International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 04 Issue: 04 | Apr -2017 www.irjet.net p-ISSN: 2395-0072

Integration of Speech, Image & Text Processing Technologies

Akhil S. Deshpande[1], Shreyas S. Vaidya[2], Pravin B. Swami[3], Pavan R. Jaiswal[4]

[1]Student, Computer Engineering (B.E.), P.I.C.T, Pune, Maharashtra, India
[2]Student, Computer Engineering (B.E.), P.I.C.T, Pune, Maharashtra, India
[3]Student, Computer Engineering (B.E.), P.I.C.T, Pune, Maharashtra, India
[4]Professor, Computer Engineering (B.E.), P.I.C.T, Pune, Maharashtra, India

---

Abstract - *In everyday life, speech is considered as one the most important medium of communication. While conveying the message the most widely used form can be termed to be a speech signal. The main objective behind this work is to utilize this speech signal to simplify everyday life of common people. In today's world, technology is hugely gaining popularity as it is meant to simplify the day-to-day life of people but on a deeper note it is in turn making things complicated as a person need to go through numerous number of automated client support systems. A person has to deal with huge amount of text at every moment thus the use of speech signal has become important especially in multitasking. Therefore, this paper proposes a design to satisfy general requirements which uses three methodologies namely image-to-text, text-to-speech and speech-to-text. It provides a simple working (processing) of these methodologies and a way in which these can be collaborated.*

Key Words: **Text-to-Speech, Image-to-Text, Speech-to-Text, Unit Selection Synthesis, Natural Language Processing, Image Processing, Text Processing**

1. INTRODUCTION

The most important form of communication is speech communication rather communication is defined by speech itself. In our everyday life, we need to communicate with other people in order to carry out our tasks. This is referred as man-to-man communication. But there exists one another aspect of communication known as man-to-machine communication. In this era of technology a person needs to interact with a large number of client support systems to get done with simplest of tasks. This becomes tedious and time consuming if the communication is through text. This explains why humans want to have speech as communication/interaction medium with computers as well.

1.1 Speech Input and Speech Output

In general, a speech-based user interface requires both, speech input (recognition) and speech output (speech synthesis). When we think of these two, several arguments (merits) come along as follows. Speech is convenient as it makes hands and eyes free for other activities. In addition to that communication with a machine and other humans is simultaneously possible. In this kind of system design the user is not bound to a fixed place, has freedom of movement and orientation etc. The system especially can be used by visually impaired and other disabled people (e.g. physically handicapped) gracing it a commendable social value. However, there are few thing which we need to take care of. For example, the speech input can be disturbing for the environment, the recognizers are extremely sensitive against environment noise. Above all, background speaker, for some applications (e.g. those with high security requirements) the recognition accuracy might be insufficient. Thus usually high efforts for system training are necessary.

1.2 STT, TTS AND ITT

When a user speaks to a conversational interface, the system has to be able to recognize what was said. The speech-to-text (STT) component processes the acoustic signal that represents the spoken utterance and outputs a sequence of word hypotheses, thus transforming the speech into text. The other side of the coin is text-to-speech synthesis (TTS), in which written text is transformed into speech. There has been extensive research in both these areas, and striking improvements have been made over the past decade. In the following sections, an overview of the processes of STT and TTS is provided.

One another perspective of the human machine communication is the input in the form of an image or picture instead of speech. This also replace the text as fundamental input form and plays a vital role. In this case speech input is replaced by image input but the speech output remains unaltered. Image-to-text conversion (ITT) comes into the picture. At the end of the discussion, a system design is proposed which depicts a way in which STT, TTS, ITT can be collaborated to produce a convenient work unit.

2. TEXT TO SPEECH:

TTS stands for Text-to-Speech (also written as Text to Speech) a form of speech synthesis that converts text into voice output. Text-To-Speech software basically takes the text you write and turns it into speech files that you can use. There are numerous ways to create audio from text. There is a process called Unit Selection Synthesis (USS). The process starts on both ends, voice database building language text processing that meets in the middle to produce speech. In fully or partially Customer support services, Unified Messaging Systems, specifically email reading systems, Interactive voice response systems, examples banking applications TTS system plays important role. Unfortunately, some of the educationally backward, visually impaired, illiterate people are deprived of this opportunity. To solve this issue a system is designed to speak the text in the language in which such people can understand it. This is achieved by a Text to speech system.

To develop complete text-to-speech model first we need to recognize the natural sound of human. For this purpose a voice actor is chosen (with a great sounding voice) who is fluent in certain language this actor is made to speak the whole sentences or paragraphs for that matter. This is recorded and a voice database is built. Thus a database is created that has thousands of recorded sound files. It is required to sort and organize these files; the speech units are labelled and segments by phonemes, syllables, morphemes, words, phrases and sentences. There is one dedicated unit for natural language processing. In that, text is normalized and broken down into phonetic sound. Then it is made to go through a series of analyses to understand the structure of sentences as well as to determine the context of the word for pronunciation. Now these two modules namely the voice database and the natural language processing are collaborated to produce speech. This is done as follows: Once the NLP is complete, the voice database is searched and the speech units are selected that best fit together to produce the sounds associated with the given text. This entire procedure is called Unit Selection Synthesis (USS). We use USS because it is considered to produce the most natural sounding speech. The other main speech synthesis technique used today is called HTS (HTS based speech synthesis system).

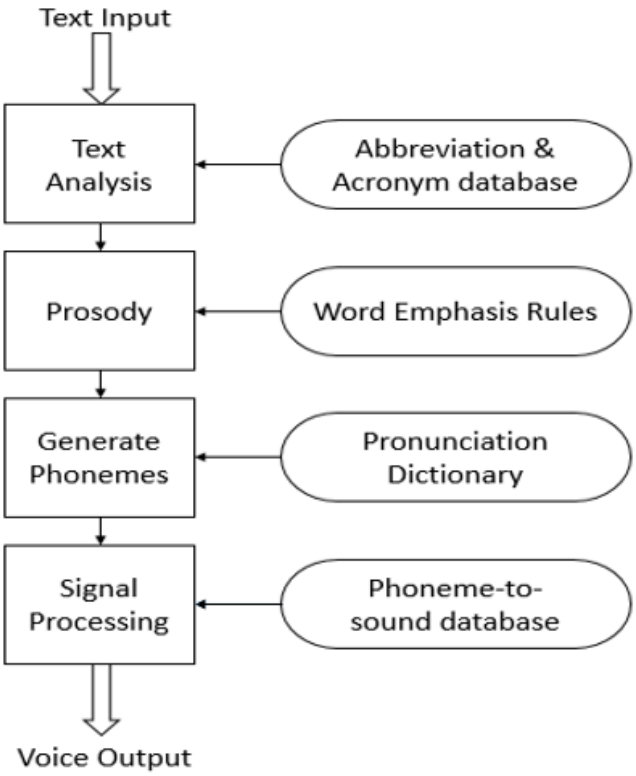

Fig-1: Text-to-speech conversion

3. SPEECH TO TEXT:

The speech-to-text basically refers the conversion of speech signal to the text format. Thus to achieve this a computer has to go through several complex steps, When a person speaks vibrations in the air are created which are translated into digital data by analog to digital convertor. This mainly comprises of sampling and digitization of the sound wave by taking precise measurements at frequent intervals. There is further processing which needs to be performed on digitized sound. One of them is filtering to remove unwanted noise and to segregate the sound into different bands (as human's pitch differs). The other one normalizes the sound that is adjusts it to a constant volume level. For example it is not always possible to all people to speak at same speed, so the sound must be adjusted to match the speed of the sound samples already stored in the system's memory. The signal is then divided into small segments. Then these segments known as phonemes (a phoneme is the smallest element of the language) are matched. Then these phonemes are examined in the context of the other phonemes around them

which requires a complex statistical model and compares them to a large library of known words, phrases and sentences.

The speech recognizer then determines what the user was probably saying and either outputs it as text or issues a computer command. This task of determining is not that simple as it appears. The recognizer has to do comparison of all the phonemes against a lexicon of pronunciation. While doing this it restricts acceptable inputs from the user. For example, most people only use relatively small grammar or certain specific word sequences. One of the technique to reduce the computation is the use of context free grammar which limit the vocabulary and syntax structure. It significantly reduces the number of generated hypotheses. There are also other techniques such as discrete dictation in which the use can speak any word out of vocabulary but must leave pauses between the words and continuous dictation in which user doesn't leaves pauses.

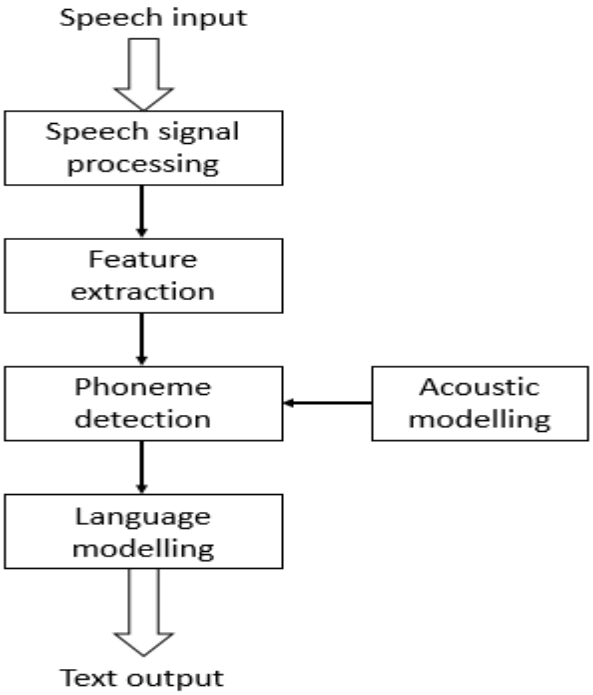

Fig-2: Speech-to-text conversion

4. IMAGE TO TEXT

Text recognition is the process of detecting text in images and video streams and recognizing the text contained therein. Once detected, the recognizer then determines the actual text in each block and segments it into lines and words. The Text API detects text in English language in real-time, on device.

The Text Recognizer segments text into blocks, lines, and words. Roughly speaking:

1. A Block is a contiguous set of text lines, such as a paragraph or column.
2. A Line is a contiguous set of words on the same vertical axis.
3. A Word is a contiguous set of alphanumeric characters on the same vertical axis.

IIT requires preprocessing of images to improve the recognition efficiency. The preprocessing includes several techniques such as De-skew, Despeckle, Binarisation, line removal, layout analysis, etc. Binarisation means converting an image from color or grayscale to black and white. It is performed in order to separate the text from the background. Segmentation of fixed-pitch fonts is accomplished by aligning the image to a uniform grid based on where vertical grid lines will least often intersect black areas.

The main phase, character recognition produces a ranked list of candidate characters. It comprises of two methods knows as matrix matching and feature extraction. Matrix matching also known as pattern recognition or image correlation, involves comparing an image to a stored glyph on a pixel-by-pixel basis. Whereas, feature extraction decomposes glyphs into features like lines, closed loops, line direction and line intersections. These are compared with an abstract vector-like representation of a character, which might reduce to one or more glyph prototypes.

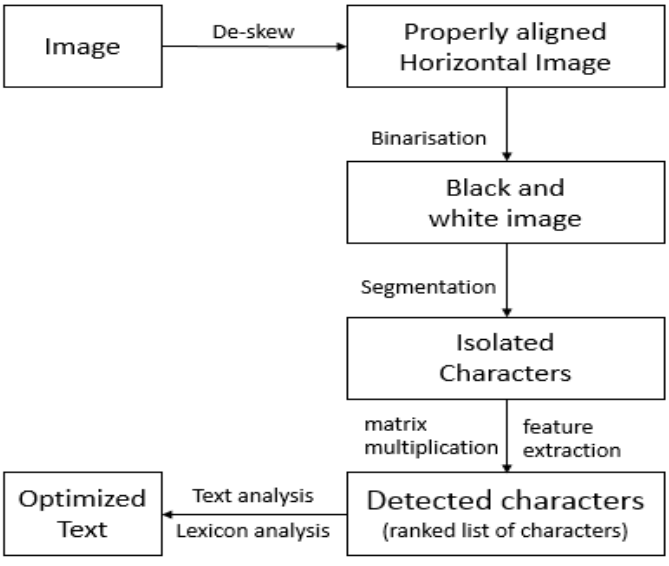

Fig-3: Image-to-text conversion

5. PROPOSED SYSTEM

Our proposed design takes input in the form of speech signal or image. This input is stored for the further processing purpose. Initial steps of processing is convert the speech or

image input to text format using speech-to-text or image-to-text conversion. Once we get the entire input in text format, it is analyzed and further processed as follows. First the text is normalized into the destination format so as to adjust it according to usage. For example, if we want to automate e-mail service using speech input format we need to take into account the constraints and restrictions imposed in case of standard e-mail addresses. We also need to segregate keywords from the normal text in some cases. This preprocessed text is given as input to the required system processing component to carry out its actual usage. The output which will be generated from these various system components will again be converted into speech signal as the final output using text-to-speech conversion.

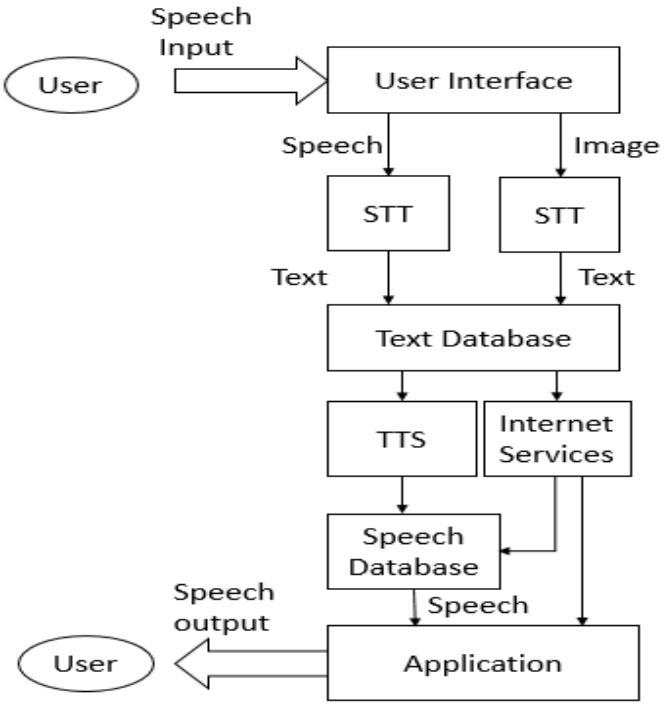

Fig-4: Proposed system architecture

6. CONCLUSION

The purpose of this paper has been the integration of several domains to implement a comprehensive proposal for the new trends design of conversational systems. In this paper the design and development of an integrated system which uses conversion techniques from image or speech to text has been discussed. Text to speech device can change text input into speech signal similarly the other two methodologies perform the job of interconversions. The system with this design will simplify the tasks of people especially those who are visually impaired, handicapped and the people who are more comfortable in working with speech than with text.

7. REFERENCES

[1] Annisa Arrahmah , Samantha Harisa , Hasballah Zakaria , Richard Mengko "Text-to-Speech Device for Patients with Low Vision", 2015 4th International Conference on Instrumentation, Communications, Information Technology, and Biomedical Engineering (ICICI-BME) Bandung, November 2-3, 2015

[2] Mouna Abdelkefi, Ilhem Kallel, "Conversational Agent for Mobile-Learning", University of Sfax, ENIS, BP 1173, Sfax, 3038, Tunisia

[3] Yogita H. Ghadage, Sushama D. Shelke, "Speech to text conversion for multilingual languages", Communication and Signal Processing (ICCSP), 2016 International Conference on 6-8 April 2016

[4] Christos Liambas, Miltiadis Saratzidis, "Autonomous OCR dictating system for blind people", Global Humanitarian Technology Conference (GHTC), 2016

[5] http://www.codeguru.com/cpp/g-m/multimedia/audio/article.php/c12363/How-Speech-Recognition-Works.htm

[6] https://en.wikipedia.org/wiki/Optical_character_recognition

[7] http://www.danvk.org/2015/01/09/extracting-text-from-an-image-using-ocropus.html