

Speech Quality Assessment Using Audio Features

¹Renuka Devi M.N,

VTU Research Center, PESIT Bangalore South Campus, Bengaluru, India
Visvesvaraya Technological University Belagavi, Karnataka, India
Dept. of CSE, Dayananda Sagar University, Bengaluru, Karnataka, India

²Gowri Srinivasa,

Dept. of Computer Science and Engineering,
PESIT Bangalore South Campus
(currently, PES University), Bengaluru, Karnataka, India.

Abstract

This paper discusses the design of features that aid in the classification of the quality of speech of a speaker. The data used in this study pertains to TED Talks. Since most TED speakers are high achievers and expert orators, we have a rich source of audio cues that define speech that is appealing to a large audience. The features used to categorize the speech quality can be the basis of analyzing the speech quality of novice speakers. Such a system can be used to draw a novice speaker's attention to specific areas of improvement, such an increase in amplitude or maintaining vocal consistency and facilitate directed effort towards improving the quality of one's speech.

We use a speaker classification technique designed and developed in house including Short Term Energy (STE), Zero Crossing Rate (ZCR), Mean power, Pitch, Magnitude and standard deviation. Finally we use an unsupervised classifying method called 'Hierarchical clustering technique' to group speakers into 6 categories.

Keywords: Mean Power, Magnitude, Short Term Energy (STE), Standard Deviation, Zero Crossing Rate (ZCR).

1. Introduction

Speech signals contain a rich repository of features specific to a speaker, such as modulation of voice, intonation, etc., in addition to the content of speech. Every speaker would have unique style of presenting with unique characteristics of speech which makes the speech a special and a popular one or an ordinary one that is not so popular. We make an attempt to understand the pattern of speech of popular speakers by studying audio features of a set of speakers. Extracting audio features is one of the basic components of designing a speech classification system. Hence we focus our effort on audio analysis from a speech video.

The primary step of efficient audio processing is the automatic segregation of speech signals. In particular, the fundamental step of segregation is separating sound signals from silence that pertains to pauses, etc. With 'sound' signals, we further categorize speech as 'voiced' and 'unvoiced' [1]. While voiced speech refers to articulated words, unvoiced could include the clearing of throat, a cough, chortle, etc.

Speech is thought of as fluctuating signals whose features vary regularly when it is recorded for a long time. Grouping speech signals in to areas of voiced, unvoiced regions accelerates the rate of recognition [2]. In the case of voiced speech, sound is produced by the vocal cords, with the aid of specific articulators. The vocal chords are tensed and vibrate periodically. Hence, voiced speech contains continual frequency tones of a certain

length, produced when vowels are pronounced and it is categorized by strong fundamental frequency. Voiced speech is most important for measuring the intelligibility of the speaker's voice.

Its periodicity relates to the fundamental frequency of such vibration in the speech signal, also called as "pitch".

Unvoiced Speech region is random in nature and the vibration of the vocal cords or the regularity of the vibration are absent. The signals produced are random and non-periodic in nature and produced by air passing through a narrow contraction in the vocal tract. Here, almost every frequency component is represented equally. In this scenario, building a model for predicting patterns is not viable and pitch estimation is not suitable.

In technical terms, some of the main features of voiced speech signals may be; Zero Crossing Rate (ZCR), Short Term Energy (STE), Mean Power, Pitch, Magnitude and standard deviation that relate to perceived attributes such as loudness, modulation of voice, etc. This systematic technical analysis improves accuracy of a classification system meant to rate the quality of speech.

2. Literature survey

In the last few decades, there has been a substantial effort by researchers to segment speech into voiced and unvoiced regions. Among these several pattern recognition-based algorithms, statistical and non-statistical, feed-forward network techniques have been proposed and used segment speech signals. Multilayer feed forward networks was used by QI and Hunt to classify voiced and unvoiced speech by using non parametric method [3]. Acoustic features were also used to classify speech segments using rule-based techniques.

Most of the techniques focus on a few audio parameters that aid in identifying features of speech. Since we are not keen on identifying the speakers themselves, rather only the quality of speech [4, 5], and our design methodology also focuses on principal audio features that characterize attributes of speech such as the volume (loudness), modulation of voice, vocal consistency, etc.

3. Methodology

In this framework, we focus on the classification of speeches collated from 'TED Talks' and study the Zero Crossing Rate (ZCR), Short Term Energy (STE), Mean Power, Pitch, Magnitude and Standard deviation and finally we use an unsupervised method called hierarchical clustering technic for classifying the speakers in to different categories.

A. ZCR

ZCR is used for separating voiced and unvoiced regions of the speech signals and also used for various speech recognition systems [6]. ZCR indicates the frequency at which energy is concentrated in the signal spectrum which is shown in Figure 1. Zero crossing is defined as rate of total number of times in a sequence that the amplitude of the speech crosses through the value of zero. Zero crossing measures the frequency of the signal to analyze and segregate the region of voice from unvoiced which is shown in figure 2.

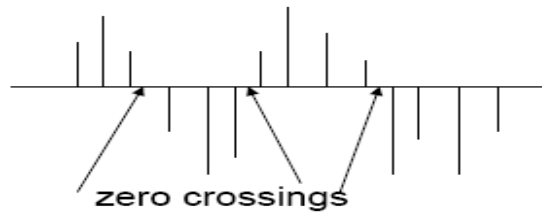


Fig.1. Description of Zero Crossing Rate

Voiced speech shows a lower zero crossing rate, in the range of 0 to 30, for a given interval of speech. Unvoiced speech has a higher concentration of energy at higher frequency and shows higher zero crossing count characteristically ranging from 10 to 100 in an equal interval in time. Only occasionally, the unvoiced regions have a smaller zero crossing rate. Therefore, the zero crossing rate alone cannot be used for classification of speech.

Steps to measure ZCR:

1. Read the audio file
Audio read ('file1.wav');
2. Normalize the data using
3. Perform Framing.
Framing (data_fs, f_d);
(where fs is the sampling frequency)
4. Calculate ZCR
 $ZCR = \sum (\text{abs}(\text{diff}(x > 0)))$;
5. Find ZCR of all frames
For $l = 1$ to r
Do $x = \text{frames}(i)$;
 $ZCR = \sum (\text{abs}(\text{diff}(x > 0)))$;
6. Plot the ZCR and remove silence frames
Plot (t,data);
Fr_ws=frames (id, :);
7. Reconstruct the speech signal without silence.

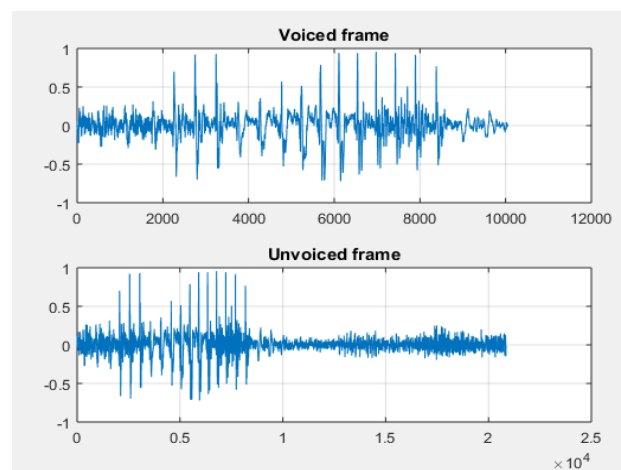


Fig.2. Separation of voiced and unvoiced frame from speech signal

B. Short Term Energy (STE)

STE is used for delineating voiced speech from unvoiced signals. STE are also effective in distinguishing speech from silence and hence used in many audio classification problems.

The energy of speech signal varies in nature. We attempt to understand the change in energy with time for short segments of the input speech. Energy connected with voiced part of the signal is high when compared to the unvoiced part. Segments that correspond to silence may have insignificant energy. Hence we use STE for classification of audio signal in to voiced, unvoiced and silence regions. Figure 3 explains the plot for STE. Red line in graph indicates detected Short term energy for the audio signal. [7]

Steps for computing the STE:

1. Read the audio file.
Audio read (file.wav);
2. Regularize the data using absolute function
Abs (data);
3. Perform framing – zero overlapping with a rectangular window.
4. Calculate the frame energy
 $STE = \sum (\text{frames})$;
5. Normalize the data
6. Plot STE with the Signal using plot function in mat lab
7. Removal segments that correspond to ‘Silence’ from the audio frames.
8. Reconstruct the speech signal without silence.

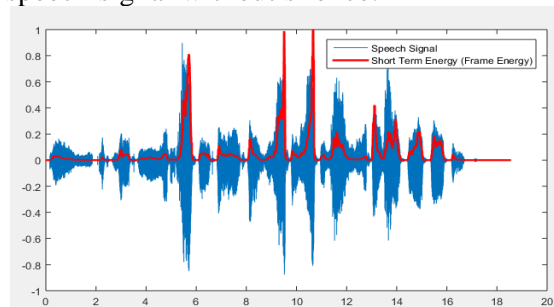


Fig.3. Graph of STE signal plotted

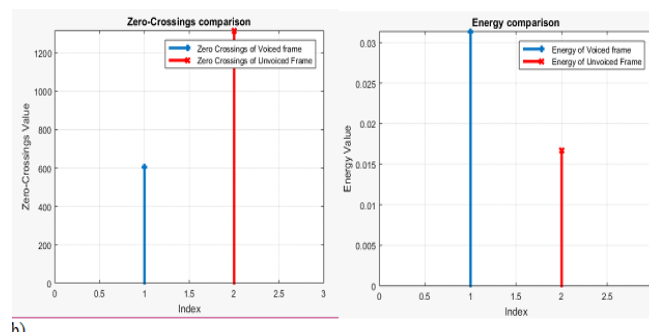


Fig.4. Comparison Graph for ZCR and STE.

Figure 4 presents a comparison of the values of ZCR and STE for voiced and unvoiced signals. Energy connected with voiced part of the signals is high compared to unvoiced part whereas ZCR value of voiced part is lower compared to unvoiced part of the audio signals. Although there may be exceptions with an occasional unvoiced data showing larger values for STE, the combination of the two features is fairly accurate.

C. Power Spectral Density (PSD)

With the power spectral density (PSD), we study the signal in the frequency domain. This provides probabilistic distributions for dynamic responses and also

indicates the frequency ranges for which variations are strong and can facilitate further analysis. [8]

PSD indicates the energy distribution over frequencies. Computation of the PSD involves performing the Fourier Transform (or for computational purposes, a faster implementation called the FFT) or computing the autocorrelation function and then transforming the same.

Figure 5 presents an input speech signal with a plot of the corresponding power spectral density.

Figure 6 provides a comparison between graphs of the normalized cross correlation and the auto correlation of a portion of the input signal.

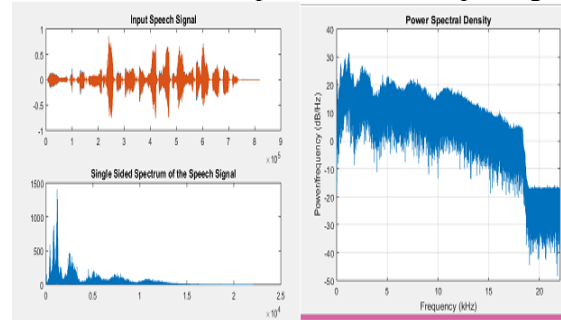


Fig.5. Input speech signal and power spectral density

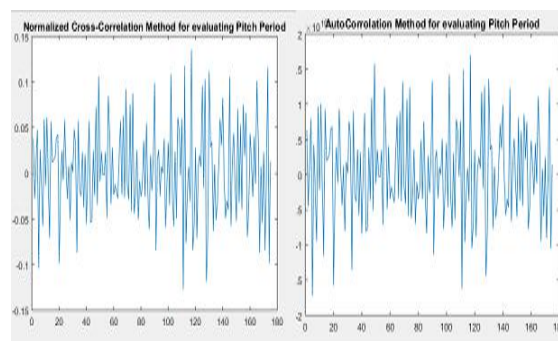


Fig.6. Normalized cross correlation method b: Auto correlation function

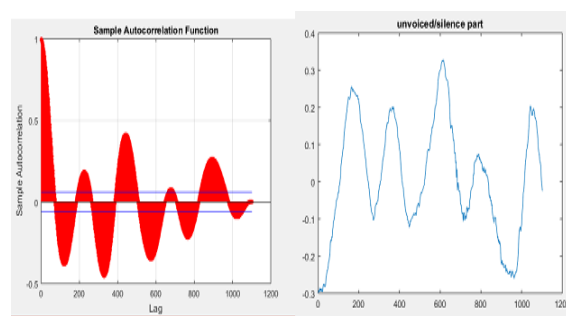


Fig.7. Sample autocorrelation function unvoiced part of the signal.

Pitch Hz = 512.7907

We use the normalized cross-correlation (NCC) method for evaluating Pitch in TED Talk Audio files as shown in figure 6 and 7.

D. Mean Power (intensity) in air

The power of a sound wave groups the rate and density of energy transfer. It is an objective quantity related to the wave. Loudness is the perceptive equivalent of the physical property of intensity. Larger the amplitude greater the intensity, louder the sound. Speech waves with high amplitudes are said to be loud. Speech waves with less amplitude are said to be soft.

Mean power measured in terms of decibels.

E. Standard Deviation (SD)

Standard deviation indicates the spread about the magnitude of an audio signal for any given interval of the audio signal. Lower the value of SD, values are closer to the mean amplitude. Higher is the value of SD, values are more spread out.

It is calculated by squaring each of the deviations before taking the average.

The standard deviation is calculated by an equation as below;

$$\sigma^2 = \frac{1}{N-1} \sum_{i=0}^{N-1} (x_i - \mu)^2 \quad (1)$$

Equation 1 is the variance of a speech signal of N points. The square root of equation (1) yields the standard deviation. The signal is stored in x_i . μ is mean. The mean of the signal is simply the sample average of the N points:

$$\mu = \frac{1}{N} \sum_{i=0}^{N-1} x_i \quad (2)$$

The signal is confined in x_0 through x_{N-1} , i is the index that runs through these values and μ is the mean

Adding the values in the signal, x_i , by allowing the index, i, track from 0 to N-1. Divide the sum by N. This is equal to the equation: $\mu = (x_0 + x_1 + x_2 + \dots + x_{N-1})/N$

F. Magnitude

The magnitude of an audio signal is a measure of the strength of the signal. While amplitude of a variable is a measure of change relative to its central position, magnitude is a measure of the squared distance (that does not take into account the direction of this change).

Magnitude of a signal = $X = |x(n)|^2$

Modulus symbol (| |) to represent the magnitude of $x_1(n)$

Table 1 explains the results of audio features extracted from speech files and implemented ZCR, STE calculations to isolate the voiced and unvoiced regions of the speech. We can observe from the table that ZCR is low for voiced portions and high for unvoiced parts of the input speech signal. Whereas, STE (Short Term Energy) is high for voiced speech and low for unvoiced speech audios. Standard deviation is closer to 1, indicating equal distribution of audio quality of speaker over average value.

In terms of interpretation in the context of speech classification, we understand higher the intensity of the air, greater is the volume and hence, the energy of the speaker. Similarly, higher the magnitude of the components of the Fourier Transform, greater the strength of those frequency components in the audio signal.

In order to categorize the speakers based on their audio features, we are using unsupervised clustering technique called Hierarchical Clustering.

Table I. Resulting Table for audio features

S	STE	SD	ZCR	M P	Mag
1.	1.00	0.18	31	79.27	0.11
2.	0.55	0.06	52	69.82	0.03
3.	0.55	0.06	53	73.88	0.04
4.	0.83	0.02	45	66.4	0.01
5.	0.63	0.09	62	73.66	0.05

S- Speakers, STE (joule/meter square), SD-Pascal, ZCR, M P (mean power), Mag (Magnitude).

Hierarchical Clustering Method.

We use unsupervised hierarchical clustering to grouping the Ted speakers in to different classes based on their audio features. This method groups the data based on rules which occurs in multilevel and creates a cluster tree called dendrograms which can be joined as clusters at next higher level. [9]

1. First, the similarity between features in the data set is computed by first forming a vector of these features and storing them in X and then computing the distances between such X's through

$X = [STE, ZCR, SD, \text{Magnitude}]$

$Y = \text{pdist}(X)$

2. Clustering the features into a hierarchical tree.

Here we link all the features which are closer to each other using the linkage function. These features are paired in to binary clusters and they are combined to higher clusters to form a hierarchical tree called dendrogram.

$Z = \text{linkage}(Y)$

Dendrogram- Binary cluster tree which is hierarchical in nature is created by the linkage function; a sample is presented below in Figure 8.

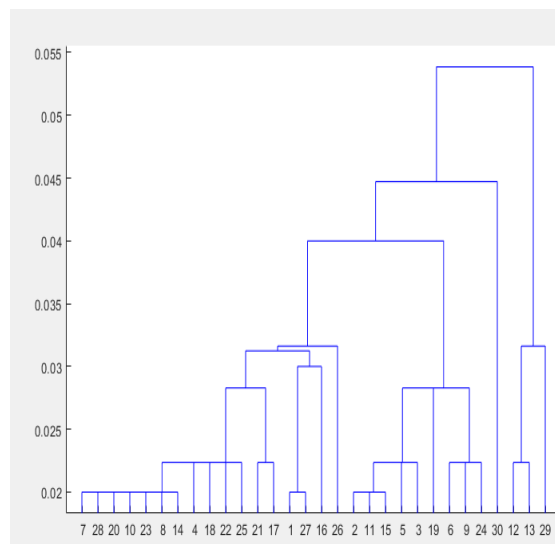


Fig.8. Dendrogram Binary Cluster Tree

3. Pruning the branches of the hierarchical tree: This step partitions the data by creating natural groupings in the hierarchical tree. We can visually verify that the distances in the tree reflect the distance between features and that natural divisions occur among the links between objects

We can further verify the similarity and dissimilarity in the hierarchical tree by observing that any two objects are linked together at some level. The elevation of the connection signifies distance between two clusters or the cophenetic distance between two objects. The cophenet function computes correlation by returning the cophenetic correlation coefficient. The clustering algorithm groups inputs with cophenetic correlation coefficients closer to 1, while separating those that are significantly different.

In the example below, Z indicates the matrix output by linkage function. Y is the distance vector output by pdist () function.

$Y = \text{pdist}(X, 'cityblock');$


```

Z= linkage (Y,'average');
C= cophenet (Z, Y)
C = 0.7402

```

We verify consistency to refine normal cluster divisions in the feature set. The cluster function uses a quantitative measure of inconsistency to regulate what parts of the dataset form a cluster [10, 11]. The distance and linkage values are recomputed again using the default settings.

```

Y= pdist(X);
Z= linkage(Y);
I = inconsistent (Z)

```

The inconsistent function outputs data as (m-1)-by-4 matrix, whose columns are categorized in the following Table II.

4. Results and Discussion

Our data comprises 215 speeches from TED Talks; these were categorized into 6 categories using ZCR, Pitch, mean, Power and energy. Inconsistency value in the below Table II indicate a speaker is not able to maintain all the detected audio features (ZCR, Pitch, mean, Power and energy,) consistently throughout the speech. Higher the inconsistency value, lower the consistency of the speaker.

Table II Resulting table for inconsistency values for a small sample of the speakers

Speaker	Mean	SD	links	IC
1	0	0	1.0	0
2	0	0	1.0	0
3	0	0	2.0	0.70
4	0	0	2.0	0.70
5	0	0	3.0	0.57

Mean-Mean of the heights of all the, SD- Standard deviation of all the links links- Number of links, IC-Inconsistency Co-efficient

Table III delineates the 215 speakers whose audio speeches were analyzed into 6 classes: The classifications are effective speaker, Mid-Level 1, 2, 3 and 4. and low level 1. Speakers having inconsistency coefficient 0 considered to have all the calculated audio features high. Mid-level speakers having average audio features and low level speaker having minimum audio features.

Table III. Final result table for classification of speakers.

Category	Inconsistency value	Number of speakers from Ted Talks
1.Effective speaker	0	139
2. Mid-level 1	0.5774	24
	0.5927	2
3. Mid-level 2	0.7071	30
	0.7478	4
	0.8066	1
4. Mid-level 3	0.9355	2
	0.9424	2
5. Mid-Level-4	1.0000	2
	1.1097	1
6. Low	1.1547	8

5. Conclusion

This paper presented a framework for categorizing speakers based on their audio quality. The categorization is based on audio features such as STE (Short Term Energy), ZCR (Zero Crossing Rate), Mean Power, Magnitude, Standard Deviation followed by the unsupervised hierarchical clustering. This classification can further be used to evaluate the features of the speech of a novice speaker. By studying the differences in various components between various categories, the novice speaker would have a clear insight on aspects of audio features or voice training that could catapult the speaker to a higher category. We can further enhance the study of characterizing speech quality assessment by including visual features such as facial appearance, hand gestures, pose of the speaker and sentiment analysis and text analysis for evaluating the quality of the content of the speech.

References

- [1]. Bachu R.G., Kopparthi S., Adapa B., Barkana B.D, "Voiced/Unvoiced Decision for Speech Signals Based on Zero-Crossing Rate and Energy", Chapter: *Advanced Techniques in Computing Sciences and Software Engineering*
- [2]. Anil Kumar C1, Shiva Prasad K. M2, Manjunatha M. B3, Kodanda Ramaiah G. N, "Analysis And Classification Of Voiced And Unvoiced Speech In Speech Processing", j. *Technological Advances and Scientific Res./eISSN- 2454-1788, pISSN- 2395-5600/ Vol. 2/ Issue 03/ July-Sept. 2016*
- [3]. Roy, Tanmoy, Tshilidzi Marwala, and Snehashish Chakraverty. "Precise detection of speech endpoints dynamically: A wavelet convolution based approach." *Communications in Nonlinear Science and Numerical Simulation* 67 (2019): 162-175.
- [4]. Huang, Lian, Yanfen Gan, and Hongke Ye. "Audio-replay Attacks Spoofing Detection for Automatic Speaker Verification System." *2019 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA)*. IEEE, 2019.
- [5]. Muckenhirn, Hannah, Mathew Magimai Doss, and Sébastien Marcell. "Towards directly modeling raw speech signal for speaker verification using CNNs." *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018.
- [6]. Afroz, Fathima, and Shashidhar G. Koolagudi. "Recognition and Classification of Pauses in Stuttered Speech Using Acoustic Features." *2019 6th International Conference on Signal Processing and Integrated Networks (SPIN)*. IEEE, 2019.
- [7]. Chen, Chia-Ping, et al. "Speaker Characterization Using TDNN-LSTM Based Speaker Embedding." *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019.
- [8]. Li, Xiaofei, et al. "Audio-noise Power Spectral Density Estimation Using Long Short-term Memory." *IEEE Signal Processing Letters* 26.6 (2019): 918-922.
- [9]. Anumanchipalli, Gopala K., Josh Chartier, and Edward F. Chang. "Speech synthesis from neural decoding of spoken sentences." *Nature* 568.7753 (2019): 493.

[10]. Khoury, Elie, and Matthew Garland. "Speaker recognition in the call center." U.S. Patent No. 10,325,601. 18 Jun. 2019.

[11]. Le, Nam, and Jean-Marc Odobez. "Improving speech embedding using crossmodal transfer learning with audio-visual data." *Multimedia Tools and Applications* 78.11 (2019): 15681-15704.

Authors



Renuka Devi M.N is currently working as Assistant Professor in Department of Computer Science and Engineering. School of Engineering, Dayananda Sagar University Bengaluru. She has submitted her Ph-D Thesis to VTU on “**Characterization and detection of behavioural pattern in videos**” under the guidance of Dr.Gowri Srinivasa. She has completed M.Tech degree in Computer Science and Engineering from Nitte Meenakshi Institute of Technology, Bengaluru. She obtained her B.E degree in Computer Science and Engineering from PESIT, Bengaluru. She has 6.6 years of teaching experience. She has worked as Assistant Professor at PESIT from July 2011 to July 2016. She has worked as Assistant Professor at BNMIT from Jan 2019 to July 2019. She has published 3 Scopus Indexed Journals paper and 6 international conferences papers (*IEEE proceedings, ICECCT, ICE2CSM2E*) and received “**BEST PAPER AWARD**” at *7th International Conference on Innovations in Computer Science and Engineering, Scopus Indexed LNNS Springer Series at Hyderabad*.



Dr.Gowri Srinivasa graduated with a bachelors in Computer Science and Engineering from PES Institute of Technology, Bangalore in 2004 and then worked with Prof. Jelena Kovacevic to obtain her PhD in Biomedical Engineering from Carnrgie Mellon University in 2008. Since August 2008, she has served as a Professor at PESIT Bangalore South Campus, India (now PES University), heading the center for Pattern Recognition. She has worked with students at the undergraduate and graduate level on various problems focusing primarily on the application area of Biomedicine/Healthcare and has designed and taught courses during the semester and over the summer in the fields of Data Analytics, image processing and analysis, machine learning, Multi resolution analysis .etc.. Gowri is a senior member of the IEEE and has served as an Associate member of the TC of Bio Imaging and Signal processing during 2009-2013 and elected as a member of the TC of BISP for two consecutive terms: 2014-2016 and 2017-2019. She holds a patent from USPTO (2014) for multimodal communication assist device that helps people who are blind communicate with deaf-mute persons.