IJCSE International Journal of Computer Sciences and Engineering Open Access

Research Paper Vol.-6, Issue-12, Dec 2018 E-ISSN: 2347-2693

Identification System for Different Punjabi Dialects Using Random Forest Technique

Ravinder Singh[1*], Anand Sharma[2]

[1,2]University College of Computer Application, Guru Kashi University, Talwandi Sabo, Bathinda, India

[*]*Corresponding Author: ravndr94@gmail.com*

Available online at: www.ijcseonline.org

Accepted: 07/Dec/2018, Published: 31/Dec/2018

Abstract— In modern era of technology every one relies on technology. From start of day to end of day humans depends on machines and machines need input signal for performing tasks. Many systems have been developed which works on native language input speech. Punjabi is also one of them, there are many speeches and dialect recognition systems are available but all have some common problems like problem with different dialects words of Punjabi is main one. In Punjabi language Majha, Malwa, Doaba are main dialect in eastern Punjab, most of time words from Majha dialect is similar to Taksali Punjabi but when we talk in Doaba and most populated dialect Malwa it is difficult for speech recognition system to understand that word and perform tasks so that was whey dialect identification system is need of hour. The aim of this paper is discuss about new proposed algorithm by authors which works on Punjabi dialects and to compare with previous algorithms with respect to accuracy.

Keywords— Dialect, Natural Language Processing, Taksali Punjabi, Automatic Speech Recognition, Artificial Neural Networks, Random forest.

I. INTRODUCTION

Punjabi language is at 12th rank from other spoken language in world and speakers of Punjabi language are distribute in Europe , North America, Australia but mainly from Asia (India and Pakistan). Punjabi speaking style or dialects are very because of religious communities and geographic distribution. A famous myth in Punjab is that "**Language Change every half mile**". In context of Punjab there are three main dialects named as Majha, Malwa and Doaba. Majha dialect is standard dialect of Punjabi language and spoken by more than half population who speak Punjabi language include western Punjab and eastern Punjab. The Majha region spreads in Punjab (Pakistan) and in Amritsar, Gurdaspur and Tarntaran .Malwai Dialect is spoken by most people in eastern Punjab. Main area of Malwa are Ferozpur, Muktsar, Faridkot, Bathinda, Barnala, Mansa, Patiala, Sangrur, Fatehgarh, Ropar, Mohali and in some area of Rajasthan and Haryana. Doaba dialect is spoken between rivers Beas and Satluj. The main areas under Doabi are Jalandhar, Nawanshahr, Kapurthala and Hoshiarpur. Every dialect have different words for representing same feeling and second term effect of other language like Hindi, English, Urdu, Persian, French, German and etc. These two factors make vocabulary of Punjabi Language more complex because integration of words from different language and there is no proper spelling for these words in Punjabi are available. There is lack of dialect identification system from natural speaking of user for Punjabi and other indo Aryan languages. Dialect identification system will help in many real time situation for identify a person."**The process of enabling a computer to identify and respond to sound produced in human speech**" definition of speech recognition by oxford. Speech recognition is part of Natural language Processing which is basically method of human and computer communication. NLP is divided into two areas named Natural language Understanding which require a lot of knowledge about human language and manipulate it. Second Natural Language Generation which convert information from database into human understandable language. The main aim of NLP is to train machines using appropriate tools and methods so they can understand Natural Language and perform desired tasks. NLP used in many specific areas like linguistics mathematics, robotics, electronics engineering and psychology. Speech recognition is also specific area of NLP in which machines works on human speech input. Because of easy communication with machines using speech rather than other medium of communication research and development on these system is increase day by day. Speech Recognition system classify into two types named speaker independent are those in which a person involve in training may or may not involve in testing and there is no effect on system and in dependent system a person involve in training must be part of testing otherwise system will not work as desired. Another classification of Speech Recognition system is speech to text system in which input signal is converted to text output, second speech to signal in which input speech works as instruction for machines to perform tasks and last speech to speech system in this input speech is filtered or refined and produce output speech. Speech recognition system consists of two parts named as system training and system testing. System training is training of system against particular words, sentences etc. System training is important part because accuracy is depends on it. System testing where accuracy of system checked against input speech with feature store in database in early phase. Speech recognition system mainly used for ease communication with machines and in authentication systems. Main example of speech recognition system which we see commonly in mobiles is "**Google Voice Search**" which makes Google suffering so easy then text input. Build accurate system for speech is complex and costly process because number of language present in world. Currently 6909 languages are being used in world and every language has thousands of words which make speech system complex and less accurate. English is an international language approximate 20% population of world understands and speaks English, but another complexity of different people from different countries has particular words and tones to represent feeling which also make speech recognition system complex and inaccurate. Not just English every language face this problem and Punjabi also suffer from this problem. In proposed system author will solve out some part of problem for Punjabi language.

II. Related Work

The [1] dialect conversion system has been developed for converting Punjabi input text from one dialect to another dialect with accuracy of 97% with 5600 words approximately. The system firstly recognizes the words from the given input parse these into individual words and translate them to dialectal form of Punjabi. The conversion engine plays a main role in translation process which is based on both bilingual dictionaries and morphological transfer rules. This research could help to increase the resources relating to dialects of Punjabi and processing tools for processing them. The results of this research work for the inter-conversion of Malwai, Majhi and Doabi dialects of Punjabi are quite promising. In future, system can be extended to handle more dialects. Also, the research can be accomplished to achieve the automatic conversion of various dialects of Punjabi. The system can be furnished to learn new transfer rules from data and work automatically to convert them. Singh [2] developed a rule based system for converting Punjabi text from one dialect to another dialect. The accuracy of system was checked 850 sentences and 30 produce wrong recognition so accuracy of system was 98.5%. According to author reason of errors are improper nouns and transfer rules mismatching. Overall result of first dialect conversion system is quite promising. The system takes input texts identify it and then translate to Dialectal Punjabi. In [3] conversion of text written in Punjabi text into its Malwai dialect and Doabi dialect equivalents using rule based approach. The developed system identifies the words from the input text, splits these into individual segments and then converts them to its equivalent Punjabi Dialect text. The development of rules is not an easy task as dialect form originates from pronunciation by native speakers and there are very less linguistic resources available from which the conversion rules are developed. Better results are achieved by increasing the size of the training data. The results square measure quite promising and show the success of 1st non-standard speech conversion system for Punjabi language. The accuracy of system was recorded on 12000 words for Malwai and Doabi, 95% accuracy for Malwai words and 94% for Doabi words. In [4] comparative study for the development of machine translation system from standard Punjabi to Malwai dialect. Since Standard Punjabi and Malwai are closely related languages so direct approach of Machine Translation shall be used. The Challenge for the development of Standard Punjabi-Malwai Machine Translation system is lack of language resources. The Malwai dialect is either used in the spoken communication or used by some of the Punjabi writers in their novels, Natak's. So it's a challenging task to collect the data and prepare a machine readable bilingual dictionary. In [5] proposed system conversion process is done by morphological analyzer and conversion engine which are main components of system. First component identify and segment the word and second convert the word using bilingual dictionary that contain rules for conversion. Dictionary contains 3247 words and 85 rules. The accuracy of this system is measured on 8460 words with 8179 right conversions, so overall accuracy is 96.7%.Kaur [6] developed a conversion of Punjabi text to its Dialectal Lahndi form. The system is developed employing a rule based mostly approach that embrace about eighty 5 rules for conversion and a dictionary consisting 3247 words. The conversion system 1st segments the sentences into words and identifies the words which require conversion so it converts the words by applying conversion rules and dictionary. This Conversion system will be extended to different dialects of Punjabi additionally. The accuracy of system was found 96.7% on 8460 different words. Kaur [7] studied the existing techniques of Dialect identification and then make corpus of Malwai, Pothohari and Pwadhi words. Authors develop algorithm that identify the dialect from given text input. When algorithm finds dialect word from these three dialects, system highlights the dialect words. This research will help to increase processing of Dialectal Punjabi resources. In future, further work on more dialect and different technique will increase accuracy and efficiency. Ghai [8] studied the efforts to develop Automatic speech recognition for Hindi, Oriya, Malayalam, Bengali, Assamese, Marathi, Urdu and Sinhala languages of Indo-Aryan languages family. Authors observed that use of techniques Cooperative Heterogeneous ANN Architecture, Maximum Likelihood Linear Regression, Extended MFCC and Learning Vector quantization are helping the researchers to get improved recognition performance of speech recognition systems. Computerized Speech Lab has also helped in speech acquisition process. Punjabi, being a widely spoken Indo-Aryan language, is still trailing in the research and development for the field of automatic speech recognition. So far the work finished Punjabi language is isolated word speech recognition victimization Acoustic example matching technique on MATLAB. In this research authors shows almost all the efforts made by various researchers for the research and development of ASR for Indo-Aryan languages have been analyzed and the applicability of techniques applied for alternative Indo-Aryan languages has been mentioned for Punjabi language. Ghai [9] used HTK 3.4.1 speech engine in implementation of ASR systems. Two approaches of acoustic modeling were used: whole word models and tri-phone models. The word recognition accuracy of isolated word speech was 92.05% for acoustic whole word model based system and 97.14% for acoustic tri-phone model based system. The word recognition accuracy of connected word speech was 87.75% for acoustic whole word model based system and 91.62% for acoustic tri-phone model based system. Yogesh [10] worked on interview speech corpus using Sphinx and Java programming for graphical user interface of system. In this system 461 sentences and 1227 words of Punjabi language were used to test the system.6 sentences and 14 words are not recognized by system in testing phase so accuracy is near about 98% combine. Harpreet [11] worked on Speech recognition for Punjabi language using MFCCs for extract features from the speech signal. System had been trained using HMM which was considered to be one of the most efficient pattern recognition techniques. Viterbi Algorithm had been used for system testing as it was found to be the one that finds the most probable sequence of path. The authors proposed the system for recognition of 90 alphabets taken from Punjabi language with 9 speakers having Punjabi as their native language from the age group of 20-50 years. The accuracy of the system was evaluated in three different scenarios. The results were found to be 80%, 100% and 55% accurate for each scenario respectively. Dua [12] worked for Punjabi speech to text system for connected words on HTK 3.4.1 speech engine on the Linux platform. The overall performance was analyzed for both class and open environment. 6 unique speakers have been asked to record test data each consisting of 30-50 words. The performance of system recorded as Word Recognition Rate 95.8% and 95.4%, Word Accuracy Rate 94.1% and 91.6% and Word Error Rate 5.9% and 8.35% in class room and open environment respectively. The performance of the system was tested against speaker independent parameter by using two types of speakers: one who are involved in training and testing both and the other who are involved in only testing. The system was tested in an exceedingly category area and in open area. A total of 6 distinct speakers were used for this and each one was asked to speak 35-50 words. The results shown reveal that the enforced system performs well with totally different speakers and in numerous environments. The average performance of the system lies within ninety four to ninety six with word error rate4% to six. Leo Breiman [16] developed random forests, random forest combination of tree predictors in which each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest. The generalization error for forests converges a.s. to a limit because the variety of trees within the forest becomes massive. The generalization error of a forest of tree classifiers depends on the strength of the individual trees within the forest and also the correlation between them. Random forests are an effective tool in prediction. The framework in terms of strength of the individual predictors and their correlations offers insight into the power of the random forest to predict. Fatemeh Noroozi [17] developed a vocal-based emotion recognition method using the Random Forest decision making algorithm to the speech signals comprising six emotion categories named as happiness, fear, disgust, sad, neutral and surprise. The corresponding emotion labels are then assigned to each voice signal by means of multi-class classification. Authors use following steps to develop vocal-based emotion recognition system. The following is accuracy table of vocal based Emotion recognition System using random forest decision making. The average accuracy of system is calculated near 66% in all emotion on average. Kaur [18] worked on name entity recognition system in Punjabi language to seek and classify words which represent proper names in text into predefined categories like location, person name, organization, date, time, designation etc. The system was good results for location, organization, designation, date/ time NER's with average accuracy of 85%. The results for person name were not good as compared to other NEs because accuracy of that was around 62%. The result of NER system is measured using precision (P), recall(R) and F-measure. The precision measures P= (number of correct NEs/total number of NEs). The recall measures R= (number of correct NEs/total number of NEs in a text).The F-measure represents mean value of preciseness and recall. F= 2RP/R+P

Table 1:-Accuracy of NER System

NE CLASS	P (%)	R (%)	F (%)
Person	74.52	62.86	65.67

Location	91.52	92.89	91.25
Organization	90.27	90.10	88.77
Designation	98.84	87.09	91.98
Date/Time	94.79	89.79	91.75
Total	89.98	84.55	85.88

Table 2:-Comparative study of Developed system for Punjabi Dialects

S.No	Authors	Year	Name	Methodology	Input type	Accuracy
1.	Anterpreet Kaur, Parminder Singh and Kamaldeep Kaur	2017	*Punjabi Dialects Conversion System for Majhi, Malwai and Doabi Dialects*	Build bilingual dictionaries and morphological transfer rules for better accuracy.	Text	97%
2.	Arshdeep Singh and Jagroop Kaur	2016	*Identification of Dialects in Punjabi Language*	Make a corpus of words of three Punjabi dialects such as Malvayi, Pothohari and Pwadhi	Text	--
3.	Parneet Kaur and Simrat Kaur	2015	*Standard Punjabi Text to Lahndi Dialect Text Conversion System*	Build bilingual dictionaries from linguistics resources and develop Morphological Analyzer and Conversion Engine.	Text	96.7%
4.	Arvinder Singh and Parminder Singh	2015	*Rule Based Punjabi Dialect Conversion System*	Build a Database of words and conversion rules then text's one to one mapping improve accuracy.	Text	96.5%
5.	Arvinder Singh and Parminder Singh	2015	*Punjabi Dialect Conversion System for Malwai and Doabi Dialect*	Build a Database of words and conversion rules then input text of one dialect convert in to second.	Text	94.5%

III PROPOSED ALGORITHM

1. First a dataset of the different speeches of different region is stored as a training set.

a. dataset=$\sum_{i=1}^{n} speech(i)$

2. Identify the features of the each speech.

a. $f(x) = a_0 + \sum_{n=1}^{x} \left(a_n \cos\frac{n\pi x}{L} + b_n \sin\frac{n\pi x}{L}\right)$

3. Classify the features extraction from each speech.

a. $\coprod_{i=1}^{n} classification\ category$

4. Input the speech from the user. This inputted speech will be that whose features are to be matched.

5. Extract the features of the inputted speech

$$f(Input) = a_0 + \sum_{n=1}^{x} \left(a_n \cos\frac{n\pi x}{L} + b_n \sin\frac{n\pi x}{L}\right)$$

6. Classify the features extracted from the speech

$$\coprod_{i=1}^{n} classification\ category$$

7. Perform the matching based on features.

IV FLOWCHART

Purposed flowchart for speech recognition mainly depends on system training because accuracy depends upon training of system. First step collect the speech samples for different region of Punjab region wise for training purpose then extract and classify main feature from that sample and save it into dataset. Second phase where an input speech is matched with speech already stored in dataset. First step of second phase is input the speech from user or use a recorded sample then extract and classify features from that sample and match with feature in Dataset if match found it prints dialect of user, else it ask no match found i.e. more training require.

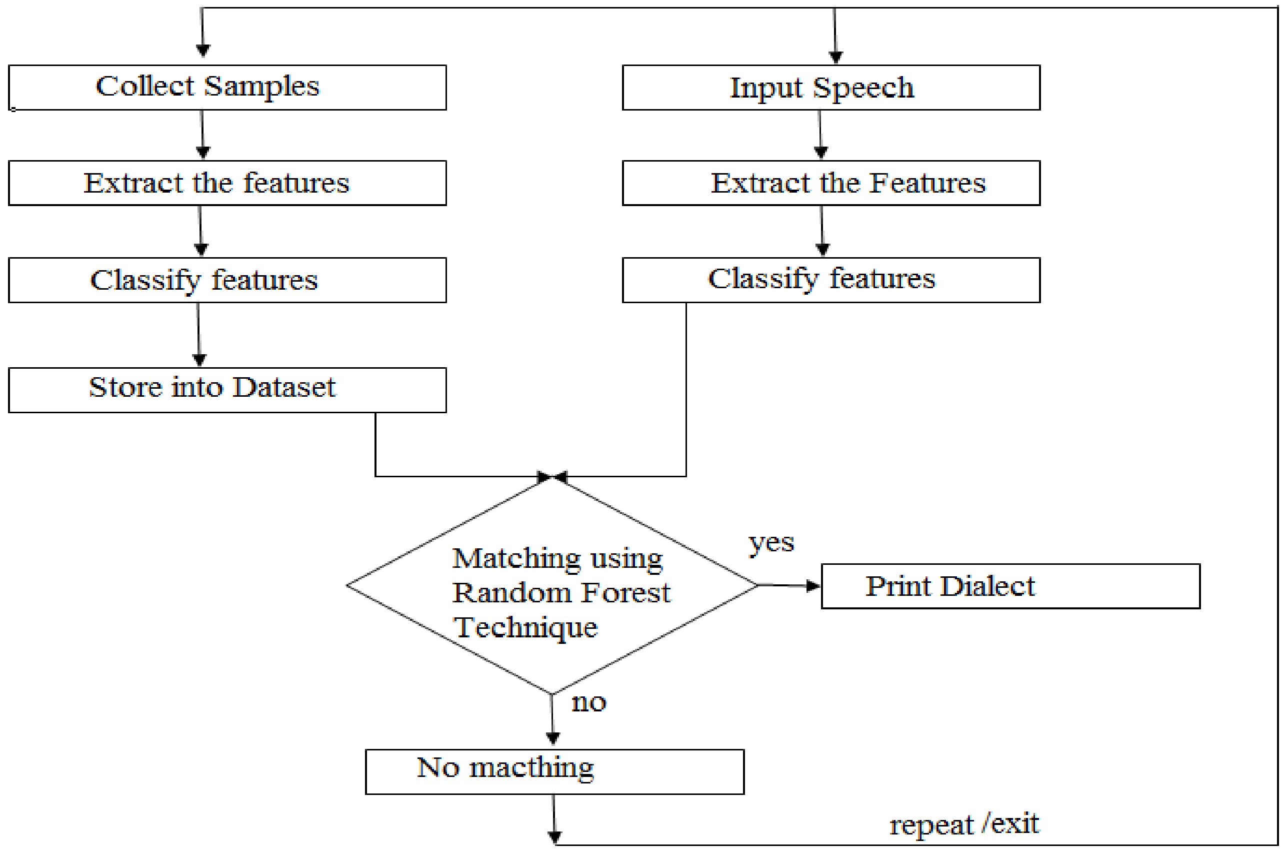

Figure 1:-Flow chart of proposed system.

V. RESULTS AND DISCUSSION

Dialect identification system is the features based system. Various features from sample set is recognized and stored for matching at later stage. Input speech features are matched with training sets. Various features like Energy entropy block, Short Time Energy, Zero Crossing rate (ZCR), Spectral Centroid, Spectral Flux, Spectral Roll Off are being used for dialect identification. Random forest based technique is used in this system for better performance. This technique sub divides the features in to the tree and identify the match based on sub divided features. The matching percentage is around 98%. Almost all the speeches are being recognized successfully.

Table 3: Results

Dialect of Punjabi	Units of Conversion (in words) Existing Morphological approach		Units of Conversion (in words) Proposed Random Forest Based approach			Comparison
	Right Conversion	Wrong Conversion	Total Conv ersion	Right Conversion	Wrong Conversion	
Majhi	96.58%	3.42	100	98.2%	1.8	+1.62
Malwai	96.48%	3.52	100	98.32%	2.68	+1.84
Doabi	97.54%	2.46	100	99.12%	0.88	+1.58

VI. CONCLUSION AND FUTURE WORK

Currently random forest based technique has been used for matching the features of the speech sample with already stored speeches features. This technique can be further enhanced by using fuzzy based system. Which can identify the level of match rather than identifying match or no match.
.

REFERENCES

[1] Anterpreet Kaur, Parminder Singh and Kamaldeep Kaur," Punjabi Dialects Conversion System for Majhi, Malwai and Doabi Dialects", International Conference on Computing Modeling and Simulation, 2017.

[2]Arvinder Singh and Parminder Singh," A Rule Based Punjabi Dialect Conversion System", International Journal for Research in Applied Science & Engineering Technology, pp.398-404, 2015.

[3]Arvinder Singh and Parminder Singh," Punjabi Dialect Conversion System for Malwai and Doabi Dialect", International Journal of Science and Technology, vol. 8(27), 2015.

[4] Harjeet Singh,"Comparative Study of Standard Punjabi and Malwai Dialect with regard to Machine Translation," An International Journal of Engineering Sciences, June 2013.

[5] Parneet Kaur and Simrat Kaur," Machine Translation of languages and dialects ", International Research Journal of Engineering and Technology, 3, 2016.

[6] Parneet Kaur and Simrat Kaur,"Standard Punjabi Text to Lahndi Dialect Text Conversion System", International Journal of Science and Research, 2015.

[7] Arshdeep Singh and Jagroop Kaur," Identification of Dialects in Punjabi Language", International Journal of Innovations & Advancement in Computer Science, vol.5, 2016.

[8] Wiqas Ghai and Navdeep Singh, "Analysis of Automatic Speech Recognition Systems for Indo-Aryan Languages: Punjabi a Case Study", International Journal of Soft Computing and Engineering, vol.2, pp.379-385, 2012.

[9] Wiqas Ghai and Navdeep Singh, "Phone Based Acoustic Modeling for Automatic Speech Recognition for Punjabi Language", International Journal of Computer Application, vol.1 (3), pp.69-83, 2013.

[10] Yogesh Kumar and Navdeep Singh, "An Automatic Spontaneous Live Speech Recognition System for Punjabi Language Corpus", International Science Press, vol.9 (20), pp. 259-266, 2016.

[11] Harpreet Kaur and Rekha Bhatia, "Speech Recognition system for Punjabi language", International Journal of Advanced Research in Computer Science and Software Engineering, vol.5, 2015.

[12]Mohit Dua and R.K Aggarwal,"Punjabi Speech to Text System for Connected Words", Fourth International Conference on Advances in Recent Technologies, 2015.

[13]Alan W Black, Prasanna kumar Muthukumar,"Random Forest for statistical Speech Synthesis", International Science Community Association, 2015.

[14]Urmila Shrawankar and Dr.VilasThakare,"Techniques for Feature Extraction in Speech Recognition System: A Comparative Study", International Journal of Computer Applications in Engineering, Technology and Sciences, 2011.

[15] Manjutha M and Gracy J,"Automated Speech Recognition System- A Literature Review", International Journal of Engineering trends and applications, vol.4, 2017.

[16] Leo Breiman,"Random Forest", springer, vol.45, 2001.

[17]Fatemeh Noroozi, Tomasz Sapiński, Dorota Kamińska, Gholamreza Anbarjafar, "Vocal based emotion recognition using random forests and decision tree", Springer Science,2017.

[18] Kamaldeep Kaur and Vishal Gupta," Name Entity Recognition for Punjabi Language", International Journal of Computer Application, vol.2, 2012.

[19]. Bhojaraj Barhate, Dipashri Sisodiya, Rakesh Deore, " Application of Speech Recognition: For Programming Languages ", International Journal of Scientific Research in Computer Science and Engineering, Vol.06, Issue.01, 2018.

[20].Rohit Katyal,"Analysis of SMO and BPNN Model for Speech Recognition System", International Journal of Computer Science and Engineering, Vol.4, 2016.