exhaling out of lungs through the trachea is not interfered periodically by the vibrating vocal folds. However, total or partial closure occurs somewhere along the length of the vocal tract starting from the glottis. This leads to complete obstruction of airflow which results in frication excitation and exhilarates the vocal tract system to produce unvoiced speech. The production of unvoiced speech does not entail the involvement of the vocal chords. Unvoiced speech tends to be more abrupt like the stop consonants /p/, /t/, /k/.[5]

1.3 Silence Region

The silence region separates the production of voiced and unvoiced speech which is successively produced by the speech production system. No speech output is observed during the silence region because no excitation is given to the vocal tract.[6]

2. Background

The speech production of a person starts from the infant's initial ramble and by the age of five is modified into fully advanced speech. There are three important levels of processing involved in manufacturing speech. The first is conceptualization, followed by formulation, and finally articulation. The process in which created speech links to a concept due to intention of expressing is called conceptualization. Formulation involves linguistic form which is essential for conveying the desired message created. It includes morpho-phonological encoding, phonetic encoding, and grammatical encoding. The third stage involves implementation of articulatory score by parts of the vocal apparatus (lungs, glottis, larynx, tongue, lips) that results in speech. This is articulation. The phonation created in the glottis, due to the pulmonary pressure provided by the lungs, results in production of normal human speech. It is then modified into different vowels and consonants by the vocal tract. Nevertheless, humans are able to enunciate words in the absence of the use of the lungs and glottis in alaryngeal speech. These are of three types, namely, esophageal speech, pharyngeal speech and buccal speech. The process involved with human speech production is depicted in figure 1.[7]

The major components involved in speech production are as follows

- Lungs: The organ which supplies sufficient air during exhalation for producing speech.
- Trachea: This connects the lungs to the glottis and is also called as windpipe.
- Glottis: This consists of vocal folds or chords and a slit-like opening between them. This obstructs the airflow to generate the required excitation signal for speech production for the specific category of speech.
- Other related organs include Pharynx (throat), Oral cavity and Nasal cavity.

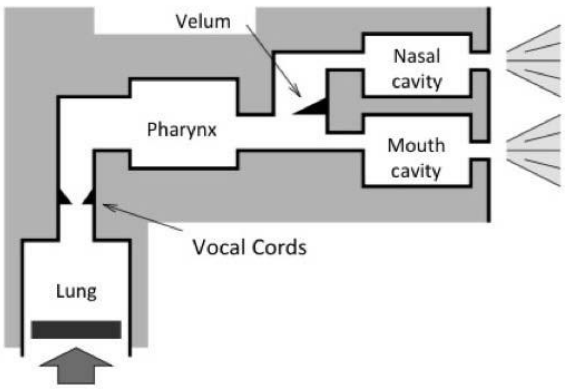

Figure 1. Human SpeechProduction – A block Diagram Perspective.

If the dominant hand is right hand the motor control for speech production relies mostly upon the left cerebral areas of the brain. These areas include the left posterior inferior frontal gyrus, the bilateral supplementary motor area, the left insula, the left primary motor cortex, temporal cortex and it also includes basal ganglia and cerebellum. While considering speech as a signal, the production of speech involves two processes where the sound is commenced in the first stage and filtered in the second stage. This is depicted in figure 2. The source signal which is generated at the glottal level is initially linearly filtered through the vocal tract. The sound resulting by this through radiation loading (lips) is emitted to the nearby air, producing speech. [8]

3. Experimental Protocol

The present protocol utilizes the concept of the energy of a sequence in order to classify speech into voiced, unvoiced frames. This is accomplished by dividing a speech signal into shorter frames and by computing short time energy (STE) of each frame. If the average energy of the speech in a particular frame exceeds a threshold level which is user-defined, then it is declared to be. Else the frame is declared to be unvoiced.

3.1 Speech Processing

Properties of speech change with time. As the sound is being produced, along with it the peak amplitude varies too. Whereas, the property of sound, like pitch, varies across and within voiced sounds. Due of the slowly fluctuating character of the speech signal, it is customary to process speech in frames over which the characteristics and properties of the speech waveform can be presumed to be relatively stable over very short (5-20 msec) intervals. Thus, Frame-by-Frame Processing is used here.

3.2 Frame-by-frame Processing

Framing is a procedure that decomposes the speech signal into a series of overlapping frames as shown in figure 3.

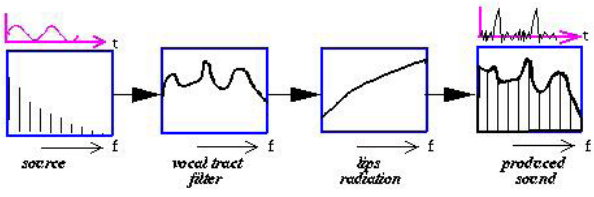

Figure 2. Production and Filtering of Speech.

The speech is processed frame-by-frame in overlapping intervals until entire region of speech is covered by at least one such frame.

3.3 Short-Time Energy

Short-Time energy is a simple short-time speech measurement and is defined by eq 1. This measurement can in a way distinguish between voiced and unvoiced speech segments due to the fact that the unvoiced speech has significantly smaller short- time energy as compared to its voiced counterpart.

$$E_n = \sum_{m=-\infty}^{m=\infty} [x(m)\omega(n-m)]^2 \qquad (1)$$

3.4 Windowing

The choice of the window in short-time speech processing determines the nature of the measurement representation. Opting for a long window will lead to obtain measurements with very little variations but the measurement with a shorter window would not be sufficiently smooth. Keeping this into consideration, Hamming window, as shown in figure 4 is used in the present protocol and is depicted by eq 2.

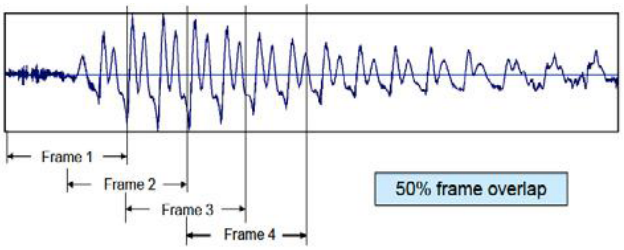

Figure 3. Frame-by-frame Processing of the Signal.

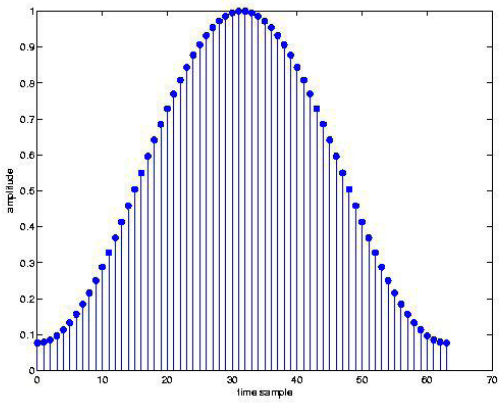

Figure 4. Hamming Window.

This experiment utilizes concepts such as segmentation and averaging to the given speech sequences in order to classify the speech into voiced and unvoiced components. Here, the recorded speech signal is broken down into short frames of 5msec each and then the short-term energy is computed. As the energy of a frame indicates itsloudness, higher short term energy value is expected for a voiced signal as compared to the unvoiced components. The threshold is appropriately defined to classify into voiced and unvoiced components as per the suitability. In the present case, the threshold value is set to be 3J.

$$\omega_H[m] = \begin{cases} 0.54 + 0.46\cos\left(\frac{\pi m}{M}\right) & -M \le m \le M \\ 0 & otherwise \end{cases} \quad (2)$$

Initially, the speech signal is segmented into non-overlapping frames with 48000 samples at a rate of 8000 Hz. The frame size is set to be 10 samples which mean that the short term energy of the speech signal is computed for every 10 samples. Hamming window is chosen over rectangular window due to its higher bandwidth in comparison [9][10]

3.5 Algorithm

Two separate tests are conducted for every subject. In the first trial, a Kannada set of alphabets are made to be uttered where as in the next trial, a predefined English set of words are made to be uttered by the individuals. The protocol given in figure 5 is run for both Kannada and English words. 6 seconds are allotted for every trial for the individual to utter the below mentioned lines during both the trials.

Kannada: ***ಯ (ya), ರ (ra), ಲ (la), ವ (va), ಶ (śa), ಷ (Ṣa) , ಸ (sa), ಹ (ha), ಳ (!a)***

English: WE STUDY BIOMEDICAL DIGITALSIGNAL PROCESSING

The speech is then sampled at 8000 Hz and is converted into non-overlapping frames of 10 samples each using hamming window. The framed signal is converted into a data frame matrix. The matrix is then made sparse. Thereafter, the short time energy is calculated. The sentence used and the graphs of the energy are then plotted and displayed along with a corresponding audio output. After this, a suitable threshold (3 in the present context) is defined and applied to the energy calculated soas to differentiate between voiced and unvoiced speech components. This process is shown in figure 5

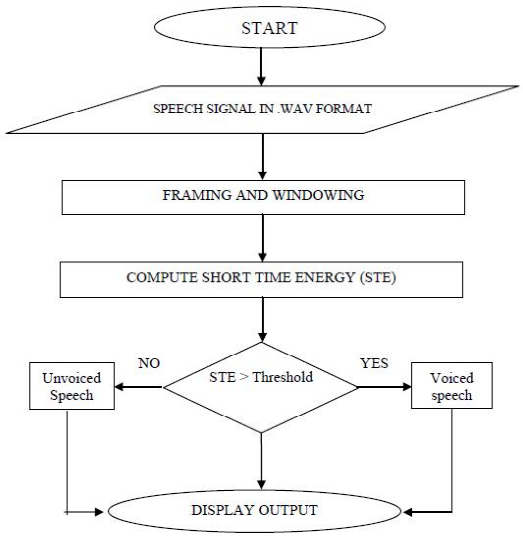

Figure 5. Experimental Paradigm used to Separate Voiced and Unvoiced Components.

4. Results

The short term energy was computed for every frame (frame length was considered to be 10 samples). With a threshold of 3, the frames were classified into voiced and unvoiced components with the latter one having a value lower than the threshold.

The experimental paradigm described in fig 5 has been used to record the speech from 44 subjects, both males and females of no known vocal pathological history. This test was run for both Kannada as well as English sentences, as mentioned prior for the same set of subjects. The speech was recorded and then converted into frames at a rate of 8000 Hz.

The statistical parameters were obtained for the short term energy of the recorded speech as well as for the voiced and unvoiced components of the speech. This is depicted in table 1, table 2 and table 3.

Figure 5 depicts a graphical representation of the statistical parameters obtained for the Kannada speech recorded from males and also the voiced and unvoiced components of the recorded speech.

Figure 6 depicts a graphical representation of the statistical parameters obtained for the English speech recorded from males and also the voiced and unvoiced components of the recorded speech.

Figure 7 represents the statistical parameters calculated for males, for both Kannada and English combined. Fig 8 depicts a graphical representation of the statistical

Table 1. short term energy attributes for the recorded speech

Short-Time Energy Attributes - Unsegregated Speech								
		min	max	mean	median	mode	std	range
Kannada	Males	0.005312	36.55	6.887727	3.312518	0.005312	8.104182	36.54545
	Females	0.025594	40.21864	7.700091	3.596405	0.025594	9.097318	40.19318
English	Males	0.008311	40.07227	4.671636	1.129041	0.008311	7.310773	40.06318
	Females	0.033801	49.30245	5.290855	1.438768	0.033801	8.221136	49.26914
Together	Males	0.006812	38.31114	5.779682	2.22078	0.006812	7.707477	38.30432
	Females	0.029698	44.76055	6.495473	2.517586	0.029698	8.659227	44.73116

Table 2. Short term energy attributes for the unvoiced components of the speech signal

Short-Time Energy Attributes - Unvoiced Speech								
		min	max	mean	median	mode	std	range
Kannada	Males	0	2.999545	0.297245	0.00659	0	0.636282	2.999545
	Females	0	1.391813	0.373709	0.061674	0	0.653441	2.999455
English	Males	0	2.9995	0.363109	0.064844	0	0.647536	2.9995
	Females	0	2.999682	1.984664	0.096324	0	0.672382	2.999682
Together	Males	0	2.999523	0.330177	0.035717	0	0.641909	2.999523
	Females	0	2.190905	1.179186	0.078999	0	0.662911	2.999568

Table 3. Short term energy attributes for the voiced components of the speech signal

Short-Time Energy Attributes - Voiced Speech								
		min	max	mean	median	mode	std	range
Kannada	Males	0	36.45182	6.336091	2.0385	0	8.238455	36.45182
	Females	0	41.02727	7.254973	3.128136	0	9.432045	41.02727
English	Males	0	38.99591	4.149227	0	0	7.306818	38.99591
	Females	0	46.1633	4.907318	0.637136	0	8.459773	49.50882
Together	Males	0	37.72386	5.242659	1.01925	0	7.772636	37.72386
	Females	0	43.59529	6.081145	1.882636	0	8.945909	45.26805

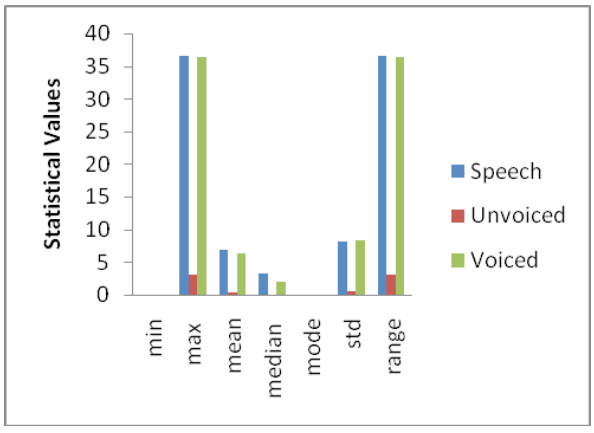

Figure 5. Statistical Parameters for the Kannada Speech of Males.

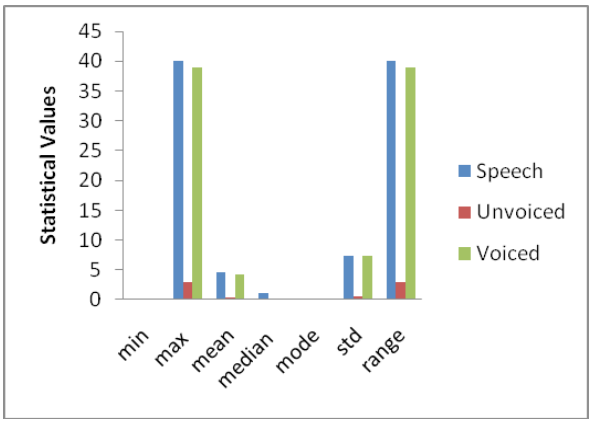

Figure 6. Statistical Parameters for the English Speech of Males.

parameters obtained for the Kannada speech recorded from females and also the voiced and unvoiced components of the recorded speech. Fig 9 depicts a graphical representation of the statistical parameters obtained for the English speech recorded from females and also the voiced and unvoiced components of the recorded speech.

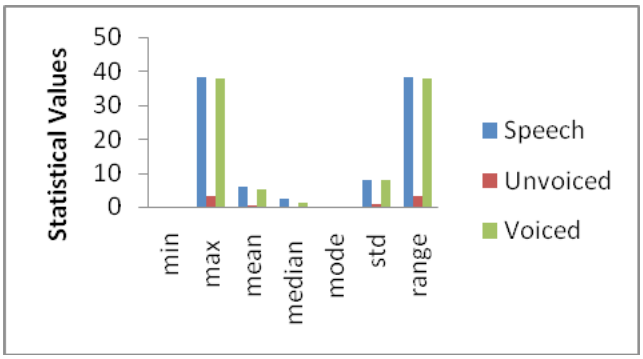

Figure 7. Statistical Parameters for Males – both Kannada and English Speech.

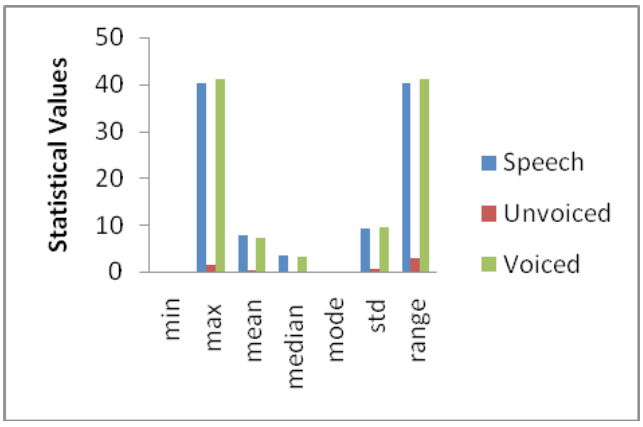

Figure 8. Statistical Parameters for Females for Kannada Speech.

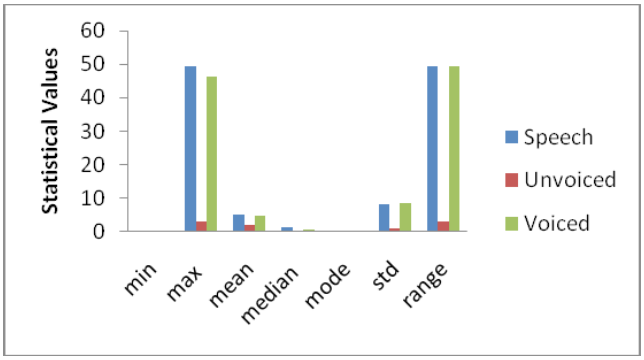

Figure 9. Statistical Parameters of the English Speech Recorded from Females along with the Voiced and Unvoiced. Components.

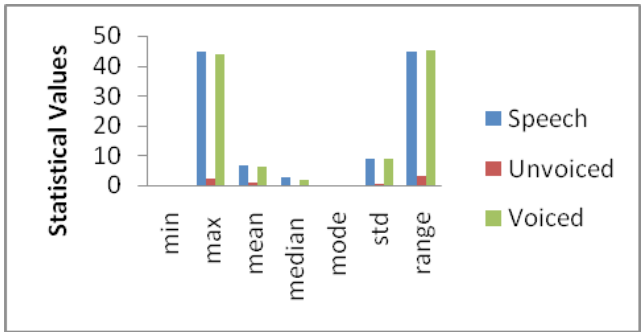

Figure 10. The Statistical Parameters Calculated for Females, for both Kannada and English Combined.

5. Conclusions

The results presented in this work provide a definite cue to prove that the presently proposed method is an efficient approach to classify speech based on sex by demonstrating clear differences between the male and female speech in terms of statistical parameters. The results show that in males the mean value is around 7 for Kannada and 5 for English whereas, for females, it is about 8 for Kannada and 5 for English. Although the Kannada speech shows a variation among males and females, the same is not true for English in case of mean values. Also in case of unvoiced speech, there is not much of a fluctuation in the mean value. But for voiced components, mean proves to be a significant parameter to differentiate between male and female speech for Kannada (variation between 6.33 and 7.25) which is not that evident in case of English (4.14 and 4.9) as shown in Table 1. In other words, one could also conclude that mean can be an efficient parameter for sex based classification for Kannada speech in an automated analysis.

In case of unvoiced components, the statistical parameters are almost similar for Kannada as well as for English. The only variation is in terms of mean wherein the females display a higher mean value than males for both Kannada as well as English speech indicating that mean can also be used to classify male and female voices in case of unvoiced components as well, as shown in table 2 For voiced components of the English and Kannada speech recorded almost all the statistical parameters mentioned in table 3 provide a clear variation in terms of sex. The females exhibit a higher value in every aspect and hence are clearly differentiable when compared to male voices.

From the present analysis, one could easily conclude on the abilities of speech classification based on statistical parameters of voiced and unvoiced components with the aid of short term energy calculated for both male as well

as female voices in case of Kannada as well as for English language. Summarizing, one could say that men have lower short term energy of speech compared to that of women. Also by suitably adjusting the threshold value (3, in the present experiment), the voiced as well as unvoiced components are easily separated. This property of speech could further be probed to explore various speech processing as well as speech recognition based applications in healthcare.

6. Acknowledgements

The present work was carried out in association with M S Ramaiah Institute of Technology, M S Ramaiah Medical College and Teaching Hospital and MedNXT Innovative Technologies, Bangalore and the authors wish to acknowledge the same.

7. References

1. Gold, Ben, Nelson Morgan, and Dan Ellis. Speech and audio signal processing: processing and perception of speech and music. John Wiley and Sons; 2011.
2. Schone, Patrick, et al. Automatic Transcription of Historical Newsprint by Leveraging the Kaldi Speech Recognition Toolkit. Electronic Imaging. 2016;17 (2016): 1–10.
3. Blanco, José Luis, et al. Improving automatic detection of obstructive sleep apnea through nonlinear analysis of sustained speech. Cognitive Computation. 2013; 5(4): 458–72.
4. O'Shaughnessy, Douglas. Acoustic analysis for automatic speech recognition. Proceedings of the IEEE. 2013; 101(5): 1038–53.
5. Gao, Yang. Unvoiced/Voiced Decision for Speech Processing. U.S. Patent Application No. 14/476,547.
6. Abrol, Vinayak, Pulkit Sharma, Anil Kumar Sao. Voiced/non voiced detection in compressively sensed speech signals. Speech Communication.2015; 72: 194–207.
7. Maxfield, Lynn, et al. Intraoral pressures produced by thirteen semi-occluded vocal tract gestures. Logopedics Phoniatrics Vocology.2015; 40(2): 86–92.
8. Flanagan, James L. Speech analysis synthesis and perception. Vol. 3. Springer Science and BusinessMedia, 2013.
9. Sanjay HS, et al. EEG based GAP perception in human beings. International Journal of Advancesin Engineering Research. 2015; 10(2): 92–101.
10. Sanjay HS, Anusha G, Lalitha BS. Auditory perception of random gaps in human beings. International Journal of BiomedicalEngineering and Consumer Health Informatics. 2012; 4(2): 29–31.
11. Dong-Ill Kim and Byung-Cheol Kim. Speech Recognition using Hidden Markov Models in Embedded Platform. Indian Journal of Science and Technology. 2015 December; 8(34):782–4
12. Venkateswarlu S,. Kamesh DBK, Sastry JKR, Radhika Rani. Text to Speech Conversion. Indian Journal of Science and Technology.2016 October; 9(38): 1–4.
13. Hairol Nizam Mohd. Shah, Mohd. ZamzuriAb Rashid, Mohd. Fairus Abdollah, Muhammad Nizam Kamarudin, Chow Kok Lin, ZalinaKamis. Biometric Voice Recognition in Security System. Indian Journal of Science and Technology.2014 February; 7(2): 104–112.
14. Chithra PL Aparna R. Performance Analysis of Windowing Techniques in Automatic Speech Signal Segmentation. Indian Journal of Science and Technology2015 November; 8(29):1–7.