

QUERY PROCESSING FOR HINDI KEYWORDS SEARCHING USING NLP

¹Anjali Arora, ²Anil Chopra

¹Research Scholar, M.Tech (CSE), MRIU, Faridabad, Haryana, India

²Professor (CSE), Faculty of Engineering & Technology Engineering, MRIU, Faridabad, Haryana, India

Abstract: People with less knowledge such as farmers, labours elderly for whom it is very difficult to identify the screen text, as they are not aware of English language and lack of computer knowledge where the keyboard and mouse may not be an appropriate means of communication between systems. So it would be a real relief to have the option to use ears to listen to the contents and use Hindi keypad to navigate and control the computer systems. Microsoft has designed an interface called SAPI (Speech Application Programming Interface) which supports dynamic speech input and output, and is integrated in our current operating systems. With the API it is possible to develop speech enabled applications without caring about the details of synthesis and recognition. In this paper, we are present a Hindi Speech Output Enabled Windows Application to demonstrate the use of speech output by processing the query in hindi as input using Microsoft SAPI in Microsoft Windows Operating Systems and DLL name is speechlib.dll.

Keywords: GB Theory, Natural language processing (NLP), Text-to-Speech (TTS), Hindi Speech Enabled Applications (HSeA)

I. INTRODUCTION

Computers are no longer a field of education for the Educational Institutes. They have changed from a large desktop computer or laptop to small pocket type PC like tablet and smart phone. In developed countries, using a computer has become the basic requirement like reading and writing a book. According to the survey done by Asia's premier monthly magazine on ICT in education, 85% of the total people on the earth will be using the Smartphone's or mobile and computer by 2020. The importance of digital information has already reached such an enormous level that many companies, corporate sector and politicians around the world try to find a way to paperless e-governance. But a question arises, is this really accessible to everybody? The growing need of a computer system and the application programs eventually made speech enabled operating systems a necessity. Most of the software coded today is only accessible through mouse and keyboard. But the expected improvements to the SAPI version included in Windows Vista may lead to a wave of new English speech enabled applications.. Full integration for speech synthesis and recognition as well as support for native and managed code could be part of the Windows operating system. So we are trying to develop an application which will search results from google and directly present the output to the user rather than number of links to navigate further and parse exact data. Output of this application is speech enabled that means user will get the direct output as audio.

II. RELATED WORK

Searching for information by voice has been part of our every day lives since long before the internet became prevalent. It was already the case thirty years ago that, if you needed information for a local business, the common approach was to dial directory assistance (411 in the US) and ask an operator for the telephone number. 800-GOOG-411 [2] is an automated system that uses speech recognition and web search to help people and and call businesses. Initially this system followed the well known model of _rst prompting the user for the \city and state" followed by the desired business. Traditional directory assistance applications are limited to a single modality, using voice as both input and output. With

the advent of smartphones with large screens and data connectivity, we could move to a multi-modal user interface with speech or text as the input modality, and maps with super-imposed business listings as the output modality. In March 2008 introduced first multi-modal speech application for Google Maps for Mobile (GMM). Johan Schalkwyk, Doug Beeferman, Françoise Beaufays, Bill Byrne, Ciprian Chelba, Mike Cohen, Maryam Garret, Brian Strope Google, Inc. In November 2008 introduced Google Mobile App (GMA) for iPhone that included a search by voice feature. GMA search by voice extended the paradigm of multi-modal voice search from searching for businesses on maps to searching the entire world wide web. Researches have also been done in order to convert text to speech and proved much more beneficial. Microsoft Speech Application Programming Interface (SAPI) is an API developed by Microsoft to allow the use of speech synthesis within Windows applications. It is possible for a 3rd-party company to produce their own Speech Recognition and Text-To-Speech engines or adapt existing engines to work with SAPI [14]. SAPI 5 however was a completely new interface, released in 2000. Various different Hindi speech enabled applications (HSeA) have been developed so far. Hindi Speech enabled word is one of the application in which user control word by speaking to it a microphone, entering text, or issuing commands to the computer. User can talk to the computer by using predefined commands and instructions. Hindi speech enabled calculator is another such application which provides the control to the calculator using speech and it is used for mathematical calculation like jama, ghata, guna, etc. So in this paper we introduce an application that executes query given by the user in Hindi and provides the output in Hindi in the form of speech.

III. NATURAL LANGUAGE PROCESSING (NLP)

Natural Language Processing (NLP) is an approach to analyze text based data or verbal stimulations through both set of theories and set of technologies. It's a very active area of research and development, there is no universal definition that would satisfy everyone, but there are various aspects, which could be a part of anyone's knowledge. It is a division of AI which includes Information Retrieval, Machine Translation and Language Analysis. The major area of concern for NLP is to enable communication between user and system without storage or memorization of complex commands and procedures. In other words, NLP can help the system understand the languages that humans normally use for conversations among themselves. Natural language may be the easiest to learn and use, it has proved to be the hardest for a real time implementation. Despite of various challenges, natural language processing is widely regarded as a promising and important attempt in the field of computer researches. The main area of concern for most computational linguists is to enhance the capability of the computer to understand and generate natural language so that in due course of time people can deal with their computers through text or speech as if they were addressing another person. The Applications that would be generated with NLP capabilities would be able to fully realize and process natural language along with translating languages correctly and in real time and also extracting information from a variety of data sources on demand of users. Language is the ability to express one's thoughts by means of a set of signs (text), gestures, and sounds.

Text-to-speech (TTS) convention transforms linguistic information stored as data or text into speech. It is most widely used in the audio reading devices for the deaf and dumb people now a days. TTS is one of the major applications of NLP. So, the main aim of this work is to study and analyse the concept of NLP and speech synthesis in Hindi. Text-To-Speech (TTS) is a technology that converts a written text into human understandable voice. A TTS synthesizer is a computer based system that can be able to read any text aloud that is given through standard input devices. In general, a TTS system can be broken down into three main parts: a linguistic, a phonetic and an acoustic part.

IV. ARCHITECTURE OF NLP

Generally NLP has following steps:

Morphological analysis: Individual words are analyzed into their components and non-word tokens are separated from the words.

Syntactic analysis: Linear sequences of words are transformed into structures that show how the words relate to each other.

Semantic Analysis: The structures created by the syntactic analyzer-are assigned meanings.

Discourse integration: The meaning of an individual sentence may depend on the sentences that precede it and may influence the meanings of the sentences that follow it..

Pragmatic Analysis: The structure representing what was said is reinterpreted to determine what was actually meant.

The system includes the following modules:

GUI: Designing the front end or the user interface where the user will enter the query in Natural Language.

Parsing: Derives the Semantics of the Natural Query given by the user and parses it in its technical form.

Query Generation: After the successful parsing of the statement given by the user, the system generates a query against the user statement in SQL and further gives it to the back end database.

Data Collection: This module collects the output of the SQL statement and places it in the User Interface Screen as a result form.

V. PROPOSED METHODOLOGY

In this paper, an attempt has been made to develop a Hindi text and Speech Synthesis application as an assistive technology to provide a solution for the Hindi speaking people.

Algorithms used in this approach:

1) Preprocessing:

- Tokenization
- Stemming
- Stop word removal

2) GB theory:

Steps involved in this approach are:

- 1) Tokenize Hindi Sentence.
- 2) Using wordnet dictionary, convert hindi word to english word.
- 3) Priorities as per G.B theory are-
 - I. Noun-priority1
 - II. Verb-priority2
 - III. Subject-priority3
 - IV. Object-priority4
 - V. Preposition-priority5
 - VI. Aux Verb-priority6
- 4) As per these priorities set meaning of hindi sentence.
- 5) Return output as english sentence.

3) Google Search API:

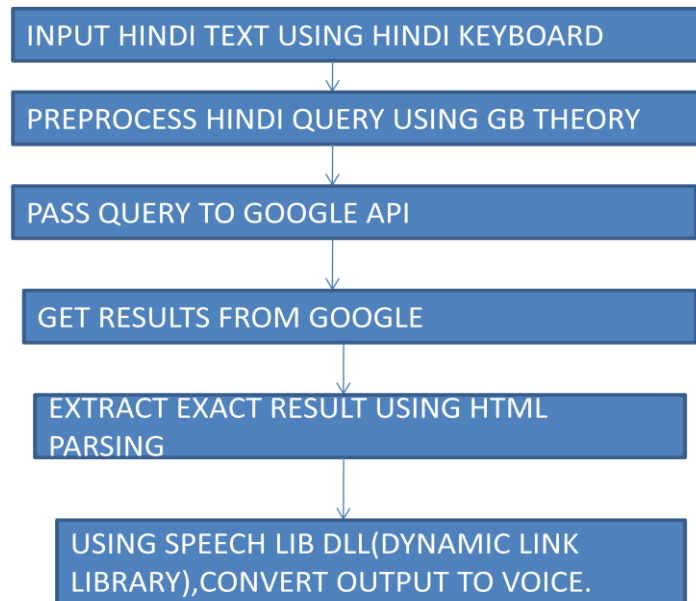
4) HTML Parsing:

Steps involved in this approach are:

- 1) Pass url as input.
- 2) Extract HTML source code.
- 3) After exploring tages reach out to final output.
- 4) Return final output.

5) Speech Lib dll for output as speech: This is the last step of the system. In this the query is converted and the output is given in the form of speech to the user.

Proposed System Design:



VI. CONCLUSION

People with less knowledge such as farmers, labours, who are not aware of English language and lack of computer knowledge where the keyboard and mouse may not be an appropriate means of communication between systems. So it would be a real relief to have the option to use ears to listen to the contents and use Hindi keypad to navigate and control the computer systems.

So therefore we are trying to develop an application that will be very useful to them and hence we propose this methodology. This application may be implemented in various other languages for different types of people as a future scope.

REFERENCES

- [1] C. Allauzen, M. Riley, J. Schalkwyk, W. Skut, and M. Mohri. OpenFst: A general and efficient weighted finite-state transducer library. Lecture Notes in Computer Science, 4783:11, 2007.
- [2] M. Bacchiani, F. Beaufays, J. Schalkwyk, M. Schuster, and B. Strope. Deploying GOOG-411: Early lessons in data, measurement, and testing. In Proceedings of ICASSP, pages 5260-5263, 2008.
- [3] MJF Gales. Semi-tied full-covariance matrices for hidden Markov models. 1997.
- [4] B. Harb, C. Chelba, J. Dean, and G. Ghemawat. Back-off Language Model Compression. 2009.
- [5] H. Hermansky. Perceptual linear predictive (PLP) analysis of speech. Journal of the Acoustical Society of America, 87(4):1738-1752, 1990.
- [6] Maryam Kamvar and Shumeet Baluja. A large scale study of wireless search behavior: Google mobile search. In CHI, pages 701-709, 2006.
- [7] S. Katz. Estimation of probabilities from sparse data for the language model component of a speech recognizer. In IEEE Transactions on Acoustics, Speech and Signal Processing, volume 35, pages 400-411, March 1987.
- [8] D. Povey, D. Kanevsky, B. Kingsbury, B. Ramabhadran, G. Saon, and K. Visweswariah. Boosted MMI for model and feature-space discriminative training. In Proc. of the Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP), 2008.
- [9] R. Sproat, C. Shih, W. Gale, and N. Chang. A stochastic finite-state word-segmentation algorithm for Chinese. Computational linguistics, 22(3):377-404, 1996.