

Detection and Identification of a required keyword within an audio content

Naresh E, Vijaya Kumar B. P, Niranjnamurthy M

Abstract: In modern era of communication, information sharing is very easy and within reach of every common man. Hence, spreading or sharing of ideology is widely possible in very quick time and creates a huge benefit in real time information sharing. With technology there could be a huge possibility of impacting people with harmful information which cannot be tracked. Data privacy is an important factor hence tapping the voice information or monitoring the information becomes illegal so we propose a method based on voice to text conversion and then performing data filtration. The proposed method converts voice to text and looks for illegal words as described by admin and reports the same with number of occurrence of the words with time stamp. The paper proposes a Smart Data Filtration (SDF) technique and extracting Mel frequency and other time domain statistical parameter associated with voice signal. The proposed system was tested on 102 samples of 20 seconds each, where the proposed methodology has shown a high efficiency in tackling the problem associated with violence and hatred speech sharing.

Index Terms: Smart data filtration, Voice to text, Mel frequency, suspicious words.

I. INTRODUCTION

Characteristic dialect preparing is a region of research from couple of years. Various strategies have been utilized for Natural Language Processing among them discourse acknowledgment is the most essential applications as discourse turns into an endless piece of our everyday life. Word by word acknowledgment is a procedure of extricating the discourse traits and grouping among the pre-recorded dataset. To perceive a word, the word must be passed on to larger amount programming for syntactic and semantic examination. It is a method of example coordinating, where sound signs are tried and surrounded into phonetics (number of words, expressions and sentences) [1]. To perform such errand one needs to record a voice test and after that change over this voice test into wave arrange. Range based parameters are acquired when a word is perceived. Different factual techniques are utilized for the examination of words which give some particular estimation of words. Words vary between in its limited scope of event. In the change of word

acknowledgment process, one of the imperative assignments is to locate the most useful parameters of the discourse flag. To perform such assignments a portion of the strategies are utilized Linear Predicted coefficients (LPC) and Mel Frequency Cepstrum Coefficients (MFCC) [2]. By utilizing such systems new range is acquired that is unique in relation to the past range of talked words. Discourse heightening is the procedure of improvement limitlessness or nature of a discourse test when it debased. Discourse improvement isn't just to decrease commotion from a discourse test however to de-resound and isolate the unconstrained signs. It is attractive to improve the discourse since when discourse is handled through any of the instruments in the lab it gets impacted with the commotion (foundation clamor or something else) and independence of the discourse changes with time which influences the entire acknowledgment process. Along these lines, it has turned into an extremely troublesome errand to discover creations that truly work in various practice situations. In any case, this paradigm assumes an imperative part in advocating the execution of the calculation with reference to quality and fathom ability. [3] Kalman Filter is a state estimator that creates an ideal gauge and limits the mean square blunder. Kalman sifting is a powerful way to deal with expels non-stationary commotion from the foundation or something else. It is a state-space show that dependably distillates the cute data from the flag which will be prepared [4]. In its contraption, a framework show is first chosen and model parameters are assessed from its past state [5]. Among proposed calculations, Kalman separating demonstrates that it the best parameter estimator [6, 7], Mathe et al [8] proposed Kalman channel for discourse improvement reason and utilized Bidirectional Kalman channel [9] for a hearty discourse acknowledgment framework.

II. RELATED WORK

Dynamic time warping coordinating calculation and mel-recurrence cepstrum coefficients include extraction which helps us in extracting the required noise reduced words within the audio content [18]. Quick fourier transform and discrete cosine transform are utilized to comprehend and obtain the vector estimate over coordinating words. This paper makes a codebook which holds tests which at that point used to coordinate with the created catchphrases [19]. The bending measure between the first example and the remade example ought to be least. fault recognition in top of the line framework must be dealt with so this paper has a calculation based execution to expand the productivity of the output [20].

Revised Manuscript Received on 30 March 2019.

* Correspondence Author

Naresh E, Research Scholar, CSE, Jain University, Dept. of Information Science and Engineering, M S Ramaiah Instituter of Technology, Bengaluru, India.

Vijaya Kumar B. P, Dept. of Information Science and Engineering, M S Ramaiah Instituter of Technology, Bengaluru, India.

Niranjnamurthy M, Department of Computer Applications, M S Ramaiah Instituter of Technology, Bengaluru, India.

© The Authors. Published by Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP). This is an [open access](https://creativecommons.org/licenses/by-nc-nd/4.0/) article under the CC-BY-NC-ND license <http://creativecommons.org/licenses/by-nc-nd/4.0/>

Detection and Identification of a required keyword within an audio content

This paper proposes another coordinating calculation to recover discourse data from a discourse database by discourse inquiry that permits persistent info. the calculation is called shift continuous dp (cdp) [21]. move cdp removes comparable areas between two discourse informational collections. two discourse datasets are considered as reference designs that are viewed as a discourse database and info discourse respectively. an ongoing discourse analyzer that identifies the nearness of discourse on the information line, and examines the discourse to give highlights proper for a word recognizer. 2.

a disengaged word recognizer that chooses which of an arrangement of words was talked. 3. a voice reaction framework to give talked summons to the client to manage the utilization of the repertory dialer framework. 4. a dialer (mimicked) to out heartbeat the coveted phone number [22]. setting subordinate (cd) show for extensive vocabulary discourse acknowledgment (lvsr) that use late advances in utilizing profound conviction systems for telephone acknowledgment [23]. we portray a pre-prepared profound neural system concealed markov show (dnn-hmm) half and half design that prepares the dnn to create a circulation over senones (tied triphone states) as its yield.

III. METHODOLOGY

The recorded discourse from telephone are put away into .wav design in Matlab neighborhood database after this stage it is important to separate discourse part utilized for the discourse acknowledgment framework. Each flag is made out of a few highlights/qualities. As per its highlights we can order the flag qualities. If there should arise an occurrence of discourse we extricate a portion of its qualities. Characteristics extraction is one of the simplest approaches to perceive the discourse. Discourse is a period shifting sign and to manage such a period fluctuating sign is a troublesome errand. So properties of discourse assume a vital part in acknowledgment. To manage a huge succession of discourse is brief time traits are gone up against Mel scale (tune of discourse). So we chose to remove brief time highlights of discourse which are MFCC. MFCC qualities for acknowledgment of each word for this framework. The word 'Mel' in the MFCCs speaks to the tune of a discourse flag. MFCC highlights depend on the human ear observation that implies human's ear's basic data transfer capacity frequencies channels the divided linearity between the high recurrence and low recurrence of each word articulated by the client. The human comprehension for various recurrence scopes of the expressed word is appeared on a nonlinear scale. Pitch time of each word is estimated with a Mel scale. Diverse systems received are introduced in the following stages

The proposed framework is demonstrated in Figure 1 which demonstrates the stages advanced in building up the framework. The voice test under thought is recorded from the telephone or the framework module created utilizing Matlab sound handling tool stash. Voice is recorded as a brief span portion of 5 seconds or 10 seconds till a greatest of 50 seconds which is preset by the administrator.

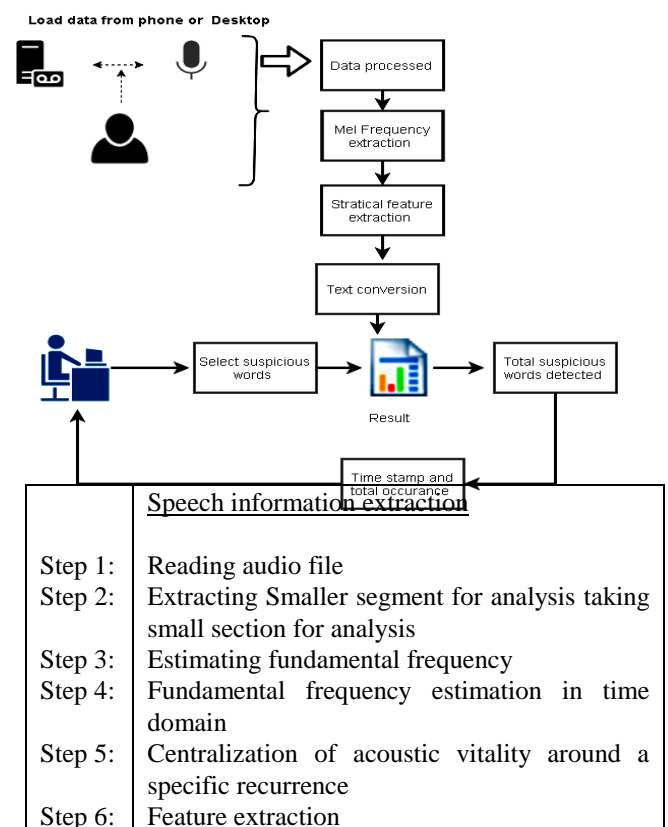
A. Preprocessing and speech information extraction

For starting investigation, the sound document taken is broken into a little fragment where we take just a segment of the flag

and in light of the fact that discourse flag isn't intermittent their crucial recurrence changes after some time. Finding littlest occasional interim gives us the crucial recurrence. Cepstrum examination helps in evaluating the prevailing recurrence which is only the Fourier of the logarithmic adequacy of the flag. On the off chance that the log plentifulness range contains numerous consistently dispersed sounds, at that point the Fourier investigation of the range will demonstrate a pinnacle relating to the dividing between the music: i.e. the central recurrence. The cepstrum in the flag searches for periodicity in the log range of the flag, though the view of discourse is mostly in light of a pitch which unequivocally identifies with periodicity. One of the easiest approaches to appraise central freq is to perform Autocorrelation. Autocorrelation essentially reflects how well flag relates inside itself for a wide range. So as we take little sections of flag we expect vowels can be assessed proficiently with relating pitch periods. Autocorrelation is inclined to specific blunders especially for female voice as formant recurrence is particularly lower.

Figure 1: Architecture diagram

Table1: Steps to perform speech information Extraction



The acoustic vitality around for specific recurrence characterizes the frequencies of words happening at a specific recurrence. At the point when excitation happens in the vocal tract excitation is impacted, however not generally henceforth a direct forecast demonstrate is created by utilizing most extreme vitality went through a recursive IIR channel

B. Smart data filtration

In smart data filtration we try to filter out those occurring words which are frequently spoken and check if there exist any words which the admin has set as the word to be filtered.

Table 2: Smart Data Filtration

Algorithm : Smart data filtration	
Step 1	Use Null delimiter to split the statement into cells
Step 2	Check Number of array elements
Step 3	Convert cells in character array to lower case characters
Step 3:	Collect all elements of array that have identical subscripts and store their sum in the location of corresponding subscript
Step 4:	Find unique values in array

C. Word Count and time stamp

Table 3: Word Count and Time Stamp

Step:1	Convert the word to its decimal format
Step 2:	Count the number of times the number is occurring
Step 3	Based on the count sort the occurrence of number
Step 4 :	Find the length of the words and find the time stamp by Each word estimate = count(total words) / duration Time_occurance=Position of occurrence / total count of word Time stamp=[Time_occurance- k : Time_occurance] Where k is an assumption estimate
Step 5 :	Report if query word exist with time stamp

Feature Extraction:

The procedure of highlights extraction of the discourse flag begins with: encircling the discourse flag and windowing it. From that point onward, one of the three strategies (Linear forecast channel coefficients(LPC), Mel-recurrence cepstral coefficients (MFCCs), and Spectrogram) will be connected to extricate the coefficients from the discourse flag.

Linear prediction filter coefficients

LPC decides the coefficients of a forward direct indicator by limiting the forecast blunder at all squares sense. It has applications in channel plan and discourse coding. [ac,gc] = lpc(xc,pc) finds the coefficients of a pth-arrange direct indicator (FIR channel) that predicts the present estimation of the genuine esteemed time arrangement xc in light of past examples.

$$\hat{xc}(n) = -ac(2)xc(n-1) - ac(3)xc(n-2) - \dots - ac(pc+1)xc(n-pc)$$

pc is the request of the expectation channel polynomial, ac=[1 ac(2) ... ac(pc+1)]. On the off chance that pc is unspecified, LPC utilizes as a default pc = length(xc)- 1. On the off chance that xc is a framework containing a different flag in every segment, LPC restores a model gauge for every section in the lines of grid ac and a segment vector of expectation blunder

differences gc. The length of pc must be not exactly or equivalent to the length of xc.

Hamming window

Hamming window is generally utilized as a part of discourse acknowledgment frameworks for windowing task. Figure 5.7 demonstrates the acquired flag from the task of duplicating the casing signal by hamming window.

$$w(na) = 0.54 - 0.46 \cos\left(\frac{2\pi na}{N-1}\right), \quad 0 \leq na \leq N-1$$

LPC coefficients can be assessed by applying a few methods on the discourse flag. These strategies began with applying autocorrelation on the windowed outlines. Each windowed outline is autocorrelated by pth arrange by applying the MATLAB code underneath:

```
p = 15;
x11 = x;
x22 = x;
No = length(x);
for k = 1:p+1
b = sum(x11.*x22);
B(k,1) = b;
x11 = zeros(No+k,1);
for i = 1:No
x11(i) = x(i);
end
x22 = zeros(No+k,1);
j = 1;
for i = k+1:No+k
x22(i) = x(j);
j = j + 1;
end
end
```

where x is an information vector of one edge, p is the request of the connection coefficients and B is the relationship coefficients. The MATLAB code above is equivalent to the inserted MATLAB work:

```
B = xcorr(xc);
B = B(N:(N+P));
```

The autocorrelation coefficients have been ascertained and afterward Levinson-Durbin calculation is connected to figure the last LPC coefficients.

Toward the starting, the principal coefficient of the main segment is computed by applying condition

$$A_{(i,i)} = [\sum_{k=0}^i B_{k+2} A_{(i-k,i)}] / E_i$$

$$E(1) = B(1)$$

where A is the network of the LPC coefficients, B is the vector of the autocorrelation coefficients, and E is the vector of the vitality of the expectation mistake.

Then El(2) – El(p) is figured by applying condition

$$El(i+1) = (I-A(i,i)^2) * El(i)$$



Detection and Identification of a required keyword within an audio content

In the second stage the second coefficient of the second segment is ascertained by condition and the rest coefficients of that (second) section are figured by condition

$$A_{(i,j)} = A_{(i,j-1)} + A_{(j,j)} * A_{(j-i,j-1)} \dots\dots [22]$$

And so on. These processes are repeated until all the coefficients are calculated.

The MATLAB code below was used to compute LPC coefficients:

```
p = 12;
for i = 1:frames
lpc_coef(:,i) = lpc(xw(:,i),p);
end
```

Where p is the order of coefficients for every frame, usually it is between 10 and 20.

IV. EXPERIMENTAL RESULTS

Our last trial comes about are accumulated with the assistance of Three techniques (LPC, MFCC, and Spectrogram) were utilized as a part of this proposal for highlight extraction, Feature extraction stage is begun in the wake of Denoising and End purposes of the discourse flag are recognized.

Result presented in the segment show process by output of the proposed algorithm where we implement on a single data time.

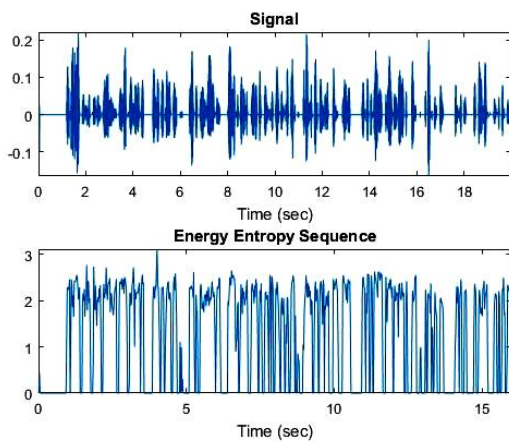


Figure2: Energy and entropy sequence

As shown in Figure 2 the variation in the speech signal is measured with respect to the energy found in the acoustic signal, it conveys variation in energy for a short frame of 5 second.

In Figure 3 the short time frame of the signal is presented which helps in widening each segment of the signal for accurate processing.

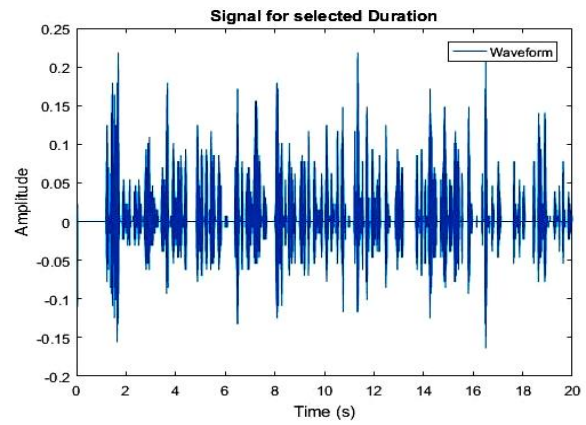


Figure 3: Short segment of signal

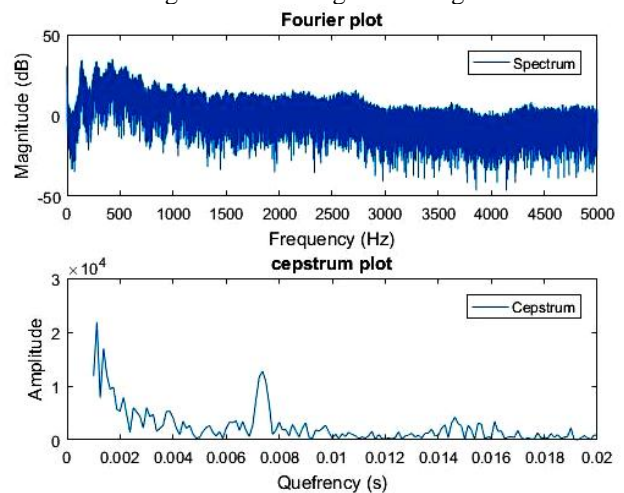


Figure 4: Fourier plot

In Figure 4 we can observe the frequency of the signal with its power spectral density, the Fourier transform represents all the frequency which lies in the signal and it becomes one of the major factor when distinguishing the voice as our vocal chord represent different frequencies corresponding to different words.

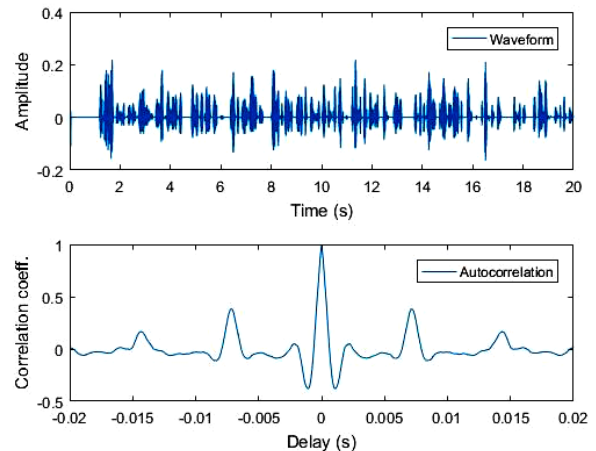


Figure 5: Autocorrelation

Figure 5 represents the autocorrelation in words and displays the delay in the signal. The Amplitude is plotted along side the time in the above Figure 5.

Words	Count
'of'	[3]
'and'	[2]
'is'	[2]
'the'	[2]
'theory'	[2]
'academic'	[1]
'among'	[1]
'attachment'	[1]
'been'	[1]
'dielectric'	[1]
'equity'	[1]
'expensive'	[1]
'field'	[1]
'for'	[1]
'friendship'	[1]
'have'	[1]
'in'	[1]
'included'	[1]
'paper'	[1]
'philosophy'	[1]
'proposed'	[1]
'psychology'	[1]
'relational'	[1]
'selecting'	[1]
'social'	[1]
'sociology'	[1]
'study'	[1]
'type'	[1]
'various'	[1]
'with'	[1]
'year'	[1]
'zoology'	[1]

Figure 6: The words and its occurrence

Figure 6 represents the words detected and find the occurrence of each words in decending order. These words are formed from the audio recording given initially. These words and their occurrence are shown in the above diagram within the square brackets.

F_x=49.6894Hz
 F_x=888.889Hz
 R_{max}=0.384828: F_x=137.931Hz
 Formant freq 1: 385.9
 Formant freq 2: 770.7
 Formant freq 3: 1738.1
 Formant freq 4: 2488.2
 Formant freq 5: 3122.5
 Formant freq 6: 3730.4
 processing Completed

Figure 7: The frequencies of the samples.

Figure 7 represents the frequencies in the recorded speech. And they are represented in hertz along with number of formaed samples which can be used to analyze the audio better.

Figure 8 finally displays the result with time frame of the identified word. When a admin gives the words to be searched for the programme automatically detects wether the word occurred in the speech and also the time frame of occurrence.

```

Do you want to look for specific word search if yes (1) else (0): 1
Specify particular word to checked: 'of'
The specified word exist

Possible occurrence of the word in time stamp could be between
0.1815 0.6190 2.3065 0.1935 0.6310 2.3185

Do you want to look another word if yes (1) else (0): 0

Exit
    
```

Figure 8: Final result with the time stamp.

V. CONCLUSION AND FUTURE WORK

The paper concludes with the observation that current method first records audio signal and the converts to speech, and finally retrieves all the words which are suspected to demonstrate violent activity with the time frame hence detecting suspicious activity without actually monitoring the speech and hence not breaching any privacy

VI. ACKNOWLEDGEMENTS

I thank to my undergraduate students Praveen Kumar D P, Chethan Kumar S, Suraj Kumar B C and Shravan Kumar M B for helping me to bring this work for publicaion and management of MSRIT, Bangalore for excellent support in completing this work at right time, and also for providing the academic and research atmosphere at the institute. A special thanks to the authors mentioned in the references.

REFERENCES

1. A. Winursito, R. Hidayat and A. Bejo, "Improvement of MFCC feature extraction accuracy using PCA in Indonesian speech recognition", *International Conference on Information and Communications Technology (ICOIACT)*, Yogyakarta, Indonesia, 2018, pp. 379-383.
2. L. Marlina *et al.*, "Makhraj recognition of Hijaiyah letter for children based on Mel-Frequency Cepstrum Coefficients (MFCC) and Support Vector Machines (SVM) method", *International Conference on Information and Communications Technology (ICOIACT)*, Yogyakarta, Indonesia, 2018, pp. 935-940.
3. C. Kumar, F. ur Rehman, S. Kumar, A. Mehmood and G. Shabir, "Analysis of MFCC and BFCC in a speaker identification system", *International Conference on Computing, Mathematics and Engineering Technologies (iCoMET)*, Sukkur, Pakistan, 2018, pp. 1-5.
4. M. Sadeghi and H. Marvi, "Optimal MFCC features extraction by differential evolution algorithm for speaker recognition", *3rd Iranian Conference on Intelligent Systems and Signal Processing (ICSPIS)*, Shahrood, 2017, pp. 169-173.
5. T. S. Gunawan, R. Husain and M. Kartiwi, "Development of language identification system using MFCC and vector quantization", *IEEE 4th International Conference on Smart Instrumentation, Measurement and Application (ICSIMA)*, Putrajaya, 2017, pp. 1-4.



Detection and Identification of a required keyword within an audio content

6. E. C. Djamel, N. Nurhamidah and R. Ilyas, "Spoken word recognition using MFCC and learning vector quantization," 2017 4th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI), Yogyakarta, 2017, pp. 1-6.
7. L. G. Singh, N. Adiga, B. Sharma, S. R. Singh and S. R. M. Prasanna, "Automatic pause marking for speech synthesis," TENCON 2017 - 2017 IEEE Region 10 Conference, Penang, 2017, pp.1790-1794.
8. N. Shokouhi, A. Ziaei, A. Sangwan and J. H. L. Hansen, "Robust overlapped speech detection and its application in word-count estimation for Prof-Life-Log data," 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), South Brisbane, QLD, 2015, pp. 4724-4728.
9. D. Weinshall et al., "Beyond Novelty Detection: Incongruent Events, When General and Specific Classifiers Disagree," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 34, no. 10, pp. 1886-1901, Oct. 2012.
10. V. Mitra, D. Garcia-Romero and C. Y. Espy-Wilson, "Language detection in audio content analysis," 2008 IEEE International Conference on Acoustics, Speech and Signal Processing, Las Vegas, NV, 2008, pp. 2109-2112.
11. Ł. Laszko, "Word detection in recorded speech using textual queries," 2015 Federated Conference on Computer Science and Information Systems (FedCSIS), Lodz, 2015, pp. 849-853.
12. V. Mitra, D. Garcia-Romero and C. Y. Espy-Wilson, "Language detection in audio content analysis," 2008 IEEE International Conference on Acoustics, Speech and Signal Processing, Las Vegas, NV, 2008, pp. 2109-2112.
13. D. M. Munoz, M. R. Zurera, F. L. Ferreras and N. R. Reyes, "Low delay audio coder using adaptive vector quantization," 2000 10th European Signal Processing Conference, Tampere, Finland, 2000, pp. 1-4.
14. S. Jiang, R. Yin and P. Liu, "Finite-state entropy-constrained vector quantiser for audio modified discrete cosine transform coefficients uniform quantisation," in IET Signal Processing, vol. 9, no. 1, pp. 30-36, 2 2015.
15. K. Xiang and R. Hu, "Low bitrates audio coding using lattice vector quantization based on scalable and high order codebook extension scheme," 2010 International Conference on Wireless Communications & Signal Processing (WCSP), Suzhou, 2010, pp. 1-4.
16. "2003 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (IEEE Cat. No.03TH8684)," 2003 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (IEEE Cat. No.03TH8684), New Paltz, NY, USA, 2003, pp.
17. R. Shadiev, Y. M. Huang, W. Y. Hwang and N. Shadiev, "Investigating the Effectiveness of Speech-to-Text Recognition Application on Learning Performance in Traditional Learning Environment," 2015 IEEE 15th International Conference on Advanced Learning Technologies, Hualien, 2015, pp. 441-445.
18. K. Dutta and K. K. Sarma, "Multiple feature extraction for RNN-based Assamese speech recognition for speech to text conversion application," 2012 International Conference on Communications, Devices and Intelligent Systems (CODIS), Kolkata, 2012, pp. 600-603.
19. S. Sultana, M. A. H. Akhand, P. K. Das and M. M. Hafizur Rahman, "Bangla Speech-to-Text conversion using SAPL," 2012 International Conference on Computer and Communication Engineering (ICCC), Kuala Lumpur, 2012, pp. 385-390.
20. Bing J. Sheu; Mohammed Ismail; Michelle Y. Wang; Richard H. Tsai, "Intelligent Speech/Audio Processing for Multimedia Applications," in Multimedia Technology for Applications , 1, Wiley-IEEE Press, 1998
21. M. V. Nicodem, I. C. Seara, R. Seara and D. dos Anjos, "Recording script design for a Brazilian Portuguese tts system aiming at a higher phonetic and prosodic variability," 2007 9th International Symposium on Signal Processing and Its Applications, Sharjah, 2007, pp. 1-4.
22. D. Sundermann, H. Hoge, A. Bonafonte, H. Ney, A. Black and S. Narayanan, "Text-Independent Voice Conversion Based on Unit Selection," 2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings, Toulouse, 2006, pp. I-1.
23. J. Picone, T. Staples, K. Kondo and N. Arai, "Kanji-to-Hiragana conversion based on a length-constrained n-gram analysis," in IEEE Transactions on Speech and Audio Processing, vol. 7, no. 6, pp. 685-696, Nov 1999.
24. Fang Chen and Baozong Yuan, "The modeling and realization of natural speech generation system," Signal Processing and Its Applications, 1999. ISSPA '99. Proceedings of the Fifth International Symposium on, Brisbane, Qld., 1999, pp. 99-102 vol.1.

AUTHORS PROFILE



NARESH. E received the M.Tech degree in software engineering from MS Ramaiah Institute of Technology, Bangalore in 2008 and pursuing PhD from Jain University, Bangalore under the guidance of Dr. Vijaya Kumar B. P. Currently, he is working as an Assistant Professor in the department of Information Science and

Engineering at MSRIT, Bangalore. His research interests include Software cost and effort estimation, empirical software engineering, and software process improvement. He is a member for ACM and Indian Society for Technical Education.



VIJAYA KUMAR B. P. received the Ph. D degree in Electrical Communication Engg., department from Indian Institute of Science (IISc), Bangalore in 2003, M.Tech degree in Computer Science and Technology from Indian Institute of Technology, Roorkee, with honors in 1992. He is currently a professor and head, in

Information Science and Engg., Dept., MS Ramaiah Institute of Technology, Bangalore, Karnataka, India, where he is involved in research and teaching UG and PG students, and his major area of research are Computational Intelligence applications in Mobile, Software validation models, Adhoc and Sensor Networks.



Niranjnamurthy M received Ph.D. Computer Science degree from JKT University, Rajasthan, INDIA in the year 2016, M.Phil-Computer Science degree from VM University, Tamil Nadu in the year 2009. MCA degree from VT University, Karnataka in the year 2007 and BCA Degree from Kuvempu University in the year 2004. He is an Assistant Professor in the

department of Computer Applications, M S Ramaiah Institute of Technology, Bangalore. His areas of interests are software testing, e-commerce and m-commerce, software engineering, web technologies, Cloud Computing, Big data analytics. He has been participating in National and International workshops/Conferences on different aspects related to Computer Applications. Guiding Research Scholars, Recognized Ph.D. research examiner National and International. Published many research Articles related to Computer Science.

