Textaloud Assistant App Development for Multilanguage

E.Kodhai, S.Abinayalakshmi, D.Pretha, D.Anjana

Abstract- Spoken words play a big role in individual lives of the people. The conversion process of message written in English text to equivalent spoken form is known as Speech Synthesis. The paper proposed is to convert text-to-speech (TTS) for regional languages. The proposed work is defined in three main aspects. The conversion of the English text to speech is the initial step. It is beneficial for the mute person, who would be able to have voice communication. It provides the text to local language speech conversion. It is beneficial for the regional persons to communicate with people. The integration of the presented system on the android platform is the final aspect. Genuineness and fluency are the two important features of the synthesized speech. The saving of the information from websites and documents that are in different language is supported by the text to speech system. The necessary stages are character recognition, database formation and text to speech conversion.

Keywords: Text-to-Speech conversion (TTS), Creation of database, character recognition.

I. INTRODUCTION

Text-to-Speech (TTS) conversion is performed on android platform environment. It is the process of automatically converting a text into speech that is similar to a natural language who is reading that text. Text-to-speech synthesizer (TTS) technology is used by computer to speak like us[3]. The TTS system accept sentences as its input and uses the system's algorithm in the TTS engine then checks the sentences, pre-processes them and the last step is to convert them to speech with some equations. The generated output is in form of sound that can be saved in audio format.

There are two major process in the text to speech synthesis processes. The text analysis is first, the transcription of input text into a phonetic sound or some other language representation, and the next step is the generation of speech and then the result is generated from the given data.

The input text file may be in any form of data from a word processor, a mobile text-message, standard ASCII from e-mail, or scanned text from any article.

The presented paper aims at developing a working model of speech synthesizer for English as well as other languages that include Tamil and French for android based mobile phones along with light weight English speech database creation for android mobiles[5]. The work will make it user friendly environment for making the application effective.

Syllable is a unit of speech that may be a full word or a single part of word which is separated by a vowel sound. For example "man" has only one syllable, whereas the word "writing" has two separate syllables and the word "syllable" has three separate syllables.

There are totally 44 sounds of pronunciation in which consonants have about 22 of them and the remaining 22 is for vowels sounds. All the sounds of pronunciation are stored in a database for future use. When the text is read the wave file for every corresponding syllable are concatenated and played.

Example:

Word: usable

Syllables: us-a-ble

The wave file for the words that are given above when played it is in form of stammering and hard to understand the sound that is being played. It is because the syllable wave sound for each of the words are played one after the other and it is not in a continuous form with the sound that was previously played. It is important in choosing the sound unit with proper length so it is natural and understandable when the sound is generated.

All the process is made easy by using the code available in the android studio which is used for converting the text-to-speech. In the proposed system we are going convert the text into multi language. First the user must login into the app then he should enter the text or choose the file or the SMS that will be converted into speech. After the conversion stage the speech is stored as an audio MP3 file in the device which can be used for later use[9]. An added feature in the proposed system is that the frequency of the sound can be adjusted according to our choice.

Conversion for characters to sound is not an impossible task.

Revised Manuscript Received on May 04, 2019.

Kodhai. E, Sri Manakula Vinayagar Engineering College, Dept. of CSE, Puducherry, India

Abinayalakshmi. S, Sri Manakula Vinayagar Engineering College, Dept. of CSE, Puducherry, India

Pretha. D, Sri Manakula Vinayagar Engineering College, Dept. of CSE, Puducherry, India

Anjana. D, Sri Manakula Vinayagar Engineering College, Dept. of CSE, Puducherry, India

Since in the language of English there are only 26 letters and each of them have a different sound.

It is not recommended at the user level to convert characters to speech while a lengthy text has to be read because it will be hard to understand the words that is being read character by character.

When each character is read the wave file can be played, the other method is to play the wave file each word that is being read though it is practically not possible to store for all the words in the dictionary. So it is necessary to have an alternative way for backup. The initial attempt was to play the sound of each syllables in the word hence the word will be played as a whole.

II. RELATED WORKS

Erik Blankinship and Richard Beck [12] performed a work on ‘tools for expressive text to speech markup’. The paper is used by the handicapped for accessing the text to speech system developed for the poets and improve their performance. The author of the paper developed an application named as the poet shop where the user would be able to graphically change the volume and the pitch contours. The work is not smart enough for knowing where to place the stress in words and it does not create a natural sounding voice. Aidan Kehoe and Ian Pitt [9] performed a work on “designing help topics for use with text to speech”. For assisting the creation and testing of the materials for help presented for the users through the speech synthesis engines there have been many ways produced. In the paper, the initial process for the creation of a system that provides help topics for the TTS users. The sound of the speech produced by this was comparatively good and the accuracy was also improved. It is not designed to get the input text from the user instead it can only be able to read the content from the help system.

EyubB.Kaise and YaregalAssable [11] published a paper on “concatenative speech synthesis for Amharic using unit selection method”. It defines the algorithms and methods that can convert general Amharic text to speech. This was designed to get the input from the user using user friendly interface. This can be used for only one language. The accuracy of this work was comparatively low. SheillyPaddaet al. [8] performed a work on “A step towards making an effective text to speech conversion system”. Text to Speech conversion for Punjabi language was discussed in this work. Here the accuracy of the work was improved with the help of huge database. This online application can only be able to synthesis the speech in regional languages.

Swati Ahlawat and Rajiv Dahiya [10] published a paper on “A novel approach of text to speech conversion under android environment” that discusses the different approaches for the purpose of text to speech conversion. In this work, Text to speech TTS conversion is performed on android platform. It was the first work to convert regional language text to English speech in an android applications. Its accuracy was good but it cannot browse for the file and could only convert the regional language to multiple languages which was a drawback.

R. San-Segundo et al. [14] paper work “Multilingual Number Transcription for Text-to-Speech Conversion” they used a trainable conversion system with three modules and they use a number transcription method was used. It was developed for the synthesis of multi-language speech. Since must store all the language details it required a large data set for training the system which was a major drawback.

In the paper work of ItunuoluwaIsewon et al. [8] “Design and Implementation of Text To Speech Conversion for Visually Impaired People” was proposed. It used a concatenating speech synthesis method on an android platform. Instead of training the system with huge set of data, after the conversion of the text to speech the file was stored as an audio MP3 file in the system for later use. Its major issue was that it cannot browse for the file.

The Paper work on “Android Based Punjabi TTS System” by Hardeep and Parminder Singh [1] was for developing a Text to speech system for Punjabi language. They had a special feature of browsing through the file in the mobile device or the SMS that is available in the device for adding it for the conversion purpose. Their drawback was the quality of the speech synthesized.

The paper “Statistical Parametric Speech Synthesis Incorporating Generative Adversarial Networks” by Saito. Y et al. [10], proposes a method that use a GAN system which is known as the generative adversarial networks. In the paper a new system called the GAN networks was introduced it has 2 neural networks a generator to deceive the discriminator and a discriminator to distinguish natural and generated samples. The natural speech and the generator speech parameters are distinguished using a discriminator. The quality of the speech synthesized is low when compared to that of the natural speech even though powerful deep neural networks techniques are applied. There are training given to the acoustic models for generating high-quality speech are studied in the existing system since they can be used for both TTS and VC. There is over-smoothing effect on the speech parameters that is generated from these models is the one of the quality degradation issues observed in the generated speech. Though deep neural networks had been used to improve the quality of the speech there are certain problems associated with the speech quality. This was implemented in a computer it is not available in mobile devices.

III. PROPOSED SYSTEM

The proposed system can convert typed text or the text from any browsed document into speech. The proposed system consider the input sentence, and analyze the sentence word by word in order to find out their corresponding phonemes.

It makes use of the phoneme audio library in order to concatenate the phonemes obtained from the previous step and finally it plays the audio which sounds similar to the speech. The proposed system is capable of producing the waveform for the generated speech sound. After the conversion process the speech is stored as an audio MP3 file in the device which can be used for later use. An added feature in the proposed system is that the frequency of the sound can be adjusted according to our choice. This application can be implemented with the help of android studio.

Fig 1 displays the system architecture of the Text-To-Speech Conversion application.

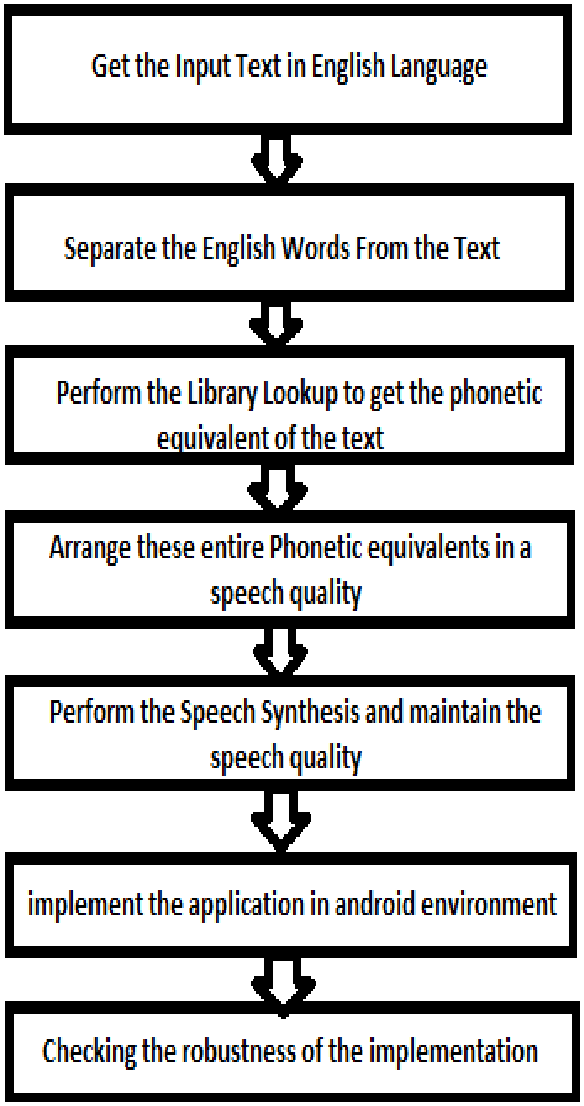
Get the Input Text in English Language
Separate the English Words From the Text
Perform the Library Lookup to get the phonetic equivalent of the text
Arrange these entire Phonetic equivalents in a speech quality
Perform the Speech Synthesis and maintain the speech quality
implement the application in android environment
Checking the robustness of the implementation

Fig .1 Architecture of the System

The proposed system has 4 basic modules they are as follows:

- Registration module
- Language Selection module
- Text processing module
- Speech synthesis module

Registration Module

The Registration is one of the primary modules which is used for signing up into the application for accessing the features of the application. It helps the user to create an account for them for accessing the content. This module requests the basic information of the user for the creation of an account. Once the user logged into the app he can enter the text or choose the file or the SMS that should be converted into speech.

It can be used for registering the users into the application and would be able to access the restricted features of the application[8]. A guest user will also be able to access this application but may not be able to access some of the features of the app. The user would be able to register only once for avoiding confusions. Once registered the user would be able to login into the app the user's id and password of his account. The most important task is the creation of database for the text to speech system which is used to save all the information about the user. The major criteria for designing the TTS system is the selection of the suitable database which serve the purpose.

Language Selection Module

Conversion for characters to sound is not an impossible task. Since in the language of English there are only 26 letters and each of them have a different sound. It is not recommended at the user level to convert characters to speech while a lengthy text has to be read because it will be hard to understand the words that is being read character by character.

In character to speech conversion it is must to generate the file that stores the waveform for each word in the sentences which is read. It is also necessary to play the wave file which is generated[2]. Clearly it is not possible to store each and every words that are present in the dictionary. Here it becomes mandatory for having some other alternative ways. The first way was to play full word that is playing the syllables of a word.

Apps may include certain resources that are very specific to a particular language. Language selection module provides the opportunity for the user to select the desired language to which the text must be converted to the speech. In this module, some of the tradition oriented strings may be included in the app that must be translated to the language chosen by the user. This module resolves language and culture-specific resources based on the chosen language. This is made possible with the help of the resource directory in the Android studio which is used for the implementation process.

After the selection of the language the most critical step involved in this module is the selection of the most suitable parts of speech that will be output of smooth utterance of sounds. There are three major phases involved in the construction of inventory.

Usually, the natural speech is recorded in order to make sure that it includes all the used contexts. The units from the spoken speech data must be labeled then it is a must to choose the appropriate units. Sample collection from natural speech is basically time-consuming task. The most important task is to select correct samples for concatenation based on certain rules.

Few users prefer a language that uses right-to-left (RTL) scripts, which include Arabic or Hebrew. Other user prefer to generate the content in a language that uses Right to left scripts, though they've set a language that uses Left to right scripts, such as English, as their UI locale.

In order to fulfil the needs of both the users, the following features are to be included:

- Need to employ the RTL UI layout for user who prefers RTL locales.
- Need to determine the direction of text data that are to be displayed. In common, this process is done with the help of method call.

Text Processing Module

The input text is obtained using a text field, the text input can be obtained from a document in the device or from the SMS. The text may be a word or a sentence. The input text is split up by character since each character has its own sound. This process will help us to reduce the time for converting the text to speech. The database system which we use for the text to voice conversion process should contain the stored alphabets(a-z) and numbers(0-9) in waveform files (.wav) format. The final strategies the creation of text file (.txt).

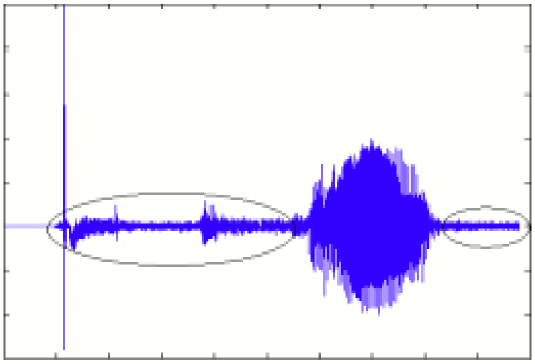

Fig .2 Default delay in the recorded sound

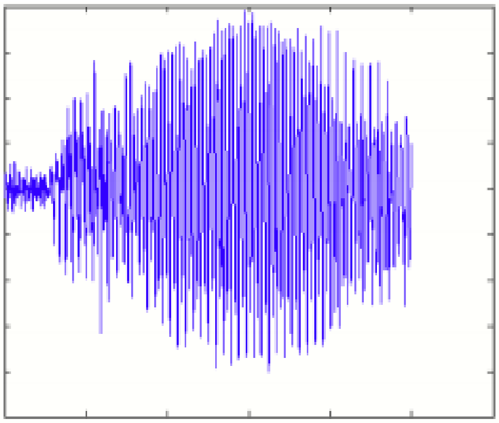

Fig .3 Sound waves after the removal of delay

Fig 2 depicts the waveform of sound being stored manually which already have an existing delay. These delays have to be deleted and the speech has to be re-written. Fig 3 shows the stage of the sound after the error is deleted.

There may be some delay of sound by default in case of recording the sound. Such kind of time consuming task has to be avoided in order to have a continuous speech utterance.

In this module, the input text that include the abbreviations, numbers and symbols are not considered. In natural language processing technique, usually the input sentence is spitted into words and their Parts of speech (POS). Finally tagger is assigned to each and every word by using method known as bigram technique. Then, each word is converted to phoneme transcription with the use of dictionary based approach. These features are considered as an input for the unit selection.

After the conversion the speech output will be stored in the device as a MP3 audio file. This is done for future purposes for referring the stored content.

Speech Synthesis Module

Speech synthesizer is based on a computer system that is capable of reading aloud the given text, when it is given directly to the computer.

Speech synthesis is defined as a generation of speech by 'grapheme to phoneme' transcription by machines automatically. The written language has a smallest part known as a grapheme which does not carry any meaning. It includes alphabets, numbers, punctuations, and the individual marks of any of the world's writing systems. A phoneme is said to be "the smallest segmental unit of sound employed to form meaningful utterances".

It is clear, that this involves some sort of input. If the input given by us is the plain text, which does not have any additional phonetic information then the system is said to be the text-to-speech (TTS) system. The TTS system translates the ASCII sentence to speech.

The most important step is the extraction of the pronunciation data of the message, so that we can obtain the string of symbols denoting sound (phonemes or allophones), edges of the words, phrases present in the given information.

The next stage includes the method for finding the similarity among the various symbols and suitable data stored in the phonetic database and combining them to generate the acoustic waveform for the voice.

To compute the output, the system need the following

- The parameter value of the sounds stored in the database.
- The various methods for generating the sound must be stored in the knowledge base.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

IV. CONCLUSION

This paper is a try made in order to implement the text to multi-language speech conversion using text to speech conversion technology. The ultimate aim is to develop an application which is easy to access, involves less cost and suitable for all real time purposes. With the help of this approach, we are given the access to read the sentence which is present in the file. It is also possible to read the text from the browser which can be used as the input to produce the output [6]. The speech is produced through the speaker present in the phone or computer. The approach which is defined in this paper has set of all policies for every alphabet, their method of pronunciation, the manner how they are used in standard dictionary. The time saving feature is added which provides access for the user to hear the generated speech simultaneously doing some other work. The users can also chose the file that as to be converted to speech from the stored files this will help people who are not able to perform write and read operation. The text present in the browsed content can also be very easily converted to speech. People who are visually challenged or completely blind may make use of this method for reading the documents and books. People who cannot speak may utilize this method to convert given text to vocal sound.

REFERENCES

1. Singh, P. (2015). Android Based Punjabi TTS System, 3(3), 233–237.
2. Trivedi, A., Pant, N., Shah, P., Sonik, S., & Agrawal, S. (2018). Speech to text and text to speech recognition systems-A review, 20(2), 36–43. https://doi.org/10.9790/0661-2002013643
3. Jadhav, A., &Patil, A. (2012). Android Speech to Text Converter for SMS Application. IOSR Journal of Engineering, 2(3), 420–423. Retrieved from http://www.iosrjen.org/Papers/vol2_issue3/H023420423.pdf
4. Myat Mon, S., &MyoTun, H. (2015). Speech-To-Text Conversion (STT) System Using Hidden Markov Model (HMM). International Journal of Scientific & Technology Khilari, P., & P, P. B. V. (2015). Implementation of Speech to Text Conversion, 6441–6450. https://doi.org/10.15680/IJIRSET.2015.0407167
5. Research, 4(06), 349–352. Retrieved from www.ijstr.org
6. Reddy, B. R., &Mahender, E. (2013). Speech to Text Conversion using Android Platform. International Journal of Engineering Research and Applications (IJERA), 3(1), 253–258. https://doi.org/10.1016/j.jseaes.2009.11.002
7. Saito, Y., Takamichi, S., &Saruwatari, H. (2018). Statistical Parametric Speech Synthesis Incorporating Generative Adversarial Networks. IEEE/ACM Transactions on Audio Speech and Language Processing, 26(1), 84–96. https://doi.org/10.1109/TASLP.2017.2761547
8. Padda, E. S., Nidhi, E., & Kaur, M. R. (2012). A Step towards Making an EffectiveText to speech Conversion System, 2(2), 1242–1244.
9. Kehoe, A., & Pitt, I. (2006). Designing help topics for use with text-to-speech. SIGDOC ’06: Proceedings of the 24th Annual ACM International Conference on Design of Communication,157–163. https://doi.org/10.1145/1166324.1166362
10. Ahlawat, S., Engineering, C., Dahiya, R., & Engineering, C. (2013). a Novel Approach of Text To Speech Conversion, 13(05), 189–193.
11. Kasie, E. B., &Assabie, Y. (2012). Concatenative speech synthesis for Amharic using unit selection method. Proceedings of the International Conference on Management of Emergent Digital EcoSystems - MEDES ’12, 27. https://doi.org/10.1145/2457276.2457282
12. Blankinship, E., & Beckwith, R. (2001). Tools for expressive text-to-speech markup. Proceedings of the 14th Annual ACM Symposium on User Interface Software and Technology-UIST’01,159. https://doi.org/10.1145/502348.502375
13. Aida-Zade, K. R., Ardil, C., &Sharifova, A. M. (2010). The main principles of text-to-speech synthesis system. International Journal of Signal Processing, 6(1), 13–19. https://doi.org/10.1007/s00394-015-1103-y
14. San-Segundo, R., Montero, J. M., Giurgiu, M., Muresan, I., & King, S. (2013). Multilingual Number Transcription for Text-to-Speech Conversion. 8th ISCA Workshop on Speech Synthesis (SSW), 65–69. Retrieved from http://ssw8.talp.cat/papers/ssw8_PS1-8_San-Segundo.pdf
15. Isewon, I. (2014). Design and Implementation of Text To Speech Conversion for Visually Impaired People. International Journal of Applied Information Systems (IJAIS), 7(2), 25–30.