Speech Recognition System Architecture for Gujarati Language

Jinal H. Tailor
PhD Scholar
Sardar Patel University,
Vallabh Vidhyanagar, India

Dipti B. Shah, PhD
Professor
G.H. Patel Post Graduate Dept. of Comp Science & Technology
Sardar Patel University,
Vallabh Vidhyanagar, India

ABSTRACT
Speech recognition is an area of Natural Language Processing and Artificial Intelligence. To achieve good accuracy and efficiency of Automatic Speech Recognition (ASR) system for Indian Gujarati language is challenging task due to its morphology, language barriers, different dialects, and unavailability of resources. This paper presents proposed architecture of ASR for Gujarati language. Raw input data have been collected from 4 male and 2 female who belongs from age between 18 to 36 years to prepare dataset for training purpose. The goal of Speech recognition system is to make machines capable enough to operate in natural languages. ASR is a system to convert vocalized form to visualized form using different computational devices. This convincing approach is useful to the people having disabilities deaf or inability to use input device. In this paper we have used Hidden Markov Model Toolkit HTK Tool to measure performance and error parameters. The system implementation analyzed WR (Word Recognition Rate) 95.9% and WER (Word Error Rate) as 5.85 % in Lab environment. For the open noisy environment calculated WR was 95.1% and WER found 7.40%.

Keywords
Acoustic Model, Hidden Markov Model, Gujarati, Speech-To-Text

1. INTRODUCTION
Gujarati is an Indo-Aryan language which is evolved from Devnagri script and it is spoken by more than 50 million people in Gujarat state. Gujarati is rich in morphology and has complex structure of syntax. Diversity of dialects in the form of pronunciation, grammar and vocabulary make speech recognition process more complex. Converting speech into text mechanism for Gujarati language includes major limitations regarding accuracy because of bulk corpora set, diversity of dialects, complex language semantics and morphological structure. The basic framework for the speech processing system requires narrow design steps to achieve accuracy and encapsulation of data through the processing. The ASR system uses Hidden Markov Model (HMM) for speech processing with detailed eight phases. The model is based upon HMM initialized values assigned to all the words and sub words. Viterbi algorithm calculates joint probability and generates best matched state.

2. BLOCK DIAGRAM OF STT
A typical Speech-To-Text system incorporates different phases that are speech pre-processing, feature extraction, acoustic analysis, modeling, decoding and filtering as shown in Fig. 1. The speech is used as system input which passes through phases and converts it into text format.

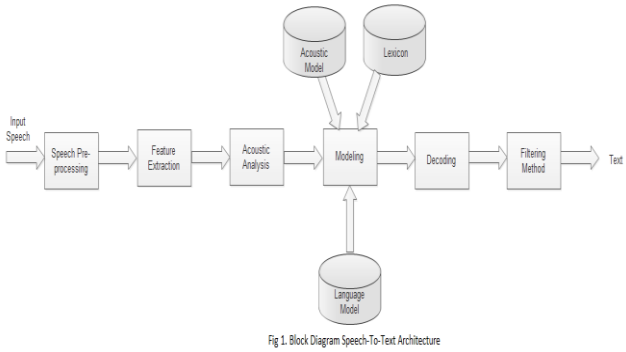

Fig 1. Block Diagram of Speech-To-Text Architecture

After decoding phase, filtering phase extract recognized word from the output transcription and represent in text format.

3. GUJARATI LANGUAGE CHARACTER SET
Gujarati language phoneme set mainly includes 34 consonants, 12 vowels, 10 digits and 106 special characters.

Gujarati Language – Consonants

Table 1. Gujarati Consonants

ક Ka	ખ Kha	ગ Ga	ઘ Gha	ઙ ṅa	ચ ca	છ Cha	જ ja	ઝ jha	ઞ Ña
ટ ṭa	ઠ ṭha	ડ ḍa	ઢ ḍha	ણ ṇa	ત ta	થ Tha	દ da	ધ dha	ન Na
પ Pa	ફ Pha	બ Ba	ભ Bha	મ Ma	ય ya	ર Ra	લ la	વ va	શ Śa
ષ ṣa	સ Sa	હ Ha	ળ ḷa	ક્ષ kṣa	જ્ઞ Jña				

Gujarati Vowels

Table 2. Gujarati Vowels

અ	આ	ઇ	ઈ	ઉ	ઊ	એ
A	Ā	I	ī	U	Ū	E
ઐ	ઓ	ઔ	અં	અઃ		
Ai	O	Au	aṃ	Ah		

Numerals

Table 3. Numerals

૦	૧	૨	૩	૪	૫	૬	૭	૮	૯
0	1	2	3	4	5	6	7	8	9

4. ARCHITECTURE OF STT

A detailed STT architecture as shown in Fig. 2 include eight phases that are

1) Speech Pre-Processing
2) Acoustic Analysis – feature extraction
3) Acoustic Model Generation – Initialization
4) Acoustic Model Generation – Re-Estimation
5) Language Model – Parsing
6) Pronunciation Model
7) Decoding
8) Filtering Methods

Speech Pre-Processing – Through the recording tool input speech data is stored in .wav format. Quality of recorded data is depending upon recording device and speed. Segmentation is used to divide raw data speech signal into evenly spaced frames. The frame size is selected in smaller size to capture rapid transitions and achieve sufficient resolution in frequency domain. The frame size can vary from 10 milliseconds to 25 milliseconds. Segmented frames are analyzed to produce feature vector which is useful for speech sound classification. The required measurement parameters for feature analysis are recording channel, environmental noise and speaker variability.

Acoustic Analysis – Feature extraction is used to extract the related information from the audio based upon features like pitch, frequency and environment for the statistical analysis. Feature vectors are computed to measure vibration of voiced sound or non-specific noise from unvoiced sound and frequency. *HCopy* tool is used in HTK toolkit for feature extraction. MFCC Cepstral analysis a standard technique is used for feature extraction. Minimum 10-12 coefficients are considered for recorded speech. Speech signal is converted to acoustic vector using MFCC technique. All related .wav files are converted into .mfcc files which includes list of cepstrum coefficients.

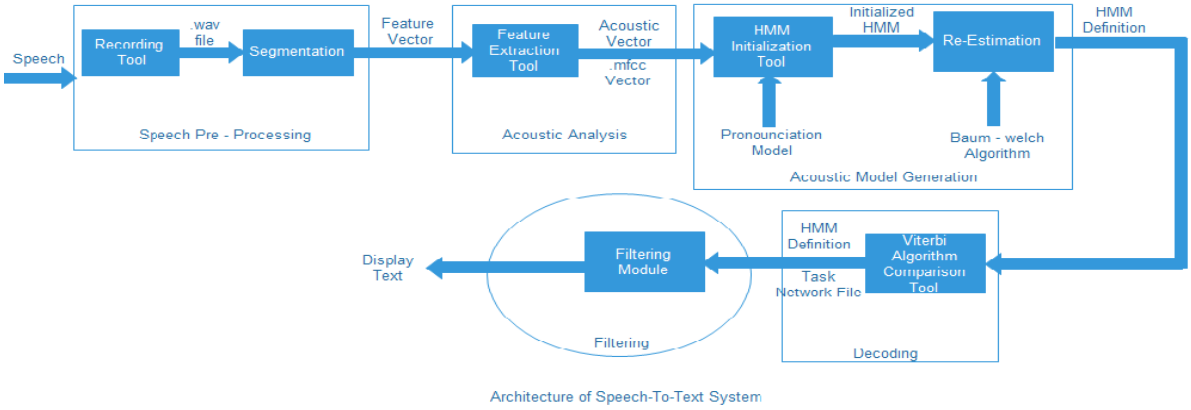

Fig 2: Architecture of Speech-To-Text System

Acoustic Model Generation – Initialization – speech recognition system uses HMM Model as statistical model for the speech generation process. HMM pattern recognition used for speech recognition and speech synthesis. Speech signals are quasi-stationary and stable only for short period of time. Stability of signals can be viewed in form of states in a HMM Topology. as speech signals are continuous it evolve quantization error which can be resolve using vector quantization that coordinate continuous units of acoustic space to discrete symbol. In acoustic model generation process HMM topology is formed with series of states. HMM prototype tool *HInit* initializes each HMM for individual words as well as for sub-words units. Resulting HMMs are concatenated according to pronunciation model for each input sentence. The result of acoustic model is considered in form of scores rather than probabilities and different heuristics used to regularize them.

Acoustic Model Generation – Re-estimation – The standard Baum-Welch algorithm is used to update HMM active state with optimal values for HMM parameters such as mean, variance and transition probabilities. Re-estimation tool generate HMM definition for all related probabilities.

Language Model – For continuous speech, language model is the essential part with acoustic model in speech processing. Acoustic model is the part of component of speech synthesizer that check the utterance match a word or sequence. Language model is the component that determines

how the word or sentence is spoken specially for words that sounds similar. *HParse* tool of HTK is used to generate task network files.

Pronunciation Model – Pronunciation model is used to develop correspondence between different HMMs to form model for each input sentence. The concurrency is checked between HMM name and variable specified in language model.

Decoding – in pronunciation model word models are created by concatenating sub word models which are composed into a decoding network. Lexicon tree algorithm is used to find best path that provides most appropriate match according to score. Viterbi algorithm is used to calculate joint probability of the observation and single best state sequence. *HVite* tool is used to generate transcription files as output. Viterbi algorithm based on BFS can be replaced by DFS for performance improvement. Decoding network generates transcription file after processing HMM definition and task network files.

Filtering Methods – filtering method use to extract recognized word from output transcription file and convert into relevant text format.

5. PERFORMANCE PARAMETERS

Word Recognition Rate (WR) = (N – D – S / N) / 100

WER is the common measure of speech recognition system performance. Through the power low correlation is measured between recognized words with spoken words from transcription file.

WER = (S + D + I / N) * 100

Where S = Number of Substitutions

D = Number of Deletion

I = Number of Insertion

N = No of Words in Reference

6. RESULT

Mainly 2 types of speakers have been selected male and female for both in noisy environment and in quite laboratory environment. Recorded speech used for evaluating performance for the speech recognition system. The implemented system tested for total 10-12 words which gives analyzed values of WR 95.9% and WER as 5.85 % in Laboratory room environment. For the noisy open environment system calculated WR 95.1% and WER 7.40%.

Speech Recognition Analysis in Laboratory Environment						
Speakers List	**N**	**D**	**I**	**S**	**WR**	**WER**
SP1	**42**	**0**	**1**	**1**	**97.6**	**4.76**
SP2	**38**	**1**	**0**	**0**	**97.3**	**2.63**
SP3	**44**	**1**	**1**	**2**	**93.1**	**9.09**
SP4	**30**	**0**	**1**	**1**	**96.6**	**6.66**
SP5	**35**	**1**	**1**	**0**	**97.1**	**5.71**
SP6	**32**	**1**	**0**	**1**	**93.7**	**6.25**
System Performance Summary					**95.9**	**5.85**

Speech Recognition Analysis in Noisy- Open Environment						
Speakers List	**N**	**D**	**I**	**S**	**WR**	**WER**
SP1	**32**	**0**	**1**	**1**	**96.8**	**6.25**
SP2	**44**	**1**	**0**	**0**	**97.7**	**2.27**
SP3	**20**	**0**	**1**	**1**	**95**	**10**
SP4	**35**	**0**	**1**	**1**	**97.1**	**5.71**
SP5	**22**	**3**	**1**	**0**	**86.3**	**18.1**
SP6	**48**	**1**	**0**	**0**	**97.9**	**2.08**
System Performance Summary					**95.1**	**7.40**

7. CONCLUSION

The accuracy and performance are the primary concern for speech-to-text system. The complexity of system can be minimize by emphasizes upon interaction between acoustic and language model using modeling technique. In natural language processing, reduce latency time in real-time application is a challenging task. To minimize latency, long words are removing from the vocabulary by decomposing into compound or pseudo-compound words. Large vocabulary set resolves the limitation of dialects variation from speakers who belongs from different regions. Optimization process can be applied to reduce redundant data and ambiguity to release the confusion between language model and acoustic model which result in minimize WER. Hence, the implemented work can be extended for sentences with large number of words and for the continuous speech.

8. REFERENCES

[1] Akila A.,E. Chandra. , - "Isolated Tamil Word Speech Recognition System Using HTK", International Journal of Computer Science Research and Application, Vol. 3, Issue 2,Pages 30-38

[2] C.Vimala, M.Krishnaveni , 2012, Continuous Speech Recognition system for Tamil language using monophone-based Hidden Markov Model , Proceedings of the Second International Conference on computational science, Engineering and Information Technology CCSEIT'12 pp 227-231.

[3] Chowdhury, S., —"Implementation of Speech Recognition System for Bangla", BRAC University, DHAKA, Bangladesh, August 2010.

[4] Daines D., - "An Architecture for Scalable, Universal Speech Recognition". PhD Thesis, School of Computer science, Carnegie Mellon University, USA, 2011

[5] Durbin M., & Cardona G., (1968) "A Gujarati Reference Grammar Language", 44(2), 411

[6] Hasnat, M., Molwa, J.,Khan, M., —"Isolated and Continuous Bangla Speech Recognition: Implementation, Performance and application perspective",2007.

[7] Hidden Markov Model Toolkit [HTK] downloaded from http://htk.eng.cam.ac.uk/

[8] Kumar, K.,Aggarwal R., ―”Hindi Speech Recognition System Using HTK”, International Journal of Computing and Business Research, ISSN (Online): 2229-6166, Volume 2 Issue 2, May 2011.

[9] Kumar R., Singh C., Kaushik S., ―”Isolated and Connected Word Recognition for Punjabi Language using Acoustic Template Matching Technique‖, 2004.

[10] Laxmi A. and Hema A Murthy,2006 A Syllable Based Continuous Speech Recognition for Tamil, INTERSPEECH – ICSLP, Pennsylvania,pp 1878-1881

[11] M. R. Hassan, B. Nath and M. Ala Uddin Bhuiyan, “Bengali Phoneme Recognition: A New Approach”, Proc. 6th ICCIT, Dhaka, 2003.

[12] Nadungodage T., Weerasinghe, R., ―Continuous Sinhala Speech Recognizer‖, Conference on Human Language Technology for Development, Alexandria, Egypt, May 2011.

[13] Pandit P., Bhatt S., - “Automatic Speech Recognition of Gujarati Digits using Dynamic Time Warping”, International Journal of Engineering and Innovative Technology, Vol. 3, Issue 12, June 2014

[14] Sarfraz H., Hussain S., Bokhari R., Raza A, Ullah I., Sarfraz Z., Pervez S., Mustafa A., Javed I., Parveen R., ―”Large Vocabulary Continuous Speech Recognition for Urdu”, International Conference on Frontiers of Information Technology, Islamabad, 2010.

[15] Syama R, Suma Mary Idikkula (2008) “HMM Based Speech Recognition System for Malayalam”, ICAI’08 – The 2008 International Conference on Artificial Intelligence, Monte Carlo Resort, Las Vegas, Nevada, USA (July 14-17, 2008)