ISSN: 2321-7782 (Online)
Volume 3, Issue 2, February 2015
International Journal of Advance Research in Computer Science and Management Studies
Research Article / Survey Paper / Case Study
Available online at: www.ijarcsms.com

A Study on Male Voice Mutation

Rinku Sebastian[1]
Electronics and Communication Engineering
Amal Jyothi College of Engineering
Kerala - India

Therese Yamuna Mahesh[2]
Computer Science and Engineering
Research Scholar, Bharath University
Chennai - India

Dr. K.L.Shunmuganathan[3]
Department of Computer Science & Engineering
R.M.K. College of Engineering
Chennai - India

Abstract: ***In this paper a study of male voice mutation using acoustic features is conducted. This paper explores and compares various acoustic features and classifies the data using multilayer perceptron. A data set is created by collecting data of children between age nine and twenty five. The speech signal is then analyzed in order to extract the acoustic parameters such as the Signal Energy, pitch, formant frequencies, Jitter and Shimmer. In this study various acoustic features are combined to form a feature set, which is used for voice classification. Voice is classified into three classes. Hence, a successful pathological voice classification enables an automatic non-invasive method that help teenagers to analyze their voice change. This method is also helpful in easy diagnosis of voice disorder that can occur during the growth period.***

Keywords: pitch, jitter, formants, shimmer, multilayer perceptron.

I. INTRODUCTION

The larynx, which is located in the throat, plays the major role in forming the human voice. Two muscles, or vocal cords, that are stretched across the larynx, act like rubber bands. When a person speaks, air rushes from the lungs and makes the vocal cords vibrate, which in turn produces the sound of the voice. The pitch of the sound produced is controlled by how tightly the vocal cord muscles contract as the air from the lungs hits them.

Larynx of a boy is pretty small and his vocal cords are kind of small and thin till he reaches puberty. That's why his voice is higher than an adult's. But as he goes through puberty, the larynx gets bigger and the vocal cords lengthen and thicken, so his voice gets deeper. Along with the larynx, the vocal cords grow significantly longer and become thicker. In addition, the facial bones begin to grow. Cavities in the sinuses, the nose, and the back of the throat grow bigger, creating more space in the face in which to give the voice more room to resonate.

During the puberty period the voice box, or larynx grows in both sexes. This growth is far more prominent in boys, causing the male voice to drop and deepen. Before puberty, the larynx of boys and girls is about equally small. Occasionally, voice change is accompanied by unsteadiness of vocalization in the early stages of untrained voices. Most of the voice change happens during stage 3-4 of male puberty around the time of peak growth. Adult pitch is attained at an average age of 15 years, although the voice may not fully settle until early twenties.[1]

In this study an automatic classification of voice using acoustic features is proposed. We are making use of various acoustic features which best describe the functioning and condition of various speech organs to analyze the voice change during puberty period. Pitch is an attribute which represents the structure and size of the larynx and vocal folds. Formants are the distinguishing or meaningful frequency components of human speech that humans require to distinguish between vowels. In general if the energy of the speech signal is higher, the volume of the output speech signal will also be higher. Using these

acoustic features an extensive number of researches are carried out and various algorithms are used for extracting these features from the speech signal. The goal of the feature extraction is to characterize an object to be recognized by measurements whose values are very similar for objects in the same category and very different for the objects in different categories leading to the idea of seeking distinguishing features that are invariant to irrelevant transformations of the input.

The patterns for training the Neural Network was obtained from the recordings of children voices between age nine and twenty-five. Multilayer perceptron is a feedforward artificial neural network model that maps input data set onto a set of appropriate outputs. A MLP consists of multiple layers of nodes in a directed graph, with each layer fully connected to the next one. The general process of classification is shown in Figure 1

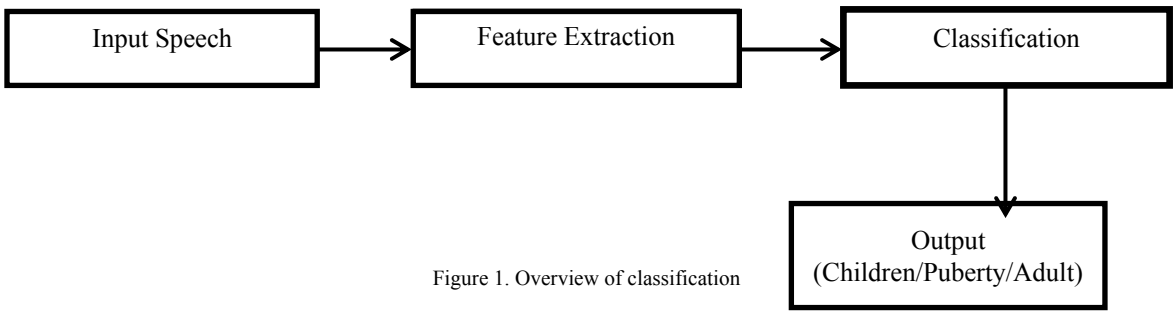


Figure 1. Overview of classification

II. Acoustic Features

A. *Pitch*

Pitch is a subjective psychoacoustic attribute of sound. Pitch is a perceptual property that allows the ordering of sounds on a frequency-related scale. The sensation of a frequency is commonly referred to as the pitch of a sound. A high pitch sound corresponds to a high frequency sound wave and a low pitch sound corresponds to a low frequency sound wave. Voiced speech signals can be considered as quasi-periodic. The basic period is called the pitch period. Pitch is simply the rate at which vibrations are produced. This is usually expressed as the number of Hz (hertz, or cycles per second). One cycle is a complete vibration back and forth. The number of Hz is the frequency of the tone. The higher the frequency of a tone, higher its pitch. Vocal folds of a boy grow 0.7 mm in length during the puberty period. This growth eventually slows down. A longer vocal fold means a deeper voice. The mean fundamental frequency decreases during the period of mutation. This correlates with physiological development, laryngeal growth, and a subsequent decrease of mean fundamental frequency. [5]

B. *Formants*

Formants are defined as the spectral peaks of the sound spectrum. The formant with the lowest frequency is called f1, the second f2, and the third f3. Most often the first two formants, f1 and f2 are enough to disambiguate vowels. These two formants determine the quality of vowels in terms of the open/ close and front/back dimensions. Formants are defined as the spectral peaks of the sound spectrum of the voice. Formant is also used to mean an acoustic resonance of the human vocal tract. It is often measured as an amplitude peak in the frequency spectrum of the sound. The acoustics of the vocal tract are often modeled using a mathematical model of a filter. The frequencies of the poles of this filter model fall close to those of the formants. As a result, some voice researchers now refer to the frequencies of the poles as formants. So, to some voice researchers, the formant refers to a peak in the spectrum, to others it refers to a resonance of the vocal tract while to a third group it refers to the pole in a mathematical filter model. Formant frequencies remain almost stable during the mutational voice change.

C. Jitter

Jitter is the deviation from true periodicity of a presumed periodic signal in electronics and telecommunications, often in relation to a reference clock source. Jitter may be observed in characteristics such as the frequency of successive pulses, the signal amplitude, or phase of periodic signals. Jitter is also known as frequency perturbation and refers to the minute involuntary variations in the timing variability between cycles of vibration. In essence, it is a measure of frequency variability in comparison to the client's fundamental frequency. Research shows that jitter values in normal voices range from 0.2 to 1 percent. Jitter values above this level indicate that the vocal folds are vibrating in a way that is not as periodic as it should be. Higher jitter levels suggest that something is interfering with normal vocal fold vibration and the mucosal wave. Jitter is assumed to increase during voice mutation.

D. Shimmer

Shimmer deals with a frequent back and forth change in amplitude in the voice. Shimmer is a measure of the percentage irregularity in the amplitude of the vocal note. It is often referred to as amplitude perturbation. Shimmer, therefore, measures the variability in the intensity of adjacent vibratory cycles of the vocal folds. Research estimates that shimmer values below 0.5 dB are normal in the human voice. Jitter and shimmer reflect the internal noises of the human body. Higher jitter and shimmer levels reflect neuromuscular problems. Measuring the cycle-to-cycle variability of vibration can allow us to detect changes in neuromuscular function or changes in the layers of the vocal folds.

III. PROPOSED METHODOLOGY

The speech from children of different age ranging from nine to twenty five was recorded. They were trained to utter different phonemes.

A. Windowing

Speech signal is non-stationary, where change in properties occur quite rapidly over time. This is completely natural thing but makes the use of DFT or autocorrelation. For most phonemes the properties of the speech remain invariant for a short period of time (5 -100 ms). Thus for a short window of time, traditional signal processing methods can be applied relatively successfully. Most of speech processing is done by taking short window of time and processing them. The short window of signal is called frame. A long signal of speech is multiplied with a window function of finite length, giving finite length weighted version of the original signal. [6]

B. Pitch Estimation

A pitch detection algorithm (PDA) is an algorithm designed to estimate the pitch or fundamental frequency of a quasi periodic. The time-domain pitch period estimation techniques use auto-correlation function (ACF). The basic idea of correlation-based pitch tracking is that the correlation signal will have a peak of large magnitude at a lag corresponding to the pitch period. The autocorrelation computation is made directly on the waveform and is a fairly straightforward computation. Autocorrelation function for a signal $x(n)$ is computed as given in

$$Øx(m) = \lim_{n\to\infty} \frac{1}{2n+1} \sum_{n=-N}^{N} x(n)x(n+m)$$

The autocorrelation function of a signal is basically a (non-invertible) transformation of the signal which is useful for displaying structure in the waveform. Thus, for pitch detection, if we assume $x(n)$ is exactly periodic with period P, *i.e.* $x(n) = x(n + P)$ for all n, then the autocorrelation function $Øx(m)$ is also periodic with the same period.[3]

C. Formant Estimation

Linear Predictive Coding analyzes the speech signal by estimating the formants, removing their effects from the speech signal, and estimating the intensity and frequency of the remaining buzz. The process of removing the formants is called inverse filtering, and the remaining signal after the subtraction of the filtered modeled signal is called the residue. A formant or resonance of the vocal tract above the vocal folds is a frequency region that will strongly pass energy in that frequency region if it receives energy at those frequencies from the glottal source (glottal flow). The formant frequencies depend upon the size and shape of the vocal tract

D. Jitter Estimation

Jitter is the measures of the cycle-to-cycle variations of fundamental frequency. The average absolute difference between consecutive periods is expressed as

$$\text{Jitter (absolute)}=\frac{1}{N-1}\sum_{i=1}^{N-1}|T_i - T_{i+1}|$$

Where, *Ti* are the extracted *F*0 period lengths, *N* is the number of extracted *F*0 periods. [2]

Jitter (relative) is the average absolute difference between consecutive periods, divided by the average period and is expressed as a percentage

$$\text{Jitter (relative)}=\frac{\frac{1}{N-1}\sum_{i=1}^{N-1}|T_i - T_{i+1}|}{\frac{1\sum_{i=1}^{N}Ti}{N}}$$

E. Shimmer Estimation

Shimmer is a measure of the percentage irregularity in the amplitude of the voice signal. The average absolute base-10 logarithm of the difference between the amplitudes of consecutive periods, multiplied by 20. [2]

$$\text{Shimmer absolute}=\frac{1}{N-1}\sum_{i=1}^{N-1}|20\log(\frac{A_{i+1}}{A_i})|$$

Where *Ai* is the extracted peak-to-peak amplitude data and *N* is the number of extracted fundamental frequency periods.

F. Classification using multilayer perceptron

A multilayer perceptron (MLP) is a modification of the standard linear perceptron. MLP can distinguish data that are not linearly separable. MLP utilizes back-propagation for training the network. An MLP is a network of simple neurons called perceptron. The perceptron computes a single output from multiple real-valued inputs by forming a linear combination according to its input weights and then possibly putting the output through some nonlinear activation function. Multilayer perceptron using a back-propagation algorithm are the standard algorithm for any supervised learning pattern recognition process, which often allows one to get approximate solutions for extremely complex problems like fitness approximation.

IV. Experiment and Results

Speech samples of various phonemes from children between nine and twenty five years of age were collected. All the speech samples were recorded in noise free environment using a microphone array. Each speech sample is pre-processed using a windowing technique. Using Autocorrelation method the fundamental frequency is estimated, and Linear predictive analysis is used to extract the formant frequencies f1 and f2. The feature vector is constructed using the peaks of Formant frequencies, average pitch period, the signal energy, mean square residual signal, reflection coefficients, jitter and shimmer. Combining all these feature vectors forms coefficient feature set. The classifiers are trained and tested.

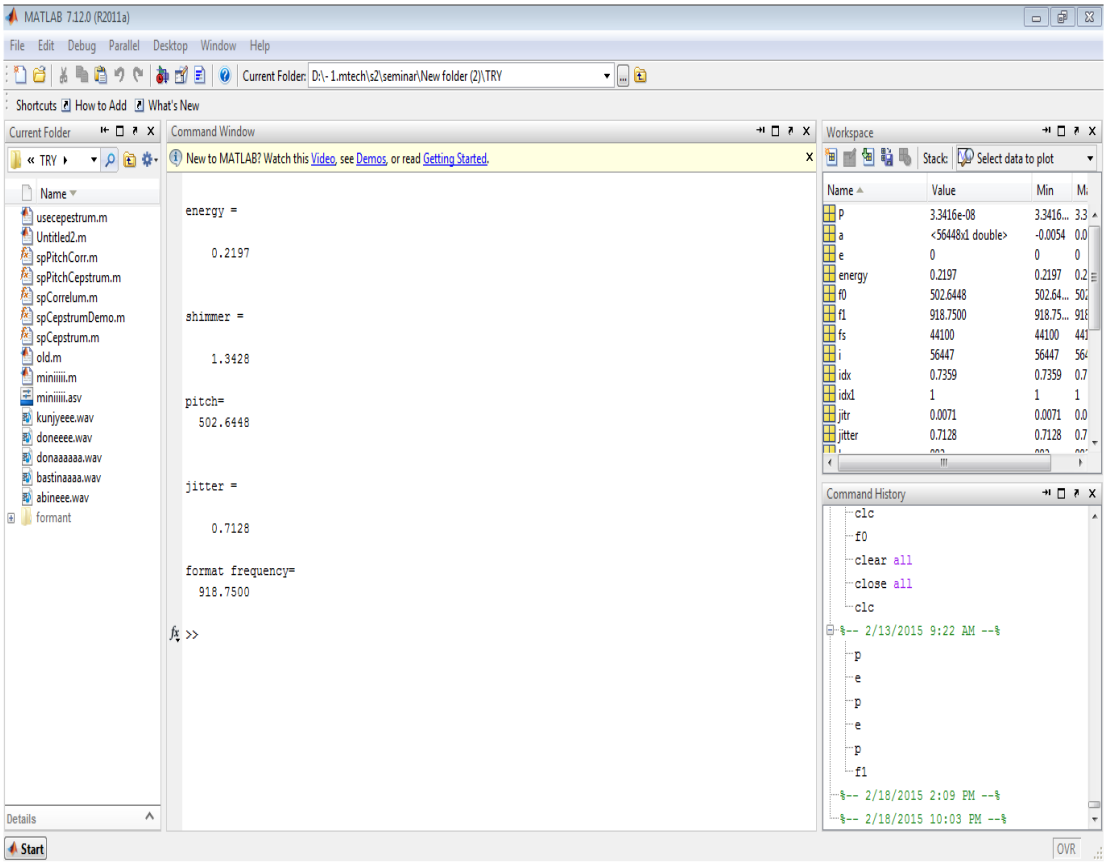


Figure 2: Matlab output of feature extraction (pitch,jitter,shimmer,formant,energy)

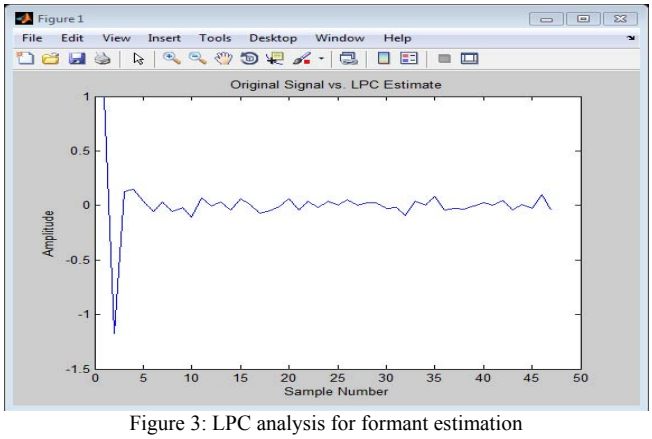


Figure 3: LPC analysis for formant estimation

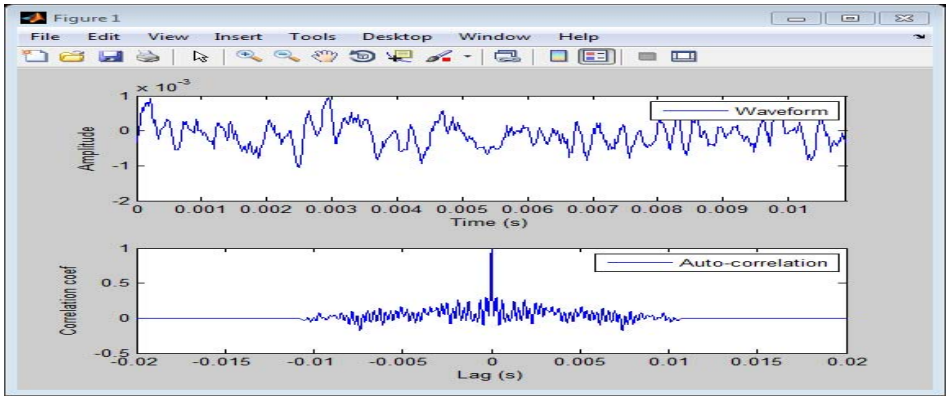
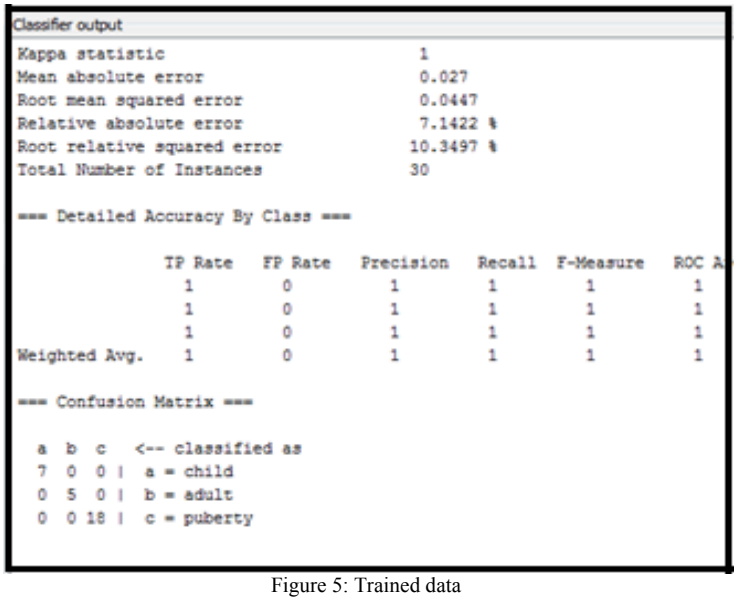


Figure 4: Autocorrelation for pitch estimation

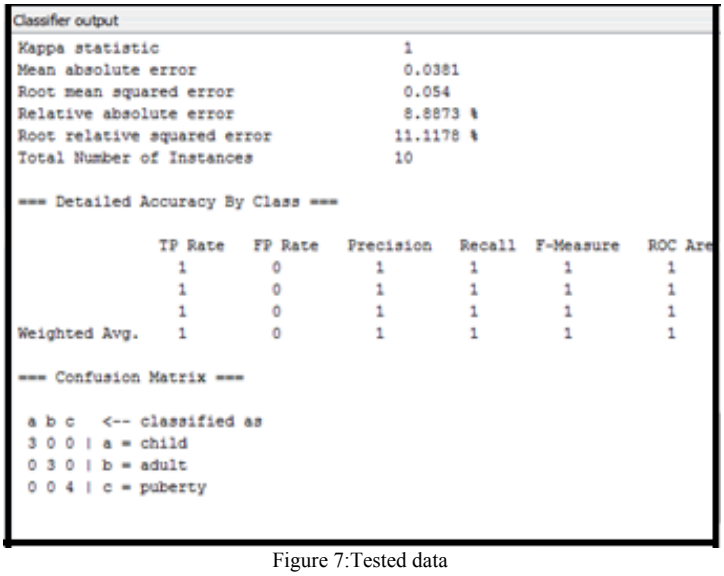
The classification was done for a set of 30 samples taken as the training set for training the multilayer perceptron neural network. The results of the classification are as shown below.



```
Classifier output

Kappa statistic                          1
Mean absolute error                      0.027
Root mean squared error                  0.0447
Relative absolute error                  7.1422 %
Root relative squared error             10.3497 %
Total Number of Instances               30

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC A
                 1         0          1         1         1          1
                 1         0          1         1         1          1
                 1         0          1         1         1          1
Weighted Avg.    1         0          1         1         1          1

=== Confusion Matrix ===

  a  b  c   <-- classified as
  7  0  0 |  a = child
  0  5  0 |  b = adult
  0  0 18 |  c = puberty
```

Figure 5: Trained data

A set of 10 samples were given as test data. All the 10 samples were correctly classified. The results of the classification are as shown below.



```
Classifier output

Kappa statistic                          1
Mean absolute error                      0.0381
Root mean squared error                  0.054
Relative absolute error                  8.8873 %
Root relative squared error             11.1178 %
Total Number of Instances               10

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Are
                 1         0          1         1         1          1
                 1         0          1         1         1          1
                 1         0          1         1         1          1
Weighted Avg.    1         0          1         1         1          1

=== Confusion Matrix ===

 a b c   <-- classified as
 3 0 0 | a = child
 0 3 0 | b = adult
 0 0 4 | c = puberty
```

Figure 7:Tested data

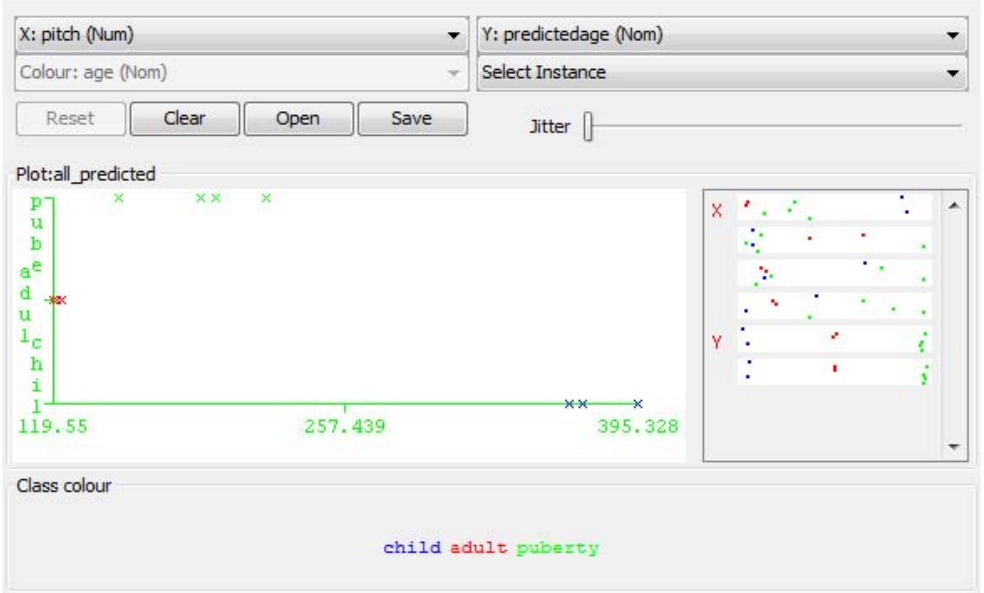


Figure 8:Classifier visualization of pitch

V. Conclusion

In this paper several techniques for extracting different acoustic parameters and feature extraction are discussed. The purpose of this methodology is to classify the voice dataset using multilayer perceptron. This method is a powerful tool which helps in determining the rate of growth of vocal cord and to diagnose the diseases related to vocal cord that can occur during voice mutation in teenagers. Higher values of Jitter and Shimmer detect changes in the neuromuscular function or changes in the layers of the vocal folds.

Acknowledgement

The authors would like to thanks St George School and Amal Jyothi College for rendering support for data collection. We also express our gratitude to all our friends who helped us with voice samples in preparation of this paper.

References

1. Arnold E Aronson, Diane M Bless, clinical voice disorder, thieme publishers fourth edition,page 16-21
2. V.Sellam, J. Jagadeesan, Classification of Normal and Pathological Voice Using SVM and RBFNN, Journal of Signal and Information Processing, 2014, 5, 1-7.
3. A. R. Rabiner On the Use of Autocorrelation Analysis for Pitch Detection, , IEEE Transactions on Acoustics, Speech, and Signal Processing, Vol. 25, No. 1, 1977, pp.24-33.
4. Roy C Snell, Fausto millinaso, Formant location from LPC analysis of data, , IEEE Transactions on Speech, and Audio Processing, Vol. 1, No. 2, 1993.
5. L. R. Rabiner and R. W. Schafer, Digital Processing of Speech Signals, Pearson Education India.
6. P. Dhanalakshmi, S. Palanivel and V. Ramalingam, "Classification of Audio Signals Using SVM and RBFNN," Expert Systems with Applications, Vol. 36, No. 3, 2009, pp. 6069-6075.

Author(s) Profile

Ms.Rinku Sebastian, pursuing Masters Degree in Communication Engineering at Amal Jyothi College of Engineering, Kottayam, Kerala, India. She has obtained her Bachelor Degree from Mahatama Gandhi University, Kerala, India.

Ms. Therese Yamuna Mahesh, M. Tech, works as Assistant Professor at Amal Jyothi College of Engineering, Kottayam, Kerala, India. She is also a research scholar at Bharath University, Chennai. She has more 15 years of teaching experience and her areas of interest include Image processing, Pattern recognition and Computer networks.

Dr. K. L. Shunmuganathan, B.E, M.E., M.S., Ph.D., works as the Professor and Head of the CSE Department of RMK Engineering College, Chennai, Tamil Nadu, India. He has more than 18 years of teaching experience, and his areas of specialization are artificial intelligence, computer networks and DBMS.

