Dialectal Variations of Isolated Word Recognition

Shipra J. Arora[1] Dr. Rishipal Singh[2]

jkshipra22@gmail.com, pal_rishi@yahoo.com

[1]*CSE Department, Guru Jambheshwar University of Science & Technology Hisar – Haryana (India)*

[2]*CSE Department, Guru Jambheshwar University of Science & Technology Hisar – Haryana (India)*

Abstract. Speech can be considered as most important aspect of communication among living creatures. A lot of work has been done in the past in the area of speech processing but it has a wide variety of applications such as speech recognition, speaker identification, speech synthesis, machine translation, information retrieval system and others. The objective of this paper is to discuss the performance of dialectal variations of isolated word recognition. Hidden Markov Model (HMM) technique is used for implementation of isolated speech recognition system. MFCC technique is used for feature extraction. Speech corpus consists of 125 isolated words, spoken by 100 speakers i.e. 30 males and 30 female speakers of Malwi dialect, 10 males and 10 female speakers of Majhi dialect and 10 males and 10 female speakers of Doabi dialect. System performance is tested in computer lab environment by male speakers and female speakers. Speech recognition system is also tested in two modes i.e. by speakers involving in both training and testing phase and by speakers involving in testing only. In the first mode, Speech recognition accuracy is 82.56% by Malwi dialect male speakers, 78.80% by Malwi dialect female speakers and 80.68% by Malwi dialect mixed (males and female) speakers. Speech recognition accuracy is 68.80% by Majhi dialect male speakers, 87.28% by Majhi dialect female speakers and 78.04% by Majhi dialect mixed (males and female) speakers. Speech recognition accuracy is 86.40% by Doabi dialect male speakers, 81.68% by Doabi dialect female speakers and 84.04% by Doabi dialect mixed (males and female) speakers. In the second mode, Speech recognition accuracy is 83.12% by Malwi dialect male speakers, 79.52% by Malwi dialect female speakers and 81.32% by Malwi dialect mixed (males and female) speakers. Speech recognition accuracy is 63.68% by Majhi dialect male speakers, 83.60% by Majhi dialect female speakers and 73.64% by Majhi dialect mixed (males and female) speakers. Speech recognition accuracy is 83.28% by Doabi dialect male speakers, 79.60% by Doabi dialect female speakers and 81.44% by Doabi dialect mixed (males and female) speakers.

Keywords: Speech Corpus, Dialects, Hidden Markov Model, MFCC, Speech Recognition

(Received May 14th 2017 / Accepted March 3rd, 2018)

1 Introduction

Speech is the most natural means of communication. One of the most important challenges for researchers is accuracy of speech recognition. The speech recognition system concentrates on problems with its basic building blocks, features extraction and performance evaluations. Human machine interaction requires hardware interfaces such as mouse and keyboard but we need a smart device to communicate in natural means. An alternative solution to overcome hardware interfaces is software interface i.e. speech recognition. It is one of the most extensively emergent areas in present scenario. The main goal of speech recognition system is to convert speech signal into text independent of speaker, atmosphere and machine.

The motivational factors for creation of Punjabi Corpora are a) physically handicapped as well as illiterate persons communicate in their natural language with smart device. b) to overcome the problems of hardware interface such as mouse, monitor and keyboard.

According to 8th Schedule of Languages, Punjabi and Mandarin are tonal languages out of 22 official languages of India. Punjabi language is a state

language of Punjab. It is the first official language in east Punjab. Punjabi language script is Gurumukhi which means from the mouth of Guru. It has 10 vowels and 41 letters consisting of 38 consonants and 3 basic vowel sign bearer. Out of 38 consonants, there are 6 consonants with a dot below the consonant which are used to represent borrowed words from other languages. It is also recognized as "Painti" because of combination of 32 consonants except 6 consonants with a dot below the consonant which are used to represent borrowed words plus three basic vowel sign bearer. It has 3 conjunct vargs and also 3 signs i.e tippi, bindi and adhak.

2 Related Work

This section will represent work related to present study. In 1920, Speech recognition came into existence . The first machine i.e. a toy named Radio Rex was manufactured to recognize voice.

In 1950[27], for recognizing speech, most of systems examines the vowels spectral resonances of each utterances. At Bell Labs, in 1952, K.H.Davis et.al. designed an isolated digit recognition system for a single speaker [28] .

In 1990, Carnegie Mellon University's Harphy system [29] was designed to recognize speech with vocabulary size of 1011 words.

K. Kumar et al. [16] designed a connected-words SR system for Hindi language. The system was implemented using HTK toolkit and vocabulary size was 102 words.

R. Kumar [19] implemented an isolated word recognizer for the Punjabi language. He extended his work for comparing the speech recognition system performances for small vocabulary using the HMM and Dynamic Time Warp (DTW) technique.

Al-Qatab et al. [20] implemented an Arabic automatic speech recognition engine using HTK. The engine recognized both continuous speech as well as isolated words. The developed system used an Arabic dictionary built manually by the speech-sounds of 13 speakers and it used vocabulary of 33 words.
The main aim of the present study is to develop Speech recognition system for Punjabi language in agriculture domain to work in real time environment.

3 Text and Speech Corpora Collection

The corpus consists of 125 isolated phonetically rich words. These words have been recorded by 100 native speakers in the age group of 15 to 18 years. It consists of 30 males and 30 females speakers of Malwi dialect, 10 males and 10 females speakers of Majhi dialect and 10 males and 10 females speakers of Doabi dialect. Recordings have been prepared in two recording mediums, these are desktop mounted microphone and mobile phone using PRAAT software having sampling rate 16khz/16bit. Recording, in the present circumstances has minimal background noise and is clean and clear. Wave surfer is used for transcription of these words. A speech corpora of 12500 isolated words of 100 speakers have been organized entirety.

4 Research Method

Speech Recognition comprises of following constituents pronunciation dictionary, language model, feature extraction acoustic analysis, Hidden Markov Model (HMM) based acoustic model. Dictionary in general, is a book or electronic resource which provides alphabetically sorted list of words. Dictionary provides mapping between words in task grammar and acoustic model. It is also required to train HMM networks. It has been created using HDMan tool of HTK toolkit. In a formal language, grammar defines set of production rules for strings which describes how to construct it using alphabets of language. We can also check the validity of string according to the syntax of language. Grammar contains a list of words which are listened by Speech Recognition system as input and give text representation of words as output. HTK provides a grammar definition language to prepare task grammar.

4.1 Feature Extraction

Davis and Mermelstein introduced MFCC(Mel Frequency Cepstral Coefficients) in the 1980 AD. There are various methods for features extraction such as vector quantization, LPC (Linear Predictive Coding), LPCC(Linear predictive Cepstral Coefficients), LDB (Local Discriminant Bases) etc. But MFCC is the efficient method for speech recognition and speaker identification systems. It is standard feature extraction method in speech recognition system. Following are the steps concerned in MFCC are:

Pre-Emphasis Isolated word sample is conceded through a filter in this step which emphasizes higher frequencies.

$y_n = x_n - k*x_{n-1}$ where y_n is the output signal and value of k lies between 0.9 and 1.0.

During human sound production mechanism, high frequency part is concealed. In order to balance that part, pre-emphasis is used. At higher frequencies, energy of signal is amplified.

Framing As speech is a continuously time varying signal, so framing is required. Frames are usually blocks of 10 to 20 ms with overlap of 50 percentage of frame size. In order to smooth the progress of Fast Fourier Transform, frame size is equal to power of 2. Sample rate is16 KHz and frame size is 256 sample points.

Frame duration is 256/16000 = .016 ~ 16 ms.

50% overlap means 128 sample points.

Frame rate = 16000/128 =125 frames/second. To generate continuity with in frame, overlapping is done.

Hamming Window In order to generate continuity of initial and final points in frame, each frame is multiplied with hamming window. If x_n is the frame signal then

$z_n = x_n * w_n$ where w_n is the hamming window defined by

$w_n = 0.54 - 0.46 \cos(2\pi n/(N-1))$ where $0 \leq n \leq (N-1)$.

Fast Fourier Transform It converts time domain speech signals into frequency domain speech signals. FFT yields magnitude frequency response of each frame.

Mel Filter bank In order to get a flat magnitude frequency response, magnitude frequency response produced by FFT is multiplied by 26 triangular band pass filters. It also decreases size of involved features.

Discrete Cosine Transform In order to obtain L Mel Cepstral Coefficients, we apply DCT to the output of N triangular filters by using the formula

$$C_n = \sum_{k=1}^{n} E_k * \cos[m*(k-0.5)*\pi/N] \quad , m= 1,2, \ldots, L$$

These are called mel scale cepstral coefficients. Usually N=20 and L=12. We can also calculate feature i.e energy with in a frame. So L = 12 MFCC + energy .

In our system, the target parameters are to be MFCC using C0 as the energy component, the frame period is 16msec, the output should be saved in compressed format. The FFT should use a Hamming window and the signal should have first order pre-emphasis applied using a coefficient of 0.97. The filter bank have 26 channels and 12 MFCC coefficients be output.

Delta Ceptrum There are some other features which can be obtained by calculating (MFCC + energy)single and double time derivatives which yields velocity and acceleration. Total MFCC features are 39.

4.2 HMM Based Acoustic Model

Hidden Markov Model (HMM) is referred to as probabilistic functions of Markov Chain. It is a process of establishing statistical representation for speech waveform computed feature vector sequences. It is based on the modelling of features variations. Therefore, it provides an illustration of how speaker produces sounds. There are various acoustic models such as segmental models, neural networks, maximum entropy models etc. One of the most common acoustic model is HMM. Each word is represented by its HMM, which consists of six states out of which four are observation function and one initial state and one final state. Observation functions are described by the mean vectors and covariance matrices. By parameter estimation from the feature vectors, a preliminary model can be obtained. After the estimation, HMM parameters are re-estimated for each observation function. This process is iterated until model parameters converge. Re-estimation iterations are repeated three times in our system.

5 Results and Analysis

System training is done in two phases. In first phase, system is trained by using speakers data from Malwi, Majhi and Doabi dialects. Testing is done by using male and female speakers involving in training also. Following are the obtained results:

Dialects	Sex	Environment	No. of Spoken words	No. of Recognized words	Recognition Accuracy (%)
Malwi	Male	Computer Lab	1250	1032	82.56%
Malwi	Female	Computer Lab	1250	985	78.80%
Malwi	Male & Female	Computer Lab	2500	2017	80.68%
Majhi	Male	Computer Lab	1250	860	68.80%
Majhi	Female	Computer Lab	1250	1091	87.28%
Majhi	Male & Female	Computer Lab	2500	1951	78.04%
Doabi	Male	Computer Lab	1250	1080	86.40%
Doabi	Female	Computer Lab	1250	1021	81.68%
Doabi	Male & Female	Computer Lab	2500	2101	84.04%

TABLE 1
Recognition in Computer Lab Environment by Males and Females Speakers (Involving Both in Training & Testing)

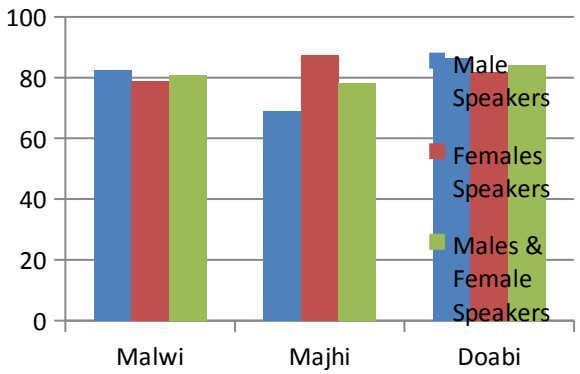

Fig.1 Recognition accuracy in phase 1

It has been found that for Malwi dialect speakers, recognition accuracy is highest for male speakers and lowest in female speakers. For Majhi dialect speakers , result is just reverse of Malwi dialect speakers. It is highest for female speakers and lowest for male speakers. For Doabi dialect speakers, result is same as Malwi speakers i.e. highest for male speakers and lowest in female speakers. Recognition accuracy lies in the order

Malwi Dialect Speakers:

Males> Female> Males & Females

Majhi Dialect Speakers:

Females>Males & Females > Males

Doabi Dialect Speakers:

Males> Males & Females> Females

In second phase, system is trained by using Malwi dialect speakers and tested using Majhi and Doabi dialect male and female speakers. 60% data is used for training and 40% data is used for testing. Following are the results.

Dialects	Sex Male / Female (M/F)	Environment	No. of Spoken words	No. of Recognized words	Recognition Accuracy (%)
Malwi	M	Computer Lab	1250	1039	83.12 %
Malwi	F	Computer Lab	1250	994	79.52 %
Malwi	M& F	Computer Lab	2500	2033	81.32 %
Majhi	M	Computer Lab	1250	796	63.68 %
Majhi	F	Computer Lab	1250	1045	83.60 %
Majhi	M& F	Computer Lab	2500	1841	73.64 %
Doabi	M	Computer Lab	1250	1041	83.28 %
Doabi	F	Computer Lab	1250	995	79.60 %
Doabi	M& F	Computer Lab	2500	2036	81.44 %

TABLE 2
Recognition in Computer Lab Environment by Males and Females Speakers (Involving Only in Testing)

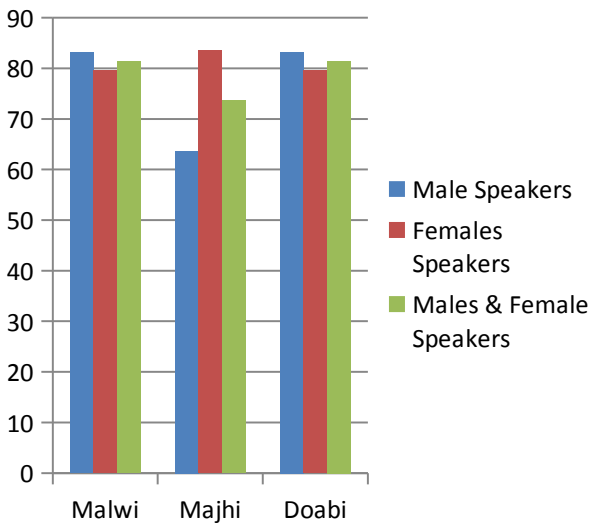

Fig. 2 Recognition accuracy in phase 2

In both phases, It has been found that for Malwi dialect speakers, recognition accuracy is highest for male speakers and lowest for females speakers. For Majhi dialect speakers, It is highest for female speakers and lowest for male speakers. For Doabi dialect speakers, result is same as Malwi speakers i.e highest for male speakers and lowest in female speakers. Recognition accuracy lies in the order

Malwi Dialect Speakers:

Males> Males & Females> Females

Majhi Dialect Speakers:

Females>Males & Females > Males

Doabi Dialect Speakers:

Males> Males & Females> Females

6 Conclusions

The present study emphasized implementation of Punjabi Language Isolated Words speech recognition system in Malwi, Majhi and Doabi dialects. In both phases, recognition accuracy resulted in highest for males Malwi and Doabi dialect speakers and lowest for females Malwi and Doabi dialect speakers. But result is reverse for Majhi dialect speakers.

Recognition accuracy is more in phase 2 than phase 1 for Malwi and Doabi dialect speakers but it is less in phase 2 than phase 1 for Majhi speakers. Variations in results due to unique features of Punjabi Language dialectal variations i.e. most crucial tonal feature. This work may be extended for continuous speech recognition system and large size of vocabulary. A detailed analysis of dialectal variations will be discussed in upcoming paper. We hope that this system will serve as a baseline for machine translation and in order to improve accuracy of speech recognition system for various Indian languages.

ACKNOWLEDGEMENTS

I would like to thanks Late Sh. S.B. Panihar Insaan, Dr. Chanchal Insaan & Ms. Neetu Insaan of SSJITM, Sirsa and Shah Satnam Ji Girls' School, Sirsa for providing me valuable support and environment for recording speech corpus.

REFERENCES

[1] Kadyan, Virender, Archana Mantri and R.K.Aggarwal, " Refinement of HMM Model Parameters for Punjabi Automatic Speech Recognition(PASR) System" IETE Journal of Research (2017), pp 1-16

[2] Mittal Shama, Kaur Rupinderdeep, "Implementation of phonetic level speech recognition system for Punjabi language; 1st India International Conference on Information Processing (IICIP) 2016 pp 1-6

[3] Arora Vaibhav, Sood Pulkit, Keshari Kumar Utkarsh , ""A stacked sparse autoencoder based architecture for Punjabi and English spoken language classification using MFCC features" 3rd International Conference on Computing for Sustainable Global Development (INDIACom) 2016 pp 269 - 272

[4] Harpreet Kaur & Rekha Bhatia, "Speech Recognition System for Punjabi Language", International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 5, Issue 8, ISSN: 2277-128X, Aug. 2015.

[5] Arora Shipra J. and, Singh Rishipal; "Acoustic and Phonological Analysis of Homophones of Punjabi Language" International Journal of Computer Science Engineering and Information Technology Research (IJCSEITR) ISSN (P): 2249-6831; ISSN (E): 2249-7943 Vol. 4, Issue 1, Feb 2014, pp 95-102

[6] LI Jinyu et al., “An overview of noise robust Automatic Speech recognition”, IEEE/ACM Transactions on Audio, Speech and Language processing 22.4(2014) , pp 745-777

[7] Cini Kurian, “A Survey on Speech recognition in Indian Languages”, International Journal of Computer Science and Information Technology, Vol. 5(5), ISSN: 0975-9646, 2014.

[8] Nareshkumar et.al., “Database interaction using Automatic Speech Recognition”, International Journal of Innovative Research in Science, Engineering and Technology, vol.3(3) ISSN(Online) :2319-8753, ISSN(Print), pp 2347-6710, Mar. 2014.

[9] Arora Shipra J. and, Singh Rishipal; “Automatic Speech Recognition: A Review” International Journal of Computer Applications (0975 – 8887) Volume 60– No.9, Dec 2012, pp 34-44

[10] Dhanjal Surinder and Bhatia S.S.; “A New Corpus for the Punjabi Speech processing”; International Symposium on frontiers of research on Music and Speech(FRSM-2012), KIIT Gurgaon, India; January 18-19, 2012, pp 223-227.

[11] Agrawal S.S., Sinha Shweta, Singh Pooja and Jesper Olsen; “Development of Text and Speech Database for Hindi and Indian English specific to Mobile Communication Environment”, Proceeding of International Conference on the Language Resources and Evaluation Conference, LREC, Istanbul, Turkey, 2012.

[12] Saon, George and Jen-Tzung Chien. “Large Vocabulary continuous speech recognitions systema: A look at some recent advances” IEEE Signal Processing magazine29.6 (2012) pp18-33

[13] Acero, Alejandro, “Acoustical and Environmental Robustness in Automatic Speech Recognition” Vol. 201. Springer Science & Business Media, 2012

[14] Dua Mohit et. al., “Punjabi Automatic Speech recognition using HTK”, International Journal of Computer Science Issues”, Vol. 9 Issue 4 No. 1 , ISSN(online) 1694-0814,July 2012.

[15] Aggarwal Rajesh Kumar and Mayank Dave, “Acoustic Modelling problem for Automatic Speech Recognition System:advances and refinement (Part II). International Journal of Speech technology 14.4(2011):309-320

[16] K.Kumar and R.K.Aggarwal, “Hindi Speech Recognition System using HTK “, International Journal of Computing and Business Research, Vol. 2(2), May 2011.

[17] Anusuya M.A et.al, “Front end analysis of Speech Recognition: A Review”, International Journal of Speech Technology, Springer, Vol. 14, pp 99-145, 2011.

[18] Bhaskararao Peri; “Sailent phonetic features of Indian Languages in Speech Technology”; Sadhana Academy Proceedings in Engineering Sciences; Indian Academy of Sciences, Banglore, India; Volume 36, Number 5, October 2011, pp 587-599.

[19] R. Kumar “Comparison of HMM and DTW for Isolated Word Recognition of Punjabi Language” In Proceedings of Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications, Sao Paulo, Brazil. Vol. 6419 of Lecture Notes in Computer Science (LNCS), pp. 244– 252, Springer Verlag, November 8-11, 2010.

[20] B. A. Q. Al-Qatab and R. N. Ainon, “Arabic Speech Recognition Using Hidden Markov Model Toolkit (HTK)”, Paper presented at International Symposium in Information Technology (ITSim). Kuala Lumpur, June 15-17, 2010.

[21] Dhanjal Surinder and Bhatia S.S; “Punjabi Bhasha da Takneekee Bhavikh”; Silver Jublilee International Punjabi Developmet Conference; Punjabi University, Patiala, February 3-5, 2009.

[22] Dhanjal Surinder and Bhatia S.S.; “Computerization of the Punjabi Language”; 2nd World Punjabi Conference; Punjab University, Chandigarh; February 24-25, 2009.

[23] Baker, Janet M. et al., “ Developments and Directions in Speech Recognition and Understanding, Part 1 [DSP Education]” IEEE Signal processing magazine 26.3(2009)

[24] Furui Sadaoki, “40 years of progress in Automatic Speech Recognition”, Advances in Biometrics (2009),pp 1050-1059.

[25] Agrawal S.S., Samudravijaya K, Arora K.; “Recent advances of Speech Database Development Activities”, International Symposium on Chinese Spoken Language Processing (ISCSLP 2006), 2006.

[26] Agrawal S.S and Samudravijaya K.(Chief Editors); “Text and Speech Corpora

Development in Indian Languages"; Proceedings of the International Symposium on Speech technology and Processing Systems (ISTEPS-2004 and Oriental COCOSDA-2004); vol. –II, CDAC, New Delhi, India, November 17-19, 2004; pp 21-27.

[27] Samudravijaya K, P.V.S.Rao and S.S.Agrawal; "Hindi Speech Database"; Proceeding International Conference on Spoken language processing (ICSLP00), Beijing, China, October 2000.

[28]. K.H.Davis, R.Biddulph, and S.Balashek, Automatic Recognition of spoken Digits, J.Acoust.Soc.Am., 24(6):637-642,1952.

[29]. B.Lowrre, The HARPY speech understanding system, Trends in Speech Recognition, W.Lea, Ed., Speech Science Pub., pp.576-586, 1990.