HiTEK Multilingual Speech Identification Using Combinatorial Model

Naveenkumar T. Rudrappa(✉) and Mallamma V. Reddy

Department of Computer Science, Rani Channamma University, Vidyasangama, Belagavi, India
trnphd2019@gmail.com

Abstract. Speech is a common form of communication as it expresses the feelings, thoughts, and intentions between human beings either verbally or non-verbally. Our research focuses on verbal communication as India is a language diverse country with more than 19500 spoken languages, considered as mother tongue. The diversity in spoken language understanding leads to Speech Processing. Speech retrieval and translation is a subfield of speech processing by which spoken sentences are recorded, stored and retrieved to identify the languages which is a major challenge in natural language processing. This paper presents MFCC-GNN combinatorial model that includes speech segmentation, morphological analyzer and generator, part of speech tagger for language identification. Multilingual speech dictionary is created and consists of 250 spoken sentences for each language. There are ten most spoken languages in India namely Bengali, Gujarati, Hindi, Kannada, Malayalam, Marathi, Odia, Urdu, Tamil and Telugu. This research considers the identification of multilingual speech particularly for Hindi, Telugu, English and Kannada. Once the language being spoken is identified the future scope is the analysis of Morphological structure for each language and then translation. Translation is conversion of the meaning of a source language speech to a target language speech.

Keywords: Phoneme · Phone · Syllable · Speech processing · Articulatory Phonetic

1 Introduction

Human beings express their ideas, feelings and thoughts to one another orally through the movement of speech organ that modifies the voice into an understandable sound. Speech is produced by the muscle coordination of stomach, chest, neck and head. Speech development is a slow and steady process and it improves over years to produce understandable speech. Communication of speech from man-to-man called spoken languages or from man-to-machine called machine readable low level languages. The 8th Schedule [1] of Constitution has declared 22 official languages namely Nepali, Marathi, Manipuri, Malayalam, Konkani, Kashmiri, Kannada, Hindi, Gujarati, Bengali, Assamese, Oriya, Punjabi, Sanskrit, Sindhi, Tamil, Telugu, Urdu, Bodo, Santhali, Maithili and Dogri. As per 2011 Census Ten most Spoken languages in India [2] with the count are as shown in Table 1:

R. Manza et al. (Eds.): ACVAIT 2022, AISR 176, pp. 286–303, 2023.
https://doi.org/10.2991/978-94-6463-196-8_23

Table 1. Ten most spoken languages in India as per 2011 Census

Sl. no	Language	No of Speakers in Crores
1	Hindi	52.83
2	Bengali	9.72
3	Marathi	8.3
4	Telugu	8.11
5	Tamil	6.9
6	Gujarati	5.54
7	Urdu	5.07
8	Kannada	4.37
9	Odia	3.75
10	Malayalam	3.48

As per 2021 census [3] English is widely spoken across the globe with a count of 1.35 billion people and the default langugae specified by the world wide web is English. Among the above languages the research focuses on four langugaes namely Hindi, Telugu, English and Kannada referred to as HiTEK languages. When users learn a language some skills are essential for complete communication, they are usually 1. Learn to listen 2. to speak 3. to read 4. to write. These are called the four "Communication Skills" which help to communicate between human beings as shown in Fig. 1:

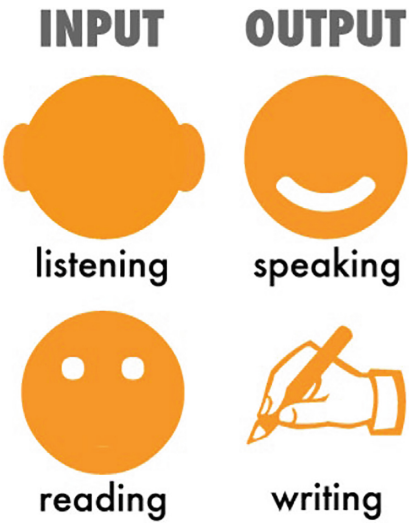

Fig. 1. Skills of Communication

1.1 Listening: It is the act of hearing a language by human ears. It involves identification of speech sounds (letters, stress, rhythm, pauses), a process of converting it into words, sentences and later human brain converts these into messages that convey the meaning.
1.2 Speaking: It is the delivery of language through the mouth. To speak users create sounds using many parts of human organs including the lungs, vocal tract, vocal chords, tongue, teeth and lips.
1.3 Reading: It is a process that involves motivation, recognition of a word, comprehension and fluency. Readers integrate these processes to make meaning from a printed copy.
1.4 Writing: It is the process of production of symbols, alphabets and punctuations in the brain first and then communicates the same thoughts and ideas onto a printed copy [4].

Linguistics is a systematic analysis of language skills which may be written or spoken. It studies the three viewpoints of a language namely language formation, its meaning and the context in which it is used. Phonetics is a study of acoustic and articulatory properties of speech [5]. Spoken speech are categorized into two sets namely set of vowel and set of consonant. Vowel is a sound formed by pronouncing [6] with an open vocal tract and hence there exists no air pressure at any spot over the glottis. On the other hand consonants are the sounds constructed by restricting the vocal tract that reduces flow of air in and out of the lungs. Place of articulation is the point where the airflow is restricted in the vocal tract. Some of the human speech production terminologies [7] are shown in Table 2.

Human speech processing inculcates Combinatorial Models. This research paper focuses on 1. HMM-GMM, 2.HMM-ANN and 3. HMM-DNN combinatorial model for analysis and experimentation on HiTEK languages for better results.

Table 2. Human speech production terminologies

Sl. no	Terminologies	Explanation
1	Respiration	Breathing is the air pressure inside the lungs that helps in human speech production and to control vocal intensity and loudness.
2	Phonation	It is the determination of how voiced sounds are produced.
3	Articulation	It is the action of producing a speech word clearly.
4	Resonance	Sound produced as it goes through the mouth called oral resonance or nose called nasal resonance.
5	Prosody	It reflects the features of speaker utterance that may be a question or a command or the presence of irony or emphasis, contrast and focus, may reflect elements of language not considered by grammar or vocabulary

2 Literature Review

Cuiling [8] proposed that English Speech Recognition system consists of four steps 1. Voice Acquisition 2. Speech modelling 3. Speech Recognizing 4. Results. English language database utilized was Aurora 2. Hidden Markov model (HMM) was applied for four different noisy environments like subway, babble, Car and Exhibition Hall and utilized Computer Assisted Language Learning (CALL).

Chao [9] proposed English speech recognition by searching the most suitable word sequence depending on a segment of English speech utilizing HMM based Semi-Non Parametric method to enhance performance and accuracy. Word sequences are trained and Probabilistic transition frequency profile matrix and average probabilistic emission matrix calculated. He has elaborated on speech recognition for a cross subject involving, digital signal processing pattern recognition, linguistics, acoustics, information theory and optimization theory. He suggested that signal to noise ratio lies in the range of -5 to 20.

Santosh kumar [10] suggests that speech recognition works for Multlingual environment by combining language specific acoustic models. He has used cross language transfer in addition to cross language adaptation for Monolingual system. Training for English and Tamil languages was carried out separately in bilingual system acoustic models. The combinatorial model used decision tree clustering. Experiments conducted demonstrated that acoustic modelling can be carried out on multiple languages. This reduces computational cost on the search engine because we utilize one acoustic model for multiple languages.

Ling [11] studied and concluded that DNN-HMM is superior than GMM-HMM method. 40 MFCC features were extracted and the tool used was Kaldi toolkit. Signal preprocessing consists of premphasis, subframe windowing and end point detection. In Speech recognition Deep Neural Network is utilized for training the acoustic model. The input to DNN are the acoustic characteristics of current frame for the calculation of each possible HMM state. HMM for Speech Recognition is one way, from left to right, self-ring and can be spanned topologically. HMM works on a composition of multiple phonemes, words and silences. Parameter clustering is carried out using top down method and decision tree. Cool Edit software utilizes a frequency of 16 kHz for sampling using 16 bit encoding.

Trivedi [12] elaborated on the types of Speech, Speech Recognition, S2T conversion, T2S conversion and Speech Translation. Dynamic Time Warping models and HM Models with neural network perform well for classifying of phonemes, recognition of isolated words recognition and adaptation of speaker. Synthesizing of speech performs well for conversion of tokenized words to artificial human speech. Speech production components are phonation, fluency, intonation, pitch variance and respiration. Speech recognition system classification can be performed on the basis of speaker dependency, vocal sound and vocabulary. Commonly used feature extraction methods are Linear Predictive Coding (LPC), Mel- Frequency Cepstrum Coefficients (MFCC) and Dynamic Time Warping. Various pattern classification methods used are template based, Knowledge based, Neural network based and statistical based. The methods utilized Hidden Markov Model and ANN based Cuckoo Search Optimization for S2T conversion. T2S conversion involves processing of text, various speech synthesis techniques namely

Articulatory, Formant and Concatenative. Some of the language translation models are Rule Based, Statistical, Example Based and hybrid machine translation.

Kumar and Aggarwal [13] have proposed Hindi language continuos Automatic Speech Recognition (ASR) system utilizing Recurrent Neural Network (RNN) based Language Modelling (RNN-LM) which uses Maximum Likelihood Linear Regression (MLLR) with Constrained Maximum Likelihood Linear Regression (C-MLLR) by training with Maximum Mutual Information (MMI) and Minimum Phone Error (MPE) methodologies with Two Fifty Six Gaussian Mixture per Hidden Markov Model (HMM) state.

Gopal [14] proposed K-Means clustering algorithm and logistic regression to improve accuracy. Noise reduction was performed using Butterworth low pass filters. Recognition of Hindi Speech utilized Selected Time Delay Neural Network (STDNN) and modeling of acoustics was carried out with i-vector adaptation. Hindi syllables have longer units of acoustics, faster decoding due to reduction in contextual effects and irregularities caused due to phonemes. K-Means clustering is used for segregation of inaudible low quality audio and hence detect human voice and silences.

Jewani [15] talks about speaker dependent and speaker independent models, types of Hindi speech like connected words, isolated words, continuos speech and spontaneous speech and proposes whole word matching and subword matching techniques using Mel-Frequency Cepstrum (MFC) and HMM. MFC and distance minimum algorithm can be combined to improve overall efficiency. Dynamic Time Warping for speech pattern comparison.

Shobha and Anurag [16] have proposed improvement of HMM using hybridization of units like Phones, Syllables which are the acoustic units to improve nasal sounds and Domain Syntactic specific structures that reduce the search space of the recognizer and hence improve performance and are tested for both Speaker Dependent and Speaker Independent Systems.

Sharada and Vijaya [17] have elaborated on Kannada Speech Recognition using tri-state Hidden Markov Model with each state represented by Gaussian Mixture Model. Three approaches for Kannada speech recognition have been identified i.e. Acoustic Phonetics, Pattern Recognition and Artificial Intelligence. They are of the opinion that speech is context dependent and the occurring of a phoneme is dependent on preceding and succeeding phonemes which lead to the development of triphone clustering model. MFCC represents speech parameters better, DTW and HMM are best classifier methods and Viterbi search algorithm is better for pattern matching. Specific language models predict the occurrence of words one after another, which helps to narrow down the search process using Unigram (Normal Search), Bigram (gives statistics of occurrence of words given previous words), Trigram depends on two previous words. Acoustic models represent each distinct sound that make up a word.

Hemakumar and Punitha [18] have elaborated on speech signal segmentation by decomposing a signal into basic phonetic units like phoneme, syllable and subword. Proposed method consists of pre-processing stage, detection of voiced section, feature extraction, model building and testing of an unknown signal.

Prashanth and Ananthakrishna [19] have emphasized on Maximum a Posteriori (MAP) and Gaussian Probability Density Function (GPDF). Baum Welch Forward-Backward algorithm is used for training.

Akhila and Kumarswamy [20] have justified that phoneme level search is effective for searching words/phrases. DBN is used and 16 MFCC features extracted from each speech frame. Conventional acoustic modelling techniques like Multilayer Feed Forward Neural Network (MFFNN) and Support Vector Machines (SVM) utilized. They concluded that performance of any network is affected by the size of phonemes used for training and testing.

Anand and Jangamashetti [21] have focused on speech signal preprocessing to frames and then extracting features using Linear Predictive Coding. MFCC and Euclidean distance is used for isolated word recognition in the first case. MFCC with SVM classifier was used to remove silence in the second case. Gaussian Multivariate Model was utilized in the recognition of an unknown phoneme. Confusion Matrix inferred the performance of classifiers.

Priya and Soumya [22] have used MFCC to extract features and using HMM with triphone acoustic modelling. Baum welch algorithm was utilized for model reestimation to obtain good results in offline recognition mode. They have derived 39 cepstral parameters from speech signal.

Pradeep and Srinivasa [23] have performed a comparative analysis of speech recognition using HMM-GMM, ANN, DNN for various recording modes like reading, lecturing and conversation.

Praveen and Neerudu [24] have recognized Telugu speech speaker independent data using Teager energy operator Delta Spectral Cepstral Coefficients (TDSCC) which is a feature extraction technique and Deep Neural Networks (DNN) feature classification technique. Isolated speech recognition performed using 2 stage Deep Learning Neural Network (2DNN). Stressed speech can be recognized by (TDSCC). Recorded speech consisting of noise is preprocessed using Computation Auditory Scene Analysis (CASA). Artificial Neural Network and Deep Neural Network are used for feature classification.

Jeetendra [25] analyzes speech through signal processing and linguistic processing. Linear Predictive Coding (LPC) and Cepstral Analysis are used for feature extraction of Telugu language and to design speaker independent system. Discrete Fourier Transform (DFT) and Fast Fourier Transform (FFT) is used for calculations. Linguistic processing involves conversion from speech to text or generate speech from text. The basic units involved are allophones string of phonemes and set of string of phonemes called morphophonemes. Morphophonemes are matched with words in the dictionary or various prefixes/suffixes. Wide band spectogram and Narrow band spectrogram are used for the analysis of speech.

Kodali [26] processed continuos speech using open source speech recognition and Kaldi tool kit Static Vector Machine(SVM) and Binary Static Vector Machine(BSVM) were used for Automatic Speech Recognition (ASR) that included Language Models(LM) and Acoustic Model(AM).Training process included Monophone Hidden Markov Model (HMM) for training, aligning training dataset using monophone model

and triphone HMM training. Sentimental Analysis was carried out by identifying positive, negative and neutral conversations. Categorized noise into five types namely cough, laugh, noise, breath and background noise. Developed a Multi-Modal that consists of both speech and text.

Sunitha and Kalyani [27] proposed a model that includes five phases namely syllable extraction, building a tri state model for each syllable, a Trie structure through morphological analysis of Telugu language, marking rough boundary of the syllable and syllable recognition. Morphological analysis is carried out by removing prefix, suffix, infix or crucifix from the stem and identification of inflectional and derivational words. Speech segmentation is carried out using linguistic rules. Syllable recognition is carried out using Mahalanobi's distance measure. Trie structure places all words with common prefix under the same path.

Praveen and Ratnadeep [28] focused on continuos speech recognition in two modes namely speaker dependent and speaker independent systems using Melfrequency Cepstral Coefficients (MFCC), Discrete Wavelet Packet Decomposition (DWPD) and Discrete Wavelet Transformation (DWT) for noise removal. Features were classified through Hidden Markov Model (HMM) and Deep Neural Networks (DNN). Word based model was used to recognize continuos data. Vitteri algorithm was used for recognition. Feature classification could be performed through pattern recognition, vector analysis and Artificial Neural Networks.

3 Challenges

Language identification by a human involves listening to spoken speech by another human called as man-man communication, analyze the vocal transcription in the neurons of human brain and then decide the language spoken by the other user. In comparison, language identification by an electronic machine a computer is still more complex as the machine should be trained with different language datasets to identify the spoken speech appropriately. This is performed by given a possible set of languages, their rules and names machine applies classification and comparison techniques so that the exact language is identified by the machine. System accepts input speech and then classifies the language into its predefined class. Hence language identification is a classification problem of data mining. Some of the challenges involved in the identification of spoken speech languages are as shown in Table 3.

Language identification by a machine involves creation of a hybrid HiTEK speech/text dictionary. This involves recording of human spoken sentences in multiple languages by a number of users using a microphone for input and further transforming and storing these recordings in a wave file format in the back end file system. This recorded file should be free from different background noise which poses a major challenge in building a speech dictionary which are detailed as shown in Table 4.

Table 3. Challenges in Speech Identification of HiTEK Languages

Sl. no	Challenges	Explanation
1	Human Speaker Characteristics	Each person has a different set of vocal characteristic features and hence feature extraction for speaker independent speech recognition systems is difficult.
2	Spoken Accent	Each person has a different accent of speaking and hence pattern matching process needs calibration as the process should take into account the non-linear nature of spoken words. Ex: Person A: Hi- 2 s, Person B: Hai- 3 s
3	Linguistic Variation	Each language has a different linguistic pattern and hence difficult to design a generalized model.
4	Sandhi Rules	Different for Hindi, Telugu, Kannada and English has no Sandhi rules but has comparative and superlative degrees for a spoken word.
5	Acoustic characteristics	A thorough knowledge of phonetic units should be known for each language.
6	Syllable Structure	Different and complex for each language.
7	Speech Segmentation	Segments created after segmentation process are of variable size.
8	Noise Removal	Regular Noise: Fan Rotation in the background.
		Irregular Noise: Cough, Laugh, Breath, Wind Speed, and Vehicle Horn.
		An efficient speech recognition system should remove both regular and irregular noise from the recorded speech.
9	Speech Types	1. Isolated Speech: Individual spoken words analyzed, Highest Matching Accuracy.
		2. Connected speech: Combination of two or more isolated words that run with a slight pause. Medium matching accuracy due to the problem of incorrect segmentation at the word boundaries.

(*continued*)

Table 3. (*continued*)

Sl. no	Challenges	Explanation
		3. Continuos Speech: Human speaks continuously without any gap. Lower matching accuracy due to the problem of incorrect segmentation at the word boundaries as there are no pause and silence between words.
		4. Spontaneous Speech: Human speaks without a written script and hence there are irregular silences, pauses, cough, laugh etc. due to the search of words through intelligence in human brain. Least matching accuracy as the spoken words are unpredictable.
10	Pattern Recognition Techniques	Use indirect methods and hence more time consuming.
11	Data set size	1. Small: Highest Accuracy
		2. Medium: Acceptable Accuracy.
		3. Large: Acceptable accuracy with maximum time for retrieval.

4 Methodology

Multilingual Speech Identification (MLSI) is a process of identifying multiple languages in the pre-recorded speech file. Complexity of MLSI lies in mapping and translation of speech/phone/textual-word with supervised HiTEK dictionary. To improvise matching across various spoken languages it inculcates translation rules for specified languages such as of Hindi [29], Telugu [30], English [31] and Kannada [32].

This research utilizes a combinatorial model consisting of 1. HMM-GMM, 2.HMM-ANN and 3. HMM-DNN for improving speech matching accuracy. The characteristic features of the three models are elaborated as shown.

4.1 Hidden Markov Model- Gaussian Mixture Model

This model is dependent on phoneme recognition while emission distribution is modeled using GMM. Readings are obtained by calculating mean and covariance in Gaussian Mixture. HMM-GMM increases the probability of fetching a sequence of phonemes.

4.2 Hidden Markov Model- Artificial Neural Networks

HM Model is used to obtain the probability of the data under observation for an HMM state that corresponds to a specific sound. ANN training produces posterior probabilities of HM Model state given the speech data.

Table 4. Challenges for building Speech Dictionary

Sl. no	Challenges	Explanation
1	Speech Clarity	Speaking in a way that could be clearly understood by the listeners.
2	Speech Projectivity	Speaking aloud such that every listener can hear the utterance.
3	Speech Enunciation	Clearly pronouncing each syllable with exact emphasis.
4	Speech Pronouncing	Proper word utterance.
5	Expression	Speaking with vocal variation so that listeners are engaged and interested.
6	Speaking Pace	Uttering at a rate that could be clearly heard by the listeners.
7	Filler	Using words that distract listeners. Ex: “Um”, “ah” and “you know”, “nothing but”
8	Slang	Language understood by a specific group. A listener not part of that group cannot understand the meaning.
9	Buzzword	Frequently used word in a specific context. Ex. “game changer” and “think outside the box” etc.
10	Acronym	Abbreviations used for some phrases CIO: Chief Innovative Officer.
11	Active Listener (Hardware)	Microphone used in recording should be active for the duration till the speaker clicks on stop recording button.
12	Human Speaker Location	Maintaining a correct distance for recording from the microphone.

4.3 Hidden Markov Model- Deep Neural Networks

HM Model is used for phoneme recognition while training of a DNN is carried out in 2 phases as follows:

Phase 1- Unsupervised Pretraining
Phase 2- Supervised fine tuning

MLSI imbibes a Hybrid/HiTEK dictionary consisting of multi lingual Speech/Text. The identification of multilingual speech processing is shown in Fig. 2 and the detailed steps are elaborated below.

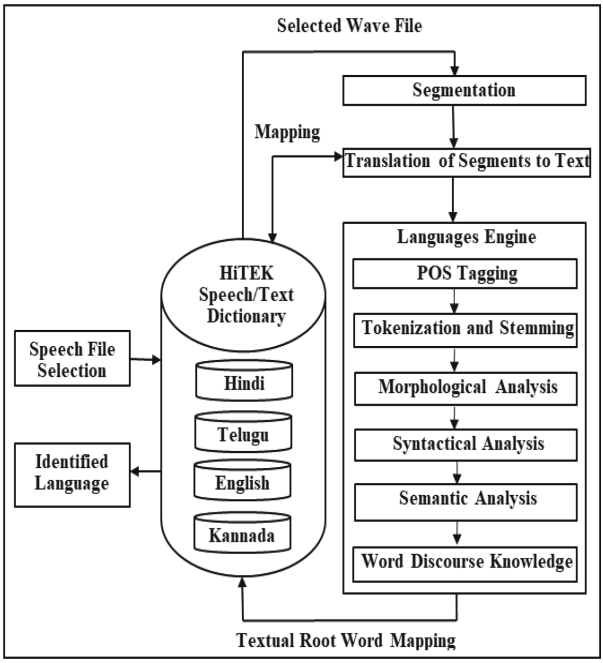

Fig. 2. Block Diagram of Language Identification

Step 1: Speech File Selection Module
The first step is to select a pre-recorded input speech file from Hybrid/HiTEK dictionary. Pronounces out the selected wave file by the activation of speaker and displaying its path.

Step 2: Hybrid/HiTEK dictionary Module
Stores pre-recorded sentences as audio files and textual words.

Step 3: Segmentation Module
This module identifies the boundaries between words, syllables, phones or phonemes in spoken languages to give a proper meaning for a word or a sentence. The challenge during segmentation is to detect boundaries by analyzing the pause between phones which is a minimum one second in our architecture.

Step 4: Translation of Segments to Text Module
The split phones are mapped with the corresponding textual words and if the words are not present in the dictionary transliterated to form the text.

Step 5: Language Engine Module
This is the heart of language identification system and it consists of various steps involved in identifying a language.

4.4 POS Tagging

It is a process of annotating a word to a specific part of speech based on the context, relation with adjacent words within a paragraph, sentence or phrase. Ex In English Vocabulary POS consist of Noun Pronoun, Verb, Adverb, Adjective, Conjunction, Preposition etc.

4.5 Tokenization and Stemming

1. Tokenization is a task of dividing a textual sentence into a predefined set of tokens. It may also break the text on whitespace characters such as a space, tab, or punctuation. Ex: A sentence can be divided into words and a paragraph can be divided into sentences, here words and sentences act as tokens.
2. Stemming is extraction of root word. Ex Making is converted to Make by removing the suffix ing and attaching e as suffix.

4.6 Morphological Analysis

Morphology is the process of formation of words from the smallest primitive chunks by finding a meaningful sub part within the word. These sub-words are called Morphemes as shown in Table 5.

4.7 Syntactical Analysis

Syntactical analysis checks for words, grammar in a sentence and their arrangement among the words by applying grammatical rules. These rules vary from language to language.

4.8 Semantic Analysis

It is an algorithmic activity to understand, merge and analyze the meaning of words, integrate words to form phrases and unite phrases to form a well-structured sentence. This is the most important phase in the full Language Engine as it checks whether the meaning of an input statement is valid in real world as the statement may be syntactically correct but semantically to be validated by the engine as shown in Table 6.

Table 5. Word Conversion to Morpheme

Sl. No	Word	Chunks	Morpheme Count
1	Like	Nil	1
2	Unbreakable	Un + break + able	3

Table 6. Syntactical and Semantical Validity of a Statement

Sl. No	Textual Statement	Syntactical Validity	Semantical Validity in Real World
1	Cat eats Rat	Correct	Valid
2	Rat eats Cat	Correct	Invalid

4.9 Word Discourse Knowledge

The meaning of one sentence depends upon other sentences or may also depend on the immediate succeeding sentence. Ex: She needed it depends on previous discourse context.

Step 6: Textual Root word mapping in HiTEK Speech/Text Dictionary
The obtained root words from the language engine are mapped to an appropriate language dictionary.

Step 7: Identified Language Module
It displays the language of identified sentence if all words are in the same language dictionary or it displays each individual word and its language if the sentence is made up of more than one language.

5 Experimental Setup and Results

Graphical user interface based application is developed with python programming as a front end and file system as back end. The system gives faster results in case of higher hardware system configurations in our case Intel CORE I5 10TH GEN processor and extended RAM Size of twenty Giga Bytes and a file system of One Terra Byte hard disk to store HiTEK pre-recorded Speech/Text dictionary. The expected results are displayed in the last line of each screenshot as shown in Figs. 3, 4, 5 and 6 that indicates the language spoken in a stored wav file which is obtained after pre-training the system using the ISO-639 standard library fused with combinatorial model and stored in a HiTEK Speech/Text dictionary consisting of Hindi, Telugu, English and Kannada languages.

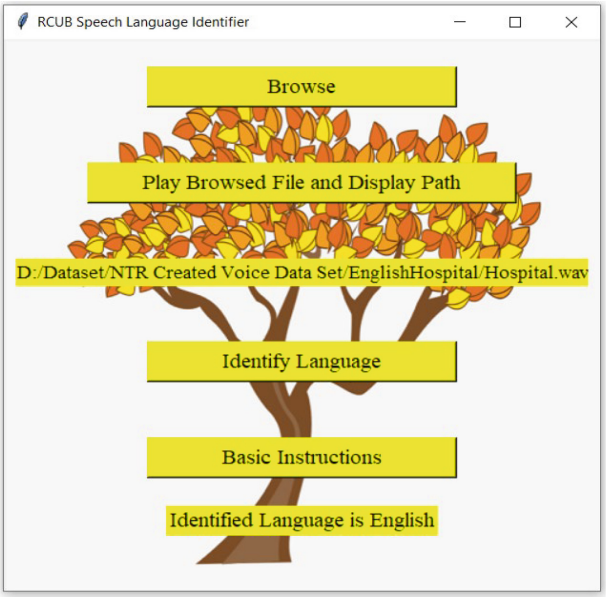

Fig. 3. English Speech Language Identification

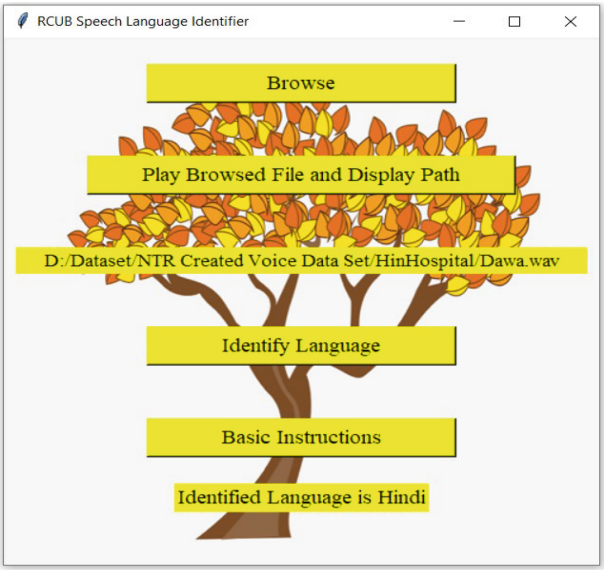

Fig. 4. Hindi Speech Language Identification

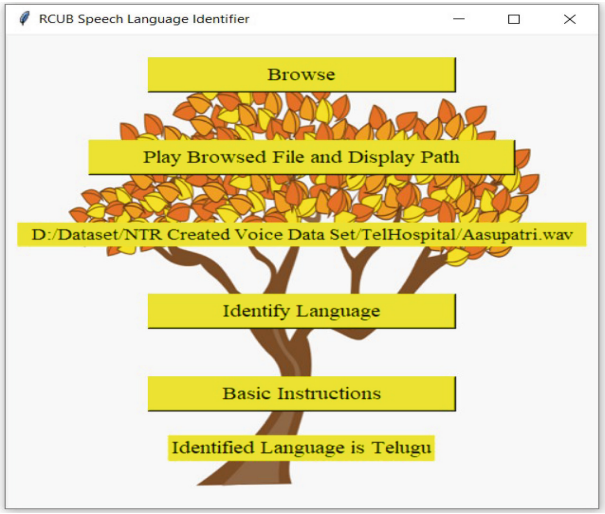

Fig. 5. Telugu Speech Language Identification

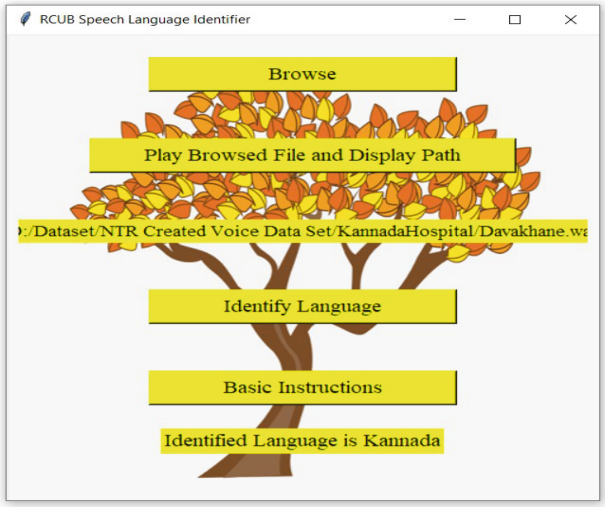

Fig. 6. Kannada Speech Language Identification

6 Conclusion and Future Scope

Variations in speaking accents, speech pronunciation, high dimensional speech feature parameters, computational and evaluation complexity and large speech dataset mandate the need for high end hardware and softwares to train the system to identify the language spoken. This paper focuses on a combinatorial model which is an outcome of combination of HMM-GMM, HNN-ANN and HMM-DNN to convert a speech signal into digital format after extracting various features and then converting to words. We are developing

a system that identifies the content of spoken speech and identifies the language of pre-recorded and stored wav file which may be Hindi, Telugu, English or Kannada. Future scope lies in the fact that the system can be used to identify other regional languages of India.

Acknowledgment. Authors thank Rani Channamma University, Belagavi Karnataka for their support to issue a separate lab for Research Scholars. Authors would like to thank NFST wing, Ministry of Tribal Affairs for selecting me as a Research Fellow and providing me the necessary fellowship grant. A special thanks to Dr. Mallamma V. Reddy for her constant support, mentoring and guidance. Special thanks to the entire family of Department of Computer Science, Rani Channamma University, Vidyasangama, Belagavi, Karnataka. I also thank my family members, research scholars and friends for their motivation, moral support and encouragement. Finally the authors would like to thank Dr. Babasaheb Ambedkar Marathwada University, Aurangabad (MS) India for providing the publication support.

References

1. https://indianexpress.com/article/india/more-than-19500-mother-tongues-spoken-in-india-census-5241056/
2. https://www.jagranjosh.com/general-knowledge/most-spoken-languages-in-india-by-number-of-speakers-1541764100-1
3. https://www.statista.com/statistics/266808/the-most-spoken-languages-worldwide/
4. https://www.englishclub.com/learn-english/language-skills.htm
5. Jakobson, Roman, Gunnar Fant, and Morris Halle. "Preliminaries to Speech Analysis: The Distinctive Features and their Correlates", MIT Press. 1976
6. Kingston, John. "The Phonetics-Phonology Interface", in the Cambridge Handbook of Phonology (ed. Paul DeLacy), Cambridge University Press. 2007
7. Neeshali R. Nandarge, Mallamma V. Reddy, Suman Gouda,Gayatri Patil, "Kannada Phonetic Transcription: NLP," Proceedings of 35th IRF International Conference, Bengaluru, India, 2017, pp. 19–21
8. L. Cuiling, "English Speech Recognition Method Based on Hidden Markov Model," 2016 International Conference on Smart Grid and Electrical Automation (ICSGEA), 2016, pp. 94-97, https://doi.org/10.1109/ICSGEA.2016.63.
9. C. Xue, "A Novel English Speech Recognition Approach Based on Hidden Markov Model," 2018 International Conference on Virtual Reality and Intelligent Systems (ICVRIS), 2018, pp. 1-4, https://doi.org/10.1109/ICVRIS.2018.00009.
10. C. S. Kumar and Foo Say Wei, "A bilingual speech recognition system for English and Tamil," Fourth International Conference on Information, Communications and Signal Processing, 2003 and the Fourth Pacific Rim Conference on Multimedia. Proceedings of the 2003 Joint, 2003, pp. 1641–1644 vol.3, https://doi.org/10.1109/ICICS.2003.1292746.
11. Z. Ling, "An Acoustic Model for English Speech Recognition Based on Deep Learning," 2019 11th International Conference on Measuring Technology and Mechatronics Automation (ICMTMA), 2019, pp. 610–614, https://doi.org/10.1109/ICMTMA.2019.00140.
12. A Trivedi, N Pant, P Shah, S Sonik and S Agrawal, "Speech to text and text to speech recognition systems-Areview", IOSR Journal of Computer Engineering, 2018, e-ISSN: 2278–0661,p-ISSN: 2278–8727, Vol. 20, Iss. 2,Ver. I, pp 36–43, www.iosrjournals.org
13. Kumar, A and Aggarwal, "Discriminatively trained continuous Hindi speech recognition using integrated acoustic features and recurrent neural network language modeling". Journal of Intelligent Systems. 2021; Vol: 30(1), pp 165–179, https://doi.org/10.1515/jisys-2018-0417

14. Anuj Gopal, "Automated Recognition of Hindi word Audio clips for Indian children using Clustering-Based Filters and Binary Classsifier", Proceedings of The Fourth International Conference on Natural Language and Speech Processing, Trento, Italy, Publisher Association for Computationl Linguistics 2021,pp 204–208
15. Kajal J , Shreesh Rao, Prashant D, Ronit D and Mrudali B, " Hindi Speech Recognition " International Journal of Advanced Science and Engineering", 2018, Vol 7, Issue No 1, pp-50–55, Available online at www.ijarse.com, ISSN- 2319–8354
16. Bhatt, Shobha & Jain, Anurag & Dev, Amita, "Monophone-based connected word Hindi speech recognition improvement". Journal Sadhana Indian Academy of Sciences (2021) 46: 99, https://doi.org/10.1007/s12046-021-01614-3
17. S. C. Sajjan and Vijaya C, "Continuous Speech Recognition of Kannada language using triphone modeling," 2016 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET), 2016, pp. 451–455, https://doi.org/10.1109/WiSPNET.2016.7566174.
18. P. Punitha and G. Hemakumar, "Speaker Dependent Continuous Kannada Speech Recognition Using HMM," 2014 International Conference on Intelligent Computing Applications, 2014, pp. 402-405, https://doi.org/10.1109/ICICA.2014.88.
19. P. Kannadaguli and A. Thalengala, "Phoneme modeling for speech recognition in Kannada using Hidden Markov Model," 2015 IEEE International Conference on Signal Processing, Informatics, Communication and Energy Systems (SPICES), 2015, pp.1–5, https://doi.org/10.1109/SPICES.2015.7091382.
20. Akhila K S and R. Kumaraswamy, "Comparative analysis of Kannada phoneme recognition using different classifiers," 2015 International Conference on Trends in Automation, Communications and Computing Technology (I-TACT-15), 2015, pp. 1–6, https://doi.org/10.1109/ITACT.2015.7492683.
21. A. H. Unnibhavi and D. S. Jangamshetti, "LPC based speech recognition for Kannada vowels," 2017 International Conference on Electrical, Electronics, Communication, Computer, and Optimization Techniques (ICEECCOT), 2017, pp. 1–4, https://doi.org/10.1109/ICEECCOT.2017.8284582.
22. K. Jeeva Priya, S. S. Sree, T. Navya and D. Gupta, "Implementation of Phonetic Level Speech Recognition in Kannada Using HTK," 2018 International Conference on Communication and Signal Processing (ICCSP), 2018, pp. 0082–0085, https://doi.org/10.1109/ICCSP.2018.8524192.
23. R. Pradeep and K. S. Rao, "Deep neural networks for kannada phoneme recognition," 2016 Ninth International Conference on Contemporary Computing (IC3), 2016, pp. 1–6, https://doi.org/10.1109/IC3.2016.7880202.
24. A. P. Kumar, N.U. Maheshwari, Y.Sangeetha and P. Jyothi, " Isolated Telugu Speech Recognition On T-DSCC And DNN Techniques", International Journal of Innovative Technology and Exploring Engineering (IJITEE), 2019, Vol.8, pp. 3419–3422 ISSN:2278–3075, https://doi.org/10.35940/ijitee.K2544.09811119
25. P. Jeethendra, M. Chandrashekar "Linear Predictive Coding and Cepstral Analysis for Telugu Speech Recognition". International Journal of Computer Trends and Technology (IJCTT) V47(1):50–60, May 2017. ISSN:2231–2803. www.ijcttjournal.org. Published by Seventh Sense Research Group.
26. Rohith Gowtham Kodali, 2Durga Prasad Manukonda, 3Rajaraman Sundararajan, Speech and Text Based Analytics in Telugu Language, © 2019 JETIR March 2019, Volume 6, Issue 3 www.jetir.org (ISSN-2349-5162)
27. Dr. K V N Sunitha and N Kalyani. Article: Isolated Word Recognition using Morph Knowledge for Telugu Language. International Journal of Computer Applications 38(12):47–54, February 2012. https://doi.org/10.5120/4765-6940

28. Archek Praveen Kumar, Ratnadeep Roy, Sanyog Rawat and Prathibha Sudhakaran,"Continuous Telugu Speech Recognition through Combined Feature Extraction by MFCC and DWPD Using HMM based DNN Techniques", International Journal of Pure and Applied Mathematics, Volume 114 No. 11 2017, 187–197 ISSN: 1311–8080 (printed version); ISSN: 1314–3395 (on-line version)
29. https://www.optilingo.com/blog/hindi/everything-about-hindi-language/
30. https://omniglot.com/writing/telugu.htm
31. https://www.englishmirror.com/englishgrammar/vowels-and-consonants.html
32. https://omniglot.com/writing/kannada.htm