I[st] International conference on New Paradigms in Engineering Technology and Management
(IETE) Institution of Electronics and Telecommunication Engineers, Pune, India ICNPETM-2017
05[th] November 2017, www.conferenceworld.in ISBN: 978-93-86171-75-7

Emotion Classification Technique in Speech Signal for Marathi

P.S.Deshpande[1], J.S.Chitode[2]

[1,2] *Department of Electronics, Bharati Vidyapeeth College of Engineering,Pune, India)*

ABSTRACT

Our earnest attempt, here, is to launch a novel emotions classification method in speech signal by supplementing emotions in Marathi. The speech signals are, initially, extracted from the database and hence there is good chance of signal being contaminated with noise pollution. These issues are tackled by denoising the input signals by means of Gaussian filter and features such as MFCC, peak, pitch spectrum, mean & standard deviation of the signal and minimum & maximum of the signal are estimated from the denoised signal. The evaluated features are then furnished to the popular classifier like Feed Forward Backpropogation Neural Network (FFBNN) to accomplish the guidance task. The execution of the envisaged method is assessed by furnishing further additional number of speech signals to the well guided FFBNN. Thereafter, the efficiency of our innovative approach is analyzed and contrasted with those of the parallel methodologies.

Keywords: Mel Frequency Cepstral Coefficients (MFCC), Peak, Pitch, Gaussian Filter

I.INTRODUCTION

Speech is the principal mode of communication between humans, both for transfer of information and for social interaction. Consequently, learning the mechanisms of speech has been of interest to scientific research, leading to a wealth of knowledge about the production of human speech, and thence to technological system to simulate and to recognize speech electronically [1]. Nowadays speech synthesis systems have reached a high degree of intelligibility and satisfactory acoustical quality. The goal of next generation speech synthesizers is to express the variability typical to human speech in a natural way or, in other words, to reproduce different speaking styles and particularly the emotional ones in a reliable way [4]. The quality of synthetic speech has been greatly improved by the continuous research of the speech scientists. Nevertheless, most of these improvements were aimed at simulating natural speech as that uttered by a professional announcer reading a neutral text in a neutral speaking style. Because of mimicking this style, the synthetic voice results to be rather monotonous, suitable for some man-machine applications, but not for a vocal prosthesis device such as the communicators used by disabled people [5].

In the last years, progress in speech synthesis has largely overcome the milestone of intelligibility, driving the research efforts to the area of naturalness and fluency. These features become more and more necessary as the synthesis tasks get larger and more complex: natural sound and good fluency and intonation are mandatory for understanding a long synthesized text [6]. A vital part of speech technology application in modern voice application platforms is a text-to-speech engine. Text-to-speech synthesis (TTS) enables automatic converts any

available textual information into spoken form. With the evolution of small portable devices has made possible the porting of high quality text-to-speech engines to embedded platforms [2] [3]. It is well known that speech contains acoustic features that vary with the speaker's emotional state. The effects of emotion in speech tend to alter pitch, timing, voice quality and articulation of the speech signal [7] [8]. Expressive speech synthesis from tagged text requires the automatic generation of prosodic parameters related to the emotion/style and a synthesis module able to generate high quality speech with the appropriate prosody and the voice quality [9].

Furthermore, adding vocal emotions to synthetic speech improves its naturalness and acceptability, and makes it more 'human'. We provide the user with the ability to generate and author vocal emotions in synthetic speech, using a limited number of prosodic parameters with the concatenative speech synthesizer [10]. The voice plays an important role for conveying emotions. For example, rhythm and intonation of the voice seem to be important features for the expression of emotions [11] [12]. Adding emotions to a synthesized speech means that the latter can verbalize language with the kind of emotion appropriate for a particular occasion (e.g. announcing bad news in a sad voice). Speech articulated with the appropriate prosodic cues can sound more convincing and may catch the listener's attention, and in extreme cases, it can even avoid tragedies [16]. An improved synthesized speech can also benefit from other speech-based human-machine interaction systems that perform specific tasks like read-aloud texts (especially materials from the newspaper) for the blind, weather information over the telephone, auditory presentation of instructions for complex hand free tasks [13].

The rest of the paper is organized as follows: Section 2 reviews the related works with respect to the proposed method. Section 3 discusses about the proposed technique. Section 4 shows the experimental result of the proposed technique and section 5 concludes the paper.

II. RECENT RELATED RESEARCHES: A REVIEW

Mumtaz Begum *et al*. [14] have presented the findings of their research which aims to develop an emotions filter that can be added to an existing Malay Text-to-Speech system to produce an output expressing happiness, anger, sadness and fear. The end goal has been to produce an output that is as natural as possible, thus contributing towards the enhancement of the existing system. The emotions filter has been developed by manipulating pitch and duration of the output using a rule-based approach. The data has been made up of emotional sentences produced by a female native speaker of Malay. The information extracted from the analysis has been used to develop the filter. The emotional speech output has undergone several acceptance tests. The results have shown that the emotions filter developed has been compatible with FASIH and other TTS systems using the rule-based approach of prosodic manipulation. However, further work needs to be done to enhance the naturalness of the output.

Zeynep Inanoglu *et al*. [15] have described the system that combines independent transformation techniques to provide a neutral utterance with some required target emotion. The system consists of three modules that are each trained on a limited amount of speech data and act on differing temporal layers. F0 contours have been modeled and generated using context-sensitive syllable HMMs, while durations are transformed using phone-based relative decision trees. For spectral conversion which is applied at the segmental level, two methods have been investigated: a GMM-based voice conversion approach and a codebook selection approach. Converted test

data have been evaluated for three emotions using an independent emotion classifier as well as perceptual listening tests. The listening test results have shown that perception of sadness output by their system has been comparable with the perception of human sad speech while the perception of surprise and anger has been around 5% worse than that of a human speaker.

Syaheerah L. Lutfi *et al*. [16] have concerned the addition of an affective component to Fasih1, one of the first Malay Text to- Speech systems developed by MIMOS Berhad. The goal has been to introduce a new method of incorporating emotions to Fasih by building an emotions filter that is template-driven. The templates have been diphone-based emotional templates that can portray four types of emotions, i.e. anger, sadness, happiness and fear. A preliminary experiment that focused on has shown that the recognition rate of Malay synthesized speech is over 60% for anger and sadness.

Al-Dakkak *et al*. [17] have discussed that many attempts have been conducted to add emotions to synthesized speech. Few are done for the Arabic language. They have introduced a work done to incorporate emotions: anger, joy, sadness, fear and surprise, in an educational Arabic text-to-speech system. After an introduction about emotions, they have given a short paragraph of their text-to-speech system, then they have discussed their methodology to extract rules for emotion generation, and finally they have presented the results and tried to draw conclusions.

Syaheerah L. Lutfi *et al.* [18] have presented the pilot experiment conducted for the purpose of adding an emotional component to the first Malay Text-to-Speech (TTS) system, Fasih. The aim has been to test a new method of generating an expressive speech via a template-driven system based on diphones as the basic sound. The synthesized expressive speech could express four types of emotions. However, as an initial test the pilot experiment has focused on anger and sadness. The results from this test have shown an impressive recognition rate of over 60% for the synthesized speech of both emotions. The pilot experiment has paved the way for the development of an emotions filter to be embedded into Fasih, thus allowing for the possibility of generating an unrestricted Malay expressive speech.

III. PROPOSED SPEECH EMOTION CLASSIFICATION TECHNIQUE

In this research work, we have proposed a novel emotion classification technique in speech signal by adding emotions. Our innovative technique consists of three stages namely,

i) Denoising,

ii) Feature Mining and

iii) Recognition

Initially, the speech signals consist of declarative sentences and interrogative sentences gathered from the database which are denoised with the help of Gaussian filter. Then features such as MFCC, peak, pitch spectrum, mean & standard deviation of the signal and minimum & maximum of the signal are extracted from the denoised signal. Subsequently, the extracted features are given to FFBNN to attain the training process. By giving more speech signals to the trained FFBNN, the performance of the projected technique is analyzed. The architecture of the new technique is given in figure 1.

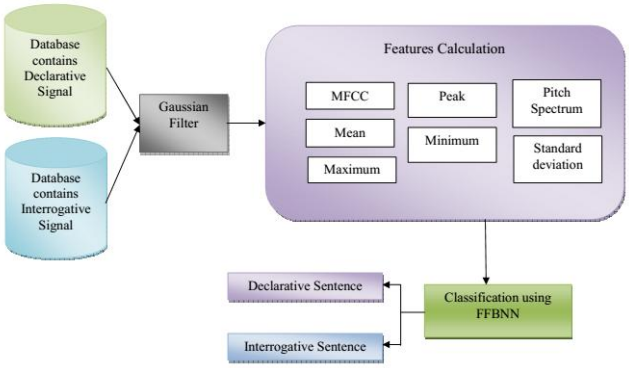


Figure 1: Architecture of our proposed Emotion Classification Technique

3.1. Denoising Confrontation

Let us consider two databases D1 and D2 which house the declarative and interrogative speech signals correspondingly. These signals are likely to be contaminated with noise pollution, which has the effect of bringing down the classification precision of the speech. With a view to remove this, Gaussian filter is employed which discharges the task of denoising. In the case of signal processing, a Gaussian filter is a filter whose impulse response tends to be a Gaussian function. Gaussian filters are designed in such a way as to block overrun to a step function input, simultaneously decreasing the interval for the rise and fall. This tendency is very much linked to the fact that the Gaussian filter causes least group delay in this regard. The system receives the input signal and it is furnished to the preprocessing phase, where the signal noise is eliminated by this Gaussian filter, resulting in the achievement of noise free output signal. Usually, a 1D Gaussian filtering is employed for the noise exclusion procedure, which is defined as

$$G(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-x^2}{2\sigma^2}} \quad (1)$$

Now, the input speech signal is furnished to the Gaussian filter, which leads to the decrease of noise in the input speech signal, in addition to realizing a superior quality speech signal for additional processing. The preprocessed speech signals from database for both declarative and interrogative signals are symbolized as,

$$D_1 = \{s_1', s_2', \cdots s_r'\}, r = 1, 2, \cdots n \quad (2)$$

$$D_2 = \{s_1', s_2', \cdots s_t'\}, t = 1, 2, \cdots m \quad (3)$$

$$s_r' = \{u_1^r', u_2^r', \cdots u_i^r'\} \quad (4)$$

$$s_t' = \{u_1^t', u_2^t', \cdots u_j^t'\} \quad (5)$$

3.2 Feature Extraction

The preprocessed signal is then subjected to feature extraction process where the features such as MFCC, peak, pitch spectrum, mean & standard deviation of the signal and minimum & maximum of the signal are extracted.

I) Mel Frequency Cepstral Coefficients (*MF*)

At this juncture, the exact features are mined from the input noise free speech signals so as to attain the preferred speech processing functions. The mining of the finest parametric illustration of acoustic signals is a fundamental function to usher in superb detection efficiency. The effectiveness of this stage is crucial for the accompanying stage. Mel frequency cepstral coefficients (MFCC) is one of the most triumphant trait representations in speech recognition linked functions, and the coefficients are attained by means of a filter bank investigation. The major steps constituting the features mining are detailed below:

(i) Pre-Emphasis

The preprocessed speech signals of both databases are furnished to the MFCC trait mining pre-emphasis stage. Pre-emphasis is a procedure, meant for enhancing the dimension of certain frequencies in relation to the dimensions of parallel frequencies. At this time, the processed speech signals are sent through a filter for emphasizing higher frequencies. This procedure tends to enhance the energy of speech signal at higher frequency. The speech signal is first pre-emphasized by a first order FIR filter with pre-emphasis coefficient κ . The first order FIR filter transfer function in the z domain is,

$$F(z) = 1 - \kappa z^{-1} \qquad (6)$$

The pre-emphasis coefficient κ lies in the range $0 \le \kappa \le 1$.

$$p(u_i^r{}') = \vartheta(u_i^r{}') - \kappa\,\vartheta(u_i^r{}'-1) \qquad (7)$$

$$p(u_j^t{}') = \vartheta(u_j^t{}') - \kappa\,\vartheta(u_j^t{}'-1) \qquad (8)$$

(ii) Frame Blocking

The statistical features of a speech signal are not subjected to any alterations only for minute time periods. Now, the pre-emphasized signal is blocked into frames of f_N samples (frame size), with adjoining frames being alienated by f_M samples (frame shift). If the l^{th} frame of speech is $x_l(u_i^r{}')$, $x_l(u_j^t{}')$ and there are L frames within the overall speech signal, then

$$x_l(u_i^r{}') - (f_{M^r} + u_i^r{}'), \quad 0 \le u_i^r{}' \le f_{M^r} - 1, \quad 0 \le l \le f_{M^r} - 1 \qquad (9)$$

$$x_l(u_j^t{}') - (f_{M^t} + u_j^t{}'), \quad 0 \le u_j^t{}' \le f_{M^t} - 1, \quad 0 \le l \le f_{M^t} - 1 \qquad (10)$$

(iii) Windowing

Subsequently, we carry out the procedure of windowing, in which every frame is windowed with a view to decrease the signal stoppages at the beginning and finish of the frame. The window is so selected as to tape the signal at the edges of every frame. If the window is defined as,

$$w(u_i^r{}'), \quad 0 \le u_i^r{}' \le f_{M^r} - 1 \qquad (11)$$

$$w(u_j^t'),\ \ 0 \le u_j^t' \le f_{M^t} - 1 \qquad (12)$$

Then the outcome of windowing the signal is furnished by:

$$x_l(u_i^r') = x_l(u_i^r')w(u_i^r'),\quad 0 \le u_i^r' \le f_{M^r} - 1 \qquad (13)$$

$$x_l(u_j^t') = x_l(u_j^t')w(u_j^t'),\quad 0 \le u_j^t' \le f_{M^t} - 1 \qquad (14)$$

Hamming window is a fine selection in speech detection, which includes the entire closest frequency lines. The Hamming window equation is furnished as,

$$w(u_i^r') = 0.54 - 0.46\cos\left(\frac{2\pi u_i^r'}{f_{M^r} - 1}\right) \qquad (15)$$

$$w(u_j^t'), = 0.54 - 0.46\cos\left(\frac{2\pi u_j^t'}{f_{M^t} - 1}\right) \qquad (16)$$

(iv) Filter Bank Analysis

The filter bank analysis is carried out to change every time domain frame of f_N samples into frequency domain. The Fourier Transform is performed to alter the intricacy of the glottal pulse and the vocal tract impulse response in the time domain into frequency domain. The frequency range in FFT spectrum is exceedingly extensive and voice signal does not toe the line of the linear scale. A group of triangular filters are utilized to estimate a weighted sum of filter spectral components in such way that the yield of procedure approximates to a Mel scale. Each filter's magnitude frequency response is triangular in form and equivalent to unity at the centre frequency and decreased linearly to zero at centre frequency of two adjoining filters. Thereafter, every filter yield is the sum of its filtered spectral components. The mel scale is defined as,

$$M^f = 2595\log_{10}\left(1 + \frac{f}{700}\right) \qquad (17)$$

The filters are jointly known as a Mel scale filter bank and the frequency response of the filter bank replicate the perceptual processing executed within the ear.

(v) Logarithmic compression

At this point, the logarithmic function compacts the filter outputs attained from filter bank analysis. The $f_m th$ filter logarithmically compressed yield is described as,

$$X_{f_{m^r}(\ln)} = \ln(X_{f_{m^r}}),\qquad 1 \le f_{m^r} \le f_{M^r} \qquad (18)$$

$$X_{f_{m^t}(\ln)} = \ln(X_{f_{m^t}}),\qquad 1 \le f_{m^t} \le f_{M^t} \qquad (19)$$

(vi) Discrete Cosine Transformation

Thereafter, Discrete Cosine Transform (DCT) is performed on the filter outputs and a certain number of initial coefficients are clustered jointly as a feature vector of a definite speech framework. The L^{th} MFCC coefficient in the range $1 \le L \le C$ is furnished as,

$$MF_k^r(u_i^{r\,\prime}) = \sqrt{\frac{2}{f_{M^r}}} \sum X_{f_{m^r}(\ln)} \cos(\pi l(f_{m^r} - 0.5) f_{M^r}) \quad (20)$$

$$MF_k^t(u_j^{t\,\prime}) = \sqrt{\frac{2}{f_{M^t}}} \sum X_{f_{m^t}(\ln)} \cos(\pi l(f_{m^t} - 0.5) f_{M^t}) \quad (21)$$

Where, C is the degree of the mel scale cepstrum.

II) Peak (*P*)

The highest echelon in a signal is known as a peak. The peak is mined by means of the MATLAB task termed 'PeakFinder'. But, the phase-wise computation of peak tracing technique is haunted by the vexed issue in which the false signals tend to be recognized as peaks, in the event of the signal being contaminated with noise. However, this task is found to adopt a special character of derivate in addition to the user defined threshold to trace the local maxima or minima in peak recognition. This task is capable of locating local peaks or valleys (local maxima) in a sound-polluted vector by means of a user defined magnitude threshold to assess whether every peak is predominantly greater or lesser than the data surrounding it.

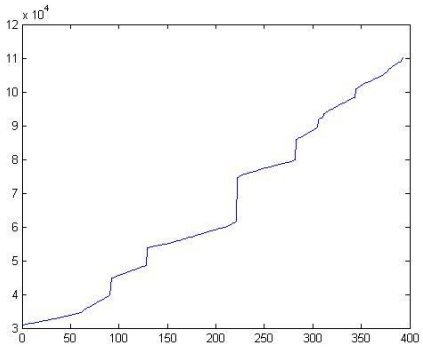


Figure 2: Output of the peak detection process

III) Pitch Spectrum (*PS*)

Pitch is the minimum frequency module of a signal that motivates a verbal mechanism. Pitch period is considered as the minimum repeating signal which varies in inverse proportion to the basic frequency. Pitch period is employed to demonstrate the pitch signal entirely. The YAAPT (Yet another Algorithm for Pitch Tracking) is a basic frequency (Pitch) tracking algorithm [19], which is intended for significant precision and high robustness in terms of excellent quality and telephone communication. The YAAPT algorithm proceeds through the following five phases:

1)Preprocessing

In this task, two types of signals such as original signal and absolute value of the signal are generated and every signal gets band pass filtered and center clipped.

2) Pitch candidate Selection Based on Normalized Cross Correlation Function (NCCF)

The correlation signal has a peak of huge magnitude at a delay analogous to the pitch period. If the magnitude of the leading peak is found to be greater than that of the threshold (about 0.6), then the framework of speech is uttered typically.

3)Candidate Refinement Based on Spectral Information

The candidate achieved in the earlier stage is adapted according to the universal and local spectral data.

4)Candidate Modifications Based on Plausibility and Continuity Constraints

A smooth pitch track is achieved by adapting the refined candidate by means of Normalized low Frequency energy Ratio (NLFER).

5)Final Path Determination Using Dynamic Programming

Pitch candidate matrix, a merit matrix, an NLFER curve (from the original signal), and the spectrographic Pitch track achieved through the phases mentioned elsewhere are employed to locate the minimum cost pitch track from among the entire accessible candidates by the use of dynamic programming.

IV) Mean and Standard deviation of the Signal

Mean (μ) is the average value of the signal which is achieved by totaling all the signals and dividing it by the total number of the signals. The mathematical expression is shown below.

$$\mu = \frac{\sum_{i=0}^{N-1} ss_i}{N} \qquad (22)$$

Here, N - Total number of values in the signal, ss_i - values in speech signal s_i

The standard deviation is analogous to the mean deviation and is obtained by squaring every one of the variances before calculating the average. At last, the square root is calculated to recompense for the preliminary squaring. The standard deviation is determined as per equation given below.

$$\sigma = \frac{\sum_{i=0}^{N-1} (s_i - \mu)^2}{N-1} \qquad (23)$$

V) Minimum and Maximum of the SignalThe minimum value (frequency) in the signal is known as the minimum of the signal (*min*) and the highest value of the signal is termed as the maximum of the signal (*max*). These determined features are thereafter furnished as input to the FFBNN with a view to analyze and categories the speech signal into interrogative or declarative cases.

4.3 Classification by FFBNN

4.3.1 Training

With the intent to analyze and categorize the speech into declarative or interrogative cases, Feed Forward Back Propagation Neural Network (FFBNN) is guided by means of the features like MFCC, peak, pitch spectrum, mean & standard deviation of the signal and minimum & maximum of the signal mined from the

preprocessed signal. The neural network is well guided by utilizing these mined features. The neural network comprises 7 input units, h concealed units and a solitary output unit.

The RProp algorithm is a supervised learning method for training multi layered neural networks, first published in 1994 by Martin Riedmiller. The idea behind it is that the sizes of the partial derivatives might have dangerous effects on the weight updates. It implements an internal adaptive algorithm which focuses only on the signs of the derivatives and completely ignores their sizes. The algorithm computes the size of the weight update by involving an update value which depends on the weights. This value is independent from the size of the gradients.

1. Assign weights randomly to all the neurons except input neurons.
2. The bias function and activation function for the neural network is described below.

$$X(q)=\beta+\sum_{a=0}^{h-1}\begin{pmatrix} w_{qa}MF_{qa}+w_{qa}P_{qa}+w_{qa}PS_{qa} \\ +w_{qa}\mu_{qa}+w_{qa}\sigma_{qa}+ \\ w_{qa}\min_{qa}+w_{qa}\max_{qa}\end{pmatrix} \quad (21)$$

$$X(A)=\frac{1}{1+e^{-X(q)}} \quad (22)$$

In bias function $MF_{qa}, P_{qa}, PS_{qa}, \mu_{qa}, \sigma_{qa}, \min_{qa}$ and $\max_{qa}$ are the calculated features such as MFCC, Peak, Pitch Spectrum, Mean of the signal, Standard Deviation of the Signal, Minimum of the Signal and maximum of the Signal respectively. The activation function for the output layer is given in Eq. (22).

3. Find the learning error.

$$E=\frac{1}{h}\sum_{a=0}^{h-1}d_a-a_{na} \quad (23)$$

E is the FFBNN network output, d_a *and* a_a are the desired and actual outputs and h is the total number of neurons in the hidden layer.

4.3.2 Error Minimization

Weights are allocated to the hidden layer and output layer neurons by randomly chosen weights. The input layer neurons have a constant weight.

1. Determine the bias function and the activation function.
2. Calculate error for each node and update the weights as follows:

$$w_{(qa)}=w_{(qa)}+\Delta w_{(qa)} \quad (24)$$

$\Delta w_{(qa)}$ is obtained as,

$$\Delta w_{(qa)}=-\left(\frac{\partial E}{\partial w_{(qa)}}\right)\Delta_{qa} \quad (25)$$

In equiv. (25), Δ_{qa} is an update value. The size of the weight change is exclusively determined by this weight-specific ,update value. Δ_{qa} evolves during the learning process based on its local sight on the errorfunction E, according to the following learning-rule.

$$\Delta^{t}{}_{qa} = \begin{cases} \chi^{+} \times \Delta^{t-1}{}_{qa,} & if \quad \frac{\partial E^{(a-1)}}{\partial w_{(qa)}} \times \frac{\partial E^{(a)}}{\partial w_{(qa)}} > 0 \\ \chi^{-} \times \Delta^{t-1}{}_{qa}, & if \quad \frac{\partial E^{(a-1)}}{\partial w_{(qa)}} \times \frac{\partial E^{(a)}}{\partial w_{(qa)}} < 0 \\ \Delta^{t-1}{}_{qa}, & otherwise \end{cases} \qquad (26)$$

The weight update Δ_{qa} follows the simple rule:

If the derivative is positive (increasing error), the weight is decreased by its update value. If the derivative is negative, the update value is added.

$$\Delta w_{(qa)} = \begin{cases} -\Delta_{qa}, & \frac{\partial E}{\partial w_{(qa)}} > 0 \\ +\Delta_{qa}, & \frac{\partial E}{\partial w_{(qa)}} < 0, \\ 0, & otherwise \end{cases} \qquad (27)$$

But it has one exception that is, if the partial derivative changes sign, i.e. the previous step was too large and the minimum was missed, the previous weight update is reverted.

$$\Delta w^{t}{}_{(qa)} = -\Delta w^{(a-1)}{}_{(qa)}, \; if \quad \frac{\partial E^{(a-1)}}{\partial w_{(qa)}} \times \frac{\partial E^{(a)}}{\partial w_{(qa)}} < 0 \qquad (28)$$

3. Then repeat the steps (2) and (3) until the error gets minimized.
4. The error gets minimized to a minimum value the FFBNN is well trained for performing the testing phase.

Then the result of the neural network (Y) is compared with the threshold value $(\tau 1)$. If it satisfies the threshold value it is recognized.

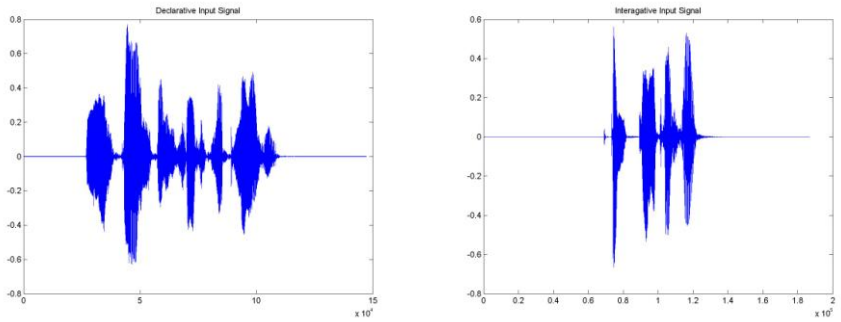
$$result = \begin{cases} recognized, Y \geq \upsilon, \\ not\ recognized, Y < \upsilon \end{cases}$$

V. RESULTS AND DISCUSSION

The proposed Emotion classification technique in Speech Signal for Marathi is implemented in the working platform of MATLAB with machine configuration as follows

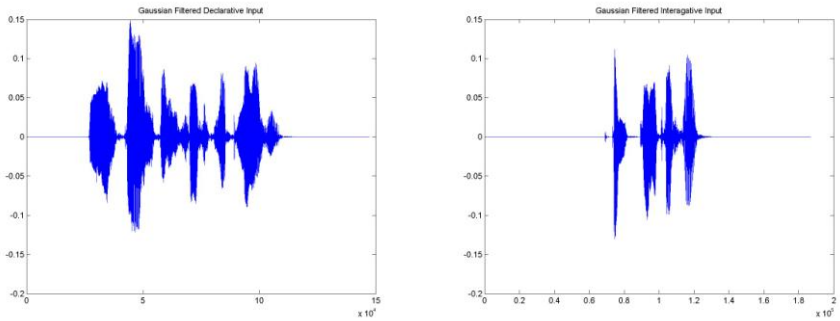
5.1 Performance Analysis

The efficiency of our projected Emotion classification method in speech signal for emotions supplemented text in Marathi is subjected to evaluation by means of the statistical measures which are furnished in [20]. The execution of the novel RP technique is contrasted with the performance of similar optimization methods like the CGP (Cartesian Genetic Programming), GD (Gradient Descent), GDM (Gradient Descent with Momentum), FFBNN- LM (Levenberg-Marquardt) and SCG (Scaled Conjugate Gradient). Moreover, the statistical measures of our innovative scheme are furnished with those of the conventional techniques in Table 1, the statistical measures being TP (True Positive), TN (True Negative), FP (False Positive), FN (False Negative) which is calculated for the novel RP technique and parallel training algorithms like CGP, GD, GDM, LM and SCG. Figure 2, 3 and 4 exhibit the input signal, denoised signal and peak detection in the signal correspondingly.



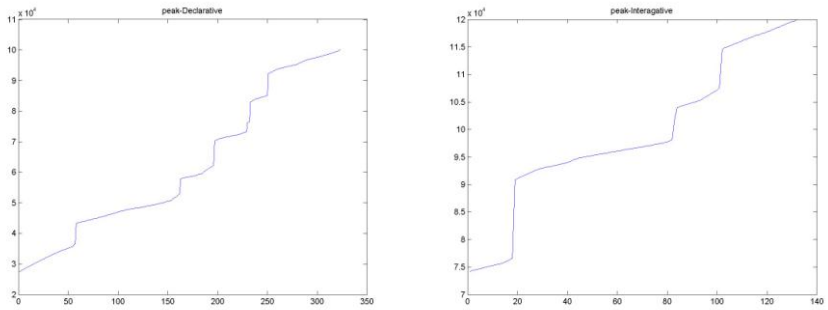
(a) (b)

Figure 2: Input signals (a) Declarative Sentence, (b) Interrogative Signals



(a) (b)

Figure 3: Gaussian Filtered Signals (a) Declarative Sentence, (b) Interrogative Sentence



(a) (b)

Figure 4: Peak (a) Declarative Sentence, (b) Interrogative Sentence

By using the TP, TN, FP and FN values the other statistical measures like accuracy, sensitivity and specificity values are calculated and it is given in table1.

Table 1: Performance of our proposed technique and other optimization techniques such as CGP, GD, GDM, LM, GDX and RP

Measures	No of hidden neurons	**Proposed RP**	CGP	GD	GDM	LM	GDX	SCG
Accuracy	2	83.87	77.42	77.42	77.42	83.87	70.97	77.42
	4	87.1	77.42	74.19	80.65	83.87	80.65	80.65
	6	83.87	80.65	77.42	70.97	80.65	61.29	83.87
	8	83.87	80.65	67.74	51.61	80.65	80.65	80.65
	10	87.1	80.65	74.19	58.06	83.87	77.42	54.84

Table2 contains statistical measures such as accuracy, sensitivity and specificity, which are estimated by changing the number of concealed neurons for all the method. The accuracy of the novel RP technique is found to be an average of 85.162% whereas CGP, GD, GDM, LM, GDX and SCG have 79.358%, 74.192%, 67.742%, 82.582%, 74.196%, and 75.486% of accuracy correspondingly. When contrasted with the projected RP technique, these parallel technique fares poorly, with CGP,GD, GDM, LM,GDX,SCG achieving lower accuracy levels to the tune of 4%, 10%s 21%, 2%, 10% and 9% respectively. This underscores the fact that our novel method is competent to realize higher levels of accuracy vis-à-vis traditional techniques. The feat of our ambitious approach is analyzed and contrasted with the parallel techniques by altering the guiding neurons. The accuracy value is altered by changing the no of neurons. In our innovative method, we have achieved an accuracy of 87.1% by setting 4 and 10 concealed neurons. This enables us to state that if no of concealed neurons increases, the result accuracy tends to increase. Though LM and SCG techniques have achieved an appreciable accuracy exceeding 83%, our innovative technique definitely has been able to achieve the best accuracy surpassing the others.

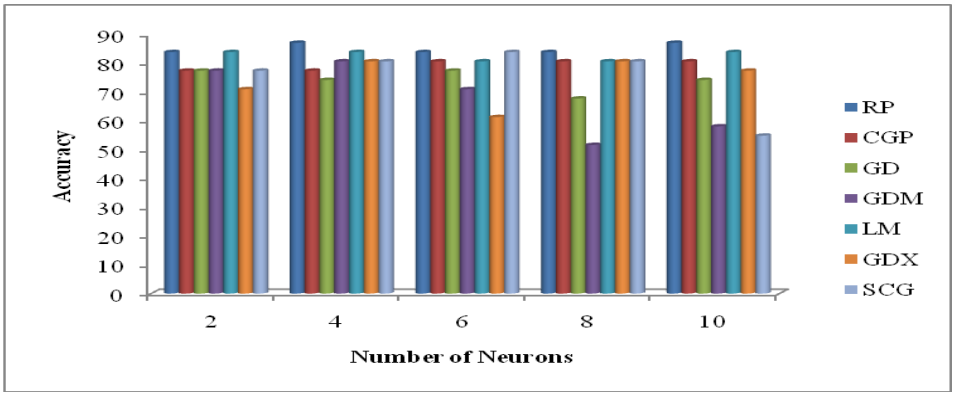


Figure 4: Proposed, CGP, GD, GDM, LM, GDX and SCG techniques performance outcomes in terms of Accuracy.

Figure 4 demonstrates the assessment and contrast of the projected method vis-à-vis parallel techniques. A close observation of the graph establishes the fact the precision of our ambitious method is appreciably superior to those of techniques such as the CGP, GD, GDM, LM, GDX and SCG. Our well-conceived RP speech integration method has been able to yield a rich harvest in precision to the tune of 85.162%. It is a clear pointer to the fact that our epoch-making speech integration technique has come out with flying colors by exhibiting superior efficiency in relation to the parallel techniques. The assessment outcome exhibits the fact that our ambitious RP speech integration technique is far superior to other systems.

In our innovative speech integration mechanism, analysis of the speech is carried out to assess the nature of the speech signal whether declarative or interrogative. With a view to carrying out efficiency evaluation, the SNR value of our Gaussian filter is assessed and contrasted with the Digital filter for both declarative and interrogative cases. The table 2 and 3 exhibit the SNR values of Gaussian filter and Digital filter in both declarative and interrogative cases.

Table 2: SNR value of Gaussian filter and Digital Filter in declarative case

SNR	
Gaussian	Digital
3.219873986	0.231704095
3.21661632	0.270859754
3.039045538	0.236144396
2.935575448	0.140020041
3.522470595	0.363386238

Table 3: SNR value of Gaussian filter and Digital Filter in interrogative case

SNR	
Gaussian	Digital
3.678722666	1.522145204
3.310926724	0.498244048
3.497103601	0.981783979
3.445375785	1.177605884
3.471421082	0.806302515

A close observation of Tables 2 and 3 makes it crystal clear that the SNR values of Gaussian filter employed in our dream scheme are smaller than those of the Digital filter in both declarative and interrogative cases. Signal-to-noise ratio (SNR) is a dimension of signal power comparative to the noise in the environment. It is usually measured in decibels (dB).The SNR value of Gaussian filter is estimated at 3.1867 and the Digital filter at 0.2484 in declarative case for the entire number of concealed neurons. On the contrary, the Gaussian value is found to be 3.480 in average and the Digital value is observed to be 0.9972 in interrogative for the entire number of concealed neurons. Thus the Digital filter is less by 2.9383dB in the declarative case and by 2.4828dB in the interrogative case. This unequivocally establishes the fact that our magnum opus mechanism shows superb strength when contrasted with the Digital filter. The obtained outcomes also underscore the unassailable fact that the SNR values are not all affected in any way by the concealed neurons. In addition, as per our envisaged mechanism, the SNR value of Gaussian filter is identical for both declarative and interrogative cases, in contrast to the modern Digital filter where the values in declarative and interrogative cases differ considerably. Solidly backed by the cheering outcomes, we are gratified that our epoch making technique for speech integration is well-equipped to achieve amazing efficiency superbly superior to those of the current parallel methods.

VI. CONCLUSION

Through this document, we have tried our level best to successfully launch an innovative emotion classification approach in speech signal by supplementing emotions in Marathi by means of FFBNN technique. The projected technique is performed and a mammoth group of test data is employed to assess the proficiency of the projected emotions classification method. The accomplishment of the innovative emotion classification technique is assessed and contrasted vis-a-vis several guidance techniques employed in the FFBNN. The assessment outcomes clearly exhibit the fact that our well-conceived emotion classification method backed by RP as the guidance algorithm in FFBNN has been able to attain higher echelons of efficiency in performance in relation to the parallel guidance algorithms. Our newly launched emotion classification method in Speech signal significantly aided by FFBNN is able to achieve an amazing precision of a whopping 85.162%. Thus we are pleased to declare that with the solid backing of the FFBNN, our novel emotion classification method in speech signal rich with emotions in Marathi is capable of attaining superb efficiency in the appropriate classification of the emotions.

REFERENCES

[1] Iain R. Murray and John L. Arnott, " Synthesizing Emotions In Speech: Is It Time To Get Excited?", In *Proceedings of the Fourth International Conference on ICSLP, Philadelphia, USA, Vol. 3, pp. 1816-1819, 1996.*

[2] Jerneja Zganec Gros, Ales Mihelic, Nikola Pavesic, Mario Zganec and Stanislav Gruden, "Slovenian Text-to-Speech Synthesis for Speech User Interfaces", *World Academy of Science, Engineering and Technology, Vol.11, No.1, pp.1-5, 2005.*

[3] M.L. Tomokoyo, W.A. Black and K.A. Lenzo, "Arabic in my hand: small footprint synthesis of Egyptian Arabic," In *Proceedings of the Eurospeech'03, Geneva, Switzerland, pp. 2049-2052, 2003*.

[4] Enrico Zovato, Stefano Sandri, Silvia Quazza and Leonardo Badino, "Prosodic analysis of a multi-style corpus in the perspective of emotional speech synthesis", In *Proceedings of the 8th International Conference on Spoken Language Processing, pp. 1897-1900, 2004*.

[5] Montero, Gutierrez-Arriola, Palazuelos, Enriquez, Aguilera and Pardo, "Emotional Speech Synthesis: From Speech Database to TTS", In *Proceedings of the 5th international conference on spoken language processing, Sidney, pp. 923-926, 1998.*

[6] Ibon Saratxaga, Eva Navas, Inmaculada Hernaez and Iker Luengo, "Designing and Recording an Emotional Speech Database for Corpus Based Synthesis in Basque", In *Proceedings of the fifth international conference on Language Resources and Evaluation (LREC), pp. 2126–2129, 2006.*

[7] Selma Yilmazyildiz, Wesley Mattheyses, Yorgos Patsis and Werner Verhelst, "Expressive Speech Recognition and Synthesis as Enabling Technologies for Affective Robot-Child Communication", In *Proceedings of 7th Pacific Rim Conference on Multimedia, Springer Lecture Notes in Computer Science, Hangzhou, China, Vol. 4261, pp. 1-8, 2006*

[8] Cynthia Breazeal and Lijin Aryananda, "Recognition of Affective Communicative Intent in Robot-Directed Speech", *Journal Autonomous Robots, Vol. 12, No. 1, pp. 83-104, 2002*.

[9] Ignasi Iriondo, Joan Claudi Socoro and Francesc Alias, "Prosody Modelling of Spanish for Expressive Speech Synthesis", *In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Honolulu, HI, pp. 821-824, 2007*

[10] Caroline Henton and Peter Litwinowicz, "Saying and seeing it with feeling: techniques for synthesizing visible, emotional speech", *In Proceedings of the 2nd ESCA/IEEE workshop on Speech Synthesis, pp. 73-76, 1994.*

[11] Enrico Zovato and Jan Romportl, "Speech synthesis and emotions: a compromise between flexibility and believability", In *Proceedings of Fourth International Workshop on Human-Computer Conversation, Bellagio, Italy, 2008*

[12] Klaus R. Scherer, "Vocal communication of emotion: a review of research paradigms", Journal Speech Communication - *Special issue on speech and emotion, Vol. 40, No. 1-2, pp. 227–256, April 2003*

[13] Andy Tams and Mark Tatham, "Intonation for Synthesis of Speaking Styles", Seminar on State-Of-The-Art in Speech Synthesis, London, UK, 2000

[14] Mumtaz Begum, Raja N. Ainon, Zuraidah M. Don and Gerry Knowles, "Adding an Emotions Filter to Malay Text-To-Speech System", In *Proceedings of International Conference on Signal Processing and Communications, Dubai, pp. 1007-1010, November 2007*

[15] Zeynep Inanoglu and Steve Young, "A System for Transforming the Emotion in Speech: Combining Data-Driven Conversion Techniques for Prosody and Voice Quality", In *Proceedings of 8th Annual Conference of the International Speech Communication Association, pp. 490-493, 2007*

[16] Syaheerah L. Lutfi, Raja N. Ainon, Salimah Mokhtar and Zuraidah M. Don, "Adding Emotions to Malay Synthesized Speech Using Diphone-based Templates", *Journal of Information Integration and Web-based Applications and Services, pp. 269-276, 2005*

[17] Al-Dakkak, N. Ghneim, M. Abou Zliekha and S. Al-Moubayed, "Emotion Inclusion In An Arabic Text-To-Speech", *In Proceedings of 13th European Signal Processing Conference, 2005*

[18] Syaheerah L. Lutfi, Raja Noor Ainon, Salimah Mokhtar and Zuraidah Mohd Don, "Template-driven Emotions Generation in Malay Text-to- Speech: A Preliminary Experiment", In Proceedings of *CITA, pp. 144-149, 2005*

[19] Kavita Kasi and Stephen A. Zahorian, "YET ANOTHER ALGORITHM FOR PITCH TRACKING", *The Journal of the Acoustical Society of America, Vol. 123, pp. 4559-4571*

[20] http://en.wikipedia.org/wiki/Sensitivity_and_specificity