iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 9 **Issue:** II **Month of publication:** February 2021

DOI: https://doi.org/10.22214/ijraset.2021.33011

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue II Feb 2021- Available at www.ijraset.com

Automatic Speech Recognition (ASR) of Isolated Words in Hindi low resource Language

Suvarnsing Bhable[1], Ashish Lahase[2], Santosh Maher[3]
[1, 2, 3]*Department of Computer Science & Information, Dr. Babasaheb Ambedkar Marathwada University, Aurangabad*

Abstract: ***Speech is the natural voice and primary means of speech. Communication. Communication. Speech is easy, hand-free, quick and it doesn't require anything. Any technical know-how. Communicating with your computer using Speech is a simple and comfortable way for human beings. Speech to the This was made possible by the recognition system. Language and acoustics There is a model for this method, but mostly in English language. There are so many groups in India that can't Comprehend English or speak it. So the device of speech recognition in To these people, the English language is of little use. Here we have implemented Isolated method of recognition of Hindi words, which is part of System for Automated Speech Recognition (ASR)]. The primary purpose of The ASR system recognizes a voice through a device or microphone and To perform the necessary process, it is translated into text. In this article, As a feature extraction technique, we used Mel frequency cepstral coefficients (MFCC), Gaussian Mixture Model(GMM) Vector quantization (VQ) for Recognition of words separated from Hindi. We have practical research for***
The Hindi word speech dataset of various males and men was prepared.
Keywords: ASR, GMM, LM, AM, MFCC, FFT, VQ

I. INTRODUCTION

The method of converting sequences of spoken words into text is automatic speech recognition. Communicating with a computer using speech is a simple and comfortable way for human beings, rather than using the other medium, such as a keyboard and a mouse, as it requires some skill and good coordination in hand-eyeing[1]. Using computers is difficult for visually impaired people or blind people. Speech recognition tackles all these questions[2].

In a speech recognition system, there are two main modes: the training mode and the recognition testing mode. In training mode, the system processes all speaker utterances and finds the corresponding feature vectors for each utterance using feature extraction techniques such as LPC, MFCC, LDA and RASTA, etc[3]. In this way, the training vector is generated by the speech signal of the user. The training vector has spectral characteristics that distinguish between different words based on its class. The training vector is used in the test mode. The external expression for which the machine is trained is used in the test mode. For that word, the test pattern is generated. These test patterns are finally tested against training patterns using a variety of pattern classification techniques such as HMM, KNN, SVM, VQ and ANN, etc[4]. If the word pattern of the test matches the word pattern of the training word, it indicates that the training mode recognizes unique patterns and that the corresponding pattern is shown as the output pattern[5].

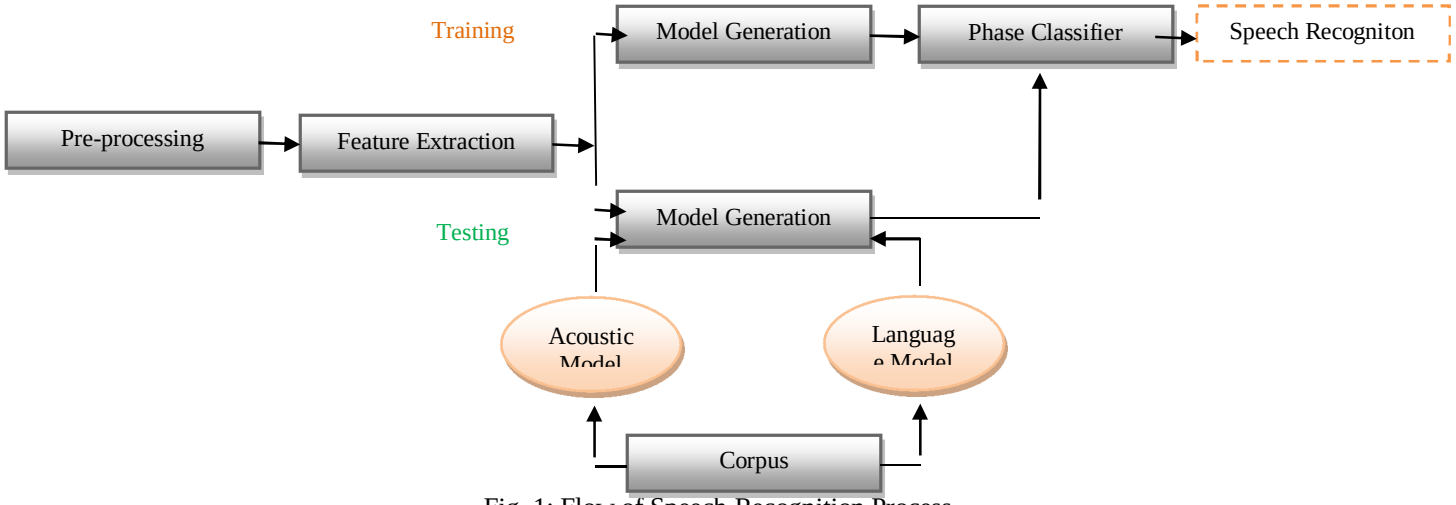


Fig. 1: Flow of Speech Recognition Process

©IJRASET: All Rights are Reserved

II. LITERATURE SURVEY

Speaker-dependent isolated word recognition device in Hindi. For feature extraction, the Linear Predictive Cepstral coefficient is used at the front end and HMM is used for identification at the back end. A system for 2 male speakers has been planned. The vocabulary of recognition consists of Hindi digits[6].

Established an isolated Hindi speech recognition device using 94.63 percent high performance limited vocabulary. This device is independent of the speaker. 5 males and 3 females will be used for preparation. The system's vocabulary size is 30 words. MFCC is used as a feature extraction technique at the front end and HMM at the back end is used for recognition[7]. Hindi ASR method for the identification of related digits. 40 separate speakers are used to create this speaker-independent system, of which 23 are female and 17 are male speakers. Any noise is artificially introduced after recording. Various features of extraction techniques are used in this paper, such as MFCC, BFCC, RPLP, and PLP at the front end and HMM at the back end. Comparative Hindi language analysis in Independent Word Comprehension[8]. MFCC is used here as a technique for feature extraction and KNN as a classifier of patterns. For 300 vocabulary info, MFCC and KNN have provided us 89 percent of the recognition rate. You may use the extra ANN classifier. Hindi Automatic recognition of speech using HTK. The framework is being developed using the Hidden Markov Model Toolkit (HTK). It uses the acoustic word model to distinguish isolated words. There are 113 Hindi words educated in the method. Training data was obtained from nine speakers[9].

III. HINDI LANGUAGE

Devanagari, as the language of Sanskrit literature, became the most commonly used script in India in the 11th century. The languages written in Devanagari include Hindi, Marathi, Gujarati, Bengali and Nepali as well as Tibetan and Burma. Hindi is spoken by 258-422 million Indians for their honor as a national language. Unlike other languages, Devanagari, as the language of Sanskrit literature, became the most frequently used script in India in the 11th century. The languages written in Devanagari are Nepali, Marathi, Bengali, Gujarati and Hindi, as well as Tibetan and Burma. Hindi is spoken by 258-422 million Indians to honor them as a national language. In relation to other languages[10].

This sequence of vowels also contains ॐ'which is called a graph. It refers to two or three sounds with concatenated phonemes. Vowel matras were often used as vowel signs to reflect this vowel sound rather than tacit. There is constriction in the form of the vocal tract over its length when pronouncing consonantsConsonants are classified according to the location of articulation (POA) and the manner of articulation (MOA). The Hindi language consonant range is illustrated in Table II. In the Hindi language, the consonants are 5 varg and 9 non-varg[11]. Each varg includes 5 consonants, the last of which is a nasal consonant. The first four consonants of each varg are primary and secondary pairs. Main consonants are voiceless and secondary consonants are articulated. The secondary consonants of each pair are the suction component of the first consonant. Each varg includes 5 consonants, the last of which is a nasal consonant. The first four consonants of each varg are primary and secondary pairs. Main consonants are voiceless and secondary consonants are articulated. The secondary consonants of each pair are the suction component of the first consonant[10].

Vowels	अ	आ	इ	ई	उ	ऊ	ऋ	ए	ऐ	ओ	औ	अं	अः
	a	ā	i	ī	u	ū	r	e	ai	o	au	aṁ	aḥ
Gutturals (कवर्ग)	क	ख	ग	घ	ङ								
	ka	kha	ga	gha	ṅa								
Palatals (चवर्ग)	च	छ	ज	झ	ञ								
	ca	cha	ja	jha	ña								
Cerebrals (टवर्ग)	ट	ठ	ड	ढ	ण								
	ṭa	ṭha	da	dha	ṇa								
Dentals (तवर्ग)	त	थ	द	ध	न								
	ta	tha	da	dha	na								
Labials (पवर्ग)	प	फ	ब	भ	म								
	pa	pha	ba	bha	ma								
Semi-Vowels	य	र	ल	व									
	ya	ra	la	va									
Sibilants	श	ष	स										
	sa	pa	sa										
Aspirate	ह												
	Ha												

Table I. Vowel & Consonant set

©IJRASET: All Rights are Reserved

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue II Feb 2021- Available at www.ijraset.com

IV. METHODOLOGY

It is processed into three major stages after receiving a voice signal from the microphone. Pre-processing is conducted on the signal in the first stage. MFCC function vectors are derived from the utterance of the speech signal in the second stage. These derived feature vectors are eventually matched to the database features using the pattern recognition algorithm. Fig. 2 displays the Isolated Word Recognition system's flow chart [12].

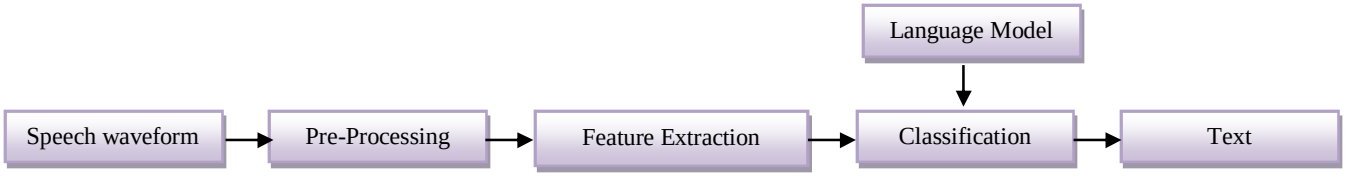


Fig.2 flow chart of Isolated Word Recognition

A. Speech Waveform

Speech will be transmitted by a microphone attached to a PC and will be an individual word that needs to be understood and shown on a PC.

B. Signal Pre-Processing

End point detection is used to isolate the spoken word by identifying the border of the word. First of all, we set a certain threshold value and we mark the beginning point of the term if the energy reaches the set threshold value and the end point is defined when the energy drops below the set threshold. Lower frequencies are increased during speech development, while higher ones are suppressed, This triggers a loss of signal information. The High Pass FIR filter is used before the feature extraction in the speaker verification and speech recognition system to prevent this data loss and to preserve speech signal features. This mechanism is known as pre-emphasis[13]. Before the is defined as—

$$y(n) = x(n) – a.x(n-1) \text{ ; where } 0.9 \leq a \leq 1$$

C. Mel frequency cepstral coefficients Feature Extraction

The Mel frequency cepstrum (MFC) can be defined as the short-time power spectrum of a speech signal, which is measured on a non-linear Mel frequency scale as the linear cosine transform of the log power spectrum. The coefficient obtained in the MFC representation is MFCCs. The distinction between MFC and cepstrum is that in the case of MFC, The Mel frequency cepstrum (MFC) can be defined as the short-time power spectrum of a speech signal, which is measured on a non-linear Mel frequency scale as the linear cosine transform of the log power spectrum. The coefficient obtained in the MFC representation is MFCCs. The difference between MFC and cepstrum is that in the case of MFCC, the frequency bands on the Mel scale are equally spaced[14].

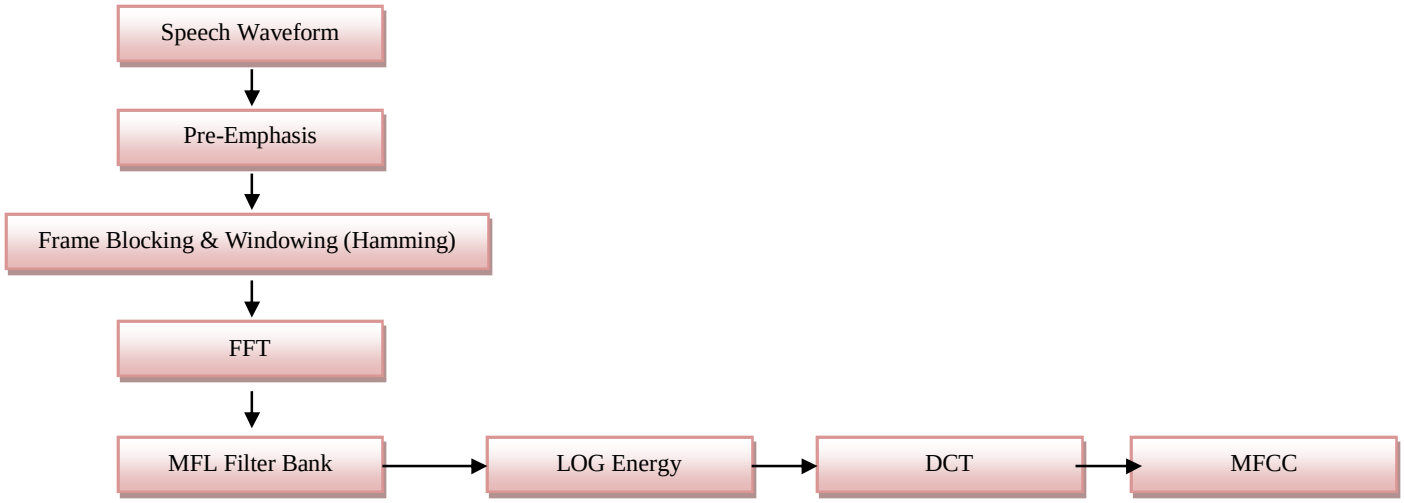


Fig.3: Block Diagram of MFCC

©IJRASET: All Rights are Reserved

The measures for obtaining MFCC acoustic vectors are:

1) *Framing:* As they differ slowly with respect to time, an audio signal is considered quasi-stationary. Because the properties of the speech signal are stationary for short periods of time, we take the signal over a short period of time to simplify this[15]. That's why can frame the signal in 20-30m frames.
2) *Windowing:* To prevent aliasing effect and to eliminate discontinuities, per frame is multiplied by window function after the speech signal is divided into frame. Three window choices are primarily available- Rectangular, Hanning and Hamming screens. The most widely used window is a hamming window for speech recognition as a hamming window that is consistent with the Mel scale and offers the highest results. Hamming window has the maximum attenuation of the side lobe and a wider phase range of $8\pi/M$ where M is the order of the filter. The Hamming window is used to eliminate the Gibbs phenomena. The Hamming window is described as a

$$W[n] = 0.54\ 0.46\cos(2\pi n/N),\ 0 \leq n \leq N\text{-}1\ 0,\ \text{otherwise}$$

Each frame is multiplied by a hamming window to ensure consistency of the first and last frame points and to avoid sudden shifts at the endpoint. [16].

3) *Fourier Transform:* When windowing is completed, Discrete Fourier transformation is applied to each frame to translate samples from the time domain of each frame to the frequency domain. Fast Fourier transformation algorithm is often used to increase Discrete Fourier transformation calculation rate by a hundred times, as Fast Fourier transformation is the fastest Discrete Fourier transformation calculation algorithm[9]. The spectrum log is taken to describe the amplitude of spectral lines in dB. Log of Fast Fourier transformation reveals that the fundamental frequency corresponds to quick fluctuations and the vocal tract parameters correspond to the slowly changing envelope[17].
4) *Mel Frequency Warping:* The mel scale is based on how frequencies are interpreted by human ears. By setting 1000 mels equal to 1000 Hz as a reference point, it was established. And listeners were asked to change the physical pitch until they saw it as two-fold, ten-fold and half, and the frequencies were labelled 2000 Mel, 10000 Mel, and 500 Mel, respectively. The resulting scale was called the Mel scale and is approximately linear below 1000 Hz and above logarithmic. The Mel frequency can be approximated by the following equation.

$$\text{Mel}(f) = 2595 * \ln(1 + f/700)$$

Where Mel (f) denotes perceived frequency and f is actual one [18].

5) *Discrete Cosine Transform:* The filtered outputs are calculated using the logarithmic procedure, and Discrete Cosine Transform is applied to it, resulting in MFCC coefficients[19].
6) *Calculation of delta and its coefficients:* In order to get Mel Frequency Cepstrum Coefficient (MFCC) from the speech signal, the log Mel spectrum is converted to time. These coefficients are used to integrate the effects of time evolution. Cepstral coefficient derivatives of the first and second order are referred to as delta coefficients and delta delta coefficients respectively. Somehow, the Delta coefficient tells us the speech rate, the deltadelta coefficient gives us something similar to the acceleration of speech. Fig.4 demonstrates the MFCC features extracted[20].

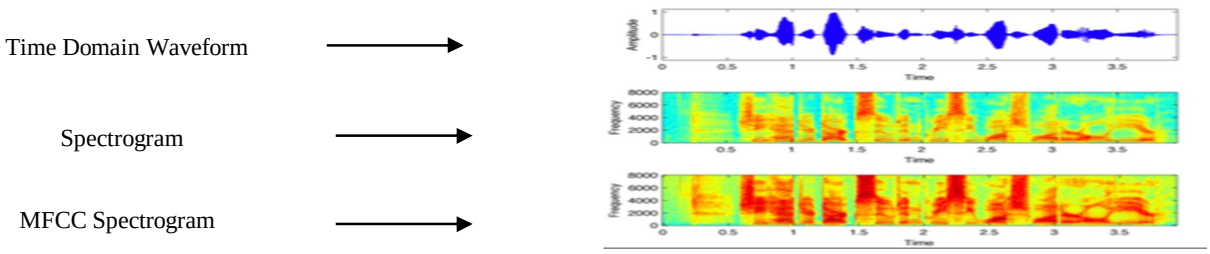


Fig 4.Extracted MFCC features:

©IJRASET: All Rights are Reserved

7) *Classification:* It's nothing but a matching feature/pattern algorithm. The method of recognition is performed using the Vector Quantization algorithm. Vector quantization (VQ) is a procedure used to take a large number of function vectors and to produce a smaller set of function vectors representing the distribution centroids, i.e. distributed points, in order to minimize the average distance to each other [21].

a) *Euclidean Distance Measure:* To measure the similarities or the dissimilarity of two spoken words, Euclidean Distance is used and takes place in its codebook after quantizing a spoken word. Here, we measure the Euclidean distance between the vector features of the unknown terms in the dataset to fit an unknown word to the codebook of known words. In order to classify the elusive word, the aim is to locate the codebook which has the minimum distance measurement. For example, the Euclidean distance between the vector and codebook features for each spoken word is calculated in the testing or identifying session and the word with the smallest average minimum distance is calculated[22].

b) *Language Model:* The language model is nothing but a database that has stored voice samples for our implementation, which is the most critical block in the Isolated Word Recognition framework[23].

V. DATABASE

First of all, we are building Hindi databases to run this framework. By documenting Hindi isolated words, the database is obtained. Recording is carried out at 16 KHz for both male and female voices. In room settings, recording was done. The speech files recorded were in the .wave format. Therefore, the archive has a limit of 60 samples. This database is used for educational purposes and for research.

VI. RESULT

An independent word recognition system in Hindi is introduced in this article. We used MFCC as a function extraction tool here, and KNN for pattern matching. MFCC functionality is derived from the .wave format. Test characteristics of both preparation and research are classified by the KNN classifier. Since 6 terms exist, 6 groups are created. The KNN classifier speech recognition system is made more stable and effective according to the groups of corresponding wave files and hence increases the efficiency of the speech recognition system. For isolated Hindi words, the outcome is satisfactory. The rate of speech recognition depends on four parameters: True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN). Using an uncertainty matrix, these parameters may be measured.

Precision = (TP/(TP + FP)) * 100;
Recall = (TP/(TP+FN))*100;
Accuracy= ((TP+TN)/(TP+FN+FP+TN))*100;

Method	H	S	I	N	Correct rate
Sentence	1697	103	-	1800	94.28%
Word	5297	103		5400	98.09%

VII. CONCLUSION

Hindi is the official language; it is used for contact by the majority of Indians. It is therefore necessary to introduce an effective and fast Hindi language ASR system so that most Indians can use advanced speech-operated gadgets. Project study has been carried out on the identification of isolated words in Hindi language. We used MFCC as a function extraction tool here, and KNN for pattern matching. This introduction would be an important way of education and connectivity with sophisticated technological devices such as laptops, smartphones and home appliances for disabled persons as well as for illiterate people.

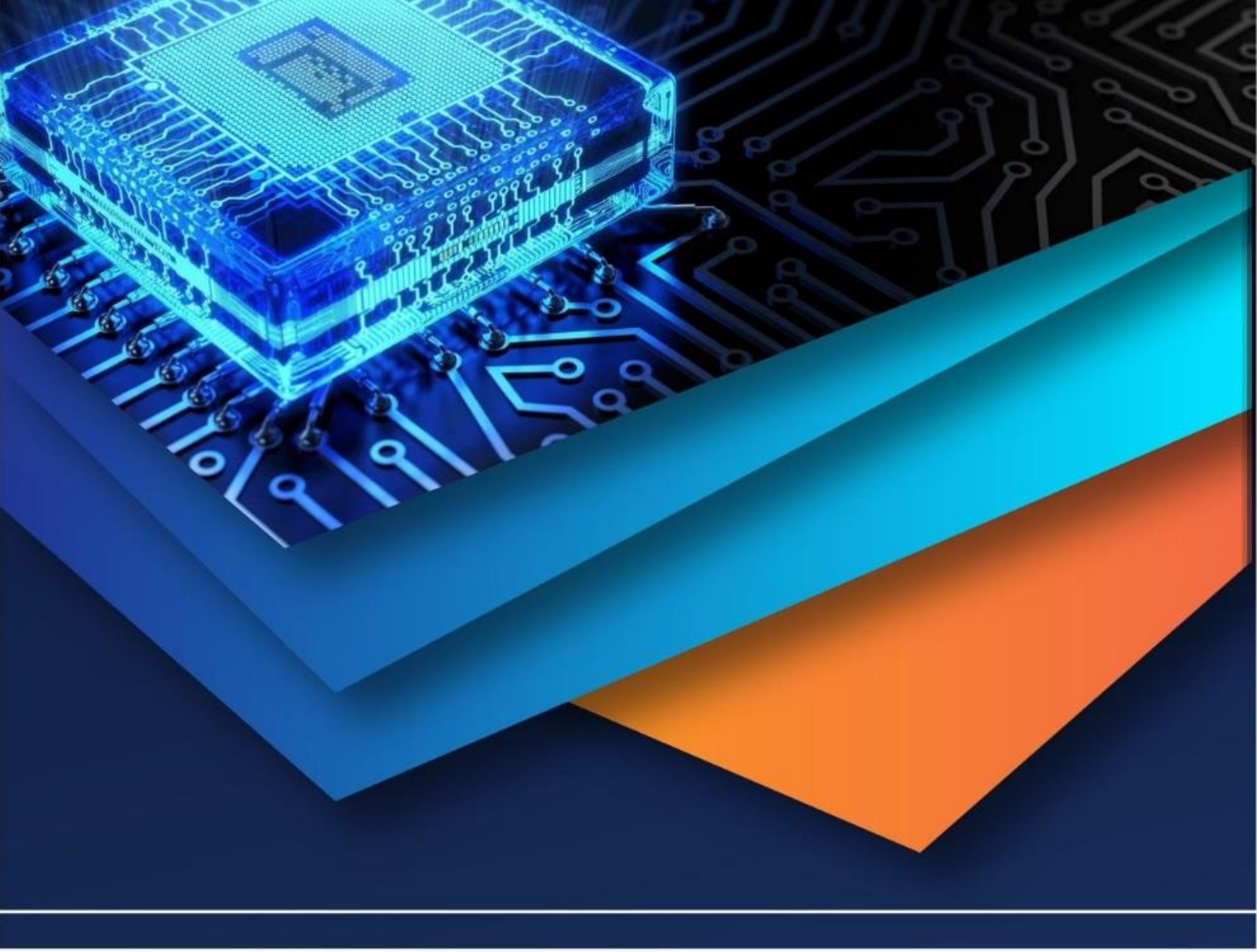
REFERENCES

[1] B.H. Juang, and S. Furui, "Automatic Recognition and Understanding of Spoken Language–A First Step Toward Natural Human Machine Communication, Proc. IEEE, 88, No. 8, 2000, pp. 1142-1165.
[2] Mohammed Fakrudeen, Sufian Yousef, Dr Mahdi H. Miraz, "Exploring the use of speech input by blind people on mobile devices", Proceedings of the 15th International ACM SIGACCESS Conference on Computers and Accessibility, Oct 2013.
[3] Priyanka Wani et al., "Automatic speech recognition of isolated words in Hindi language", International Conference on Computing Communication Control and automation (ICCUBEA), 2016.
[4] Yang Gu et. al., "Optimizing Corpus Creation for Training Word Embedding in Low Resource Domains: A Case Study in Autism Spectrum Disorder (ASD), AMIA Annu Symp Proc, 2018, PP 508–517
[5] B Yegnanarayana, "Artificial neural networks for pattern recognition", Sadhan & Vol. 19, Part 2, April 1994, pp. 189-238
[6] Gurpreet Kaur, Mohit Srivastava & Amod Kumar, "Analysis of Feature Extraction Methods for Speaker Dependent Speech Recognition, International Journal of Engineering and Technology Innovation, vol. 7, no. 2, 2017, pp. 78 – 88.

©IJRASET: All Rights are Reserved

[7] Ms.Vimala.Ca & Dr.V.Radha, "Speaker Independent Isolated Speech Recognition System for Tamil Language using HMM", International Conference on Communication Technology and System Design, Elsevier Procedia Engineering 30 (2012) , pp.1097 – 1102.
[8] Musaed Alhussein et. al. , "Automatic Gender Detection Based on Characteristics of Vocal Folds for Mobile Healthcare System", Volume 2016 , Article ID 7805217, https://doi.org/10.1155/2016/7805217.
[9] A. Moosavian⁎, H. Ahmadi, A. Tabatabaeefar and M. Khazaee, "Comparison of two classifiers; K-nearest neighbor and artificial neural network, for fault diagnosis on a main engine journal-bearing", IOS Press, Shock and Vibration 20 (2013) , pp.263–272.
[10] Sanskriti and language, http://home-tutor.in/resources/flip/ncert/6/fess1dd/files/basic-html/page48.html.
[11] Rajesh Aggarwal and Mayank Dave, "Using Gaussian Mixtures for Hindi Speech Recognition System", International Journal of Signal Processing, Image Processing and Pattern Recognition Vol. 4, No. 4, December, 2011
[12] Sabur Ajibola Alim and Nahrul Khair Alang Rashid, "Some Commonly Used Speech Feature Extraction Algorithms, Gastech Hydrogen, December 12th 2018.
[13] Lori F. Lamel, E, Lawrence R. Rabiner, Aaron E. Rosenberg & Jay G. Wilpon, "An Improved Endpoint Detector for Isolated Word Recognition", IEEE Transactions On Acoustics, Speech, and Signal Processing, Vol. Assp-29, No. 4, August 1981,pp.777-785
[14] Mahmoud I. Abdalla and Hanaa S. Ali, "Wavelet-Based Mel-Frequency Cepstral Coefficients for Speaker Identification using Hidden Markov Models", Journal Of Telecommunications, Volume 1, Issue 2, March 2010,pp.16-21.
[15] Sam Kwong and Qianhua He, "The Use of Adaptive Frame for Speech Recognition", EURASIP Journal on Applied Signal Processing 2001:2, pp.82–88.
[16] Andreas Spanias Ted Painter Venkatraman Atti, "AUDIO SIGNAL PROCESSING AND CODING", WILEY-INTERSCIENCE A John Wiley & Sons, Inc., Publication 2007,pp.1-486.
[17] C. Zhu et al. "On-line vibration monitoring and diagnosing of a multi-megawatt wind turbine gearbox", SceinceDirect 2017.
[18] Mayur Babaji Shinde and Dr. S. T. Gandhe, "Speech processing for isolated Marathi word recognition using MFCC and DTW features", International Journal of Innovations in Engineering and Technology (IJIET), Vol. 3 Issue 1 October 2013, pp.109-114.
[19] Md. Afzal Hossan, Sheeraz Memon and Mark A Gregory, "A novel approach for MFCC feature extraction", Signal Processing and Communication Systems (ICSPCS), 2010 4th International Conference, January 2011.
[20] Haytham Fayek, "Speech Processing for Machine Learning: Filter banks, Mel-Frequency Cepstral Coefficients (MFCCs) and What's In-Between" Apr 21, 2016.
[21] Balwant A. Sonkamble and D. D. Doye, "Speech Recognition Using Vector Quantization through Modified K-meansLBG Algorithm", Computer Engineering and Intelligent Systems, Vol 3, No.7, 2012, pp.137-144.
[22] E. G. Rajan, "Vector Quantization Approach for Speaker Recognition using MFCC and Inverted MFCC", International Journal of Computer Applications 17(1), March 2011.
[23] Andreas Nautscha el at., "Preserving privacy in speaker and speech characterization", Computer Speech & Language Volume 58, November 2019, pp. 441-480.

©IJRASET: All Rights are Reserved



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)