Journal of Signal and Information Processing, **2014, 5, 1-7**
Published Online February 2014 (http://www.scirp.org/journal/jsip)
http://dx.doi.org/10.4236/jsip.2014.51001

Scientific Research

Classification of Normal and Pathological Voice Using SVM and RBFNN

V. Sellam, J. Jagadeesan

Department of Computer Science and Engineering, SRM University, Chennai, India.
Email: sellamveera@gmail.com, hod.cse@rmp.srmuniv.ac.in

Received October 18th, 2013; revised November 18th, 2013; accepted November 25th, 2013

ABSTRACT

The identification and classification of pathological voice are still a challenging area of research in speech processing. Acoustic features of speech are used mainly to discriminate normal voices from pathological voices. This paper explores and compares various classification models to find the ability of acoustic parameters in differentiating normal voices from pathological voices. An attempt is made to analyze and to discriminate pathological voice from normal voice in children using different classification methods. The classification of pathological voice from normal voice is implemented using Support Vector Machine (SVM) and Radial Basis Functional Neural Network (RBFNN). The normal and pathological voices of children are used to train and test the classifiers. A dataset is constructed by recording speech utterances of a set of Tamil phrases. The speech signal is then analyzed in order to extract the acoustic parameters such as the Signal Energy, pitch, formant frequencies, Mean Square Residual signal, Reflection coefficients, Jitter and Shimmer. In this study various acoustic features are combined to form a feature set, so as to detect voice disorders in children based on which further treatments can be prescribed by a pathologist. *Hence, a successful pathological voice classification will enable an automatic non-invasive device to diagnose and analyze the voice of the patient.*

KEYWORDS

Terms—Pitch; Formants; Jitter; Shimmer; Reflection Coefficients; SVM; RBFNN

1. Introduction

In the past 20 years, a significant attention has been paid to the science of voice pathology diagnostic and monitoring. The purpose of this work is to help patients with pathological problems for monitoring their progress over the course of voice therapy. Currently, patients are required to routinely visit a specialist to follow up their progress. Moreover, the traditional ways to diagnose voice pathology are subjective, invasive methods such as the direct inspection of the vocal folds and the observations of the vocal folds by endoscopic instruments are done. These techniques are expensive, risky, time consuming, discomfort to the patients and require costly resources, such as special light sources, endoscopic instruments and specialized video-camera equipment. In order to circumvent the above problems, non-invasive methods have been developed to help the ENT clinicians and speech therapists for early detection of vocal fold pathology and can improve the accuracy of the assessments. The voice disorders are caused due to defects in the speech organs, mental illness, hearing impairment, autism, paralysis or multiple disabilities.

Clinically a number of guidelines and methods are used in practice for detection of voice disorders in children. In this study an automatic classification of pathological voice disorder using acoustic features is proposed. Acoustic features, which are used to identify voice disorders, best describe the functioning and condition of various speech organs. Pitch is an attribute which represents the structure and size of the larynx and vocal folds. Pitch is closely related to frequency, but the two are not equivalent. Formants are the distinguishing or meaningful frequency components of human speech that humans

require to distinguish between vowels. The formant with the lowest frequency is called f_1, the second f_2, and the third f_3. Most often the first two formants, f_1 and f_2 are enough to disambiguate the vowel. These two formants determine the quality of vowels in terms of the open/close and front/back dimensions. LPC is generally used for speech analysis and re-synthesis. During speech synthesis the values of the reflection coefficients are used to define the digital lattice filter which acts as the vocal tract in this speech synthesis system. 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 extractor 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 SVM and RBFNN were obtained from the recordings of children voices with normal voice and children with pathological voice. Since there are different types of classifiers, we are cross-validating different classification methods to find the best hyperparameters and best classifier. The basis of SVM approach is the projection of low-dimensional training data in a higher dimensional feature space, because it is easier to separate input data. RBFNN is a type of neural network consisting of an input layer, an output layer and a hidden layer. RBFNN is associated with radial basis function which trains faster than multi-layer perceptrons and hence it classifies the normal and pathological voice in a better and faster way. The general process of classification is shown in **Figure 1**.

2. Related Works

In the recent works of speech pathology discrimination, researchers are mostly concentrating in the implementation of feature extraction techniques and pattern classification techniques. [1] proposes a classification technique which focuses on the acoustic features of the speech using wavelet analysis and multilayer neural network. [2] proposes a system that determines the pitch using Autocorrelation method. [3] classifies the normal and pathological voice using 27 features and are incorporated using PCA and SVM (RBF). Here the audio signals are classified using a non-linear classification technique RBFNN and they are concentrating on classification part rather than feature extraction [4]. [5] compares various kernel functions and helps to identify Laryngeal disorder. [6] evaluates the computational time and hence feature extraction is carried out using MFCC and classified using

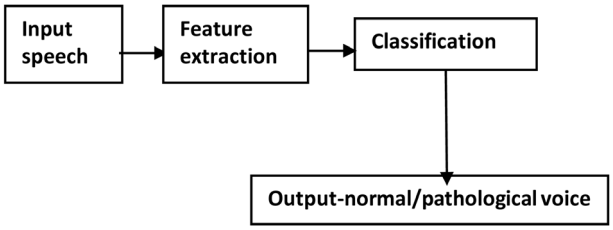

Figure 1. Overview of classification.

GMM. This paper deals with the extraction of acoustic parameters from the residue signal and diagnoses 21 different voice disorders [7]. [8] analyzes the speech signal and the feature set is optimized using Genetic Algorithm and classified using different kernels of SVM.

3. Acoustic Feature Extraction

3.1. Signal Energy

The energy level of unvoiced segments is noticeably lower than that of the voiced segments. The higher the energy, the higher the volume of the output speech signals and higher the amplitude. The short-time energy of speech signals reflects the amplitude variation and is defined using the equation below as in [1].

$$E(i) = \frac{1}{N}\sum_{n=1}^{N}\left|x_i(n)\right|^2 \quad (1)$$

N is the length of the sample.

In order to reflect the amplitude variations in time (for this a short window is necessary), and considering the need for a low pass filter to provide smoothing, $h(n)$ was chosen to be a hamming window. It has been shown to give good results in terms of reflecting amplitude variations hamming window powered by 2. It has been shown to give good results in terms of reflecting amplitude variations.

In voiced speech the short-time energy values are much higher than in unvoiced speech, which has a higher zero crossing rate (**Figure 2**).

3.2. Pitch

Voiced speech signals can be considered as quasi-periodic. The basic period is called the pitch period. The average pitch frequency (in short, the pitch), time pattern, gain, and fluctuation change from one individual speaker to another. For speech signal analysis, and especially for synthesis, identifying the pitch is extremely important. A well-known method for pitch detection is given in [9]. It is based on the fact that two consecutive pitch cycles have a high cross-correlation value, as opposed to two consecutive speech fractions of the same length but different from the pitch cycle time. **Figure 3** describes a vocal phoneme, in which the pitch marks are denoted

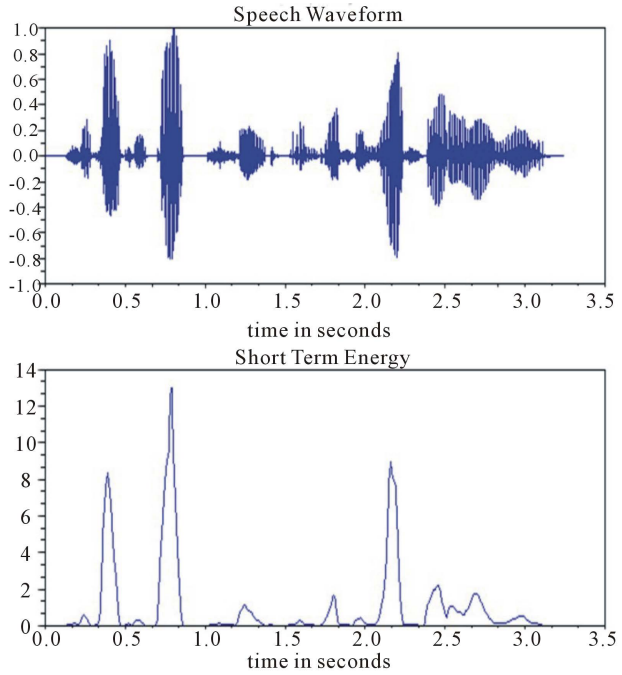

Figure 2. A speech signal (a) speech waveform (b) short term energy.

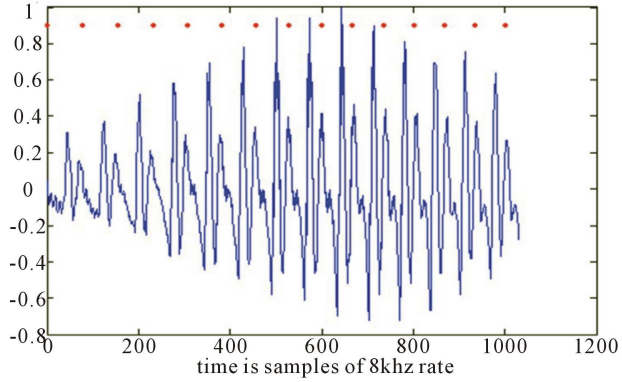

Figure 3. A phoneme with its pitch cycle marks (in red).

in red.

The pitch detector's algorithm can be given by the equation as below.

$$\rho_{\tau} = \frac{(x,y)}{\|x\| \cdot \|y\|}; \|x\| = \sqrt[2]{(x,x)} \tag{2}$$

4. Pattern Classification

4.1. Support Vector Machine

Support vector machine (SVM) is based on the principle of Structural Risk Minimization (SRM). Like RBFNN, support vector machines can be used for pattern classification and nonlinear regression. SVM constructs a linear model to estimate the decision function using non-linear class boundaries based on support vectors. The basic SVM takes a set of input data and predicts, for each given input, which of two possible classes forms the output, making it a non-probabilistic binary linear classifier.

Given a set of training examples, each marked as belonging to one of two categories, an SVM training algorithm builds a model that assigns new examples into one category or the other. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall on. (**Figure 4**) A support vector machine constructs a hyperplane or set of hyperplanes in a high- or infinite-dimensional space, which can be used for classification, regression, or other tasks. Intuitively, a good separation is achieved by the hyperplane that has the largest distance to the nearest training data point of any class (so-called functional margin), since in general the larger the margin the lower the generalization error of the classifier. A linear support vector machine is composed of a set of given support vectors z and a set of weights w. The computation for the output of a given SVM with N support vectors z_1, z_2 …, z_N and weights w_1, w_2, …, w_N is then given by:

$$F(x) = \sum_{i=1}^{N} Wi\, k(Zi, x) + b \tag{3}$$

4.2. Radial Basis Functional Neural Network

The radial basis function is so named because the radius distance is the argument to the function. Euclidean distance is computed from the test point being evaluated to the mean center of each neuron, and a radial basis function is applied to the distance to compute the weight for each neuron. The farther a neuron is from the test point being evaluated, the lesser the influence it has. Feature-response decreases monotonically with distance from a central point. RBFNN have one hidden layer and it requires more hidden units (**Figure 5**).

RBFNN are more robust to novel data and trains faster but suffers from the cause of dimensionality. RBFNN consists of an input layer with n_i units for n_i dimensional input vector fully connected to hidden layer. Hidden layer with n_h units fully connected to output layer. Output layer with n_c units for n_c number of classes. Hidden layer implements the Gaussian radial basis function and the activation function of the i^{th} hidden unit for an input vector x_j is characterised by the mean vectors and covariance matrices:

$$g_i(x_j) = \exp\left(\frac{-\|x_j - \mu_i\|^2}{2\sigma_i^2}\right) \tag{4}$$

where x_j = input vector; μ_i = mean vector(centers); σ_i^2 = variance.

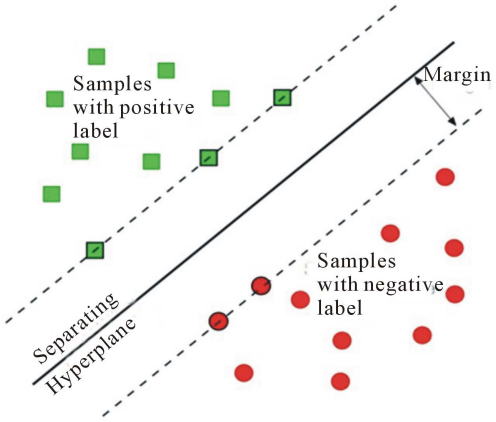

Figure 4. Principle of SVM.

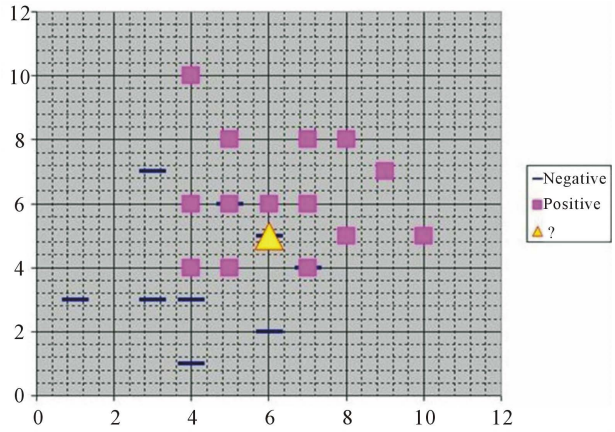

Figure 5. Principle of RBFNN.

5. Proposed Methodology

The speech from the pathological voiced children and normal children was recorded. They are trained to utter same set of phrases and the silences in between speech utterances are clipped off using a silence removal algorithm (Figure 6).

5.1. Silence Removal

Silence removal is considered to be one of the efficient dimensionality reduction processes. The signal energy and spectral centroid are used for silence removal in speech signal. The segments are decided based on the threshold value, which is extracted from the feature sequences of the input signal.

Signal Energy of the i^{th} frame is defined using the formula

$$E(i)=\frac{1}{N}\sum_{n=1}^{N}\left|x_i(n)\right|^2 \tag{5}$$

N is the length of the sample.

The spectral centroid, $\boldsymbol{C_i}$ is defined as the center of gravity of the spectrum.

$$C_i=\frac{\sum_{k=1}^{N}(k+1)x_i(k)}{\sum_{k=1}^{N}x_i(k)}\cdot x_i(k), \tag{6}$$

where $x_i(k)$ is the DFT of the i^{th} frame.

5.2. Windowing

Speech is non-stationary signal where properties change quite rapidly over time. This is completely natural and nice thing but makes the use of DFT or autocorrelation as such impossible. 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 in fact is done in this way: by taking short windows (overlapping possibly) and processing them. The short window of signal like this is called *frame*. In implementation view, the windowing corresponds to what is understood in filter design as window-method: a long signal (of speech for instance or ideal impulse response) is multiplied with a window function of finite length, giving finite length weighted (usually) version of the original signal and is shown below in Figure 7.

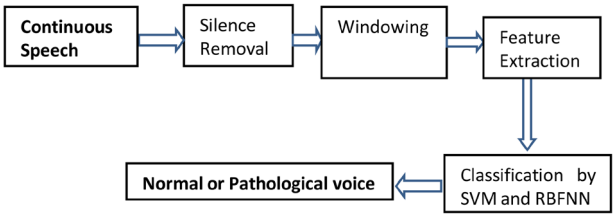

Figure 6. System overview for classifying the normal and pathological voice.

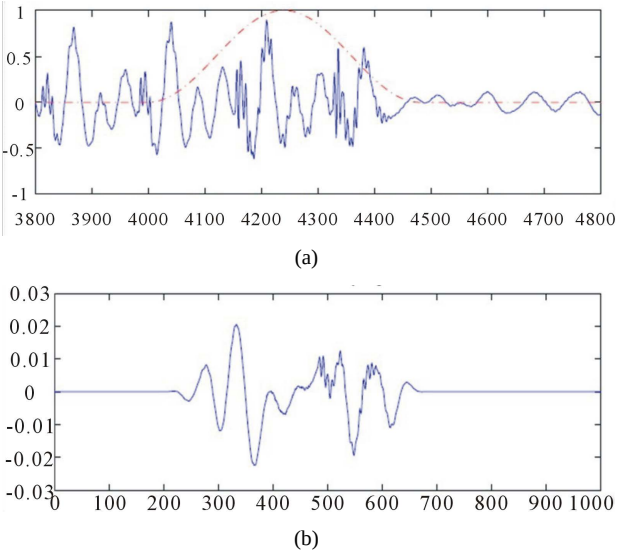

Figure 7. (a) Original signal (b)Windowed signal.

5.3. Fundamental Frequency Estimation

A pitch detection algorithm (PDA) is an algorithm designed to estimate the pitch or fundamental frequency of a quasiperiodic or virtually periodic signal, usually a digital recording of speech or a musical note or tone. Fundamental Frequency (f_0) or pitch voice corresponds perceptually to the number of times per second the vocal folds come together during phonation. Fundamental frequency has long been difficult parameter to reliably estimate from the speech signal. Previously it was neglected for number of reasons, including large computational burden required for accurate estimation, the concern that unreliable estimation would be a barrier achieving high performance, and the difficulty in characterizing complex interactions between and suprasegmental phenomena. 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 [1].

The information about pitch period "T_0" is more pronounced in the autocorrelation sequence of voiced speech compared to the speech segment itself. Since autocorrelation sequence is symmetric with respect to zero lag, only positive lag values are considered. The "T_0" information is more pronounced in the autocorrelation sequence compared to speech. By that, the second largest peak is the autocorrelation sequence, represents pitch T_0 and can be picked up easily by a simple peak picking algorithm compared to finding "T_0" from the speech segment itself. Hence autocorrelation method is preferred over other direct methods of pitch estimation from speech.

Autocorrelation function for a signal $x(n)$ is computed as given in [1]:

$$\varnothing x(m) = \lim_{n\to\infty} \frac{1}{2N+1} \sum_{n=-N}^{N} x(n)\, x(n+m) \qquad (7)$$

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 $\varnothing x(m)$ is also periodic with the same period.

$$\varnothing x(m) = \varnothing x(m+P)$$

5.4. 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 modelled 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.

In autoregressive coding of speech, it is essential that the LPC model contain accurate information about the first three formants; specifically, that the LPC spectrum should reproduce the correct formant frequencies and the corresponding bandwidths. The modified linear predictive coder (MLPC) is superior to the widely used linear predictive coder (LPC) when the data frames are short. Perception of these syllables critically depends on accurate detection of the rapid frequency changes in the first milliseconds of voicing (formant transitions). Inaccurate detection of these formant transitions inevitably interferes with the identification of the phonological cues that are typical for spoken language. The resonant (formant) frequency of a uniform tube, which is a model of the vocal tract [4] is given by the equation below:

$$F_n = (2n-1)c/4L \qquad (8)$$

where

F_n—n^{th} formant frequency [Hz]

c—sound velocity [m/s]

L—vocal tract length [m].

The aim of linear prediction is to estimate the transfer function of the vocal tract from the speech. The signal model can be defined as:

$$s(n) = \sum_{i=1}^{N_{LP}} \propto_{LP} (i)\, s(n-i) + e(n) \qquad (9)$$

where N_{LP}, α_{LP} and $e(n)$ represent**,** respectively, the number of coefficients in the model the linear prediction coefficients and the error in the model. The above equation can be written in Z-transform notation as a linear filtering operation:

$$E(z) = H_{LP}(z) \cdot s(z) \qquad (10)$$

5.5. Jitter Estimation

Jitter deals with varying loudness in the voice. Jitter is said to be the interval between the maximum effects or minimum effects of a signal characteristic that changes regularly in time. The average absolute difference between consecutive periods is expressed as in [8]:

$$\text{Jitter}(\text{absolute}) = \frac{1}{N-1} \sum_{i=1}^{N-1} |T_i - T_{i+1}| \qquad (11)$$

where T_i are the extracted F_0 period lengths and N is the number of extracted F_0 periods.

Jitter (relative) is the average absolute difference be-

tween consecutive periods, divided by the average period and is expressed as a percentage [8]:

$$\text{Jitter}(\text{Relative}) = \frac{\frac{1}{N-1}\sum_{i=1}^{N-1}\left|T_i - T_{i+1}\right|}{\frac{1}{N}\sum_{i=1}^{N} T_i} \quad (12)$$

5.6. Shimmer Estimation

Shimmer deals with a frequent back and forth change in amplitude in the voice. The average absolute base-10 logarithm of the difference between the amplitudes of consecutive periods, multiplied by 20 [8]:

$$\text{Shimmer}(\text{absolute}) = \frac{1}{N-1}\sum_{i=1}^{N-1}\left|20\log\left(A_{i+1}/A_i\right)\right| \quad (13)$$

where A_i is the extracted peak-to-peak amplitude data and N is the number of extracted fundamental frequency periods.

Shimmer (relative) is defined as the average absolute difference between the amplitudes of consecutive periods, divided by the average amplitude, expressed as in [8]:

$$\text{Shimmer}\ (\text{Relative}) = \frac{\frac{1}{N-1}\sum_{i=1}^{N-1}\left|A_i - A_{i+1}\right|}{\frac{1}{N}\sum_{i=1}^{N} A_i} \quad (14)$$

5.7. Reflection Coefficients

A reflection coefficient calculated from the cross-sectional areas of vocal tubes expresses the rate of reflection. Let the cross-sectional area of the left tube be S_n and of the right tube be S_{n+1}. The reflection coefficient k_n is defined as follows:

$$k = \frac{s_n - s_{n+1}}{s_n + s_{n+1}} \quad -1 < k_n < 1 \quad (15)$$

5.8. Classification

5.8.1. Support Vector Machine

The SVM algorithm can construct a variety of learning machines by use of different kernel functions [4]. Three kinds of kernel functions are usually used

Linear Kernel

The Linear kernel is the simplest kernel function. It is given by the common inner product <*x*, *y*> plus an optional constant *c*. Kernel algorithms using a linear kernel are equivalent to their non-kernel counterparts.

$$k(x, y) = x^T y + c \quad (16)$$

Polynomial Kernel

The polynomial kernel is a non-stationary kernel. It is well suited for problems where all data is normalized.

$$k(x, y) = \left(\propto x^T y + c\right)^d \quad (17)$$

Gaussian Kernel

Gaussian kernel is one of the most versatile kernels. The width parameter of the Gaussian kernel controls the flexibility of the resulting classifier.

$$k(x, y) = \exp\left(-\frac{\|x - y\|^2}{2\sigma^2}\right) \quad (18)$$

5.8.2. Radial Basis Functional Neural Network

A radial basis function network is an artificial neural network which uses radial basis functions as activation functions by which the output of the network is determined by a linear combination of radial basis functions of the inputs and neuron parameters (**Figure 8**).

A typical Gaussian RBF is given as:

$$h(x) = \exp\left(\frac{-\|x - c\|^2}{2\sigma^2}\right) \quad (19)$$

where x = input; c = mean centre; σ = spread.

6. Experiments and Results

The speech signal is recorded from 20 children (10 normal, 10 pathological) at the rate of 8000 samples per second and a dataset is created. All the speech samples were recorded in noise free environment using a microphone array. Each speech sample is pre-processed using a silence removal algorithm and a windowing technique. Using Autocorrelation method the fundamental frequency is estimated, and Linear predictive analysis is used to extract the formant frequencies F_1 and F_2. The two first formants, f_1 and f_2, are enough to disambiguate the normal and pathological voices. Since the number of formants is same for all the utterances the peaks of the formant frequencies are found using a magnitude threshold

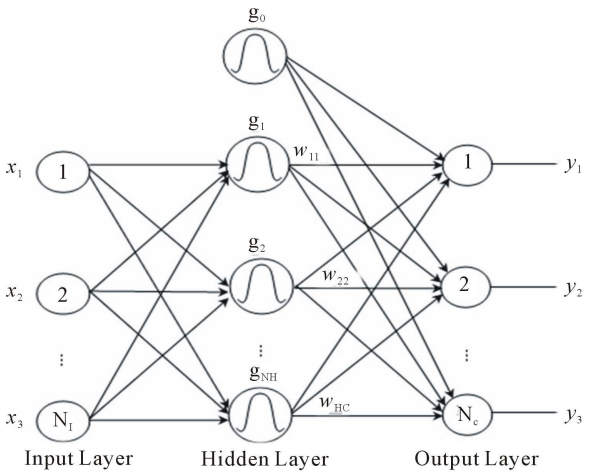

Figure 8. Architecture of RBFNN.

Table 1. Precisions of both classifiers.

Results	Classification Accuracy
RBFNN	91%
SVM	83%

based peak detection algorithm. 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 a 16 coefficient feature set. The classifiers are trained and tested with the same pre-condition like the same data set and the same characteristic parameters.

The experimental results for the classification of pathological voice are shown in **Table 1**. The **Table 1** shows the classification accuracy of the various Classifiers used for classification. Among the 2 classifiers RBFNN outperforms SVM in terms of Classification Accuracy.

7. Conclusion

In this paper several acoustic techniques for extracting different acoustic parameters and providing a hybrid approach of feature extraction are presented. The purpose of this methodology is to classify the voice dataset into normal and pathological voice and to compare the classification performance based on classification accuracy using Support Vector Machine and Radial Basis Functional Neural Network. The future work will be based on extracting different feature sets by combining the derived features like Linear Predictive Co-efficients (LPC), Linear Predictive Cepstral Co-efficients (LPCC), etc., with the currently extracted raw features. Considering all the features, a combined feature set is constructed for measuring their performance. Further this feature set will be used to implement different classification models so as to compare different classifiers based on classification accuracy and also to design a new pattern classification model.

REFERENCES

[1] L. Salhi, M. Talbi and A. Cherif, "Voice Disorders Identification Using Hybrid Approach: Wavelet Analysis and Multilayer Neural Networks," World Academy of Science, Engineering and Technology, 2008.

[2] 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.

[3] C. Pend, Q. J. Xu, B. K. Wan and W. X. Chen, "Pathological Voice Classification Based on Features Dimension Optimization," *Transactions of Tianjin University*, Vol. 13, No. 6, 2007.

[4] 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.

[5] E. Vaiciukynas, A. Gelzins, M. Bacauskiene, A. Verikas and A. Vegiene, "Exploring Kernels in SVM-Based Classification of Larynx Pathology from Human Voice," Department of Electrical and Control Instrumentation, Kaunas University of Technology, Lithuania.

[6] D. Pravena, S. Dhivya and A. Durga Devi, "Pathological Voice Recognition for Vocal Fold Didease," *International Journal of Computer Applications* (0975-888), Vol. 47, No. 13, 2012..

[7] M. de Oliviera Rosa, J. C. Pereira and M. Grellet, "Adaptive Estimation of Residue Signal for Voice Pathology Diagnosis," *IEEE Transactions on Biomedical Engineering*, Vol. 47, No. 1, 2000.

[8] M. Farra, J. Hernando and P. Ejarque, "Jitter and Shimmer Measurements for Speaker Recognition," TALP Research Center, Department of Signal Theory and Communications, Universitat Politecnica de Catalunya, Barcelona.

[9] L. R. Rabiner and R. W. Schafer, "Digital Processing of Speech Signals".