iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 9 **Issue:** IX **Month of publication:** September 2021

DOI: https://doi.org/10.22214/ijraset.2021.38094

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 IX Sep 2021- Available at www.ijraset.com

A Fundamental of Automatic Speech Recognition and Speech Database

Sonal Anilkumar Tiwari[1], Manasi R. Baheti[2]
[1, 2]*Department of Computer Science & IT, Dr. Babasaheb Ambedkar Marathwada University, Aurangabad, Maharashtra, India*

***Abstract:** This can be quite interesting when we think that we commanding something to in-animated objects. Yes it is possible with the help of ASR systems. Speech recognition system is a system that can make humans to talk with machineries. Nowadays speech recognition is such a technique that without it, a person cannot do any of his work properly. People get addicted of it. And it has become a habit for humans like we use mobile phones but when we want to type something, then we immediately can pass the voice commands. With which our Efforts are reduced, as well as a lot of our time.*
***Keywords:** Speech, Speech Recognition, ASR, Corpus, PRAAT*

I. INTRODUCTION

A. *Fundamentals of Speech Recognition*

Linguistic is also known as science of language. Linguistic involves and explore every aspect of language as well as methods for studying and modelling them. One of the sub branches of Linguistic is Computational Linguistic which uses computer science to model languages and speech. Speech is mode of communication among humans. There are various modes of communication like verbal communication which includes speech, non verbal communication which includes facial expressions, postures, eye contact, hand movement etc. Speech involves a spontaneous exchange of information or we can say that with the help of speech a person expresses his or her feelings to another person. Human speech is a sound wave that is generated by a well defined physical system.

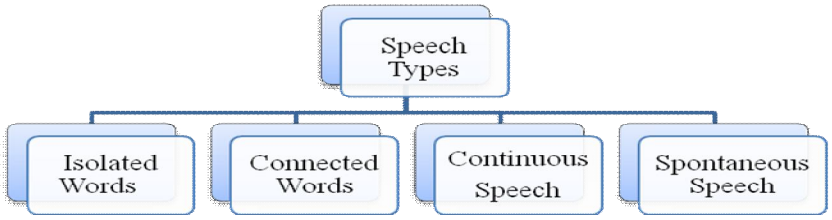


Fig. 1 Speech Types

1) *Isolated Words:* Isolated word has separation (or a pause) in between each word's pronunciation. Isolated word recognizers have many words but each word must be separated by another word by means of a pause. However it becomes easy to understand when we replace isolated words to the word isolated Utterance.
2) *Connected Words:* connected word systems (or more correctly 'connected Utterances') are similar to isolated words, but allow separate utterances to be 'run-together' with a minimal pause between them.
3) *Continuous Speech:* Continuous speech recognizers enable users to speak nearly naturally, while the system determines the content. Basically, it's computer dictation.
4) *Spontaneous Speech:* At a basic level, it can be thought of as speech that is natural sounding and not rehearsed. "ums" ,"ahs" and slight stutters words also run together in ASR system which has ability to handle variety of natural speech features.

B. *Terminologies used for Speech Database*

1) *Corpus:* Speech Database additionally contains speech signals, its annotations and Documents.

Basic Structure of Speech Corpus
DATA - Contains all signal Files
ANNOT- Contains all Annotation files
META- Contains all Meta data Files
DOC- Contains all documentation
LEX- Contains the Lexica
TOOLS- Contains software to access signal

2) *Corpora:* Plural term used for Corpus.

C. Automatic Speech Recognition

Human communicates with other humans in many ways, including body gestures, printed text, pictures, drawings and voice. But voice communication is used commonly in day to day life. Speech is an efficient way for humans to express their ideas and desires. So it should come as no surprise that we have always wanted to communicate with our machines and control them by voice.

D. Elements of Voice Processing Technology

There are five broad technology areas-

1) *Voice Coding:* Voice Coding gives a concept that over a smaller bandwidth, how we can transfer a sound signal. Therefore we can say that the process of reducing the information regarding sound file, so that it can easily stored or used over smaller bandwidth channel called voice coding.
2) *Voice Synthesis:* An electronic synthesizer that generates and combines basic elements of sound to produce simulated speech, used in computer systems, etc. It creates a synthetic replica of a sound signal.
3) *Speech Recognition:* Speech recognition is the process by which a machine can recognize a spoken command given by human. Simply we can say that a machine follows or take some actions on the basis of our oral commands.
4) *Speaker Recognition:* In Speaker recognition, Speakers Identification is done on the basis of individuals voice characteristics. Speaker's recognition is required for the security purposes like restricting unwanted access to vital areas.
5) *Spoken Language Translation:* This technique is also helpful when two persons who want to communicate but don't speak the same language. Spoken language translation recognizing a person's speech in one language, translate it to another language and producing the appropriate message for the other person in a second language.

E. Speech Recognition System

Speech recognition systems are often classified according to the scope of their capabilities.

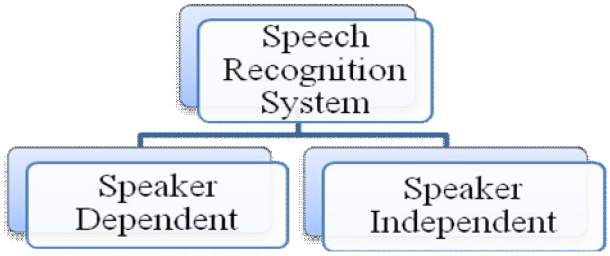


Fig. 2 Classification of speech Recognition System

Speaker Dependent systems must be "Trained" on the speech of an individual user, while Speaker- independent systems attempt to cope with the variability of speech among speakers.

F. Elements of a System for Human-Machine Communication by Voice-

1) Requires microphone to pick up the human voice and
2) Another requirement is a speaker or headphone to deliver a synthetic voice from system to human ear.

G. Advantages

1) Speech is the natural medium of communication for human being.
2) Voice control is particularly interesting when the human's hands or eyes are otherwise busy.
3) Voice communication with machines is potentially very useful to handicapped persons.

H. Applications

1) *Voice User Interfaces:* People can interact with the Machine like Computer, smart TV, Mobile, Washing machine etc with the voice commands.

   Example – Apple's siri, Amazon's Alexa, Google's Assistant, and Microsoft's Cortana etc are the examples of voice user interfaces.

2) *Call Routing:* Call Routing is used in contact centre. As speech recognition enables devices to understand and respond to spoken commands, or queries. Connecting callers to the right advisor frees up valuable time and saves customers frustration of having to repeat their verification information or explain their reason for calling to more than one agent.
3) *Search Keywords:* Search keyword allows the user to use a voice command to search the internet, a website, or an app.
4) *Simple Data Entry:* Simple data entry simply means giving a voice dialing command.
   Example - "call sonu" , "call home" etc.
5) *Medical Documentation:* In the health care sector, speech recognition can be implemented in front – end (provider dictates into a speech recognition engine) or back – end (recognized words are displayed as they are spoken) of the medical documentation process.
6) *Speech/ Telephone/Communication:* Telephony, where speech recognition is used for spoken dialogue systems for entering digits, recognizing words to accept collect calls, finding out airplane or train information, and call-routing etc.
7) *Physically Handicapped:* For Physically Handicapped person, it is helpful to give voice command for their work.
8) *Hand-busy or Eyes-Busy applications, such as where the user has objects to manipulate or equipment to control.*
9) *Dictation, that is, transcription of extended monologue by a single specific speaker.*

I. *Limitations*

Speech Recognition has severe limitations-

1) Ambient Noise Sensitivity
2) Privacy
3) Being obtrusive in shared environment
4) Interfere with other cognitive tasks.
5) Difficult to correct Errors.

II. CORPORA CREATION IN INDIAN LANGUAGES

A. *Speech Corpus*

A Speech corpus is a large collection of audio recording of spoken language with additional text files containing transcriptions of the words spoken and the time each word occurred in the recording.

When we conduct research on speech we can either

1) *Record our own data or*
2) *Use ready-made speech corpus*: There are many speech corpus available for research in linguistic department.

Speech Corpora mainly can be divided into two types-

a) Read Speech
b) Spontaneous Speech

- Dialogues and Meetings
- Narratives

B. *Marathi Language*

Marathi is a member of Indo-European language family's Indo – Aryan Branch. It has some similarity with Hindi and Punjabi language. Maharashtra is Ranked second most populated state of India. As per the projection, Population of Maharashtra in 2021 is 12.62 crore i. e. 126.2 Million [10]. Out of 12.62 crore population, 8.3 crore Marathi people spokes Marathi language in Maharashtra [11]. Marathi is the co-official language of Maharashtra state. Other than Maharashtra state, Union territory in the India sub- Continent- Dadra and Nagar Haveli and Goa state also uses Marathi language. it is spoken in Israel and Mauritius also.
Despite variations among the dialects in pronunciation and vocabulary, they are, for the most part, mutually intelligible. Marathi language uses Devanagari script. Marathi language consists of 14 vowels, 36 consonants and 2 sound modifiers.

1) *Sound System:* Like Sanskrit and other Indo Aryan languages Marathi also has similar phonological properties. The use of consonant group is extremely limited, even in borrowed words.
2) *Vowels:* Marathi Language has 14 Vowels (including two new vowels ॲ and ऑ). After those two vowels are added in Marathi Barakhadi it becomes Marathi Chaudakhadi. These vowels would be represented by using the international Alphabet of Marathi Transliteration (IAMT).

Table I
Vowels and their transliterate Forms

Vowels	अ	आ	इ	ई	उ	ऊ	ए	ऐ	ओ	औ	अं	अः	ॲ	ऑ
Transliterated	a	ā	i	ī/ee	u	ū/oo	e	ai	o	au	aṃ	aḥ	æ/ê	ô

3) Vowels Forms: There are two vowel forms-
a) Standalone Form and
b) Matra Form
Using matra form consonants gets modified because in it vowels can pair with the syllable.

Example –

Table II
Matra vowel form

क	का	कि	की	कु	कू	के	कै	को	कौ	कं	कः	कॅ	काॅ
ka	kā	ki	kī	ku	kū	ke	kai	ko	kau	kan	kaha	kê/kæ	kô

Table III
The vowel

	Front	Central	Back
Close	i, ī		u, ū
Mid	e,ē	ə	o, ō
Open		a, ā	

- /i/ = *ea* in *peat*
- /e/ = *e* in *pet*
- /ə/ = *a* in *ago*
- /a/ = *a* in *bar*
- /u/ = *oo* in *too*
- /o/ = *o* in *token*

4) Consonants: There are 36 consonants in Marathi Language, that are-

क ख ग घ ङ
K kh g gh ṅ

च छ ज झ ञ
c chh j jh ñ

ट ठ ड ढ ण
ṭ ṭh ḍ ḍh ṇ

त थ द ध न
t th d dh n

प फ ब भ म
p ph b bh m

य र ल व
y r l w (v)

श ष स
ś ṣ s

ह ळ क्ष ज्ञ
h ḷa kṣa dnya

- There is a contrast between aspirated and unaspirated stops and affricates, including voiced ones, e.g., p—pʰ, t—tʰ, k—kʰ, b—bʰ, d—dʰ, g—gʰ, etc. With a strong puff of air Aspirated Consonants are produced
- There is a contrast between apical and retroflex consonants, e.g., /t/ – /ʈ/, /d/ – /ɖ/, /n/ – /ɳ/, /r/ – /ɽ/. when tip of the tongue touching the roof of the mouth Apical consonants are produced, whereas when the tongue gets curled retroflex consonants are produced, so that the undersurface of tongue get in touch with the upper surface(roof) of the mouth.

Table IV
The consonant

		Bilabial	Alveodental	Retroflex	Post-alveolar/Palatal	Velar	Glottal
Stops	Unaspirated voiceless	pʰ	T	ʈ			
	Aspirated voiceless	pʰ	tʰ	ʈʰ		kʰ	
	Unaspirated voiced			ɖ			
	Aspirated voiced	bʰ	dʰ	ɖʰ		gʰ	
Fricatives	Voiceless				ʃ		
Affricates	Unaspirated voiceless				tʃ		
	Aspirated voiceless				tʃʰ		
	Unaspirated voiced				dʒ		
	Aspirated voiced				dʒʰ		
Nasals				..ɳ	ɲ	ŋ	
Laterals				..ɭ		...	
Flap or Trill				ɽ			
Approximant		ʋ					

- /ʃ/ = *sh* in *shop*
- /tʃ/ = *ch* in *chop*
- /dʒ/ = *j* in *job*
- /ɲ/ =first *n* in *canyon*
- /ŋ/ = *ng* in *song*
- /ʋ/ is often realized as /v/
- /j/ = *y* in *yet*

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

III. PRAAT SOFTWARE

PRAAT is very flexible and free tool to work on acoustic speech signal. It is used to analyze, reconstruct the speech signal. PRAAT offers a various procedures like spectrographic analysis, articulatory synthesis and neural networks to analyze sound signal as in [6].

The following task will be performed using PRAAT-

A. Create a speech object
B. Process a signal
C. Label a waveform
D. General analysis (waveform, intensity, sonogram, pitch, duration)
E. Spectrographic analysis
F. Intensity analysis
G. Pitch analysis
H. Using Long Sound files

IV. SPEECH DATABASE

Speech Databases (SDBs) are collections of language that we spoke, as in [5] it contains-

1) *The speech signal data:* The speech signal is in the form of audio tape, audio CD or an audio file etc which is in a reproducible form.
2) *A symbolic description of the speech signal:* A Symbolic description of speech signal contains phonemic or phonetic transcriptions, prosodic labeling, etc. Generally, it consists of annotation and technical description regarding speech signal. An annotation consists of an orthographic transcription of the speech signal and a lexicon.
3) *Contracts on ownership and rights to use:* This point includes the legal contracts, a copyright notice by the owner of the data and terms under which the Speech Database may be used and distributed.

A. SDB Classification

Speech Database can be described broadly on the basis of Language that we spoke, Number of speakers and utterances, content of the speech material, Type of speech, quality of recording, level of annotation, lexicon and license and terms of the distribution as in [5].

V. SPEECH LEXICON

The lexicon is that the link between the acoustic-level representation and the word sequence output by the speech recognizer it plays the vital role in ASR. The role of the lexicon can be consider in two steps-

A. What words or lexical things are known by the system is specified by the lexicon;
B. Lexicon builds acoustic models for every entry. Lexical design focuses on two important parts – first definition and selection of the vocabulary items and then representation of each pronunciation entry.

Lexicon development has 2 main aspects: output of the recognizer which representing basic units of written language and input of the recognizer which describes the spoken form of the language. To achieve the best performing system, attention must be given to both the parts during development.
In Speech to Text conversion or training the recognizer, lexicon plays an important role. The acoustic and language models, the lexicon, and the search engine are the main components of the recognizer as in [5].

VI. PHONETIC TRANSCRIPTION

The visual representation of speech sound is known as Phonetic Transcription or Phonetic script or phonetic notation. IPA i.e. International Phonetic Alphabet is the most known type of phonetic transcription. IPA uses phonetic alphabet. IPA Primarily based on Latin script. It is declared as a standardized representation of speech sound in written form by IP Association. The IPA represents qualities of speech which are part of lexical sounds such as Phone, syllable, Phonemes etc.

Phonetic transcription is that the task to convert the speech signal into some *symbolic morph* such as symbols that captures the linguistic information sent by the speech signal. Here, we tend to used International Phonetic Alphabet (IPA) symbols in transcription as in [7]. IPA symbols are characterized by speech production mechanism. In addition, it is enriched by different *diacritic* marks which are referred in the exceptional cases as in [7].

Table V
Marathi Language words and their equivalent IPA

Devnagari	Translated in English	International Phonetic Language(IPA)
आज	Today	ɑːʄə
मला	Me	məlɑː
तुला	You	t̪ulɑː
आहे	IS	ɑːɦeː
पाणी	Water	pɑːɳiː

Table VI
Equivalent IPA for Marathi Digits

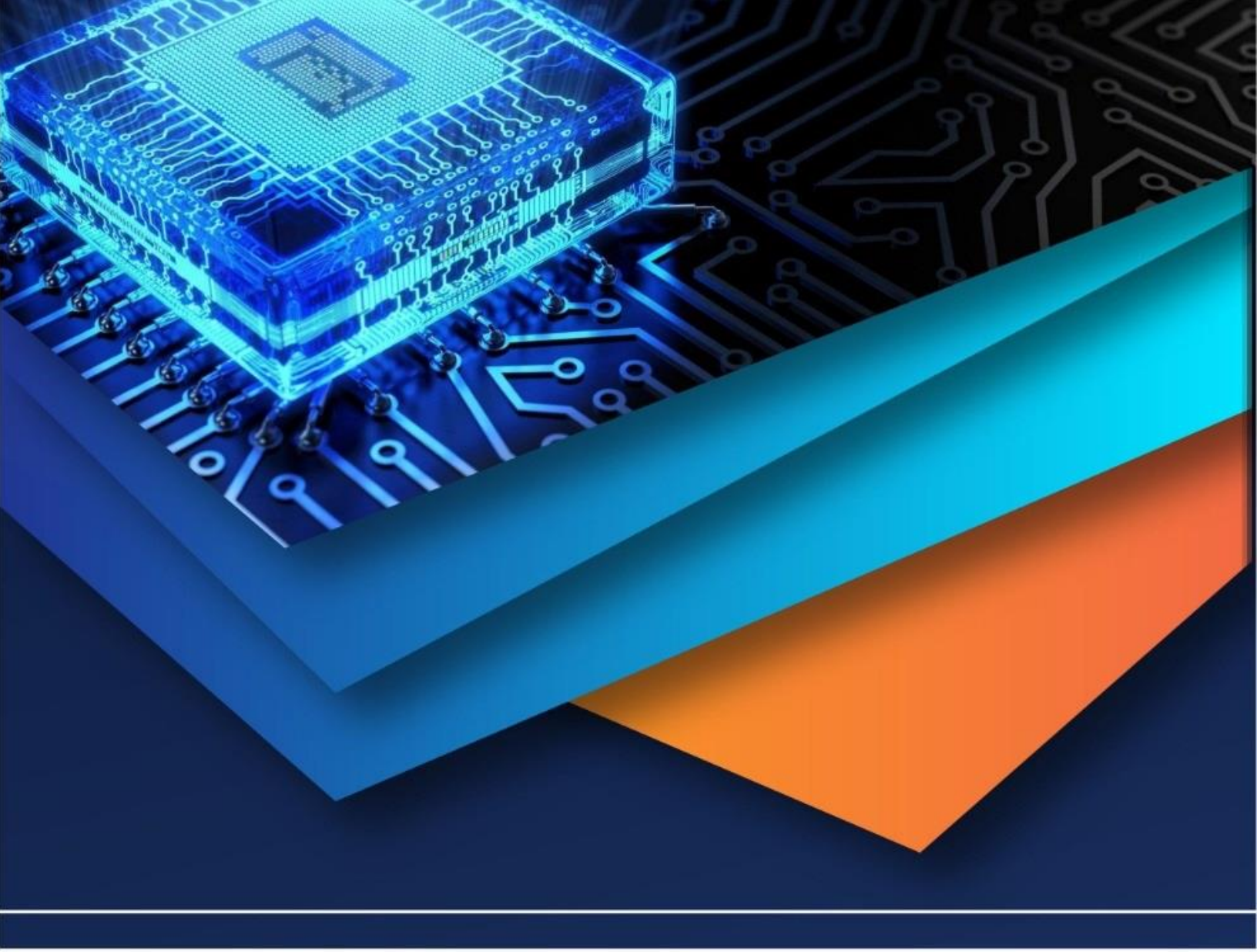
Devnagari Digit	Translated In English	International Phonetic Language(IPA)
एक	One	eːkə
दोन	Two	d̪oːnə
तीन	Three	t̪iːnə
चार	Four	cɑːrə
पाच	Five	pɑːcə
सहा	Six	səɦɑː
सात	Seven	sɑːt̪ə
आठ	Eight	ɑːʈʰə
नऊ	Nine	nəuː
दहा	Ten	d̪əɦɑː

VII. CONCLUSION

In this paper I have covered quite a few basic points of speech recognition. Because we know that speech recognition is proving too much although a lot of work has been done on this, but it is very important to know its fundamentals to move forward. I have included very basic things which are fundamental aspects of speech recognition because every big success starts with basic. Speech recognition has worked to a great extent and sometimes people are using it a lot and are getting lazy day by day. But at the same time its positive impact is very much and we cannot neglect that. To give machines human like power to talk and perceive speech there remains lot of to be learned regarding how structure and meaning in language are encoded in the speech signal and about how this knowledge can be incorporated into usable systems.

REFERENCES

[1] National Research Council 1994. Voice Communication between Humans and Machines. Washington, DC: The National Academies Press. https://doi.org/10.17226/2308
[2] https://science.jrank.org/computer-science/Applied_Linguistics.html
[3] https://web.stanford.edu/dept/linguistics/corpora/material/X_Speech_Corpora.pdf
[4] M.R. Baheti, Bharti W. Gawali, S.C. Mehrotra ,"Marathi_Interactive_Voice_Response_System_IVRS_using_MFCC_and_DTW",International Journal of computer Application, volume-125,No. 15, Sept. 2015
[5] Draxler C. (2000) Speech Databases. In: Van Eynde F., Gibbon D. (eds) Lexicon Development for Speech and Language Processing. Text, Speech and Language Technology. Springer, Dordrecht. https://doi.org/10.1007/978-94-010-9458-0_6
[6] http://www.fon.hum.uva.nl/praat/
[7] K. D. Malde, B. B. Vachhani, M. C. Madhavi, N. H. Chhayani and H. A. Patil, "Development of speech corpora in Gujarati and Marathi for phonetic transcription," 2013 International Conference Oriental COCOSDA held jointly with 2013 Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE), 2013, pp. 1-6, doi: 10.1109/ICSDA.2013.6709865.
[8] https://www.fontconverter.in/index.php?q=Devanagari-to-IPA
[9] https://www.census2011.co.in/census/state/maharashtra.html
[10] https://www.quora.com/How-many-languages-are-there-in-India
[11] https://www.census2011.co.in/census/state/maharashtra.html



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY