International Journal of Signal Processing, Image Processing and Pattern Recognition
Vol. 10, No. 10 (2017), pp.1-16
http//dx.doi.org/10.14257/ijsip.2017.10.10.01

Speaker Recognition System Using Symbolic Modelling of Voiceprint

Shanmukhappa A. Angadi[1] and Sanjeevakumar M. Hatture[2*]

[1]Department of Computer Science and Engineering, Centre for PG Studies, Visvesvaraya Technological University, Belagavi-590018, Karnataka State, India
[2] Department of Computer Science and Engineering, Basaveshwar Engineering College, Bagalkot- 587103, Karnataka State, India*
[1]vinay_angadi@yahoo.com, [2]smhatture@yahoo.com*

Abstract

Voice biometric trait is used in speaker recognition system due to its combined behavioral and physiological characteristics. This paper presents a symbolic inference system for text-dependent speaker recognition system by exploring the physiological characteristics embedded in the user utterance. These characteristics also capture the user behaviour. The symbolic data object is constructed using different voiceprint features namely the inter-lexical pause position, complementary spectral features such as spectral entropy, spectral centroid and spectral flatness, pitch, loudness and formants. These features are explored in this work as inter-lexical pause position provides the articulation capability of user vocal tract. The spectral characteristics model the functional properties of the human ear and loudness feature provides the strength of ear's perception. The relation between physical and perceptual properties of sound is estimated through pitch whereas formants provide the acoustic reverberation of the human vocal tract. The variability in features of user/speaker utterance of words is represented with symbolic data. The speaker identification is performed using span, content and position symbolic similarity measures [6], modified for the current work. The proposed method is evaluated on 100 users of voice corpus of VTU-BEC-DB multimodal biometric database. The experimental results demonstrate an overall identification rate of 90.56%. Experimental results show that the symbolic data representation of voice features provides better speaker recognition.

Keywords*: Voice Biometric, Symbolic object, Symbolic similarity measure, complementary spectral features, Speaker identification*

1. Introduction

Voice is a biometric trait which exhibits combined behavioral and physiological characteristics used to alleviate the problem of spoof attack in biometric system. Human beings are able to recognize a person just by hearing him or her talk. So, few seconds of speech is sufficient to identify a familiar voice. Voice of an user is unique as the knowledge/codes used in the utterances are user specific. The physical differences in users' voiceprints are characterized by measuring the amplitude, frequency, duration and spectral distribution. The information embedded in the voice signal is extracted at six different levels such as spectral, prosodic, phonetic, idiolectal(*i.e.* syntactical), dialogic and semantic [1].

The functional properties of the human ear are mimicked with spectral characteristics. Different users will have different spectra (location and magnitude of peaks) for similar

Received (May 19, 2017), Review Result (August 14, 2017), Accepted (August 23, 2017)
* Corresponding Author

ISSN: 2005-4254 IJSIP

sounds [2]. Spectral analysis will measure the amount of acoustic energy present at different frequencies in a sound [3]. Prosodic features are a measure of accent, intonation and stress. The Prosodic features are estimated by calculating the pitch, energy, and duration information from the user's voice [4]. The idiolectal (*i.e.* syntactical) features are the measure of the way of using the word utterance *i.e.* repetition of the user's "favourite" words. The dialogic features extract the conversational patterns of a speaker. The semantics, pronunciation, diction and idiosyncrasy are the learned traits associated to education, socioeconomic status and birth place of a user/speaker and are also used for speaker recognition, but are difficult to extract [5].

Speaker recognition systems can be classified into text-dependent and text-independent, based on the text used in the testing phase. Text-dependent systems are further divided into fixed-phrase and prompted-phrase systems. Fixed-phrase systems are trained on the phrase that is also used for testing. Prompted-phrase systems ask the user/claimant to utter a word sequence (phoneme sequence) not used in the training phase or in previous tests. Further in text-independent systems, the speech used for testing is unconstrained. In voice biometrics the speaker voiceprint may vary due to variations in the health, environmental conditions and additive noise. The features extracted in such situations form the speaker voiceprints are varying in nature. The symbolic object representation is employed to represent such variability in the features of voiceprints during speaker recognition in a robust manner.

Symbolic objects are extensions of classical data types. The real world objects are better described with symbolic objects [6]. The feature extracted from the real world objects are usually represented by complex data. The knowledge embedded in the complex data is easily extracted by representing them into symbolic data structure. Symbolic data appears in the form of continuous ratio, discrete, absolute, interval, probability distributions, random variables and multi-valued data. In pattern recognition, the variability inside classes of individuals is easily expressed by symbolic data. Symbolic objects offer a better alternative for organizing and summarizing abstract data. Symbolic objects are of three different types, assertion object, hoard object and synthetic objects [7]. An assertion object is a conjunction of events pertaining to a given object. An event is a pair which links feature variables and feature values. A hoard object is a collection of one or more assertion objects, whereas a synthetic object is a collection of one or more hoard objects [6, 7]. In this work voiceprints are represented as assertion symbolic object. This representation of the symbolic object accommodates the variability in features of speaker voiceprint and is one of the novel contributions of the proposed speaker recognition system.

In the proposed work, the text-dependent speaker recognition system is presented, in which the speaker utterance is represented as symbolic object. The object will cover the features of the speaker voice utterance such as inter-lexical pause position, complementary spectral features such as spectral entropy, spectral centroid and spectral flatness, pitch, loudness and formant frequencies. These features are employed in this work as inter-lexical pause position provides the articulation capability of user vocal tracts. The spectral characteristics model the functional properties of the human ear and loudness feature provides the strength of the human ear perception. The relation between physical and perceptual properties of sound is estimated through pitch and formants that provide the acoustic reverberation of the human vocal tract. The intra-speaker variations in the features are captured in a symbolic data structure. This representation of speaker utterance into symbolic objects is a novel technique used by the proposed system. The symbolic knowledge bases for the phrases of English number utterance namely "Twenty One (21)" to "Twenty Nine (29)" are constructed separately for 100 users. Further, speaker identification is performed using span, content and position symbolic similarity measures adopted from [6,7]. The experimentation is performed on voice corpus of VTU-BEC-DB multimodal biometric database. The experimental results show that the proposed

method offers a overall correct identification rate of 90.56% for user recognition using voice biometric trait.

The rest of the paper is organized as follows: section 2 presents the recent developments in text-dependent speaker recognition approaches. Section 3 describes the proposed model of user identification using voice symbolic objects. The experimental results and analysis are provided in section 4. Finally, section 5 concludes the work and enlists the future directions.

2. Review of Related Work

Voice is versatile, simple to use, non-intrusive and has high user acceptance as a biometric trait. The voice-tract and accent characteristics are difficult to duplicate even if obtained from a recorded voice. Speaker recognition systems can be classified into text-dependent and text-independent, based on the text used in the testing phase. Some of the works related to the voice biometric are described in the following. For text-dependent applications, whole phrases or phonemes may be modeled using multistate left-to-right hidden Morkov models [8]. Neuro-Genetic Hybrid algorithm with cepstral based features have been used to improve the performance of the text dependent speaker identification system under noisy environment [9]. The mel-frequency cepstral coefficients (MFCC) feature has been used for designing a text dependent speaker identification system. The extracted acoustic features (MFCC) of a speaker are quantized to a 23 number of centroids using vector quantization algorithm [10]. The log Mel spectrum is converted into time domain using discrete cosine transform (DCT) coefficients referred as acoustic vectors [11]. The use of discrete wavelet transformation (DWT) in multi-level decomposition helps to represent the meaningful features of the human voice, in low size coefficients [12]. Also Linear Prediction Coefficients (LPC), LPC-Cepstral (LPCC) and perceptual linear prediction (PLP) coefficients have been used to extract the features [13]. The gammatone-filter-based method for auditory feature extraction algorithm for robust speaker identification is presented in [14]. The features are tested under white noise, babble noise, tank noise and F-16 cockpit noise mismatched conditions, at different SNR levels. A joint factor analysis (JFA) for text-dependent speaker recognition with random digit strings is presented in [15].

The logistic regression method and extraction of phonetically aware Baum–Welch statistics are employed for speaker recognition. An utterance-level speaker representation is performed with attention network by combining the frame-level features extracted with convolution neural network in [16]. The use of i-vector/ Probabilistic linear discriminant analysis with hidden Markov model (HMM) for text-dependent single utterance task is presented in [17]. The verification performance of 1.22% equal error rate (EER) is achieved. A lexicon-based local representation algorithm for text-dependent i-vector speaker verification system is presented in [18].The speaker recognition system based on Gaussian mixer model-based support vector machine (GMM-SVM) and the nuisance attribute projection (NAP) technique for channel compensation is presented in [19]. Time alignment of different utterances is a serious problem for distance measures and small shift would lead to incorrect identification in text-dependent speaker recognition. In such cases time alignment can be done with Dynamic time warping (DTW) algorithm.

The different key challenges in voice biometrics are robustness, portability, adaptation, language modeling, confidence measures, out-of-vocabulary words, spontaneous speech, prosody and modeling dynamics. In order to develop a robust speaker recognition system there is a need of accurate and efficient representation techniques for recognizing voice biometric patterns. Hence, there is scope to develop an efficient representation technique for voiceprint features. In the proposed speaker recognition system the speaker utterance is represented as symbolic object to express variability in the features of voiceprints. The detailed description of the proposed methodology is given in the next sections.

3. Proposed Model for Speaker Recognition Using Symbolic Modelling

In this work a new model for text-dependent speaker recognition using symbolic object modelling and similarity is proposed. The steps involved in the speaker recognition using proposed method are, voice signal pre-processing *i.e.* framing, windowing and voice activity detection, feature extraction, symbolic object representation, construction of the knowledge base and finally recognition with the help of symbolic similarity measure. The overall methodology is depicted in Figure 1.

Firstly, the voiceprints are pre-processed for detecting and eliminating the silence portion in user utterance. In pre-processing the input user utterance is segmented with framing and windowing technique. Further, the silence portion is detected and eliminated by analyzing the segmented frames by voice (speech) activity detection *i.e.* speech detection technique. To characterize the speaker, features are extracted from the silence removed signal. The inter-lexical pause position *i.e.* start and end points of pause, is located by tracing the amplitude spectrum of the silence removed signal. The complementary spectral features *i.e.* spectral entropy, spectral centroid and spectral flatness features, pitch and formant feature are computed from the power spectrum of the silence removed signal. The loudness feature is computed from the silence removed signal. The utterance of the keywords "Twenty One (21)" to "Twenty Nine (29)" of English language from 100 users of VTU-BEC-DB voice corpus is considered for experimentation. The extracted features are used to construct symbolic data structure. Further, the symbolic knowledge base is constructed for 100 users by selecting five utterances of each English number from every user. Finally, the identification of the user is performed with the help of span, content and position symbolic similarities. The detailed description of each module of the proposed system is presented in the following subsections. The VTU-BEC-DB voice corpus is described in the following subsection.

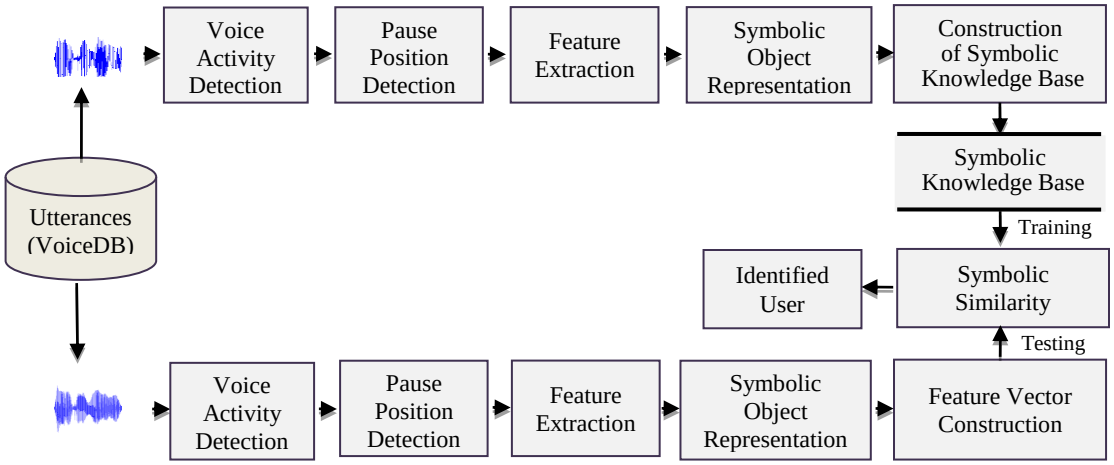

Figure 1. Speaker Recognition System Using Symbolic Modeling of Voice Features

3.1. Description of VTU-BEC-DB Voice Corpus

The proposed speaker recognition system uses voice utterances from VTU-BEC-DB Voice Corpus. The database is collected at the, Department of Computer Science and Engineering, Basaveshwar Engineering College, Bagalkot affiliated to Visvesvaraya Technological University, Belagavi, Karnataka, India. The voice corpus contains total of 31000 number utterance of Zero(0) to Thirty(30) collected separately in English and Kannada languages from the 100 persons (*i.e.* 36 males and 64 females) with age range

between 18 and 50 years collected in two sessions over a period of one year. The voice samples are collected using Sony ICD- UX533F digital Voice Recorder. The voice samples are recorded with sampling frequency of 44.1KHz stereo in Linear PCM (LPCM) format with bit rate of 16 bit wave files. The Microphone Sensitivity is set to Medium and the NCF(Noise Cut) recording filter is used. Further, the silence removed signal is obtained by processing the recorder voiceprint with steps described in next subsection.

3.2. Voice Detection and Silence Portion Elimination

During the acquisition of voiceprint sometimes the silence portion will occur due to delayed utterance of the speaker or due to improper handling of the acquisition device. The first step in audio signal processing and speaker recognition is to eliminate the unwanted/ silence information from the voiceprints. In the proposed work, to extract the stationary information of the input user voiceprint, it is partitioned into 'N' number of frames with each frame of 20 msec length, by using framing and windowing technique. To avoid the effect of discontinuities among the frames the 'hamming windowing' is used to align the edges of frames. The shorter size of the frame (*i.e.* 20 msec) is selected since the short-time magnitude spectrum contain most of the information about the voiceprint. The presence of silence portion in the frames is detected with the help of voice activity detection *i.e.* speech activity detection or speech detection technique[20]. In voice activity detection the entire voiceprint is considered as one-dimensional Gaussian distribution and its statistical properties (*i.e.* mean and standard deviation) are computed. Further the volume associated with i^{th} frame is calculated by measuring the one-dimensional 'Mahalanobis distance' from i^{th} frame to the mean of the entire voiceprint and dividing by its standard deviation using equation (1)

$$\underset{1\le i\le N}{V_i(n)} = \left(\sum_{k=1}^{n} |F_i(k) - \left(\sum_{i=1}^{N}\sum_{k=1}^{n} F_i(k)\right)/N|\right)/\sigma \quad (1)$$

where F(i) is the i^{th} frame, N = Total number of frames (N= Duration of Utterance / Duration of the Frame), and σ is standard deviation of entire voice signal. The frame is identified as either voiced and unvoiced frames based on the volume-threshold as given in equation (2)

$$V(i) = \begin{cases} \text{Voiced} & \text{if } V(i) \ge \text{volume-threshold} \quad 1 \le i \le N \\ \text{UnVoiced} & \text{Otherwise} \end{cases} \quad (2)$$

Based on the experimentation the volume-threshold is taken as 5 mv. Finally all the voiced frames are combined to obtain the silence removed signal, SR(n). Further, the silence removed signal is used to extract the features as described in next subsection.

3.3. Pause Position Detection and Feature Extraction

The pauses in user utterance will occur due to weak respiration, low muscular tone and slow articulatory rate. The pauses are broadly classified into physical and linguistic and psychological and psycholinguistic [21]. Inter-lexical pauses will appear between utterance of two words in the voice signal. The articulation information embedded in the user utterance is explored by detecting the pause position in the frames of the voice signal. In the proposed work , in order to locate the inter-lexical pause position the silence removed signal SR(n) is partitioned in to 'M' (*i.e.* M=40) non-overlapping frames as each frame contains the stationary information of SR(n). The number of samples per frame 'n' in each frame is computed by sampling frequency of 'fs' using

$$n = fs/M \quad (3)$$

Further, the sliding window (w) with a size 'wl' is devised to locate the pause portion in the frame. As the inter-lexical pause exist for shorter duration, the smaller size of the sliding window is required. The size of the sliding window is selected empirically by computing equation (4)

$$wl = \lceil (n/10) \rceil \tag{4}$$

The sliding window is initially positioned at the left of the i^{th} frame (i=1) to be analyzed and sum of the amplitude of the overlapped samples of the frame is computed. The process is repeated over the entire frame by shifting the sliding window right by one sample each time till the window reaches to the right edge of the frame. This traversing of the sliding window is repeated for all remaining M-1 frames. The sum of the amplitude (AW) of the overlapping sliding window is computed for M frames as given in the equation (5)

$$AW(i) = \sum_{j=1}^{j+wl-1} F_i(j) \qquad \forall\ 1 \le i \le M \tag{5}$$

The corresponding starting and ending position of the sliding window over the frames is calculated by equations (6) and (7) and are stored for further processing.

$$Win\_Pos\_S(i) = \sum_{j=1}^{j+wl-1} \frac{1}{fs} * j \qquad \forall\ 1 \le i \le M \tag{6}$$

$$Win\_Pos\_E(i) = \sum_{j=1}^{j+wl-1} \frac{1}{fs} * (j+wl-1) \qquad \forall\ 1 \le i \le M \tag{7}$$

The pause portion is identified by finding the minimum value of the AW(i) as given by equation (8)

$$Min\_AW(i) = \min (AW(i)) \qquad \forall\ 1 \le i \le M \tag{8}$$

The starting and ending position values associated with frame samples with Min_AW(i) viz. Win_Pos_S(i) and Win_Pos_E(i) are used to trace the start point (S) and end point(E) of the inter-lexical pause position. The window is traversing from the position of Min_AW(i) towards the backward and forward portion of the frame respectively by computing the variations in the amplitude values of AW(i). The frames with minimum amplitude in silence removed signal are shown in Figure 2. Finally, the start point (S) and end point(E) of the inter-lexical pause are located as shown in Figure 3.

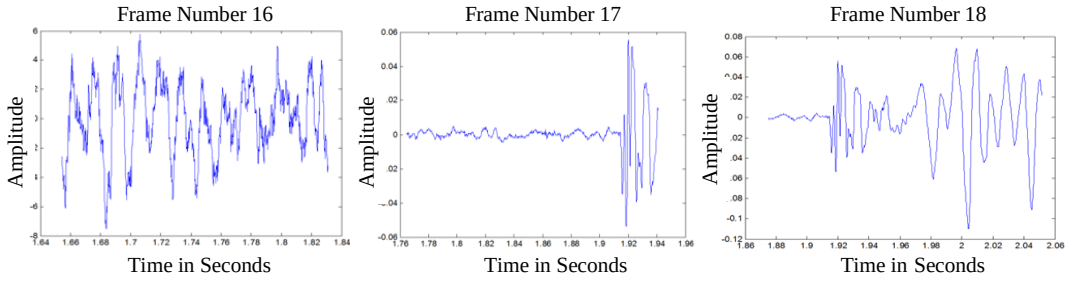

Figure 2. Frames with Minimum Amplitudes for Pause Portion Detection

Further in order to extract the complementary spectral features (*i.e.* spectral entropy, spectral centroid and spectral flatness) and pitch features the amplitude spectrum (*i.e.* FFT) of the silence removed signal is normalized. The normalization will reduce the error

during comparing the spectrum values. The normalization of the amplitude spectrum values are obtained by dividing each value of the spectrum by the maximum value of the spectrum. The normalized amplitude spectrum AS(k) is partitioned into P (*i.e.* P = 40) overlapping frames as each frame contains the stationary information of the amplitude spectrum where $1 \leq k \leq r$ *i.e.* r = number of spectral samples per frame (*i.e.* r = fs / P).

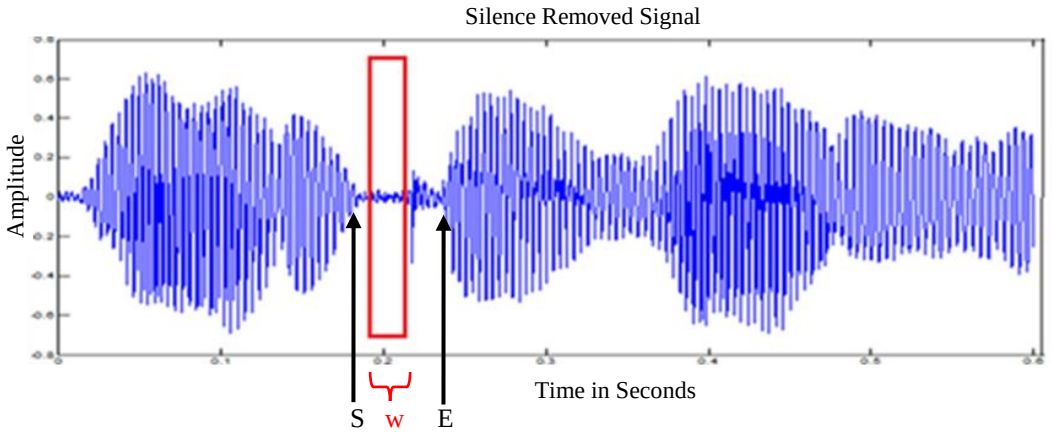

Figure 3. Pause Portion Detection with Sliding Window

Let $AS_i(k)$ be the amplitude spectrum of the i^{th} frame in spectral domain. The power spectrum of i^{th} frame $PS_i(k)$ is obtained by equation (9)

$$PS_i(k) = \sum_{k=1}^{r} |AS_i(k)|^2 \quad (9)$$

Let, the power spectrum of i^{th} frame $PS_i(k)$ be segmented into 'R' non-overlapping sub-bands/ blocks to consider each block as regularly repeating function with shorter period of time and spectrum becomes locally stationary (*i.e.* R= 10 in this work). Where each block 'b' with a size 'bw' is computed using equation (10)

$$bw = \lceil (r/10) \rceil \quad (10)$$

Further, complementary spectral features for i^{th} frame is computed from the power spectrum $PS_i(k)$ as described in the following.

3.3.1. Spectral Entropy or Shannon Entropy:

Spectral entropy or Shannon entropy (SE) describes irregularity, complexity or level of uncertainty of a signal. The nature of the probability distribution of voice components is described using the spectral entropy value *i.e.* wide and flat (higher value) or narrow and peaked (lower value)[22].

The power spectral density of each block of the i^{th} frame of power spectrum is calculated by normalizing the power spectrum of the i^{th} frame by the R number of blocks according to equation (11)

$$PS_{i,b\_Norm}(k) = \frac{PS_i(k)}{R} \quad (11)$$

The normalized energy of the each block of the i^{th} frame of the power spectrum is treated as a probability distribution for calculating entropy. The probability distribution of the each block of the i^{th} frame, $Pr_{i,b}(k)$ is computed using equation (12)

$$Pr_{i,b}(k) = \frac{PS_{i,b\_Norm}(k)}{\sum_{b=1}^{R} PS_{i,b\_Norm}(k)} \quad (12)$$

The spectral entropy of the power spectrum of silence removed signal is calculated using equation (13)

$$SE = -\sum_{i=1}^{P}\sum_{b=1}^{R} Pr_{i,b}(k)\ \log_2\ Pr_{i.b}(k) \quad (13)$$

The description of the spectral centroid feature is given below.

3.3.2. Spectral Centroid

The spectral centroid (SC) measures the spectral shape and median (*i.e.* centre of gravity) of a sound spectrum. It is an average frequency weighted by the values of the normalized energy of each frequency component in the power spectrum [23]. The spectral centroid of each block of the i^{th} frame of the power spectrum is calculated by equation (14)

$$SC_{i,b} = \frac{\sum_{k=1}^{bw} k * PS_i(k)}{\sum_{k=1}^{bw} PS_i(k)} \quad (14)$$

The spectral centroid of the power spectrum of silence removed signal is calculated using equation (15)

$$SC = \sum_{i=1}^{P}\sum_{b=1}^{R} SC_{i,b}(k) \quad (15)$$

Further, the description of the spectral flatness feature is given below.

3.3.3. Spectral Flatness or Wiener Entropy

The spectral flatness or Wiener entropy estimates the uniformity in the frequency distribution of the power spectrum. Spectral flatness is measured in decibels by computing the ratio between the geometric and the arithmetic mean of a block of power spectrum[24]. The spectral flatness of each block of size 'bw' in the i^{th} frame of the power spectrum is calculated by equation (16)

$$SF_{i,b} = \frac{(\prod_{k=1}^{bw} PS_i(k))^{1/bw}}{\frac{1}{bw}\sum_{k=1}^{bw} PS_i(k)} \quad (16)$$

Further, the spectral flatness of the power spectrum of silence removed signal is calculated using equation (17)

$$SF = \sum_{i=1}^{P}\sum_{b=1}^{R} SF_{i,b}(k) \quad (17)$$

Spectral flatness value categorizes the sound into noise-like sounds (high value) and more tonal sounds (low value). After extracting the complementary spectral features, the pitch frequency is estimated which is described in the following.

3.3.4. Pitch

Pitch is the fundamental frequency (F0) of audio signals which is the number of glottal cycles that occur per second. The relation between physical and perceptual properties of sound is estimated through pitch. The auto-correlation method is employed for pitch estimation, since the correlation functions have sharp peaks at the pitch period [25]. The auto-correlation sequence of the entire frames of the normalized amplitude spectrum of the signal is computed by equation (18)

$$AC = \sum_{y=0}^{P-1} \sum_{x=1}^{P} AS_i(x) * AS_i(x+y) \quad (18)$$

The maximum value *i.e.* peak of the auto-correlation(AC) sequence is computed and its associated sample number ($Max_{\_AC\_Sam}$) is obtained. The pitch frequency (PF) is estimated by equation (19)

$$PF = fs / Max_{\_AC\_Sam} \quad (19)$$

The pitch also depends on loudness and spectrum. The computation of loudness feature is described in the following.

3.3.5. Loudness

Loudness describes the strength of the ear's perception of a sound. It is associated with the sound level, sound quality, frequency and the duration of the sound. Different models exist in literature to estimate loudness based on the type of sound *i.e.* stationary, non-stationary and impulsive sounds[26]. The transmission of the acoustic signal through the outer and middle ear by calculating the basilar membrane excitation by critical bandwidth over 24 critical bands is modelled by Zwicker [27]. A partial loudness (N') in Sone/Bark is calculated with a power law as in equation (20).

$$N' = c * (E_{Th})^{\alpha} * [(\frac{1}{2} + \frac{1}{2}(E_{Stim})/(E_{Th}))^{\alpha} - 1] \quad (20)$$

where,
'c' is the constant which is independent form frequency
'α' is the exponent (*i.e.* $\alpha < 1$) is independent form frequency
E_{Stim} is the excitation produced by the input stimulus and
E_{Th} is the excitation at threshold in quiet.

Based on the experimentation c=0.1, α =0.25 and E_{Th}= 0.32 are selected in the proposed work. The final value for loudness in 'sones' is calculated using equation (21)

$$Loudness = \sum_{z=0}^{24} N' \quad (21)$$

Where z is the critical band rate (measured in Bark). Loudness is measured in sone, a unit based on a sensory scale as shown in Figure 4.

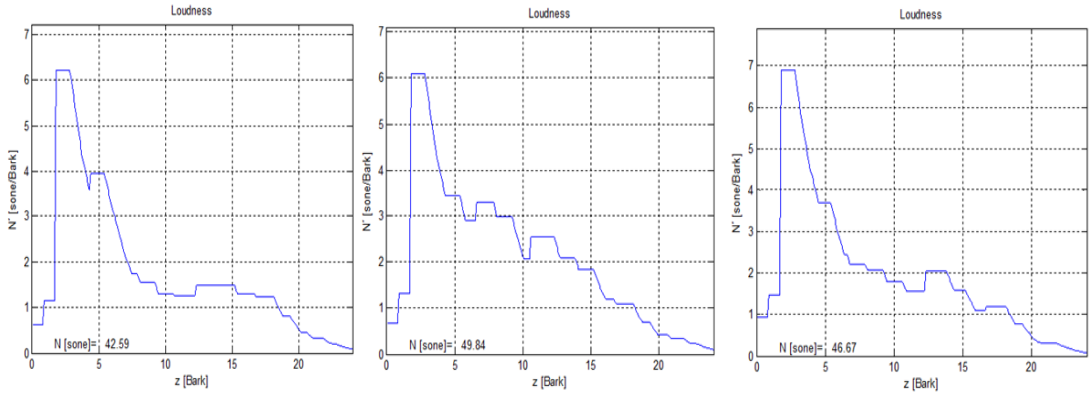

Figure 4. Loudness Feature for Voice Utterance of English Numbers "22","25" and "29"

3.3.6. Formants

Formants are characterized as the spectral peaks of sound range, of the voice, of a person. A formant frequency is the acoustic reverberation of the human vocal tract. The Linear predictive coding system (LPC) has been utilized for estimation of the formant frequencies [28]. Linear prediction estimates the vocal tract filter and shows the resonances that shape the formants of the speech signal. The order of the LPC prediction coefficients is about 4+ fs/1000 for all-pole filter (*i.e.* autoregressive) model of the voice source [29]. In the proposed work, order of the LPC prediction coefficients is 48 *i.e.* 4 + fs/1000. The calculation of the LPC coefficients obtains all members of bandwidth and formant frequencies by solving complex roots of polynomial. The frequencies of first three consecutive peaks of log spectra are extracted to represent the formants F1,F2 and F3 respectively as the frequencies of the first three formants contain sufficient information of the vowels and voiceprint. The variations in the extracted features of user voice is represented using symbolic object as described in the next subsection.

3.4. Symbolic Object Representation of Voice Signal

Symbolic objects, offer a formal methodology to represent the variability in features of user/speaker utterance. The features extracted from the user utterance *i.e.* inter-lexical pause position, complementary spectral features, pitch, loudness and formants features are described as a assertion symbolic object as given by equation (22).

$$\begin{aligned} \text{AssertionObject} = \; & [\{a_{min}, a_{max}\}, \{b_{avg}\}, \{c_{1min} - c_{1max}\}, \{c_{2min} - c_{2max}\}, \\ & \{c_{3min} - c_{3max}\}, \{d_{min} - d_{max}\}, \{e_{avg}\}, \{f_{min} - f_{max}\}, \\ & \{g_{1min} - g_{1max}\}, \{g_{2min} - g_{2max}\}, \{g_{3min} - g_{3max}\}] \end{aligned} \quad (22)$$

The knowledgebase is constructed separately for the English numbers utterances of the phrases "Twenty One" to "Twenty Nine" by selecting the five voiceprints for each phrase from every user. The symbolic object representation associated with each phrase of every individual 'x_i' associated with the i^{th} utterance *i.e.*$1 \leq i \leq 5$ is given in Table 1.

Table 1. Symbolic Object Representation of Voice Features

a_{min} = min{ Pause_Position (x_i) }	a_{max} = max{ Pause_Position (x_i) }
b_{avg} = mean{ Spectral_Entropy (xi) }	c_{1min} = min{ StdSCmin (x_i) }
c_{1max} = max{ StdSCmax (x_i) }	c_{2min} = min{ SCmin(xi) }
c_{2max} = max{ SCmin (x_i) }	c_{3min} = min{ SCmax (x_i) }
c_{3max} = max{ SCmax (x_i) }	d_{min} = min{ SF(x_i) }
d_{max} = max{ SF(x_i) }	e_{avg} = mean{ Loudness (xi) }
f_{min} = min{ PF(x_i) }	f_{max} = max{ PF(x_i) }
g_{1min} = min{ F1(x_i) }	g_{1max} = max{ F1(x_i) }
g_{2min} = min{ F2(x_i) }	g_{2max} = max{ F2(x_i) }
g_{3min} = min{ F3(x_i) }	g_{3max} = max{ F3(x_i) }

The description of the symbolic data structure for the English number utterance of the phrase "Twenty Two (22)" is depicted in Table 2.

Table 2. Sample Symbolic Data Structure for Number Utterance "22"

Symbolic Object	{ a_{min}, a_{max} }	b_{avg}	{ c_{1min} - c_{1max} }	{ c_{2min} - c_{2max} }	{ c_{3min} - c_{3max} }	{ d_{min} - d_{max} }	e_{avg}	{ f_{min} - f_{max} }	{ g_{1min} - g_{1max} }	{ g_{2min} - g_{2max} }	{ g_{3min} - g_{3max} }
Assertion Object	0.1769 - 0.2222	1.3253	0.0096 - 0.01815	0.0334 - 0.0532	0.0758 - 0.0882	5.0709e-5 - 1.3421e-4	42.35	228.49 - 238.38	271.36 - 328.51	441.52 - 750.36	1967.87 - 2287.01
Measuring Unit	Seconds	-	-	-	-	dB	Sone	Hz	Hz	Hz	Hz

Further, for every user nine assertion objects are formed separately for the English number utterances from "21" to "29" respectively. Further, the symbolic knowledge bases of English number utterances from "21" to "29" is constructed for 100 users according to the described symbolic data structure. This symbolic knowledge base is then used for user identification and the methodology is described in the following sub-section.

3.5. Methodology for Speaker Recognition

In the proposed work, the symbolic similarity measure is used for user identification. The utterance of every user is represented as symbolic data which characterizes the features of voice sample. The symbolic data object is constructed by computing the quantitative features like interval values and discrete values. The content, span and position similarity measures modelled as in [6,7] are computed to evaluate the nearest category of the user for identification. The symbolic similarity between the test feature vector (TFV) constructed for testing the utterance and the i^{th} user in the symbolic knowledge base (SKB_i) is written as equation (23)

$$S(TFV, SKB_i)=S_{Span}(TFV, SKB_i) + S_{Content}(TFV, SKB_i) + S_{Position}(TFV, SKB_i) \quad (23)$$

where the similarity component due to 'span' indicates the relative sizes of the feature values without referring to common parts between them. The similarity component due to 'content' is a measure of the common parts between two feature values. And the similarity components due to “position” indicates the distance of one object to the initial position of

other object [6]. In order to compute the similarity component between the test feature vector (TFV) and the i^{th} user object in the symbolic knowledge base (SKB_i) due to span, the following values are calculated,

$$\left.\begin{aligned} A_l &= \min(SKB_i) \\ A_u &= \max(SKB_i) \\ B_l &= \min(TFV) \\ B_u &= \max(TFV) \\ L_a &= |A_u - A_l| \\ L_b &= |B_u - B_l| \\ L_s &= |\max(A_u, B_u) - \min(A_l, B_l)| \end{aligned}\right\} \quad (24)$$

$$S_{Span}(TFV, SKB_i) = (L_a + L_b)/2L_s \quad (25)$$

Where, 'L_a' is the length of interval of feature vector, and 'L_b' represents length of interval of features of symbolic knowledge base. And 'L_s' corresponds to centered span length of both TFV and SKB_i.

To compute the similarity component between the test feature vector (TFV) and the i^{th} user in the symbolic knowledge base (SKB_i) due to content, the following values are calculated,

$$CF(TFV, SKB_i) = \begin{cases} \mathbf{1} & \text{if } |TFV - SKB_i| \le 0.05 \quad \forall 1 \le i \le N \\ \mathbf{0} & \text{Otherwise} \end{cases} \quad (26)$$

where $CF(TFV, SKB_i)$ is the common features(CF) between test feature vector and the i^{th} user in the symbolic knowledge base. For the experimentation, empirically the level of significance is assumed to be 0.05.

$$inters = \sum_{i=1}^{N} CF(TFV, SKB_i) \quad \forall 1 \le i \le N \quad (27)$$

$$Ls = (L_a + L_b - inters) \quad (28)$$

Where, N= Number of users *i.e.*100, 'L_a' and 'L_b' are computed using equation (24) and 'inters' represents the number of common features between test feature vector of the input voice sample and symbolic knowledge base. The content similarity measure is computed by equation (29)

$$S_{Content}(TFV, SKB_i) = inters/L_s \quad (29)$$

Further, to compute the similarity component between the test feature vector (TFV) of the input voice sample and the i^{th} user in the symbolic knowledge base (SKB_i) due to position is computed using equation (30)

$$S_{Position}(TFV, SKB_i) = 1 - |(A_l - B_l)| / U_k \quad (30)$$

where U_k is the length of the maximum interval of k^{th} feature of symbolic knowledge base (SKB_i) computed using equation (31)

$$U_k = |A_l - A_u| \quad (31)$$

The A_l and A_u are computed using equation (24). The symbolic similarity between the test feature vector (TFV) and the i^{th} user in the symbolic knowledge base (SKB_i) is computed by using the equation (23). As the identification is one-to-many matching, the combined similarity between the test feature vector (TFV) and the symbolic knowledge base (SKB) of all the N(*i.e.* 100) users is computed by adding the span, content and position similarity values calculated using the equation (25), equation (29) and equation

(30) respectively and represented as a 1 x N vector. Finally the best matching utterance is identified using the maximum similarity value. The experimentation is carried out for 100 users and performance is brought out, and is discussed in next section.

4. Experimentation

The performance of the proposed speaker recognition method is evaluated on voice samples from VTU-BEC-DB Voice Corpus of 100 users. The Voice Corpus consists of 31000 number utterance of Zero(0) to Thirty(30) collected separately in English and Kannada languages from the 100 persons which are acquired in two sessions. For the experimentation, the English number utterance from Twenty One (21) to Twenty Nine (29) of 100 persons are considered as they contain the inter-lexical pauses. Every utterance is modelled as assertion symbolic object by extracting the voice features namely inter-lexical pause position, pitch, loudness, formants, and complementary spectral features *i.e.* spectral entropy, spectral centroid and spectral flatness. From each user five utterances of first session are used for constructing the symbolic knowledgebase and the five utterances of second session are employed for testing. Hence, for 100 users the symbolic knowledgebase with a collection of assertion symbolic objects is constructed. During testing, the symbolic data representation of the user is compared with the symbolic knowledgebase using span, content and position symbolic similarity measure. Finally the best matching utterance in symbolic knowledge base is identified using the maximum symbolic similarity. Generally in identification mode, the performance of biometrics system is measured using genuine acceptance rate (GAR) or correct recognition rate (CRR), false acceptance rate (FAR) and false rejection rate (FRR) [30]. The GAR is the ratio of the number of authentic users accepted by the biometric system to the total number of identification attempts made. The FAR or 'type 2 error' is the ratio of the number of unauthorized users accepted by the biometric system to the total number of identification attempts made. The FRR or 'type 1 error' is the ratio of the number of authorized users rejected by the biometric system to the total number of attempts made.The system performance is given in Table 3.

Table 3. Recognition Performance with Symbolic Similarity

Sl. No.	English Number Utterance	Speaker Identification Rate (%)	Sl. No.	English Number Utterance	Speaker Identification Rate (%)	Sl. No.	English Number Utterance	Speaker Identification Rate (%)
1	21	91.60	4	24	90.60	7	27	90.40
2	22	92.20	5	25	89.80	8	28	90.00
3	23	90.60	6	26	89.60	9	29	90.20
Average of Speaker Identification Rate= 90.56 %								

The system is developed in MATLAB and experiments are conducted on Intel Core2 Duo T5550 at 1.83 GHz with 1 GB DDR2 RAM. The experimentation for the identification of user with symbolic similarity is conducted by varying the rank 'k' value in the range of $1 \leq k \leq 100$. The overall speaker identification rate of 90.56% achieved for the rank k=1. The performance of speaker identification for 100 users with English number utterance from Twenty One (21) to Twenty Nine (29) is depicted pictorially with the cumulative match characteristic (CMC) curve in Figure 5. By using the 4500 voice samples of 100 users by considering English number utterances of VTU-BEC-DB database, the overall correct identification rate of 90.56% is achieved with the help of symbolic similarity measure.

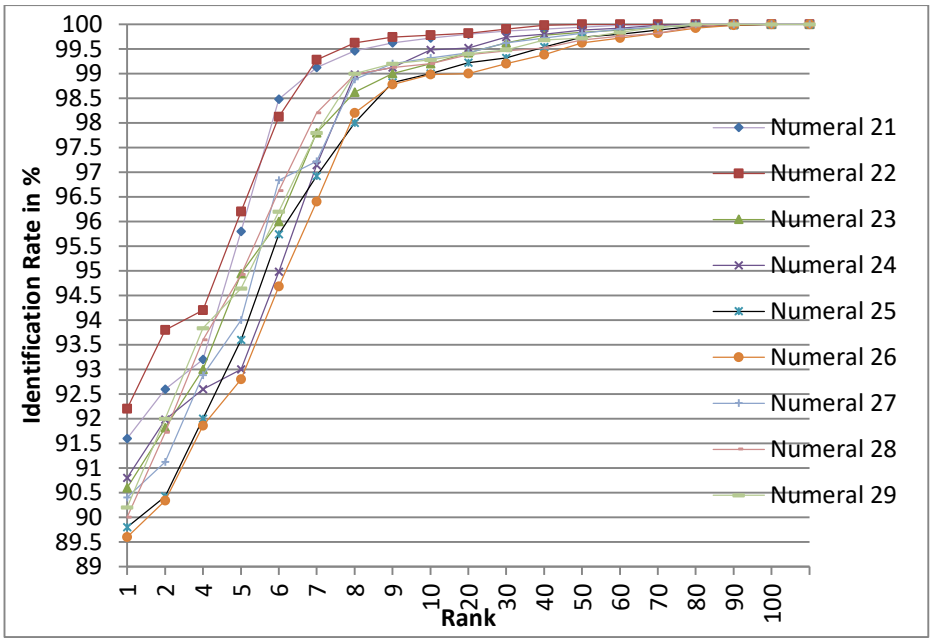

Figure 5. Identification Performance of the Proposed Speaker Recognition System

5. Conclusion

In this paper, a robust text-dependent speaker recognition system using symbolic modelling of voiceprint is presented. The symbolic data object is constructed using the voice features namely the inter-lexical pause position, complementary spectral features *i.e.* spectral entropy, spectral centroid and spectral flatness, pitch, loudness and formants. The user recognition is performed using a new symbolic similarity metric *i.e.* span, content and position. The symbolic data representation of the characteristics of user voiceprint is the novel feature of this work. The proposed method is experimented with the voice samples of the English number utterance separately by considering the phrase "Twenty One (21)" to "Twenty Nine (29)" of VTU-BEC-DB voice corpus which is constructed from the voiceprints of 100 persons. With the help of symbolic similarity metric the speaker recognition/identification rate in the range of 89.60% to 92.20% is obtained at rank-1 for the phrase utterances "Twenty Six" and "Twenty Two" respectively. The overall speaker identification rate of 90.56% is achieved at rank-1 with symbolic modelling and symbolic similarity analysis technique. In future, extensive experiments of the proposed method can be performed to make the proposed method more robust and also experimenting with other voice corpus.

References

[1] F.Z. Marcos and M.M. Enric, "State-of-the-Art in Speaker Recognition", IEEE Aerospace and Electronic Systems Magazine, vol. 20, no 5, **(2005)**, pp. 7-12.
[2] V. Tiwari, "MFCC and Its Applications In Speaker Recognition", International Journal on Emerging Technologies, vol. 1, no. 1, **(2010)**, pp. 19-22.
[3] D. Hosseinzadeh and S. Krishnan, "On the Use of Complementary Spectral Features for Speaker Recognition", EURASIP Journal on Advances in Signal Processing, vol. 2008, **(2007)**, pp. 1-10.
[4] E. S. Acevedo, M. N. Miyatake and H. P. Meana, "Evaluation of GMM Based Speaker Recognition Systems Using Dynamic Features", Científica, Mexico, vol. 10, no. 3, **(2006)**, pp. 151-156.
[5] F. Nolan, "The Phonetic Bases Of Speaker Recognition", Cambridge University Press, Cambridge, **(1983)**.
[6] K.C. Gowda, "Symbolic Objects and Symbolic Classification", Proceedings of the International Conference on Symbolic and Spatial Data Analysis:Mining Complex Data Structures Pisa, **(2004)**, pp. l-18.

[7] P. Nagabhushan, S. A. Angadi, B. S. Anami, "Symbolic Data Structure For Postal Address Representation And Address Validation Through Symbolic Knowledge Base", Proceedings of the First International Conference on Pattern Recognition and Machine Intelligence, **(2005)**, pp. 388-394.
[8] P. Jourlin, J. Luettin, D. Genoud and H. Wassner, "Integrating Acoustic And Labial Information for Speaker Identification And Verification", Proceeding of the Fifth European Conference on Speech Communication Technology, **(1997)**, pp. 1603–1606.
[9] M. R. Islam and M. F. Rahman, "Improvement of Text Dependent Speaker Identification System Using Neuro-Genetic Hybrid Algorithm in Office Environmental Conditions", International Journal of Computer Science Issues, vol. 1, **(2009)**, pp. 42-47.
[10] K. Dash, D.Padhi, B. Panda and S. Mohanty, "Speaker Identification using Mel Frequency Cepstral Coefficient and BPNN", International Journal of Advanced Research in Computer Science and Software Engineering, vol. 2, no. 4, **(2012)**, pp. 326-332.
[11] E. Erzin, Y. Yemez and A. M. Tekalp, "Multimodal Speaker Identification Using An Adaptive Classifier Cascade Based on Modality Reliability", IEEE Transaction on Multimedia, vol. 7, no. 5, **(2005)**, pp. 840–852.
[12] A. A. Aladwan, R. M. Shamroukh and A. A. Aladwan, "A Novel Study of Biometric Speaker Identification Using Neural Networks and Multi-Level Wavelet Decomposition", World of Computer Science and Information Technology Journal, vol. 2, no. 2, **(2012)**, pp. 68-73.
[13] D. K. Port and Y. Zheng, "Auditory Models Of Formant Frequency Discrimination For Isolated Vowels", Journal of the Acostical Society of America, vol. 103, no.3, **(1998)**, pp. 1654–1666.
[14] Z. Li and Y.Gao, "Acoustic Feature Extraction Method For Robust Speaker Identification", Multimedia Tools and Applications, vol. 75, no. 12, **(2016)**, pp. 7391–7406.
[15] S. Themos, M. J. Alam and K. Patrick, "Text-Dependent Speaker Recognition With Random Digit Strings", IEEE/ACM Transactions on Audio, Speech and Language Processing, vol. 24, no. 7, **(2016)**, pp. 1194-1203.
[16] S. X. Zhang, Z. Chen, Y. Zhao, J. Li and Y. Gong, "End-to-End Attention Based Text-Dependent Speaker Verification", IEEE Workshop on Spoken Language Technology, **(2017)**, pp. 171-178.
[17] O. Büyük, "Sentence-HMM State-Based i-Vector/PLDA Modelling For Improved Performance in Text Dependent Single Utterance Speaker Verification", IET Signal Processing, vol. 10, no. 8, **(2016)**, pp. 918-923.
[18] H. You, W. Li, L. Li and J. Zhu, "Lexicon-Based Local Representation for Text-Dependent Speaker Verification", IEICE Transaction on Information & Systems., vol. E100–D, no. 3, **(2017)**, pp. 587-589.
[19] H. Sun, A. L. Kong and B. Ma, "A New Study of GMM-SVM System For Text-Dependent Speaker Recognition", Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, **(2015)**, pp. 4195-4199.
[20] R. Jang and J. Shing, "End point detection", MIR Lab, CSIE Department National Taiwan University, Taiwan**, (2015)**, pp. 1-23.
[21] B. Zellner, "Pauses And The Temporal Structure Of Speech", Fundamentals of Speech Synthesis And Speech Recognition, **(1994)**, pp. 41-62,.
[22] H. Misra, I. Shajith, H. Bourlard and H. Hermansky, "Spectral Entropy Based Feature for Robust ASR", Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, **(2004)**, pp.1-6.
[23] K. K. Paliwal, "Spectral Subband Centroid Features For Speech Recognition", Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, vol. 2,**(1998)**, pp. 617–620.
[24] G. Tzanetakis and P. R. Cook, "Musical Genre Classification Of Audio Signals", IEEE Transaction on Speech Audio Processing, vol. 10, **(2002)**, pp. 293–302.
[25] J. Rouat, Y. C. Liu and D. Morissette, "A Pitch Determination and Voiced/Unvoiced Decision Algorithm For Noisy Speech", Speech Communication, vol. 21, **(1997)**, pp. 191–207.
[26] S. Esben and H. N. Soren, "Evaluation of Different Loudness Models with Music and Speech Material", Audio Engineering Society 117 Convention, USA, **(2004)**, pp. 1-34.
[27] E. Zwicker, "Program For Calculating Loudness According To DIN 45631 (ISO 532B)", Journal of the Acoustical Society of Japan, vol. 12, **(1991)**, pp. 39-42.
[28] S. C. Roy and M. Fausto, "Formant Location From LPC Analysis Data", IEEE Transaction on Speech and audio Processing, vol. 1, no. 2, **(1993)**, pp. 129-134.
[29] L.R. Rabiner and R.W. Schafer, "Introduction to Digital Speech Processing", Foundations and Trends in Signal Processing, vol. 1, no. 1-2, **(2007)**.
[30] S. A. Angadi and S. M. Hatture, "User Identification Using Wavelet Features of Hand Geometry Graph", Proceedings of the IEEE Cosponsored SAI Intelligent Systems Conference 2015 London, UK, **(2015)**, pp. 828-835.

Authors

Shanmukhappa A. Angadi, he is a Professor in the Department of Computer Science and Engineering, Centre for PG Studies, Visvesvaraya Technological University, Belagavi, Karnataka, India. He earned a Bachelor Degree in Electronics and Communication Engineering from Karnataka University, Dharwad and a Master Degree in Computer Engineering from University of Mysore, Mysore and a PhD in Computer science from the Department of Studies in Computer science, University of Mysore, Karnataka, India. He also completed PGDiploma in Opeartions Management (PGDOM), from IGNOU, New Delhi. He has completed many research and consulting projects of AICTE India under RPS, MODROBS, TAPTEC schemes and Research grant scheme of VTU Belagavi. His research areas are Image Processing and Pattern Recognition, Character Recognition, soft computing, Internet of things and Graph Theoretic techniques. He is a co-author of a book on C-programming language. He is life member of professional bodies like ISTE and IETE.

Sanjeevakumar M. Hatture, he received the Bachelor's Degree in Electronics and Communication Engineering from Karnataka University, Dharwad, Karnataka, India, and the Master Degree in Computer Science and Engineering from Visvesvaraya Technological University, Belagavi, Karnataka, India, and currently pursuing Ph.D Degree in the Department of Computer Science and Engineering at Basaveshwar Engineering College, Bagalkot affiliated to Visvesvaraya Technological University, Belagavi, Karnataka, India. His research interests include biometrics, image processing, pattern recognition, soft-computing and network security. He is life member of professional bodies like IRED, IEI and ISTE.

