Journal of Engineering Science and Technology
Vol. 14, No. 2 (2019) 726 - 746
© School of Engineering, Taylor's University

A NOVEL ALGORITHM FOR SPEECH RECOGNITION USING TONAL FREQUENCY CEPSTRAL COEFFICIENTS BASED ON HUMAN COCHLEA FREQUENCY MAP

HIMGAURI KONDHALKAR*, PRACHI MUKHERJI

Research Scholar, Department of Electronics and Telecommunication Engineering, Sinhgad College of Engineering, S. No. 44/1, Vadgaon Budruk, Pune 411041, India
*Corresponding Author: gouri.ghule@viit.ac.in

Abstract

Extraction of appropriate features from speech is the core of a speech recognition system. Appropriate features can improve the accuracy of perceiving speech. Accurate speech perception is important for implementing applications like voice search, hands free dialling, voice command and control for physically disabled people. Human auditory system performs most accurate and robust perception of sound. Front end of the auditory system, named as cochlea, processes sound with remarkable sensitivity. This paper proposes a mathematical model for cochlear frequency map. This will help to develop more accurate cochlear implant leading to high level of speech recognition for hearing impaired persons. This work primarily focusses on developing a novel algorithm for feature extraction named as Tonal Frequency Cepstral Coefficients based on human cochlea frequency map. The algorithm takes advantage of the relationship between human auditory system and Ohm's acoustic law. This novel feature extraction algorithm outperforms existing feature extraction techniques like Mel Frequency Cepstral Coefficients, Linear Predictive Coding and Gammachirp Frequency Cepstral Coefficients. A hybrid classifier using neural network and fuzzification has also been developed. The classifier recognizes spoken word accurately irrespective of the speaking rate variability. The average accuracy achieved for different datasets is 99.06%, which shows significant improvement over existing algorithms.

Keywords: False positive rate, Gammachirp frequency cepstral coefficients, Mel frequency cepstral coefficients, Human cochlea, Ohm's acoustic law, Speech recognition.

1. Introduction

Isolated word speech recognition in different languages spoken across India is an interesting area of research and has immense use especially for disabled population. For developing a speech recognition system, selection of correct features to be recognizable by any classification model and the selecting right samples for training the model are two major challenges [1]. Speech recognition algorithm starts with the collection of database followed by extracting relevant information known as feature from the recorded words facilitating the subsequent matching phase. Features are extracted from recorded speech using algorithms like Mel Frequency Cepstral Coefficients (MFCC) and Gammatone Frequency Cepstral Coefficients (GFCC). MFCC feature is derived from the cepstral analysis of mel scale filter bank and GFCC feature is obtained from the cepstral analysis of a bank of Gammatone filters [2].

In this research work, we have developed an entirely new feature extraction algorithm based on spiral shape of the cochlea and polar equation of logarithmic spiral. Since sound is perceived by human ear as a set of harmonic tones, we have named the algorithm as Tonal Frequency Cepstral Coefficients (TFCC). Human auditory system helps to perceive and analyse speech irrespective of the background sounds. Cochlea is a spiral shape tube filled with fluid located in inner ear [3]. The two human cochleae are mirror shaped almost symmetrical bony tubes [4]. Spiral shape of the cochlea enhances low frequency sounds used for communication by humans. This paper gives a mathematical model for TFCC. TFCC features are compared with MFCC and GFCC features. Comparative results show that the proposed algorithm enhances the accuracy of speech recognition system considerably over the existing techniques irrespective of gender, age and dialect variations of speakers. This research work can help hearing impaired patients to achieve better speech recognition. It can also be used in variety of applications like controlling instruments, dictation, video games, search engines, authentication, medical reporting, voice biometrics, hands free assistance etc.

The algorithm used for pattern matching plays a vital role in speech recognition. Proposed work also describes a neuro fuzzy combination (NF) classifier at the classification stage. NF logic is used to fuzzify the neural network prediction accuracy into five sets of linguistic labels and to perform a matching phase using a set of fuzzy rules. Fuzzification stage, added after the neural network classifies the speech input to improve overall classification accuracy. Different test beds are used for experimentation. They include spoken Marathi numerals dataset, English numerals dataset and vowels dataset with men, women and kids voices. Spoken Marathi numerals dataset has been recorded by the authors and a copyright is granted for the same.

A number of speech recognition algorithms have been experimented by different authors worldwide. MFCC feature extraction technique and HMM technique for speech recognition is used for isolated word recognition in Punjabi language. Speech corpus has 125 isolated words spoken by 100 speakers with different dialect. A highest classification accuracy 87, 28% is achieved for Majhi dialect female speakers [5]. Ravinder [6] suggests that Punjabi language isolated word recognition can be effectively done with dynamic time warping (DTW) compared to hidden Markov model (HMM). 94% accuracy was obtained with DTW and 91. 3% accuracy was obtained with HMM. The tests were carried out on 500 isolated words dataset. Shanon and Paliwal [7] have concluded that Mel scale and Bark scale filter banks

show equivalent performance in speech recognition. The experimentation is carried out on digit sequences available in TIMIT speech corpus. Performances of MFCC, GFCC and PLP are compared for spoken word recognition. Raheli et al. [8] have experimented automatic speech recognition systems for noisy environment using GFCC feature set. GFCC features gave best recognition accuracy for clean as well as noisy database compared to MFCC features. Gedam et al. [9] have prepared a speech recognition system for Marathi numerals using MFCC feature extraction. Feature matching is done using DTW.

DNN–HMM hybrid model is developed for Chinese-English mix lingual database to promote multilingual research. The dataset had utterances recorded from 1400 speakers. Word error rate reported was 20.09% [10]. Neuro Fuzzy classification approach is used for recognizing sign language for hearing and speech impaired people. The dataset is recorded in Chinese language. Combination of neural network and fuzzy logic resulted in 78% classification accuracy for the system [11]. A Gujarati language speech corpus is prepared with 40 speakers from south Gujarat region. Number of utterances is 650. HMM is used as a classification algorithm. The system gives 87.23% word recognition rate [12]. Dalmiya et al. [13] has discussed development of speech recognition system in Tamil language. The system is developed for mobile applications. MFCC features are classified using DTW template. The developed system is evaluated with 79% accuracy. Comparative study of different feature extraction techniques has been done by Gaikwad et al. One can choose an appropriate technique after comparing its merits and demerits [14].

Debnath and Roy [15] have suggested feature set size reduction by randomly selecting MFCC features using ANOVA and IFS techniques. An average 91.76% accuracy has been achieved for English digit database. Authors have discussed the need of further research for developing new feature as well as new feature selection method for speech recognition. Speaker dependent Arabic isolated word speech recognition system has been developed [16]. Hybrid feature extraction techniques using MFCC, PLP have been used with neural network classifier. PCA is used for dimension reduction. An error rate of 0.68% is achieved but the performance of the system degrades as the number of speakers increase. Authors have also focussed on improving training database for better recognition results. An efficient speech recognition system for arm disabled students has been developed using discrete wavelet transform and modified MFCC algorithm [17]. The system cannot be used for different languages. More number of features need to be added for better performance of the system. A Japanese word recognition system has been developed using two dimensional root cepstrum coefficients [18]. The vocabulary size of the system is limited to 54 words. It can further be extended to incorporate more words. Masood et al. [19] have discussed isolated word recognition using MFCC with some additional features like length of the word, brightness. The experimentation uses neural network classifier. Proposed work is limited to English language, which can be extended for more languages. Deep convolution neural network has been effectively used for Chattisgarh dialect [20]. This isolated word recognition uses Mel frequency spectral coefficients (MFSC) and can be evaluated for continuous speech recognition. Performance of speech recognition is affected by different factors like language, dialect, corpus size and length of the word. Proposed work identifies the gaps in the existing systems and tries to develop a language independent word recognition system.

This paper is organized as follows: methodology developed by authors is described in Section 2, feature extraction using TFCC in Section 3, training stage in Section 4 and classification in Section 5. The experimentation results are discussed in Section 6 and finally conclusion is given in Section 7.

2. Methodology

Proposed work introduces a novel algorithm to generate TFCC feature set that can be extensively used in speech recognition. We have formulated tonal frequency scale which is derived from the spiral shaped human cochlea and Ohm's second acoustical law. Human ear comprises of outer ear, middle ear and inner ear. Outer ear receives sound as pressure wave. Sound wave travels through the middle ear to the base of the cochlea, which is a part of the inner ear. Cochlea is fluid filled tube with basilar membrane running along through its length. Basilar membrane separates the cochlea into two chambers and sound wave passes down the basilar membrane. Changes in the displacement along the membrane are dependent upon frequency of the input sound wave. This is due to hydrodynamics of the cochlear fluid and stiffness of the basilar membrane [2]. Human cochlea is responsible for decoding sound information from frequency domain to spatial domain in a spiral manner. The vibrations in the fluid of cochlea are transformed into neural signals by hair cells situated all along the length of cochlea. These neural signals are passed to brain by auditory nerve. If a person has damaged inner ear hair cells, he needs to undergo cochlear implant. Tonal frequency mathematical model proposed in the next section will help to develop more accurate cochlear implant leading to high level of speech recognition for hearing impaired persons.

Tonal cutoff frequencies

Ohm's acoustical law states that musical sound is perceived by the ear as a set of number of constituent pure harmonic tones. It is also interpreted as pitch corresponding to a certain frequency can be heard only if the acoustic wave contained power at that frequency [21]. Speech signal has frequency component in audio frequency range 20 Hz to 20 kHz [22]. Humans can hear for eleven octaves within this range. Frequency is doubled every octave. The angle of rotation for cochlea associated with this range varies between 0˚ to 990˚ [23]. Cochlea is responsible for hearing over dynamic range. It has 2 and 3 quarter turns corresponding to 990˚. Figure 1 represents cross sectional view of inner ear. It shows the spiral shape of the cochlea. Ohm's law states that cochlea analyses sound by decomposing it into different frequency components [24]. We know that filter banks are used to extract frequency information contained in sound. Hence, proposed TFCC algorithm derives a filter bank named as tonal frequency filter bank that matches frequency map within the cochlea.

Fig. 1. Cross sectional view of left side human cochlea.

Proposed mathematical model formulates an expression for generating tonal sale cutoff frequencies 'f_θ'. The model is constructed under consideration that human hearing range is 20 Hz to 20 kHz. As the anatomy of cochlea resembles logarithmic spiral, the model uses equation of logarithmic spiral [25]. Figure 2 represents human cochlea as logarithmic spiral in polar coordinate system. It shows eleven octaves starting from I to XI within 20 Hz to 20 kHz [3]. 20 Hz frequency is mapped to 0˚. 20 kHz frequency is mapped to 990˚. First octave corresponds to the quadrant between 0˚ to 90˚ whereas eleventh octave corresponds to the quadrant between 900˚ to 990˚. Length of the cochlear curve starting from 0˚ to 990˚ is approximately 32-42 mm [3]. The mathematical model proposed is based on the assumption that change in radius of the cochlea is proportional to change in frequency along the basilar membrane.

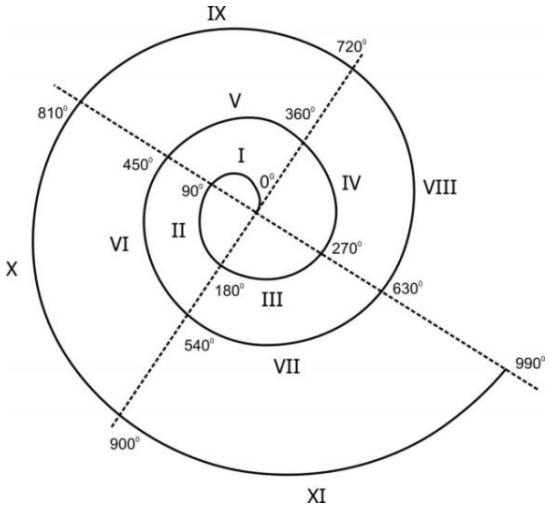


Fig. 2. Polar coordinate map of left side human cochlea.

The equation of logarithmic spiral in polar coordinates is represented by Eq. (1). According to our assumption, we replace radius by frequency in Eq. (2).

$$r = ae^{b\theta}, \ 0^\circ \leq \theta \leq 990^\circ \tag{1}$$

where a, b are constants, r represents radius of cochlea and θ is angle between radius of the cochlea and any point along the cochlear spiral.

$$f_\theta = ae^{b\theta} \tag{2}$$

For human hearing range, 20 Hz frequency corresponds to an angle of 0˚ or 0 radians. 20 kHz frequency corresponds to 990˚ or $11\frac{\pi}{2}$ radians. We place these frequencies in Eq. (2) to get Eq. (3) and Eq. (4).

$$20 = ae^{b(0)} \tag{3}$$

$$20000 = ae^{b\left(\frac{11\pi}{2}\right)} \tag{4}$$

Solving Eq. (3),

$$a = 20 \tag{5}$$

We use the value of constant 'a' obtained in Eq. (5) to solve Eq. (4),

$$20000 = 20e^{b\left(\frac{11\pi}{2}\right)} \tag{6}$$

$$1000 = e^{b(5.5\pi)} \tag{7}$$

$$b = \frac{2\ln(1000)}{11\pi} \tag{8}$$

Final expression for tonal frequencies is obtained as shown in Eq. (9)

$$f_\theta = 20e^{\left(\frac{2\ln(1000)}{11\pi}\right)\theta} \tag{9}$$

Tonal scale cut-off frequencies are generated using above expression for f_θ. In musical scale, range of audible frequencies is divided into octaves. An octave is pitch difference for two notes and frequency gets doubled after each octave. Each octave for musical frequency is divided into 12 notes that are equally spaced on logarithmic scale [26]. In case of cochlear spiral, one octave corresponds to 90˚. Therefore, 90˚ octave is divided into 12 equal angles of 7.5˚ each. To reduce the computational complexity, we have doubled the angle to 15˚. The angle θ is incremented from 0˚ to 990˚ with a lap of 15˚. This generates corresponding cutoff frequencies. Further processing is done using the cutoff frequencies below 8000 Hz, since the sampling frequency is 16000 Hz. This leads to a set of 32 Tonal scale cutoff frequencies used for our experimentation. Figure 3 shows the tonal scale curve with θ in degrees horizontal axis and the cutoff frequency f_θ generated on vertical axis.

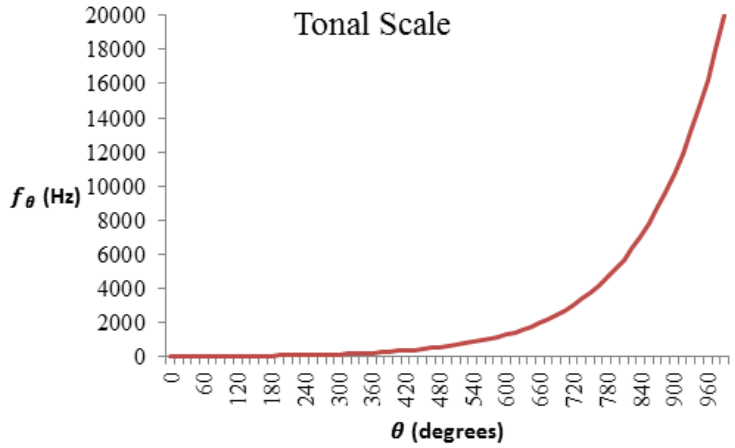


Fig. 3. Tonal scale frequency curve.

3. TFCC Feature Extraction

The performance of speech recognition system depends upon the quality of pre-processing done and the discriminating features extracted. Figure 4 represents the detailed block diagram of feature extraction using TFCC. Spoken word uttered by the speaker is recorded through a microphone. The microphone used for recording

is SM 58 LC SHURE dynamic cardioid professional vocal microphone. The recording is done at sampling frequency 16 kHz in mono mode. The speech signal thus recorded has voiced part, unvoiced part and silence portion. Pre-processing stage is a combination of voiced part detection and must be followed by preemphasis to make the speech signal ready for further processing.

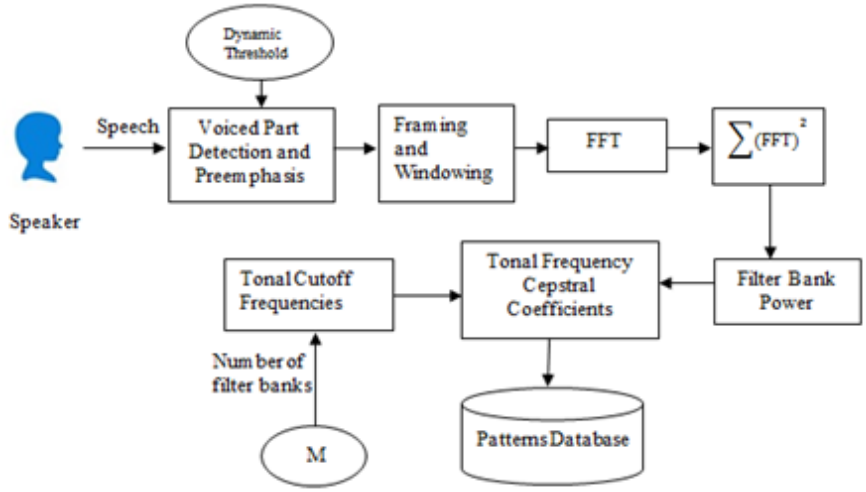


Fig .4. Block diagram of TFCC feature extraction.

Let $x(n)$ be the input speech signal with sampling frequency f_s. Let L be the length of the speech signal as in Eq. (10).

$$x(n) = \{x_0, x_1, x_2, \ldots . x_{L-1}\} \tag{10}$$

For speech processing, speech segment should be stationary. This is achieved normally by dividing it into frames of duration 20 ms to 40 ms [27]. Further processing is done after dividing $x(n)$ into number of frames denoted as F_n ass in Eq. (11). F_L represents frame length and represents the frame overlap. Here value of l is determined by considering frame size of 20 ms. Frame overlap is 50% as it is the most commonly used value. Windowing of the speech signal is done using Eq. (12). It represents Hanning window used to reduce the effect of distortions created by edges during framing.

$$F_n = \frac{(L - F_L)}{F_V} + 1, \; F_L = 0.02 f_s, F_V = 0.01 f_s \tag{11}$$

$$w(i) = 0.5\left(1 - \cos\left(\frac{2\pi i}{F_L - 1}\right)\right), 1 \le i \le F_L \tag{12}$$

Short time energy (STE) per frame is calculated as STE_{F_L} of the voiced part of a speech signal is high. The STE calculations per frame where x_i is speech signal per frame are shown in Eq. (13).

$$STE_{F_L} = \sum_{i=1}^{F_L} (x_i \, w_i)^2 \tag{13}$$

Voiced part of a speech sample can be separated for STE value greater than a threshold value. We have formulated an empirical expression for deciding dynamic threshold λ to separate out the voiced part from the signal. The expression can be effectively used for voiced part separation of a speech signal belonging to any isolated spoken word database. The amount of energy content in the speech signal is taken into account in the formulation. The dynamic threshold calculation is represented by Eq. (14). The calculation is based on the speech signal energy normalized over frame length.

$$\lambda = \frac{\sqrt{\sum_{n=1}^{L}(x_{(n)})^2}}{F_L} \tag{14}$$

Speech signal with STE greater than threshold as in Eq. (15) is considered to be voiced part of the signal, $x_1(n)$ used for further calculations.

$$x_1(n) = x(n), \;\; STE_{F_L} \geq \lambda \tag{15}$$

Preemphasis is applied to boost the amplitude of high frequency component of the speech signal. This helps to keep high frequency components between the samples n and n-1. Equation (16) represents the preemphasis of speech signal $x_1(n)$. The value of α is normally chosen as 0.97. $x_2(n)$ is preemphasized speech signal.

$$x_2(n) = x_1(n) - \alpha(x_1(n-1)) \tag{16}$$

To generate tonal scale triangular filter bank, up-slope and down-slope coefficients for the filter bank are calculated using Eqs. (17) and (18). This generates tonal scale triangular filter bank, $H(M,k)$.

Up-slope coefficients

$$t = f \geq f_c(s) \; AND \; f \leq f_c(s+1)$$

$$H(s,t) = \frac{f(t) - f_c(s)}{f_c(s+1) - f_c(s)} \tag{17}$$

Down-slope coefficients

$$t = f \geq f_c(s+1) \; AND \; f \leq f_c(s+2)$$

$$H(s,t) = \frac{f(s+2) - f_c(t)}{f_c(s+2) - f_c(s+1)} \tag{18}$$

where $1 \leq s \leq M, \quad 1 \leq t \leq K$, M is the number of triangular filter banks. f_c is the set of 30 cutoff frequencies generated through TFCC using Eq. (9). Since the cutoff frequencies are 32, they contribute to generation of 30 triangular filter banks. Hence, value of M used in the algorithm is 30. *K* is half the length of Fourier transform.

Figure 5(a) represents the MFCC triangular filter bank generated. Figure 5(b) represents GFCC filter bank and Fig. 5(c) shows the TFCC filter bank. MFCC filter bank has 8 triangular filters whereas the value is 16 for GFCC. MFCC uses mel scale and GFCC uses equivalent rectangular bandwidth (ERB) scale for filter bank generation.

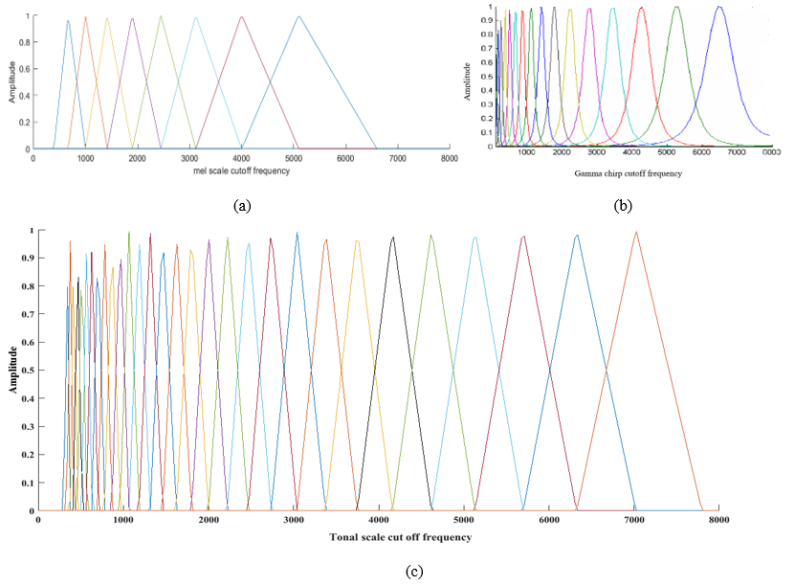


Fig. 5. (a) MFCC filter bank, (b) GFCC filter bank, (c) TFCC filter bank.

Discrete Fourier Transform $F(u)$ with length N is calculated for each frame. Here x_{2frame} is the preemphasized speech signal per frame. Equations (20) and (21) represent magnitude spectrum and power spectrum respectively.

$$F(u) = \sum_{z=0}^{N} x_{2frame}\ (z)\ e^{\frac{-j2\pi zu}{N}} \tag{19}$$

$$|F(u)| = \sqrt{real\ (F(u))^2 + \ imag\ (F(u))^2} \tag{20}$$

$$P(u) = |F(u)|^2 \tag{21}$$

Triangular filter bank $H(M,k)$ is multiplied to power spectrum to generate filter bank power vector fbp as shown in Eq. (22).

$$fbp(M) = \sum_{k=1}^{K} H(M,k)P(k,1) \tag{22}$$

A discrete cosine transform matrix ($dctm$) is created using Eq. (23) for each filter bank with number of cepstral coefficients as one of the dimensions. Number of cepstral coefficients per frame considered is 10 in order to reduce computational complexity. The number is decided with experimentation.

$$dctm(p,m) = w(p) \sum_{m=1}^{M} \cos(\frac{\pi}{2M}(2m-1)(p-1)) \tag{23}$$

$$where \quad w(p) = \begin{cases} \frac{1}{\sqrt{N}}, & p = 1 \\ \sqrt{\frac{2}{N}}, & 2 \le p \le ncep \end{cases}$$

$$ncep = number\ of\ cepstral\ coefficients$$

Filter bank power matrix is applied to discrete cosine transform matrix to get Tonal frequency cepstral coefficients vector, CC as in Eq. (25). This vector is stored in patterns database and it is used for training the neural network in the next speech recognition stage.

$$lfbp = ln(fbp) \tag{24}$$

$$CC(p,j) = \sum_{m=1}^{M} dctm(p,m)\, lfbp(m,j) \tag{25}$$

where $1 \leq j \leq M$

As mentioned above, the number of cepstral coefficients used for TFCC are 10 and the number of cutoff frequencies are 32. MFCC has 10 cepstral coefficients with 10 cutoff frequencies while GFCC has 10 number of cepstral coefficients and cutoff frequencies used is 18. Thus, the number of cepstral coefficients used for all the three feature extraction methods is same.

4. Training

Figure 6 represents the block diagram of training phase of the proposed speech recognition. Tonal frequency cepstral coefficients generated in feature extraction stage are stored in the patterns database. They are used for training artificial neural network (ANN) in the training phase. ANN used in this stage is back propagation neural network. At the output of the training process, we get trained network details, which are further used, for classification in the next section.

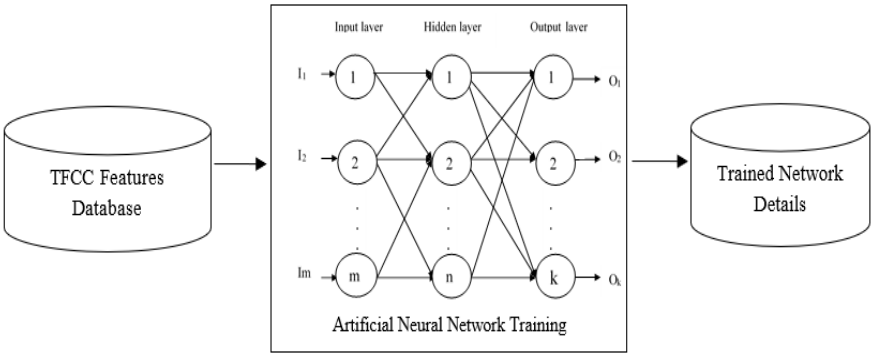


Fig. 6. Block diagram of training phase of speech recognition.

Middle block of the figure represents the architecture of back propagation neural network used for training. Network has inputs, I_1 to I_m where 'm' represents the number of cepstral coefficients generated in the previous section. It has outputs, O_1 to O_k where 'k' represents the number of output classes. The network has input layer with 'm' neurons, hidden layer with 'n' neurons and output layer with 'k' neurons. The values of 'm', 'n' and 'k' depend upon the speech dataset used for training. Mean square error (MSE) obtained after training with TFCC patterns are compared with MSE for existing techniques like MFCC and GFCC patterns. Figure 7 displays the MSE curves resulted after training spoken Marathi numerals dataset. It can be clearly seen that MSE for the proposed TFCC algorithm is minimum as compared with MFCC and GFCC algorithms.

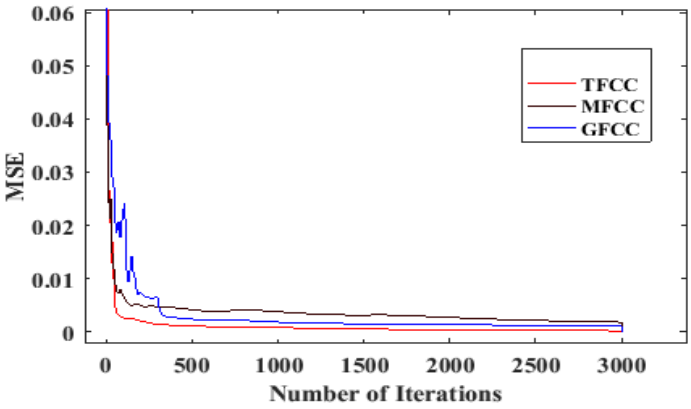


Fig. 7. Plot of MSE achieved after ANN training.

5. Speech recognition using NF classifier

Once the patterns are generated for the desired dataset, classification of the speech sample into correct class is the final stage of the proposed algorithm. Figure 8 displays a unique implementation of NF classifier proposed at this stage. Recognition rate is affected by variability in geometry of vocal organs, pitch, speaking rate among the speakers [28]. A mismatch between any of these factors in training and testing data leads to reduced recognition rate. Proposed work concentrates on speaking rate variation. To overcome the variation effects, we have varied frame length and frame overlap at the feature extraction stage. Newly generated features are used for classification at the output stage implementing NF classifier. This step is practiced only for incorrectly classified samples. This reduces computational overheads and improves the recognition accuracy considerably.

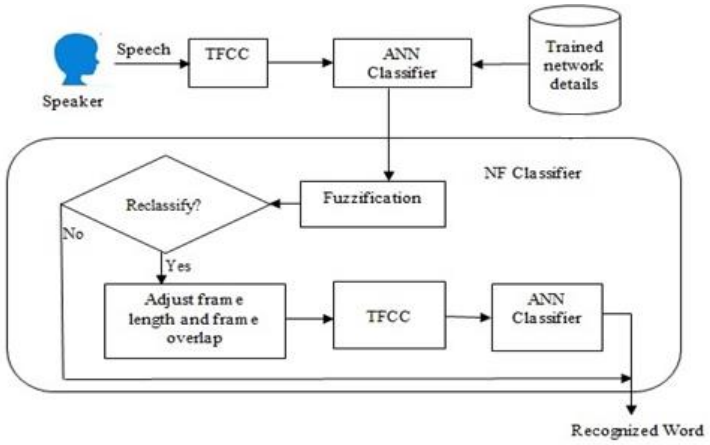


Fig. 8. Block diagram of classification stage of speech recognition process.

NF classifier is a combination of ANN and fuzzy logic. It selects the class with best matching score and generates output, which is closest to the input pattern.

TFCC features are generated for the input speech sample and these features are passed to the back propagation neural network. At the output of the neural network, prediction accuracy for different classes is obtained. As an example, for Spoken Marathi Numerals Dataset, we have ten classes at the output depending on ten digits recorded. The number of output classes will differ for each database described in the previous section. The prediction accuracy values obtained are further passed as input to next block, which is fuzzy inference system. If prediction accuracies produced by the neural network have nearly similar values, correct classification cannot be done. In such case, we need to reclassify the input pattern. Fuzzification is responsible for deciding whether we need to classify the input pattern again to improve recognition accuracy. Fuzzification defines the fuzzy membership function. Figure 9 represents the triangular membership function used in the algorithm. It has degree of membership plotted on vertical scale and prediction accuracy on horizontal scale. Fuzzy membership function maps the degree of membership value of corresponding prediction accuracy to one of the linguistic labels defined. This triangular membership function uses 5 linguistic labels. The linguistic labels are "very poor", "poor", "good", "very good" and "excellent". Linguistic labels are names used to identify membership functions. After designing membership function, fuzzy rules are defined. Following fuzzy rules are defined in the algorithm, considering P_a as prediction accuracy of each output class.

1. **If** Pa_{is} very poor for all classes, **Then** reclassification is necessary.
2. **If** Pa is poor for two or more classes, **Then** reclassification is necessary.
3. **If** Pa_{is} good for two or more classes, **Then** reclassification is necessary.
4. **If** Pa_{is} very good for two or more classes, **Then** reclassification is necessary.
5. **If** Pa_{is} excellent for two or more classes, **Then** reclassification is necessary.

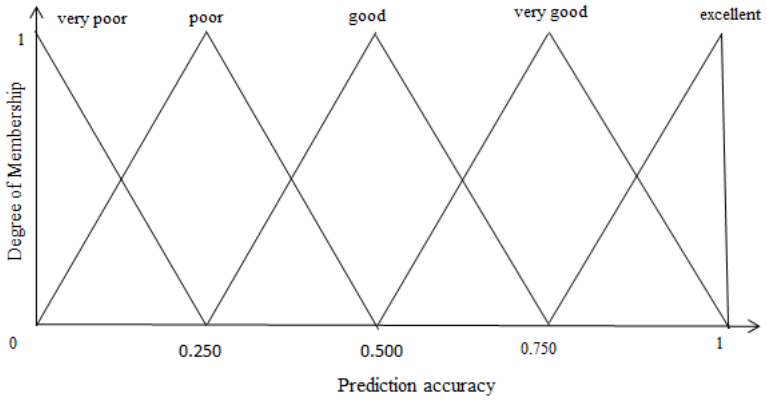


Fig. 9. Membership function.

The fuzzy inference system is limited until fuzzification and defuzzification is not used in proposed system. The reason behind this logic is fuzzification is applied only for deciding whether the neural network output has almost similar prediction accuracies that contribute to recognition error. It also decides whether the input patterns need to be reclassified. This decision is called as fuzzy output. Defuzzification converts fuzzy output to real world quantities, which is not a need of the proposed work. Once reclassification is confirmed, next step is to adjust frame length and overlap.

Speaking rate changes continuously as the speed of speaking is variable depending upon the speaker. Most speech recognition systems use fixed frame size and fixed overlap at the front end processing [29]. In order to overcome the speaking rate variability, frame length and frame overlap have been changed during the feature extraction process. Frame duration is varied between 10ms to 40ms. For 16 kHz sampling frequency, this corresponds to a frame length of 160 to 640 samples. Hence initial frame length used is 160. Figure 10 represents the pseudocode for the process of varying frame length and frame overlap used in the NF classifier. It also explains how the algorithm produces accurately recognized word at the output. The proposed NF classifier improves the recognition accuracy as well as precision significantly over other existing classifiers as explained in next section.

```
Procedure Classification
Input: Trained network details, number of output classes 'outputClasses'
Output: Recognized word
Variable:
Frame length F_L
Frame overlap F_v
Output Class count array Count [outputClasses]
Number of variations in frame length 'm'
Number of variations in frame overlaps 'n'
1: Begin
2:     Initialize F_L = 0
3:     Initialize Count [outputClasses] = 0
4:     for i = 1 to m
5:       Compute new frame length as F_L = F_L + 0.01 f_s
6:        Initial frame overlap F_v = F_L
7:         for j = 1 to n
8:           Compute TFCC feature vector
9:           Classify speech input using ANN classifier
10:             if P_s for an output class at the output of ANN ≥ 95%
11:               Increment Count [outputClasses] for the respective class
12:             end
13:            Compute new frame overlap as F_V = F_V − ( 0.2 * F_L )
14:         end
15:    end
16:    Determine the output class with maximum count in Count[outputClasses]
17:    Recognized word is the word with maximum count output class
18:    Print recognized word
19: end
```

Fig. 10. Algorithm for adjusting frame length and frame overlap for the NF classifier.

6. Results and Discussion

Proposed work uses three different datasets for experimentation. The first dataset is Spoken Marathi Numerals Dataset recorded by the authors [30]. Authors also hold a copyright for the same with registration number SR - 13543/2018. It consists of Marathi numerals where Marathi is the native language of Maharashtra, a state

of India. Closed room with provisions to minimize reverberation of sound was used for recording. The other datasets are vowels dataset and free spoken digits dataset. These datasets are available online for downloading. Vowels dataset is available on the website of Western Michigan University by Hillenbrand et al. [31]. Free spoken digits data set (DOI: 10.5281/zenodo.1136198) is an audio dataset of spoken English digits [32]. Details of the three datasets are given in Table 1. The performance of proposed TFCC feature set is evaluated for these three datasets. NF classifier is used for classification. For comparative analysis, the performance is evaluated based on accuracy, precision, sensitivity, specificity and false positive rate (FPR). FPR is the probability of a spoken word being misclassified.

Table 1. Details of the datasets used in training and recognition stages.

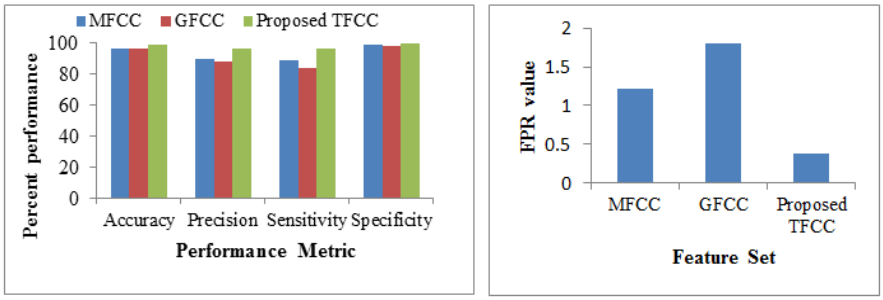
Parameter	Spoken Marathi Numerals Dataset	Vowels Dataset	Free Spoken Digits Dataset
Author	Kondhalkar [30]	Hillenbrand et al. [31]	Jakobovski et al. [32]
Number of samples	7500	1668	2000
Number of Speakers	Male: 61 Female: 39 Total : 75	Male : 45 Female : 48 Kids : 46 Total: 139	Male : 3 Female: 1 Total : 4
File format	.wav file	.wav file	.wav file
Sampling frequency	16 kHz	16 kHz	8 kHz
Vocabulary size	Marathi language numerals “Shunya” to “Nau” (0 to 9)	English language vowels ae, ah, aw, eh, ei, er, ih, iy, oa, oo, uh, uw	English language digits 0 to 9
Utterances per class	10	1	50
Output Classes	10	12	10
Training samples	50%	55%	80%
Testing samples	50%	45%	20%

Table 2 shows the results of proposed TFCC feature extraction and NF classifier algorithm for Spoken Marathi Numerals Dataset. ANN is trained for this feature set with number of hidden layer neurons as 27 and output layer neurons as 10 based on 10 classes to be produced at the output of classifier. Number of neurons for these two layers would change as per the dataset used for training. Results obtained are compared with existing feature extraction techniques like MFCC and GFCC. Figure 11(a) displays the plot for the same results and Fig. 11(b) represents the FPR value for the three techniques. It is confirmed from the results that TFCC gives highest accuracy value 99.30%.

Table 3 represents the results for Vowels dataset. Vowels dataset is a unique dataset. 139 speakers have recorded vowels where each speaker has uttered the twelve vowels only once. Thus, the training and testing datasets have completely different sets of utterances. There is no overlap between the two. As shown in Fig. 12, TFCC feature extraction generates higher accuracy than most frequently used MFCC technique for vowels dataset whereas FPR is lowest for TFCC.

Table 2. Spoken Marathi Numerals Dataset results.

	MFCC	GFCC	Proposed TFCC
Accuracy (%)	96.79	96.75	99.30
Precision (%)	89.60	87.84	96.54
Sensitivity (%)	88.98	83.75	96.53
Specificity (%)	98.77	98.19	99.61

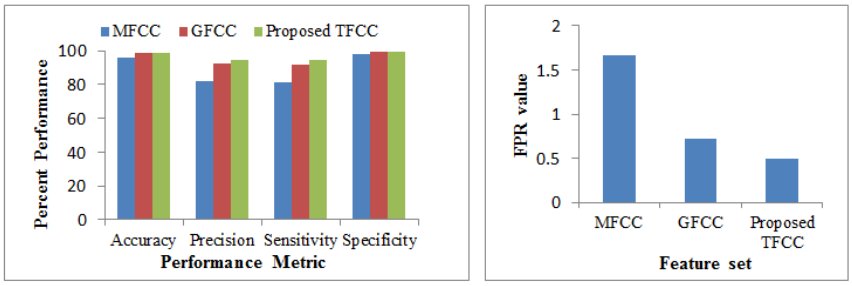


(a) **(b)**

Fig. 11. (a) Plot for the performance metrics (Spoken Marathi Numerals Dataset), (b) FPR plot.

Table 3. Vowels Dataset results

	MFCC	GFCC	Proposed TFCC
Accuracy (%)	95.94	98.68	99.09
Precision (%)	82.04	92.31	94.66
Sensitivity (%)	81.66	92.08	94.58
Specificity (%)	98.33	99.28	99.50



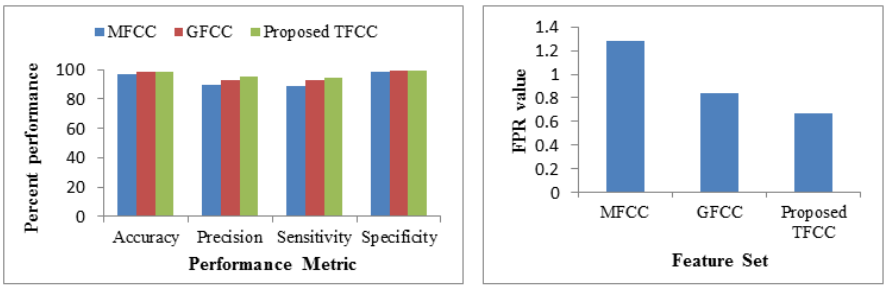
(a) **(b)**

Fig. 12. (a) Plot for the performance metrics (Vowels Dataset), (b) FPR plot.

Table 4 represents the results for Free Spoken Digits Dataset. Figure 13(a) shows graphical representation for the same results and Fig. 13(b) represents the FPR plot. It can be clearly observed that TFCC performs better for all the four parameters compared to the remaining two techniques.

Table 4. Free Spoken Digits Dataset results

	MFCC	GFCC	Proposed TFCC
Accuracy (%)	96.70	98.50	98.80
Precision (%)	89.36	92.65	94.84
Sensitivity (%)	88.50	92.50	94.00
Specificity (%)	98.72	99.16	99.33



(a) **(b)**

Fig. 13. (a) Plot for the performance metrics (Free Spoken Digits Dataset), (b) FPR plot.

These results prove that proposed TFCC feature set gives improved accuracy results for any isolated word database. Precision of TFCC feature set for all the different datasets shows a major increase over other techniques. This indicates that even for repeated testing, TFCC has stable classification results. Higher sensitivity is an indication that TFCC can correctly classify maximum speech samples into their desired class. Specificity gives us an idea that TFCC can correctly reject misclassified spoken words at the classifier. TFCC has lowest value of FPR for all the datasets. This confirms that it has lowest probability of misclassifying the spoken word.

To confirm classification consistency of the proposed NF classifier, we have classified the TFCC feature set with classifiers other than NF classifier. Table 5 shows the accuracy of different classifiers for TFCC feature set. Beside NF, we have used K-nearest neighborhood (KNN) and Support Vector Machine (SVM) classifier. It can be concluded that NF classifier performs better than existing classifiers.

Table 5. Accuracy for different classifiers.

	KNN	SVM	NF
Spoken Marathi Numerals Dataset	96.95%	98.32%	99.30%
Free Spoken Digits Dataset	97.80%	98.00%	98.80%
Vowels Dataset	96.32%	98.16%	99.09%

Table 6 shows comparative accuracy results for the different datasets used. Each row in the table represents accuracy percentage for the algorithm experimented by the corresponding author and accuracy for proposed TFCC algorithm. For all the datasets, TFCC shows considerable rise in the accuracy value. Vowels dataset [31] is generated by Hillenbrand and he has used LPC technique for speech recognition. Hence, for vowels dataset, proposed TFCC algorithm accuracy is compared with LPC technique. A number of isolated word recognition algorithms for different languages are experimented by different authors. Table 7 discusses such algorithms and corresponding accuracies achieved. Proposed TFCC and NF classifier algorithm has achieved an average accuracy of 99.06% which is highest compared with these algorithms.

Table 6. Comparative accuracy.

Dataset	Algorithm	Recognition accuracy
Spoken Marathi Numerals Dataset [30]	MFCC	96.79%
	Proposed TFCC	99.30%
Vowels Dataset [31]	LPC	95.40%
	Proposed TFCC	99.09%
Free Spoken Digits Dataset [32]	Clustering	97.80%
	Proposed TFCC	98.80%

Table 7. Recognition accuracy using existing feature extraction algorithms experimented by respective authors.

Reference Number	Methodology used	Description of database	Recognition Accuracy
[33]	Marathi numeral recognition using MFCC and LPC	100 speakers. 5000 utterances. 10 Marathi digits.	MFCC Highest- 78.9% Lowest- 13.25% LPC Highest- 66.17% Lowest- 12.32%
[34]	Kannada isolated word recognition using MFCC, LPCC feature extraction, Recognition using HTK	14 speakers 4480 utterances. 10 Kannada digits.	MFCC=90% LPCC=70%
[35]	MFCC and DTW	Spoken Arabic digits dataset. 13 utterances of each digit. 10 Arabic digits	Maximum accuracy observed was 83%
[36]	Sphinx 4 Tool	Isolated Marathi words database , 50 speakers, 2500 utterances for 10 Marathi words	60%

7. Conclusions

Proposed work provides an effective implementation of a novel speech recognition system. A unique TFCC feature extraction algorithm based on spiral shape of human cochlea and equation of logarithmic spiral has been proposed. Some concluding observations from the investigation are given below.

- TFCC algorithm has produced highest accuracy and precision for different datasets as compared to MFCC and GFCC feature extraction techniques with comparable computational complexity.
- The algorithm showed consistent performance for overlapping as well as non-overlapping databases.
- FPR for the proposed feature extraction algorithm is lowest as compared to the other techniques.
- A novel NF classifier is implemented to address the problem of speaking rate variation by varying frame length and frame overlap. This leads to a recognition rate better than KNN and SVM classifiers.
- Proposed speech recognition system with TFCC feature extraction algorithm and NF classifier produces an average accuracy 99.06% and average precision 95.34% for different datasets. Both the values are highest compared to other speech recognition systems practiced by different authors.
- Proposed TFCC model can be used for designing cochlear implants for hearing impaired patients to achieve better speech recognition. The system can also be used for applications like hands free dialling, speech activated mobile applications, automated teller machines and search engines.
- As a part of future studies, performance of the proposed TFCC algorithm can be evaluated using 'Deep learning' techniques.

Nomenclatures

CC	Tonal frequency cepstral coefficients feature vector
$dctm$	Discrete cosine transform matrix
fbp	Filter bank power
F_L	Frame length
F_n	Number of frames
f_s	Sampling frequency, Hz.
F_V	Frame overlap
f_θ	Tonal scale cutoff frequency, Hz.
M	Number of triangular filter banks
$ncep$	Number of cepstral coefficients
r	Radius of Human Cochlea, mm.
STE_{FL}	Short time energy per frame

Greek Symbols

λ	Dynamic threshold to separate voiced part of speech signal
θ	Angle between radius of the cochlea and any point along the cochlear spiral, deg.
ANN	Artificial neural network
DNN	Deep neural network

DTW	Dynamic time warping
FPR	False positive rate
GFCC	Gammachirp frequency cepstral coefficients
HMM	Hidden Markov model
KNN	k- nearest neighbour classifier
MFCC	Mel frequency cepstral coefficients
MSE	Mean square error
NF	Neuro fuzzy classifier
PLP	Perceptual linear predictive technique
SVM	Support vector machine
TFCC	Tonal frequency cepstral coefficients

References

1. Haridas, A.; Ramalatha, M.; and Sivakumar, V. (2018). A critical review and analysis on techniques of speech recognition: The road ahead. *International Journal of Knowledge-based and Intelligent Engineering Systems*, 22(1), 39-57.
2. Shao, Y.; Jin, Z.; Wang, D.; and Srinivasan S. (2009). An auditory-based feature for robust speech recognition. *Proceedings of the IEEE International Conference of Acoustics, Speech and Signal Processing*. Taipei, 4625-4628.
3. Park, A. (2003). Using the gammachirp filter for auditory analysis of speech. *Proceedings of the 18.327: Wavelets and Filterbanks*, 1-18.
4. Rask-Andersen, H.; Liu, W.; Erixon, E.; Kinnefors, A.; Pfaller, K.; Schrott-Fischer; A.; and Glueckert, R. (2012). Human cochlea: anatomical characteristics and their relevance for cochlear implantation. *The Anatomical Record: Advances in Integrative Anatomy and Evolutionary Biology*, 295(11), 1791-1811.
5. Arora, S. (2018). Dialectal variations of isolated word recognition. *Proceedings of the Eighth International Conference on Advanced Communication and Computations.* Barcelona, Spain, 38-44.
6. Ravinder K. (2010). Comparison of HMM and DTW for isolated word recognition system of Punjabi language. *Proceedings of the Bloch I., Cesar R.M. (eds) Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications. Lecture Notes in Computer Science*. 6419, Berlin, Heidelberg, Springer, 244-252.
7. Shanon, B.J.; and Paliwal, K.K. (2003). A Comparative Study of Filter Bank Spacing for Speech Recognition. *Proceedings of Microelectronic Engineering Research Conference*. Brisbane, 1-3.
8. Rahali, H.; Hajaiej, Z.; and Ellouze, N. (2014). Asr systems in noisy environment: Auditory features based on gammachirp filter using the AURORA database. *Proceedings of 22nd European Signal Processing Conference (EUSIPCO).* Lisbon, 696-700.
9. Gedam, Y.K.; Magare, S.S.; Dabhade, A.C.; and Deshmukh, R.R. (2014). Development of automatic speech recognition of Marathi numerals. *International Journal of Engineering and Innovative Technology*, 3(9), 198-203.
10. Wang, D.; Tang, Z.; Tang, D.; and Chen, Q. (2016). A Chinese-English mix lingual database and a speech recognition baseline. *Proceedings of the Conference of the Oriental Chapter of International Committee for*

Coordination and Standardization of Speech Databases and Assessment Technique. Bali, 84-88.

11. Li, W.; Hu, X.; Gravina, R.; and Fortino, G. (2017). A Neuro-Fuzzy fatigue tracking and classification system for wheelchair users. *IEEE Access*, 5, 19420-19431.
12. Tailor, J.H.; and Shah, D.B. (2018). HMM based lightweight speech recognition system for Gujarati language. *Proceedings of the Information and Communication Technology for Sustainable Development*. Singapore, Springer, 451-461.
13. Dalmiya, C.P.; Dharun. V.S.; and Rajesh, K.P. (2013). An efficient method for Tamil speech recognition using MFCC and DTW mobile applications. *Proceedings of the IEEE Conference on Information and Communication Technologies*. Jeju Island, 1263-1268.
14. Gaikwad, S.; Gawali, B.; and Yannawar, P. (2010). A review on speech recognition technique. *International Journal on Computer Applications*, 10(3), 16-24.
15. Debnath, S.; and Roy, P. (2018). Speaker independent isolated word recognition based on ANOVA and IFS. *Proceedings of the 10th International Conference on Computer Modeling and Simulation*. Sydney, Australia, 92-97.
16. Boussaid, L.; and Hassine, M. (2018). Arabic isolated word recognition system using hybrid feature extraction techniques and neural network. *International Journal of Speech Technology*, Springer, 21(1), 29-37.
17. Darabkh, K.; Haddad, L.; Sweden, S.; and Hawa, M. (2017). An efficient speech recognition system for arm disabled students based on isolated words. *Computer Applications in Engineering Education*, 26(2), 285-301.
18. Srinivas, N.; Sugan, N.; Kumar, L.; Nath, M.; and Kanhe, A. (2018). Speaker independent Japanese isolated speech word recognition using TDRC features. *Proceedings of the International Conference on Control, Communication and Computing*. IEEE, Tiruvanantpuram, India, 278-283.
19. Masood, S.; Mehta, M.; Namrata; and Rizwi, D. (2015). Isolated word recognition using neural network. *Proceedings of the Annual IEEE India Conference*. India, 1-5.
20. Londhe, N.; Kshirsagar, G.; and Tekchandani, H. (2018). Deep convolution neural network based speech recognition for Chattisgarhi. *Proceedings of the International Conference on Signal Processing and Integrated Networks*. IEEE, Noida, India, 667-671.
21. Rossing D.; Moore R.; and Wheeler P. (2014). *The Science of Sound*. (3rd ed.) : Pearson.
22. Arul, V.; and Marimuthu, R. (2014). A study on speech recognition technology. *Journal of Computing Technologies*, 3(7), 4-7.
23. Kapit, W.; Macey R.; and Meisami E. (1999). *The Physiology Coloring Book*, (2nd ed.): Pearson.
24. Zweig, G. (1989). Auditory speech preprocessors. *Proceedings of the Workshop on Speech and Natural Language*. Philadelphia, Pennsylvania, 230-235.
25. Greenwood, D.D. (1990). A cochlear frequency-position function for several species--29 years later. *The Journal of the Acoustical Society of America*, 87(6), 2592-605.

26. McDermott, J.H.; and Oxenham, A.J. (2008). Music perception, pitch, and the auditory system. *Current opinion in neurobiology*, 18(4), 452-463.
27. Stakhovskaya, O.; Sridhar, D.; Bonham, B.H.; and Leake, P.A. (2017). Frequency map for the human cochlear spiral ganglion: Implications for cochlear implants. *Journal of the Association for Research in Otolaryngology*, 8(2), 220-233.
28. Shahnawazuddin, S.; Singh, C.; Kathania, H.K.; Waquar Ahmad, W.; and Pradhan, G. (2018). An experimental study on the significance of variable frame length and overlap in the context of children's speech recognition. *Journal of Circuits System Signal Process*, 37(12), 5540-5553.
29. Chu, S.M.; and Povey, D. (2010). Speaking rate adaptation using continuous frame rate normalization. *Proceedings of the Acoustics Speech and Signal Processing (ICASSP), IEEE International Conference*. Texas, 4306-4309.
30. SR-13543/2018: (2018) Spoken Marathi Numerals Dataset.
31. Hillenbrand, J.; Getty, L.A.; Clark, M.J.; and Wheeler, K. (1995). Acoustic characteristics of American English vowels. *The Journal of the Acoustical Society of America*, 97(5), 3099-3111.
32. Jakobovski. A decoupled, generative, unsupervised, multimodal architecture for real-world agents. Retrieved August 16, 2017, from https://github.com/Jakobovski/decoupled-multimodal-learning.
33. Shrishrimal, P.P.; Deshmukh, R.R.; Janvale, G.B.; and Kulkarni, D.S. (2017). Marathi digit recognition system based on MFCC and LPC. *International Journal of Advanced and Innovative Research*, 6(6), 37-39.
34. Sneha, V.; Hardhika, G.; JeevaPriya, K.; and Gupta, D. (2018). Isolated Kannada speech recognition using HTK-A detailed approach. *Proceedings of the Process in Advanced Computing and Intelligent Engineering. Advances in Intelligent Systems and Computing*. 564, Singapore, Springer, 185-194.
35. Ganoun, A.; and Almerhag, I. (2012). Performance analysis of spoken Arabic digits recognition techniques. *Journal of Electronic, Science and Technology,* 10(2), 153-157.
36. Pratik, K. (2017). Design and development of word recognition for Marathi language. *Imperial Journal of Interdisciplinary Research*, 3(6), 30-32.