Survey on voice translation techniques

Amey Wadodkar[#1], Prof. J.R. Prasad[*2], Rahul Waghresha[#3], Siddhant Jain[#4], Pratik Tamakuwala[#5]

[#]*Computer Department, Pune University*
Vishwakarma Institute of Information and Technology, Pune, India
[1]ameywadodkar30@gmail.com
[3]rahul_waghresha@yahoo.com
[4]siddjanex@yahoo.com
[5]impratik@gmail.com
[*2]*Professor at Computer Department*
[2]vaishali_prasad@yahoo.com
Vishwakarma Institute of Information and Technology, Pune, India

***Abstract*— Language has always been the primary medium of communication whether in form of speech or text. Wide variety of language has become a hindrance in communication. Especially in a nation like India where the language and dialect changes with region, the requirement of a middle translation layer that can eliminate the linguistic barriers becomes essential. Speakers from different regional identities should be able to interact with one another without the need to understand individual languages. This paper deals with the current technologies used in building a speech translation system. As speech translation is divided into three major categories voice recognition, machine translation and speech-synthesis, we have described optimized technologies of each of the three categories available at present.**

Keywords— **Automatic speech recognizer, Speech synthesizer, FreeTTS, Java sphinx4**

I. INTRODUCTION

In today's socially booming world awareness of various languages has become vital. In present situation language should not be a hindrance in means of communication among people. It is necessary that people speaking different languages communicate with each other fluently without any communication gaps. This speech translation system aims at converting English speech to corresponding Hindi speech and vice versa thus removing the communication gap between Hindi and English speaking people residing in various corners of the world. There were hardly any commercial speech translation system few years back that can meet the requirements, but due to increasing problem of language acting as communication barriers among people of different region, there is a need of a translation system that can fill up the communication gap or remove the language barrier. In recent time's developer have developed some great speech translation system with much more accuracy of translation. The feasibility of system depends upon the scope of the translation which further depends on the dictionaries of a language. Nevertheless such system could be practical and commercial interest, as they could provide language assistance in common yet critical situation, such as people boarding abroad may need for hotel booking, ordering food, and getting direction and so on surveying major speech translation system, we found that this system varied in 3 major different technical aspects such as:

1. Speech to Text: Automatic speech recognition (ASR)
2. Text to Text: Different machine translation approaches.
3. Text to Speech: Speech synthesis techniques.

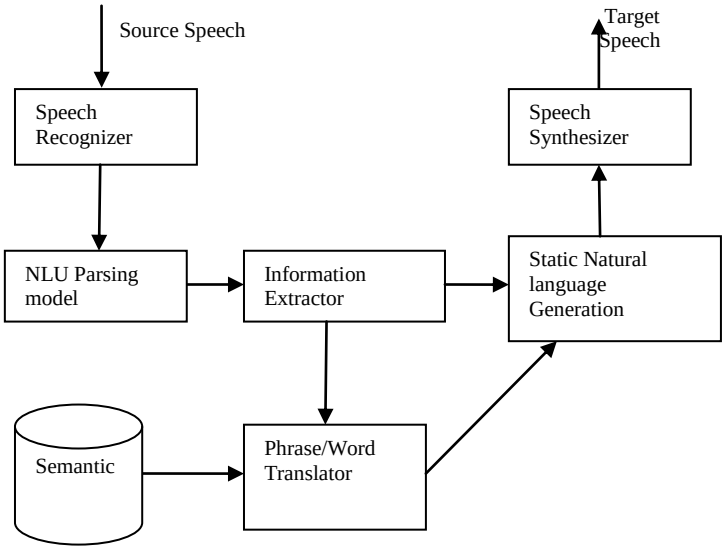

Fig 1: Architecture Diagram

II. AUTOMATIC SPEECH RECOGNITION

Speech recognition is the process of converting voice signal into corresponding text. One of the efficient technologies implemented for recognizing speech includes java sphinx developed at Carnegie Melon University, Julius has been developed as research software for Japanese LVCSR since

1997, and the work was continued under IPA Japanese dictation toolkit project (1997-2000).

A. *Java Sphinx 4*

To implement the speech recognition module, we use Sphinx 4, a speech recognition system developed at Carnegie Mellon University. Sphinx 4 is a refurbished version of the Sphinx engine which provides a more flexible framework for speech recognition, written entirely in the Java programming language. When a user calls from a mobile phone and speaks "I need to translate" on the microphone, the Sphinx module on the processing side captures the words from the audio form and converts it into text output.

Major components of sphinx4 model include:

1. Input: The user voice is taken as input from the microphone of the system.
2. Configuration Manager: The configuration file is loaded by configuration manager as the first step to set all variables. These options are loaded by the configuration manager as the first step in any program.
3. Front End and feature: The front end is constructed to get the input easily, generating feature vectors from the input using the same process used during training.
4. Decoder: The decoder constructs the search manager which in turn initializes the scorer, pruner and active list. The Search Manager uses the Feature and the Search Group to find the best path fit.
5. Result: In the final step, the result is passed back to the application as a series of recognized words. Once the initial configuration is complete, the recognition process can repeat without re-initializing everything.

The basic flow diagram for Sphinx 4 is as follows:

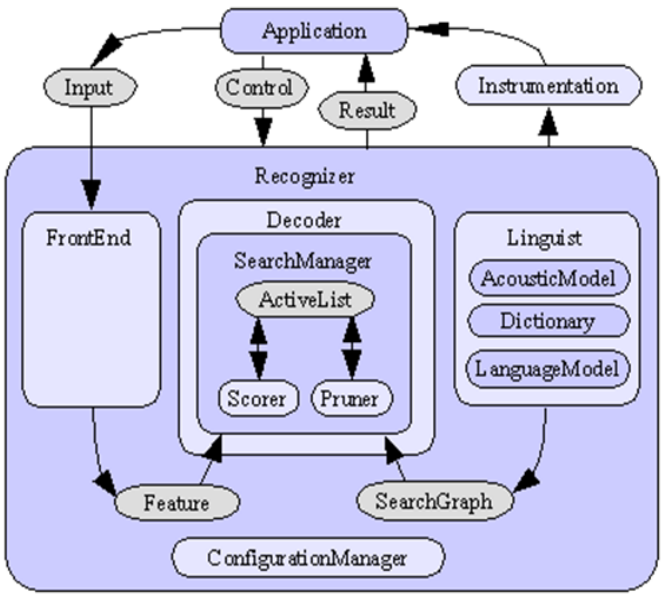

Fig 2: Sphinx flow diagram

B. *Julius*

Julius is a high-performance, two-pass large vocabulary continuous speech recognition (LVCSR) decoder software for speech-related researchers and developers. It can perform almost real-time decoding on most current PCs in 60k word dictation task using word 3-gram and context-dependent HMM. Major search techniques are fully incorporated. It is also modularized carefully to be independent from model structures, and various HMM types are supported such as shared-state trephines and tied-mixture models, with any number of mixtures, states, or phones. Standard formats are adopted to cope with other free modelling toolkit. The main platform is Linux and other UNIX workstations, and also works on Windows. Julius is open source and distributed with a revised BSD style license.

III. MACHINE TRANSLATION

Machine translation is one of the research areas under computational linguistics. Various methodologies have been devised to automate the translation process. However the objective has been to restore the meaning of original text. Over internet, online translation is offered by yahoo and AltaVista through Babelfish. Bing translator of Microsoft and Google translator of Google are tools widely used for translation. In general, there are two levels in process of translation

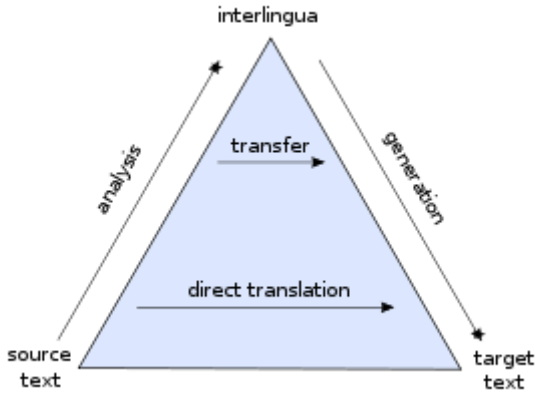

Fig 3: Machine Translation (MT)

Metaphase: Word to word translation. The translated text may not necessarily convey the meaning of original text.

Paraphrase: Translated text contains the gist of the original text. It is not a word to word translation.

Different methods of machine translation being used widely are:

1) Dictionary based machine translation
It's the first generation machine translation technique. DBMT is entirely based on entries of an electronic language dictionary. The word equivalent is used to develop the translation text. It is helpful in translation of phrases but not sentences.

2) Rule based machine translation
RBMT focuses highly on syntactic, semantic and morphological information about the source and target language .In India, Angla-Bharti is a rule based machine translation system from English to Hindi. The aim of RBMT is to convert source language structure to target language structure. It has several approaches.

3) Direct approach
The source language text is translated directly to target language text without passing through an intermediate representation. Anusarka is machine translation system based on direct approach developed at IIT, Hyderabad.

4) Transfer based approach
In this source language is transformed into an abstract, less language specific representation. An equivalent representation is then generated for the target language using language dictionary in grammar rules. It has three phases' analysis, transfer and synthesis.

List of machine translator available:

A. Google Translate Api

Google Translate Api is a paid service provided by Google Inc. With Google Translate, you can dynamically translate text between thousands of language pairs. The Google Api integrate your website and programs programmatically with Google translator.

B. Apertium

Apertium is a free/open-source machine translation platform available in the market. It is a rule-based machine translation platform. It is released under GNU General Public License.

C. Bing Translator Api

Bing translator is the sole property of Microsoft Corporation. This translator has great advantage over Google API that Microsoft is giving it for free to the developers. All translation activity are powered by the Microsoft Translator statistical machine translation platform and web services, developed by Microsoft Research ,as it's the backend translation software.

IV. SPEECH SYNTHESIS

Speech synthesis is an artificial production of human speech. Stephen hawking is one of the famous person to use speech synthesis. The computer system used for this purpose is called speech synthesizer, and can be implemented in software or hardware. A text to speech (TTS) system converts the corresponding text into speech form.

A. FreeTTS

FreeTTS is an open source speech synthesis system written entirely in the Java programming language. FreeTTS is based on Flite. Flite offers text to speech synthesis in a small and efficient way. Flite is part of the suite of free speech synthesis tools which include Edinburgh University's Festival Speech Synthesis System and Carnegie Mellon University's FestVox project, which provides tools, scripts, and documentation for building new synthetic voices. FreTTS provides support to import voice data directly from FestVox. You must first create a voice using FestVox. Festival Speech Synthesis Systems is a free software multi-lingual speech synthesis workbench that runs on multiple-platforms offering black box text to speech, as well as an open architecture for research in speech synthesis. Festival speech synthesis is completely written in C++.Festival offers a free, portable, language independent, run-time speech synthesis engine for various platforms under various APIs. Festival provides support for several languages such as Castilian Spanish, Czech, Finnish, Hindi, Italian, Marathi, Polish, Russian and Telugu. At present FestVox 2.1 and Festival 1.4.3 are available.

There are a number of steps in the synthesis process. Many of these steps need to be customized depending on the locale and the type of synthesis employed. Speech researchers also need the ability to plug in new algorithms easily. FreeTTS provides a general framework for the synthesis process that allows the various steps in the process to be customized. Figure 4 shows the overall architecture of FreeTTS.

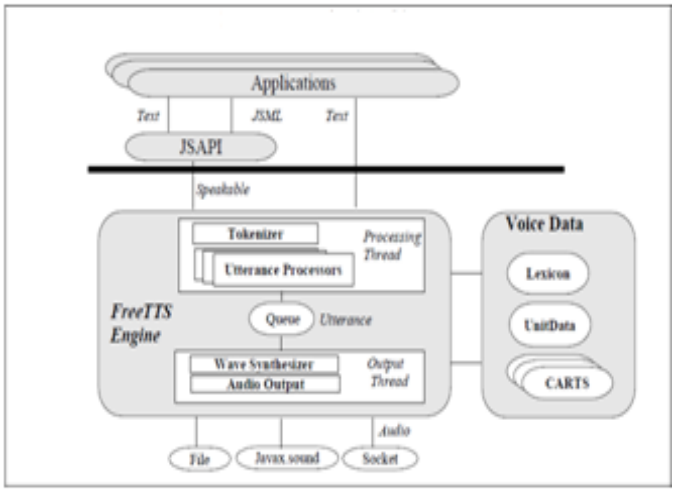

Fig 4: FreeTTS

To support a new FreeTTS voice, a developer provides a set of Utterance Processors and associated data that define the processing for the voice. An Utterance Processor takes as input an utterance, performs some processing on the utterance, and typically adds some annotation or data to the utterance as a result of this processing.

With this framework, the synthesis process becomes the following:
1. Break the input text into a series of utterances
2. For each utterance, apply each of the Utterance Processors
3. Output the generated audio data

A typical FreeTTS voice will define Utterance Processors that perform the following functions:

1) Text Normalization – Converts the input text into a stream of words. For example, the text "DR. Smith lives on 33 Garden drive" would be converted to "doctor smith lives on thirty three Garden drive" The text normalization process deals with a wide variety of cases including numbers, dates, times, titles, and place names.

2) Linguistic Analysis – Attaches semantic information to the utterance. This can include phrasing and part-of-speech information.

3) Lexical Analysis – Determines the pronunciation for each word of the utterance. Typically, a FreeTTS voice will use a lexicon to determine the pronunciation for a word. If a word is not in the lexicon, a set of sophisticated letter-to-sound rules are applied to generate a pronunciation.

4) Prosody Generation – Attaches to the utterance Information about pauses, pitch, duration, tone, stress, and amplitude. These processors will typically use classification and regression trees (CARTS) to generate this prosody information.

5) Speech Synthesis – Generates audio data for the utterance. Typically a synthesis processor concatenates speech units based on diphones or other units of speech. Synthesis can be Particularly CPU intensive since it involves a great number of floating point operations. Synthesis runs in a separate thread to reduce latency and to increase the possibility of parallelism on multi-CPU systems.

A. *Mary*

Mary is an open-source, multilingual Text-to-Speech Synthesis platform written in Java. It was originally developed as a collaborative project of DFKI's Language Technology lab and the Institute of Phonetics at Saarland University and is now being maintained by DFKI. As of version 5.0, MARY TTS supports German, British and American English, Telugu, Turkish, Russian and Italian. MARY TTS comes with toolkits for quickly adding support for new languages and for building unit selection and HMM-based synthesis voices. With the tool "EmoSpeak", MARY can synthesize emotionally expressive speech using diaphone voices.

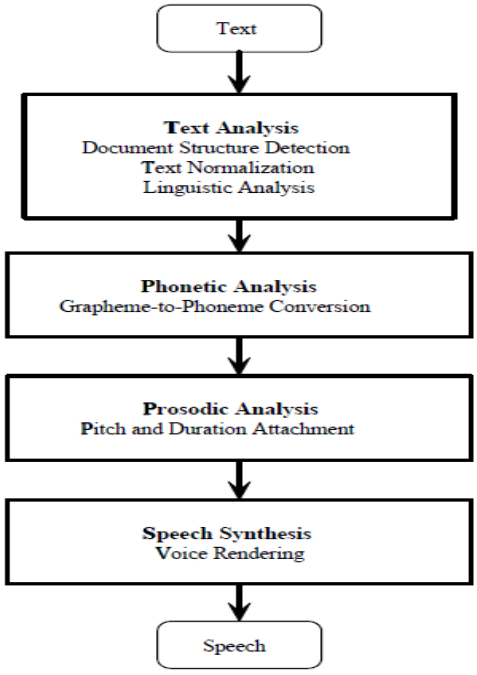

Fig 5: Speech Synthesis

V. APPLICATIONS

The translation of speech from one language to other will remove the bane of language barrier for communication. The speech translation may prove beneficial even while online calling or video conferencing. A person speaking in English can be interpreted on the other side of the call in any language by speech translation. A person giving a conference or lecture or speech in some language can be understood by people in

different languages simultaneously with the help of speech translation.

The speech translation feature is considered to be a boon for the tourism industry. Tourists visiting the country certainly don't know the local language [Hindi in most cases]. However they usually face problems in communicating with local people. Speech translation feature can help convert English speech to Hindi speech and vice versa thus a fruitful two way communication can be established without any hindrance benefitting both the host and visitors. Speech translation can be used during video calling like: on Skype, where people of different regions are communicating and the language of communication is different, than the system will translate it into their respected local language. This can be used for learning different language just by translating any language into a particular language you want to learn.

VI. Conclusion

Language has always been a barrier to effective communication. As businesses expand and technology engulfs the entire globe, reliable and real-time translation becomes imperative. While considerable progress has been done in this direction, more efforts need to be taken in order to reduce the enormous processing time involved with it. With this paper, we propose a new system model to ensure effective real- time communication between two users who do not speak a common language while ensuring minimal computing time.

References

1. Schlenker –Schulte, 1991; Perfetti et al. 2000 with respect to reading skills among deaf readers (Stinson et al.1999: Accuracy). Leitch et al.2002.
2. Zaidi Razak, Noor Jamaliah Ibrahim, Emran Mohd Tamil, Mohd Yamani Idna Idris,"Quarnic Verse Recitation Feature Extraction Using Mel-Frequency Cepstral Coefficient (Mfcc)" Department Of Al-Quran & Al-Hadith, Academy of Islamic Studies, University of Malaya.
3 D. R. Reddy, "An Approach to Computer speech Recognition by direct analysis of the speech wave", Tech. Report No.C549, Computer Science Department, Stanford University, sept.1996.
4. Cremer, Inge (1996): "Prüfungstexte verstehbar gestalten", Hörgeschädigtenpädagogik 4, 50 Jahrgang, Sonderdruck.
5. Willie Walker, Paul Lamere, Philip Kwok, Bhiksha Raj, Rita Singh, Evandro Gouvea, Peter Wolf, Joe Woelfel" Sphinx-4: A Flexible Open Source Framework for Speech recognition"SMLI TR2004-0811 c2004 SUN MICROSYSTEMS INC.
6. B Zhou, Y Gao, J Sorensen, et al. , "A hand held speech to speech translation system", Automatic Speech Recognition and Understanding, 2003.
7. IBM (2010) online IBM Research Source: - http://www.research.ibm.com/Viewed 12 Jan 2010.
8. Y. Yan and E. Bernard, "An approach to automatic language identification based on language dependant phone recognition", ICASSP'95, vol.5, May.1995 p.3511.
9. L. R. Rabiner and B. H. jaung, Fundamentals of Speech Recognition Prentice-Hall, Englewood Cliff, New Jersey 1993.
10. Aakash Nayak, Santosh Khule, Anand More, Avinash Yalgonde, Dr. Rajesh S. Prasad, "Study of various issues in voice translation", Vol2, No 2(2013): IJARCET February-2013.

