Switching Between Multiple Languages Based on Speech Recognition and Translation

[1] Salini. R ,[2] Safrin. P,[3] Shanmugapriyaa. P, [4] Sindhu. S
Department of CSE,
Panimalar Engineering College,
Chennai, India

Abstract—The most basic of all human needs is the need to understand and to be understood. That's when Communication started to play its major role. Language is the medium of Communication. Communication between people with different languages is challenging. Also people with disabilities and Dyslexia find it difficult to communicate through text. Speech to text applications and Translation applications are currently available as disunited entities. This paper is about a application that takes audio as input and converts it to text and additionally translate their text to any other language especially Indian languages. It makes use of simple java API and Translator API.

Keywords: Audio, Translator, Indian languages, Java, conversion, communication

I. INTRODUCTION

The advantage of modern means of communication is that they enable you to worry about all things around the world. People with disabilities like physical impairments often find it difficult in sending a text message or typing a statement in general. The system supports them by allowing them to speak their text through microphone and performing the conversion from audio to text. This system also helps people connect with any languages especially Indian Language from their own language. The system basically uses two techniques- Speech Recognition and Translation.

SPEECH RECOGNITION: Modern speech recognition systems use both an acoustic model and a language model to represent the statistical properties of speech. Acoustic model is used to represent the relationship between an audio signal and the phonemes that make up the speech. Speech Recognition technology allows the user to communicate to another user by talking. It has reached a high level of performance and robustness [2]. Speech Recognition is a process of decoding acoustic speech signal captured by microphone or telephone to a set of words. Figure 1 shows the working mechanism of Speech Recognition System in a brief manner. The system is a Speaker Independent System which means it is designed to recognize anyone's voice and no training is involved [15].

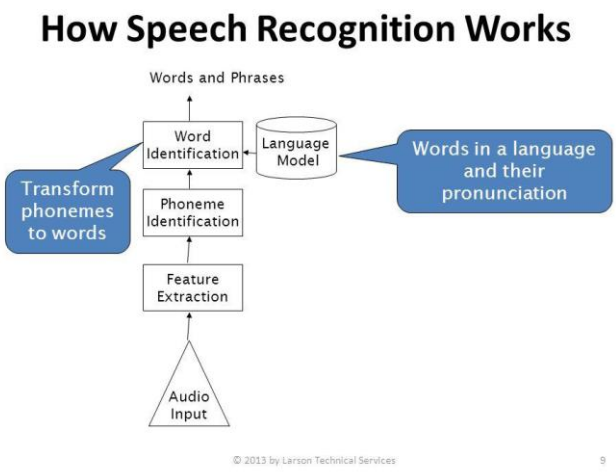

Figure 1.

Translation:

Translation from converted text to selected Language is done with the help of Statistical Machine Language. It gathers information between two Languages in such a way that a word or phrase in one language corresponds to similar thing in another language. The System uses Translator API which implements Statistical Machine Language [2]. In General when we try to teach a person a new Language, we start with alphabets, vocabulary, grammar and sentences. If the same procedure is logically applied to Computer or Machine languages we end up with multiple exceptions and alternatives. So, entirely a new way of teaching is needed for Machine languages. No Rules of Language is pre-defined; all the rules of languages are allowed to be identified by the Computer itself. They analyse Billions of documents which have been translated by human manually and design a pattern by themselves. The documents may be books, novels, websites, and text from Organizations like UN etc. Figure 2 represents the above theory in diagrammatic way. This is where the accuracy of translation is determined. If larger quantities of books are being processed for a particular Language, the translation will be of better accuracy. If the books being processed for a particular Language are of less quantity, then translation may be inappropriate.

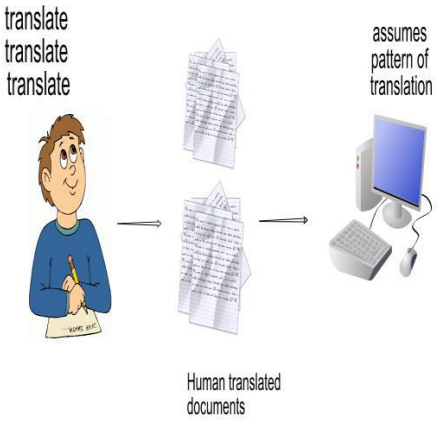

Figure 2.

How It Actually Does?

Consider an example where we need to convert a statement in English (e) into Tamil (t) [1].Now, there are many ways by which we can convert it and there are multiple solutions too. So, we assign a probability over here pr (t|e). The best way to choose is to find the one with maximum conditional probability. To know the conditional probability we use Bayes theorem, pr (e) is already known and hence it always equals to one. Now the formula becomes,

$$pr(t|e) = pr(e|t)\ pr(t)$$

To find pr (e|t) we use a translation model. To find pr (t) we use a Language model. The exact working of the above example in the different models is shown in figure 3. Using the above two models, we calculate the value of pr (t|e), based on conditional probability.

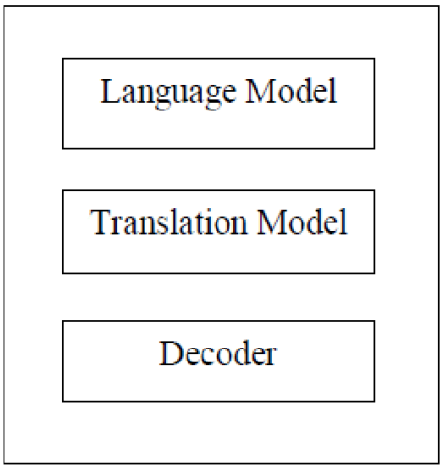

Figure 3.

II. EXISTING SYSTEM

VOX SIGMA SUITE:

VoxSigma is a set of software products which performs speech to text conversion. It is offered by Vocapia Research for platforms such as Linux x86 and x86-64[4].VoxSigma is available as a Web service. The VoxSigma software contains large vocabulary which converts speech-to-text in multiple languages. It allows transcription of noisy speech which helps the users to talk over a background sound [12]. The software is mainly used for business persons who needs to convert large quantities of audio and video files into text format, either in realtime or in batch mode.

Vox Sigma Api:

The VoxSigma REST API is used to perform speech-to-text service in any application. The important feature is that we can use this service by adding a single command line in the application script. VoxSigma API can be used with commandline HTTP clients or with HTTP client libraries. The three main processing functions of the API include:

- language identification
- speech-to-text conversion
- Speech-text alignment.

Language Identification:

Language identification is the process of recognizing the language spoken in an audio document [13]. The language identification component of the VoxSigma software can recognize one of 40 languages.

Speech To Text Conversion:

Speech to text conversion is the process of converting spoken words into text formats [13]. The speech-to-text conversion process includes the following steps,

1. Identification of the audio segments that contains speech.
2. Identification of the spoken language if it is not known a priori.
3. Conversion of the speech segments to text format.

The three models [13] which are used in speech to text conversion process include:

- an acoustic model
- a language model.
- a pronunciation model.

Speech Recognition:

Speech recognition depends on acoustics, phonetics, linguistics, semantics complex, signal processing, acoustic-phonetic models, neural networks and statistical language models. Speech recognition can make errors at times. The accuracy depends on the speaker and also on the background noise.

Speech-Text Alignment:

Speech text alignment is the process of synchronizing a speech signal with a speech transcript, providing time codes for words and sentences [14]. In this technology, time codes are assigned to each word and punctuation marks in the audio transcript and the confidence scores are also provided. When the provided transcript differs from what has really been said, the confidence scores are used in identifying the areas where the alignment is not perfect.

Applications:

- Multilingual audio indexing.
- Transcription of speeches.
- Tele-conference transcriptions.
- Subtitling.
- Telephone speech analytics.

Technical Characteristics:

PLATFORMS	Linux x86 and x86_64 (OpenSuse, Debian, Fedora, CentOS, Ubuntu, SuSE, Red Hat,...)
OPERATING MODES	batch, real-time, single or multithreaded
OUTPUTS	XML with speaker diarization, language identification tags, word transcription, punctuation, confidence measures, numeral entities and other specific entities
SUPPORTED LANGUAGES	Arabic, Dutch, English (US, UK), Finnish, French, German, Greek, Italian, Latvian, Lithuanian, Mandarin, Polish, Portuguese, Romanian, Russian, Spanish, Swedish, Turkish

Disadvantages

Template preparation and matching is expensive or impractical.

- Vocabulary size increases as we keep on adding.
- Storage and processing power is needed to perform the matching.
- Template matching is strictly speaker dependent.
- If there is noise or some other sound at the background, the number of errors will increase.
- Microphone should be placed very close to the user for better performance. More distant microphones increase the errors during speech recognition..
- Pre-processing is done after acquiring the speech. The system should recognize the correct voice of the user and so the training of user voice is required.
- At times, humans have trouble in understanding what is said by another person, so it is obvious that a computer will have the same problem more often than us.
- It involves working with foreign languages such as Greek, German, Italian etc., and not with any Indian languages.
- It is not platform independent.
- It works only on platforms such as Linux x86 and x86_64.

III. PROPOSED SYSTEM

The existing system gets input in English and it processes and then it converts into text of any other languages. Mainly, the existing system concentrates on foreign languages rather than Indian languages. But the proposed system concentrates not only on foreign languages but also on Indian languages. Also the proposed system converts the voice(or speech) of any language and processes and converts into text of any other languages. The proposed system is not restricted to certain platforms. This system directly converts the voice to text of any languages. The system first converts the voice of an language into a text format of the same language. Then it converts the text format into a text of other languages including Indian Language. For example: If the input voice is given in English, it converts it into English text and it displays the equivalent text in any other selected languages. Figure 4 clearly explains how the proposed system works.

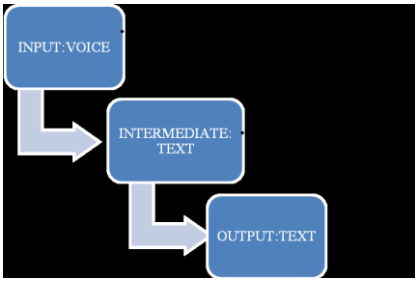

Figure 4.

IV. ADVANTAGES

The proposed system is platform independent. (i.e.,) not restricted to limited platforms.

- This system uses Cloud API for translation, thus reducing the time in displaying the text of the input language.
- Unlike the existing system, the proposed system concentrates on the Indian languages which make it stand unique among the other existing systems.
- It is not necessary to sit before keyboard and type everything.
- It is very much useful for physically handicapped people or people with dyslexia.
- Google's speech recognition engine is used for memory savings.
- It provides hands-free capability.
- It is simply faster than hand-writing and hence the documenting work can be done at a lesser time.
- It helps in providing better spelling, either in text or documents.
- Work processes become more efficient because document processing times become shorter.
- The software learns as it works if it finds the recognition errors. Recognition rate is improved even further.
- Above all, speech recognition is more fascinating and interesting that it transforms spoken words into readable text.

V. METHODOLOGY

Software Used:

The Java Development Kit (JDK) [5] in which java platform, enterprise edition, and standard edition can be *implemented in the form of binary product for java developers* on Solaris, Mac os, Linux, and Windows. Java is platform independent and users can launch the application with the single click of the mouse and if there is any further

updating JAVA Web Start Software will automatically update the installation. Python is not comfortable for every user to understand since for its complexity with minimum code. *Eclipse SDK is a combination of Java Development Tools, and the Plug-in Development Environment and several eclipse projects [6].* There are many alternatives like the following but they have some disadvantages.

1. Dragon speech SDK
2. ISpeech SDK
3. Real sense SDK

The above three has been used in big shot companies but all are payable. Dragon speech SDK is used from SIRI to Amazon along with some of the dictionaries and is not free. ISpeech SDK is light-weight as it has a small footprint but its payable. Real Sense SDK is used by INTEL for windows desktop application. We use eclipse SDK which is cheap and best. *Sphinx 4 is a Java speech recognition library [7]* .It gives quick and easy API to convert audio recordings into text using CMU Sphinx acoustic models and can be used on servers and in desktop applications. The Sphinx4 speech recognition consists of libraries like CMU Sphinx- Speech recognition toolkit which have bindings for many languages. *The Java Speech API (JSAPI) is an interface that allows command and control recognizers, dictation systems, and speech synthesizers for cross platform [8].* Java speech API has speech-to-text supporting (STT) library hence it has been used. Google Cloud Speech API has text-to-speech (TTS) supporting library and it is best since it supports 80 languages and is not free for commercial purpose. It also has features that calculate your pricing plan based on your requirements. The other API is Google speech API which supports Japanese language using simple http interface but Google does not support it. Web speech API will support both TTS and STT.

VI. ANALYSIS

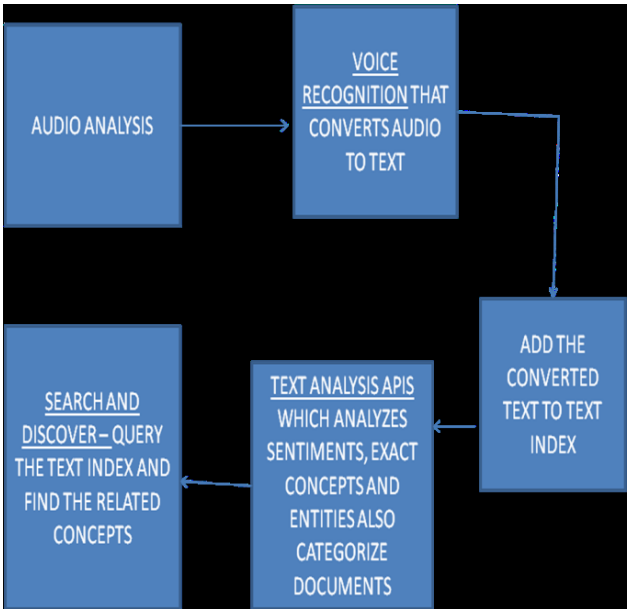

Figure 5. Voice To Text Translation

The voice to text translation is done with the help of language model and acoustic model. This conversion is shown as diagrammatic representation in figure 5.

Language Model:

Language model is a probability distribution over *sequences of words* [9]. Language model helps in distinguishing the similar words and phrases. For example, the word 'aside' and 'a side' are pronounced almost similar, but their meanings are entirely different. To overcome these confusions, language model is used along with pronunciation model and acoustic model. Language model used in information retrieval is called n-gram model. [16] When n is assigned a value 1, it is known as Unigram model and when it is assigned 2, it is known as Bigram model and so on. Let us take the example we used in Introduction section, where we need to find pr (t) using Language model. So, to calculate this we use Hidden Markov Method.

Hidden Markov Model:

The statistical model in which the system is assumed to be a markov process with hidden (as the name suggests) states [7].

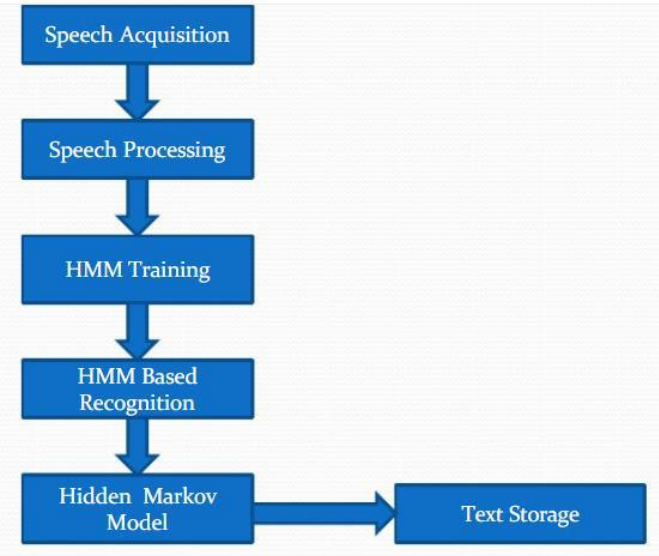

Figure 6.

In simple Markov chain, we predict the next state based on current state without focusing on the previous state. [3] .This focusing of next state rather than previous state is been shown in figure 6 as it is like a downward process. HMM is a simple markov except that the state is not directly visible to viewer. Assume the 't' is divided into t1, t2, t3,.., tn. Now, using Markov chain we get Now if we assume aour n-gram model to be an unigram model, that is, n=1, we get. pr (t1,t2,..,tn)=pr (t1)pr (t2)..pr (tn) In short, pr(t) = product of probabilities (t1, t2, ..tn).

Acoustic Model:

An acoustic model represents the relation between an audio signal and the phonemes [10]. Acoustic model contains the *statistical representation* of each and every distinct sound which forms a word. A label named phoneme is assigned to each of these statistical representations. A phoneme is a unit of sound that distinguishes one word from another. For example, the words 'pat' *and 'bat'*. In Acoustic model, data are fed as documents. These documents are manually converted audio to text. Billions of these documents

pr (t1, t2, t3... tn) = pr (t1) pr (t2|t1) pr (t3|t1, t2)...pr (tn|t1, t2 ...tn−1)

are fed to the computer. The computer designs its own pattern and predicts its own result. For this acoustic model, a huge database is been created in which pronunciation dictionary and grammar of the language is stored. The microphone converts voice signal into electrical signal persistently until pause signal is been reached. If pause is reached, the voice detector will check the statistical representation (phoneme) of voice with pronunciation dictionary stored in database. If any word matches it confirms the language and further checks with the grammar part then finally gives the exact word. In case of silence or no voice detected, the voice detection value will be 0. The acoustic model also makes use of hidden markov model discussed in the language model.

Text To Text Translation:

Cloud translation API is an interface for translating an arbitrary string into any supported language [11]. The mechanism of cloud translator is precisely explained in a diagram format on figure 7. Translation API is highly responsive and can integrate faster. It dynamically translates the source text from the source language to a specified or targeted language. For example English to Tamil. It improves by translating introducing new languages and language pairs. In case of unknown source language, language detection is done for recognizing the language

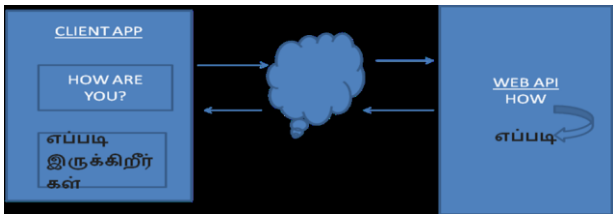

Figure 7.

Working of cloud translation API:
1.) API authentication
2.) Source Language detection
3.) Discovering targeted language
4.) Translating the text

API authentication:

- Declare a key value in the program and pass it through the URL
- Make sure your key value in URL is encrypted for security purpose
- Use JSON as your key type and create a service account
- By default, in your browser's location the service account key value will be downloaded
- The service account authentication can be provided as a bearer token or else through application default credentials (bearer token is an "access token" used as building blocks for security management of your API)
- If we use bearer token, we can directly pass service account key value for authentication
- If we use application default credentials, we need to create an environment variable for service account key value for authentication

Source language detection:

- The parameters involved is source string and an API key
- The source string will be given as a request
- In response, the requested string's language will be detected and its country code is given as a result.

Discovering targeted language:

- Only API key is used as a parameter The source string is given as a parameter and targeted language whose country's code is also given in input
- Using the country's code, targeted language will be discovered

Translating the text:

- The parameters are source string, targeted language and API key
- Source string and targeted language will be given as input
- The source language country's code will be detected and targeted string is displayed as output

VII. RESULT

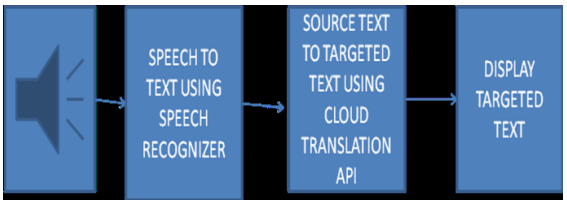

Figure 8.

The whole concept of recognizing voice and converting it into a text according to your choice of language is shown in figure 8.

VIII. CONCLUSION

It is clear that the system supports multiple language recognition and translation by making use of advanced Cloud based services and Speech Recognition System. Quite obviously it encourages localization of languages. Thus the system is beneficial for almost everyone as it saves time in typing the text and makes translation independent of linguists.

IX. FURTHER ENHANCEMENT

We can still improve this system by making using of rtificial Intelligence (AI). Artificial Intelligence is a branch of computer science which deals in simulating intelligence in machines. The ability of the computer systems to perform tasks requiring human intelligence such as speech recognition and language translation can be done using Artificial Intelligence. Better accuracy can be achieved using Artificial Intelligence. Instead of physically selecting the languages to be converted, input can be given through audio or voice. This saves a meager time and helps physically challenged people and people with dyslexia to operate the system entirely on their own.

REFERENCES

[1] Michael Nielsen,"Introduction to Statistical Machine Translation". (references)
[2] Wikipedia"SpeechRecognitionSoftware" "https://en.wikipedia.org/wiki/Category:Speech_recognition_software.

[3] Philipp Koehn " Introduction to MT Research" http://www.statmt.org/
[4] Vocapia Research "VoxSigma Software Suite" http://www.vocapia.com
[5] Wikipedia "Java Development Kit"
[6] Eclipse SDK http://www.eclipse.org/eclipse/
[7] Sphinx 4 https://cmusphinx.github.io/wiki/tutorialsphinx4/
[8] Java speech API "https://en.wikipedia.org/wiki/Java_Speech_API"
[9] Speaker Independent Continuous Speech to Text Converter for Mobile Application "https://arxiv.org/ftp/arxiv/papers/1307/1307.5736.pdf/"
[10] Acoustic model http://www.voxforge.org
[11] Cloud translation API https://cloud.google.com/translate/
[12] Vocapia Research "VoxSigma Speech to Text Software Suite" http://www.vocapia.com/voxsigma-speech-to-text.html
[13] Vocapia Research "Speech to Text Conversion" http://www.vocapia.com/speech-to-text.html
[14] Vocapia Research "Speech to Text Technology" http://www.vocapia.com/speech-to-text-technology.html
[15] CharuJoshi"SpeechRecognition" "https://www.slideshare.net/charujoshi/speech-recognition" "StatisticalLanguageModeling"https://homepages.inf.ed.ac.uk/lzhang10/slm.html