ISSN No:-2456-2165

Text to Speech Synthesis

Mantu Bera
Dept. of CSE, Bankura Unnayani Institute of Engineering, Bankura,India

Abstract:- Speech is the earliest and extremely natural means of the information exchange between human. Since ,past few years several Attempts have ben taken to develop vocally-interactive computers to understand speech/voice synthesis. Of course, this type of interface should provide big benefits. In this matter, a computer capable to synthesize text and giving out a speech. Text to Speech Synthesis is the Technique that converts written text to a spoken language which is easily realized by the end user (mainly in English language). It can be run on JAVA platform where the methodology used is Object Oriented Analysis and Development Methodology as well; while Expert System included for internal operations of the program. The design will be speeded up towards furnishing a one-way communication interface where the computer does communicate with the user by reading out the text document for quick assimilation as well as reading development.

Keywords:- Communication, Expert System, Free(Text to speech)TTS, JAVA, Text-To-Speech

I. INTRODUCTION

Speech/voice synthesis is a sector of computer science and engineering dealing with the computer systems design which really synthesize written text. This is the technology allowing a computer for convertion of a written text into speech through a microphone or telephone. As an new innovative technology, not all the developers are very familiar with speech technology. While the common functions of both speech synthesis and speech recognition needs only a few minutes to realize, there are subtle and dominant capabilities provided by computerized speech which developers want to understand and also to utilize. Automatic speech synthesis is however one of the fastest developing sector in the framework of speech science & technology. With the new generation of computer technology, it comes as a next major innovation in human-machine interaction. The general concept of text-to-speech (TTS) technology is to convert the written text input to spoken output(speech) by producing synthetic speech. There are many ways to perform speech synthesis:

- Ordinary voice recording and play on demand;
- Splitting of the speech into 30-50 phonemes (linguistic units) and their reassembly in a coherent speech pattern;
- The utilization of approx. 400 diaphones (splitting of the phrases to the centre of the phonemes and ofcourse not at the transition). The most vital qualities of modern speech synthesis systems are its naturalness and intelligibility. By the word naturalness I actually mean how the synthesized speech closely resembles real human-speech. On the other side, Intelligibility defines the relieve with which the speech can be understood. The maximization of these two concept is the primary development goal in TTS field.

II. AIM OF THE STUDY

The normal objective of the work is to develop a TTS-synthesizer for the physically impaired as well as the vocally disturbed persons using English language. The particular objectives are:

- In order to enable the deaf and dumb for the purpose of communicating and contributing to the development of an organization via synthesized voice.
- Also To enable the blind and elderly person enjoying a User-friendly computer interface.
- In order To create modern technology awareness and appreciation by computer operators.
- For the implementation of an isolated entire word speech synthesizer which is capable to convert text and respond with speech
- For validating the automatic speech synthesizer being developed during the research.

III. SCOPE OF THE STUDY

The study is mainly focused on an perfect combination of a human-like behaviour with the computer application to create a one-way interactive medium between computer and user. This application was made customized using only one(1) word sentence containing of the numeric digit 0 to 9 that can be used to operate a voice operated telephone system. Inherently, Human voice/speech is a multi modal process that includes the analysis of the uttered acoustic signal including higher level knowledge sources like grammar semantics and pragmatics. This project intending to be focused on the acoustic signal processing except the incorporation of a visual input.

IV. IMPORTANCE OF THE STUDY

This work involves practical, theoretical and methodological importance: The speech synthesizer will be consequentialy useful to any researcher. The text to speech synthesizing system enables the semi-illiterates evaluate and read via electronic documents, and thus bridging the digital divide. The technology will detect applicability in systems such as telecommunications, banking, Internet-portals, accessing PC, transports, administrative, emailing and public services and many more. Our system going to very useful to computer manufacturers and software engineer as they have a speech synthesis tool/engine in their applications.

V. TEXT TO SPEECH SYNTHESIS DEFINITION

A speech synthesis model can be defined by a system, that creates synthetic speech. It is indirectly clear, that it includes some sort of input. But the type of this input is not clear. when the input is plain text, that does not consists additional phonetic or phonological information the system is called a text-to-speech (TTS) system.

Now-a-days it may be plain text or marked-up language text e.g. HTML or something resembles like JSML (Java-Synthesis Mark-up Language).

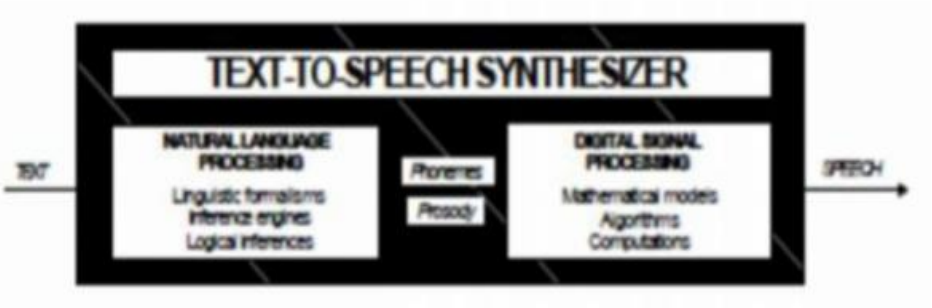

Fig 1:- General Functional Diagram of TTS System.

➢ *Representation-Analysis of Speech Signals*

Continuing speech is a set of complex audio signals producing them artificially complicated. Speech signals are normally treated as voiced or unvoiced, though in some cases those are something between these two. Voiced sounds contains fundamental frequency (F0) , its harmonic components created by vocal cords . The vocal tract does modify this excitation signal causes formant (pole) or anti-formant (zero) frequencies (Abedjieva et al., 1993). Every formant frequency has its amplitude and bandwidth as well and can be too difficult to define few of these parameters accurately. The fundamental frequency as well as formant frequencies are most probably very significant idea in speech synthesis and in speech processing. With completely unvoiced sounds, there is no actually fundamental frequency in the excitation signal and hence, none harmonic structure either and also the excitation may be considered as the white noise. The airflow is forced via a vocal tract constriction that can occur in different places between mouth and glottis. Few of the sounds are generally produced with full stoppage of airflow following a sudden release, produces an spontaneous/impulsive turbulent excitation sometimes followed by a more prolonged turbulent-excitation (Allen et al., 1987). Unvoiced sounds are ,in reality, more silent and less stable than voiced sounds.

Speech signals associated with three vowels (/a/ /i/ /u/) are represented in the time-frequency domain in Figure 2. The fundamental frequency is approx. 100 Hz in all the cases .The formant frequencies F1, F2, and F3 corresponding vowel /a/ are about. 600 Hz, 1000 Hz, 2500 Hz respectively. With the vowel /i/ the first three formants : 200 Hz, 2300 Hz, 3000 Hz, and in /u/ 300 Hz, 600 Hz, and 2300 Hz respectively

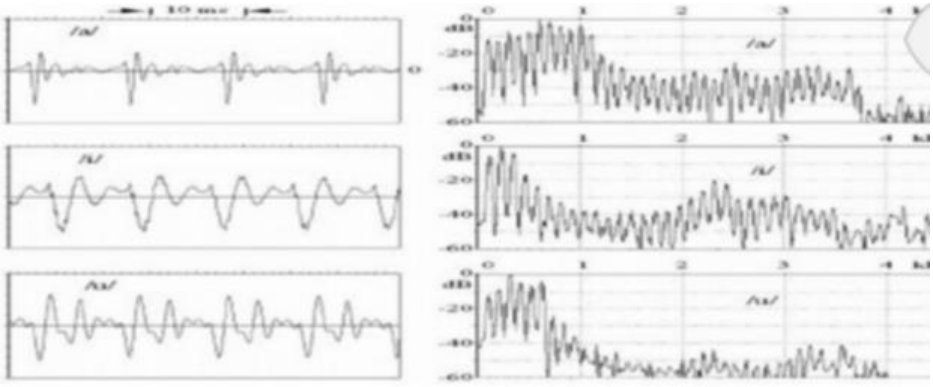

Fig 2:- Time-frequency domain representation for three vowels /a/, /i/, and /u/.

VI. SPEECH SYNTHESIS-APPLICATIONS

Synthetic speech-application is extending fast ,while the quality of TTS systems is growing rapidly. Speech synthesis seems to be more economical for common customers and making these systems appropiate for daily use.Better accessibility or availibilty of TTS systems can increase employability of people having communication difficulties. Given below are few of the applications of the proposed TTS system:

- For the Blind.
- Deafened and Vocally Handicapped
- For Educational Application.

ISSN No:-2456-2165

- Application for the Telecommunications and Multimedia.
- More Applications and Future-Directions (example: Human-Machine Interaction)

VII. SYSTEM ANALYSIS AND METHODOLOGY

A. *Analysis and Problems for the Existing Systems*

Exist systems algorithm is presented below in Figure 3. It describe that the system have no avenue to explain text to the specification for the user rather it speaking plaintext.

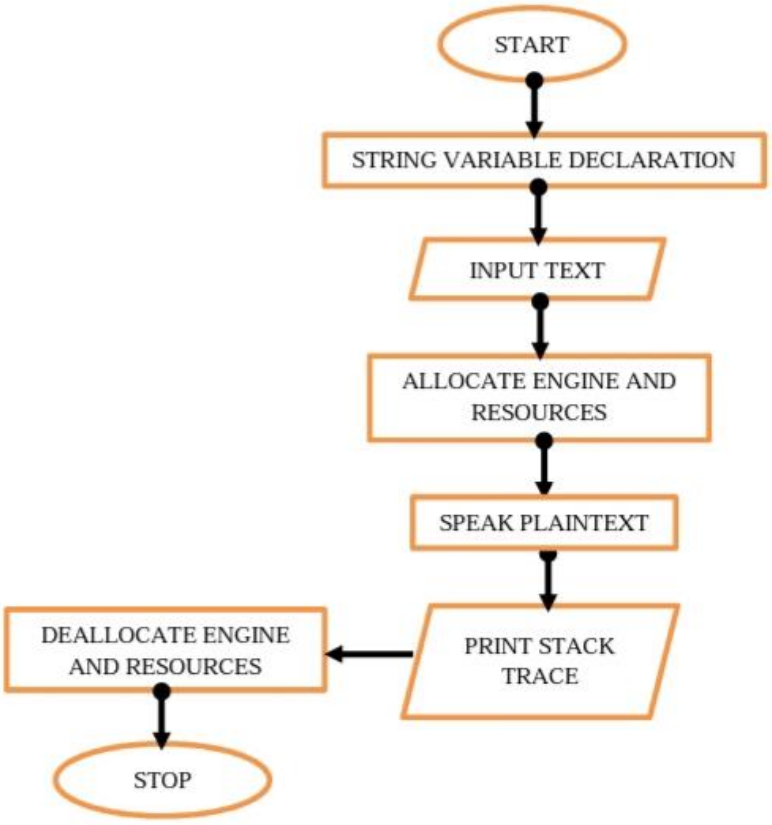

Fig 3:- Algorithm for already existing systems

Due studies has disclosed the these following inadequacies with already existed systems:

- ***Structure analysis***: formatting and punctuation never indicate where paragraphs as well as other structures begin and end. Suppose ,a example,, the last period in "P.D.P." can be misinterpreted as the ending of a sentence.
- ***Text pre-processing***: The system just produces the text which is fed into this except any pre-processing operation is being occured.
- ***Text-tophoneme conversion***:- existing synthesizer system capable to pronounce tens of thousands or hundreds of thousands of words accurately if the word(s) is/are actually not detected in the data dictionary.

B. *Expectation of the New System*

We can expect that the new system must reduce the problem occurred in the old system and also improve. The new system may includes:

- The new system involves reasoning process.
- The new system capable to do text structuring and explanations.
- The speech rate of the system should be adjusted.
- The voice-pitch also can be adjusted.
- We can select from different voices and can combine those if we want to produce a dialogue among them.
- This includes user friendly interface so that people can easily use it even with less computer education or knowledge
- It should be compatible with vocal engines 8. Compling with SSML specification.

VIII. SELECTION OF METHODOLOGY FOR THE NEW SYSTEM

Two methodologies were selected for our new system: The 1st methodology implies Object Oriented Analysis and Development Methodology (OOADM). OOADM has been chosen as the system is represented to the user in a manner which is user-friendly and reliazable by the user. Since this project is also to emulate individual behaviour, Expert system must be used for purpose of the mapping of Knowledge to a Knowledge base including reasoning procedure. Expert system has been used for the internal operations of our program, followed by the algorithm of computation on the basis of some rules. The technique has been derived from general principles reported by the researchers in knowledge engineering techniques (reMurrai et al., 1991-1996).

ISSN No:-2456-2165

The system is actually based on processes modelled in the cognitive phonetics (Hallahan, 1996 and Fagyal, 2001) that accesses various knowledge base .

IX. SPEECH SYNTHESIS MODULE

This TTS system capable to convert an arbitrary ASCII-text to speech. The 1st step includes extracting/fetching the phonetic components of message, we get a string of symbols presenting sound-units (allophones or phonemes), boundaries among words, phrases and sentences with a set of prosody markers (refering the speed, intonation, etc.). The 2nd step involves to find the matching among the sequence of symbols and particularly accurate items stored into the phonetic inventory and also binding them with each other in order to form the acoustic signal for voice output device.

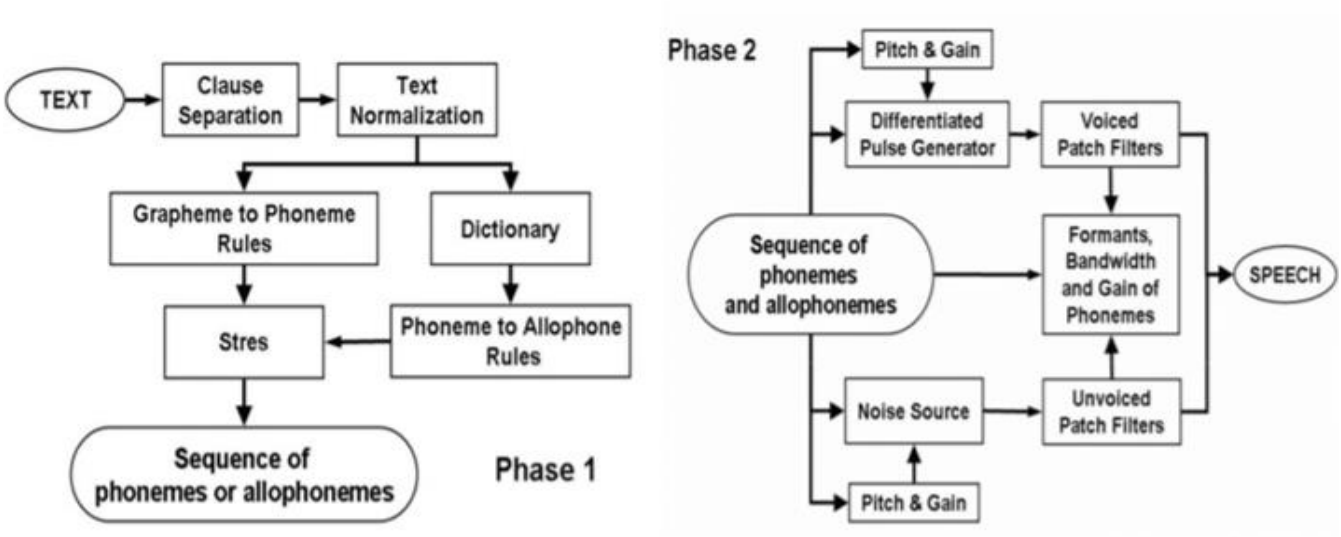

Fig 4:- Stages for TTS -synthesis process

In order to compute the output, this system may consult

- A database consists the values of parameters for the sounds inside the word,
- A knowledge base listing the options to synthesize the sounds.

Including Expert system in the internal programs makes capable the new TTS system to exhibit these characteristics:

- The system does perform at a level normally recognized as equal to a human expert
- The system is absolutely domain specific.
- The system is able to describe its reasoning process
- While the information with which it's working is really probabilistic or fuzzy, the system can rightly propagate unpredictabilities and can provide a span of sustitute solution with connected likelihood.

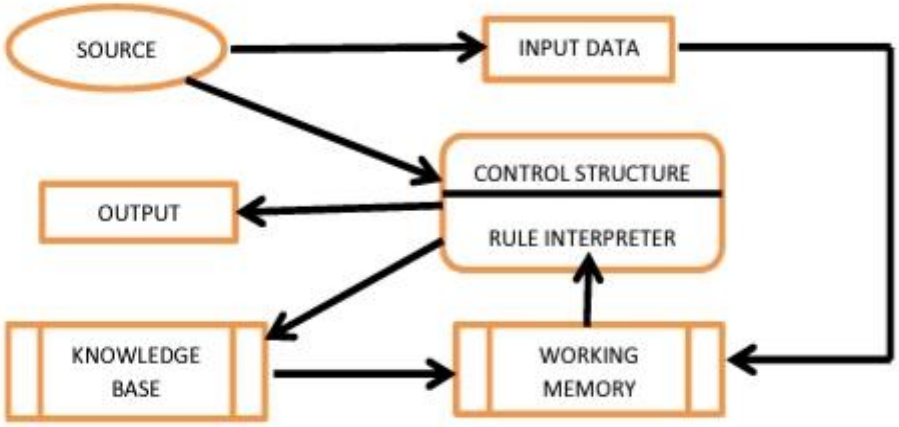

Fig 5:- Data flow-diagram for the Speech synthesis by with Gane and Sarson Symbol

ISSN No:-2456-2165

User Interface (the Source): It may be Graphical User Interface (GUI), or may be the Command Line Interface .
Knowledge Base (the Rule set): FreeTTS /system/engine. This type of source of the knowledge involves domain specific facts and heuristics utilization to solve the domain-problems. FreeTTS is an actually open source speech synthesis module/system written in the Java programming language entirely. It is based on Flite. FreeTTS is the originally implementation of Sun Java Speech API. FreeTTS carries out/supports ending-of-speech markers.
Control Structures: This rule translator/interpreter inference engine is put in application to the knowledge base information to solve the problem.
Short term memory: The working memory listed the current problem status and also history of solutions to date.

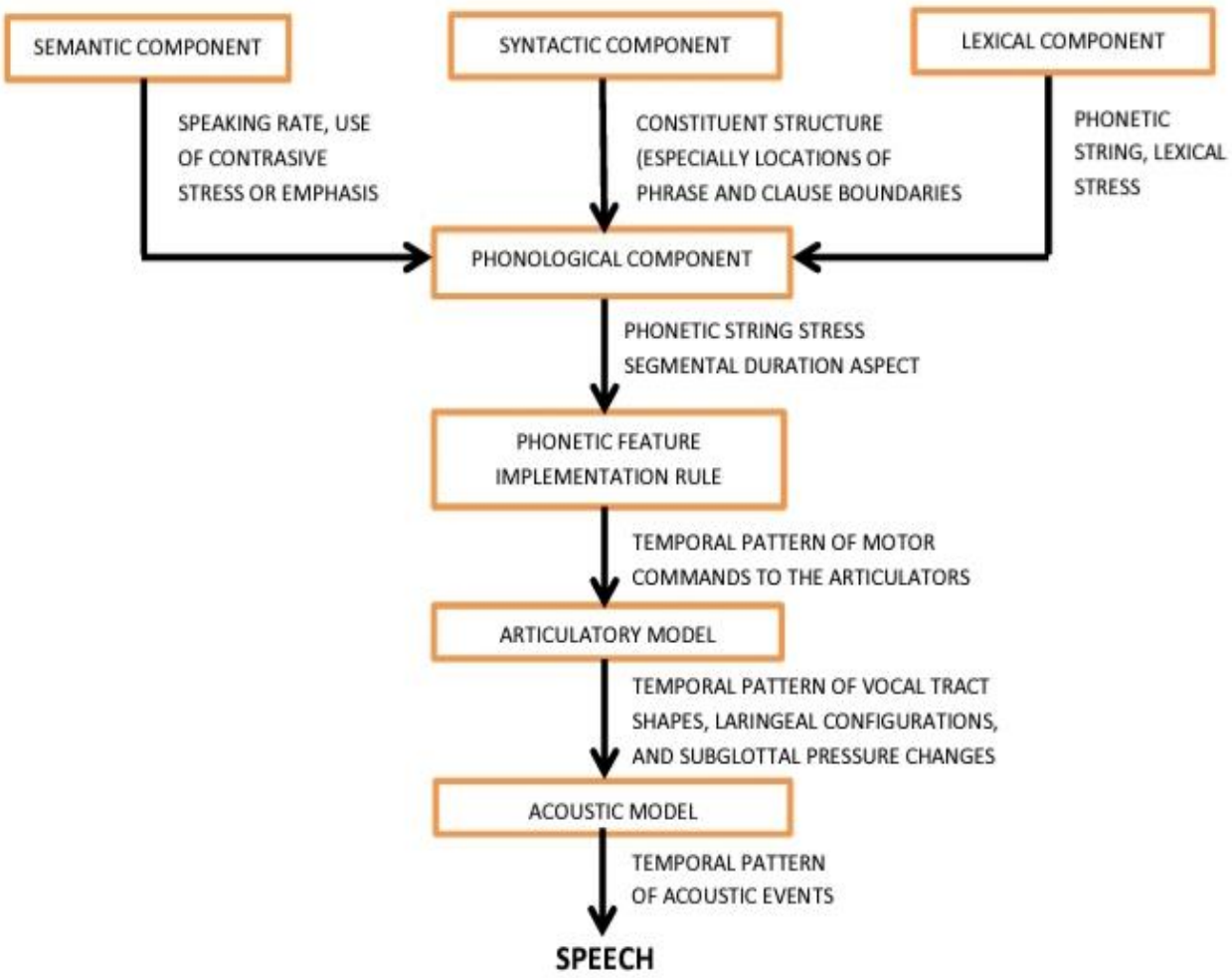

Fig 6:- High-Level-Model for the Proposed System

ISSN No:-2456-2165

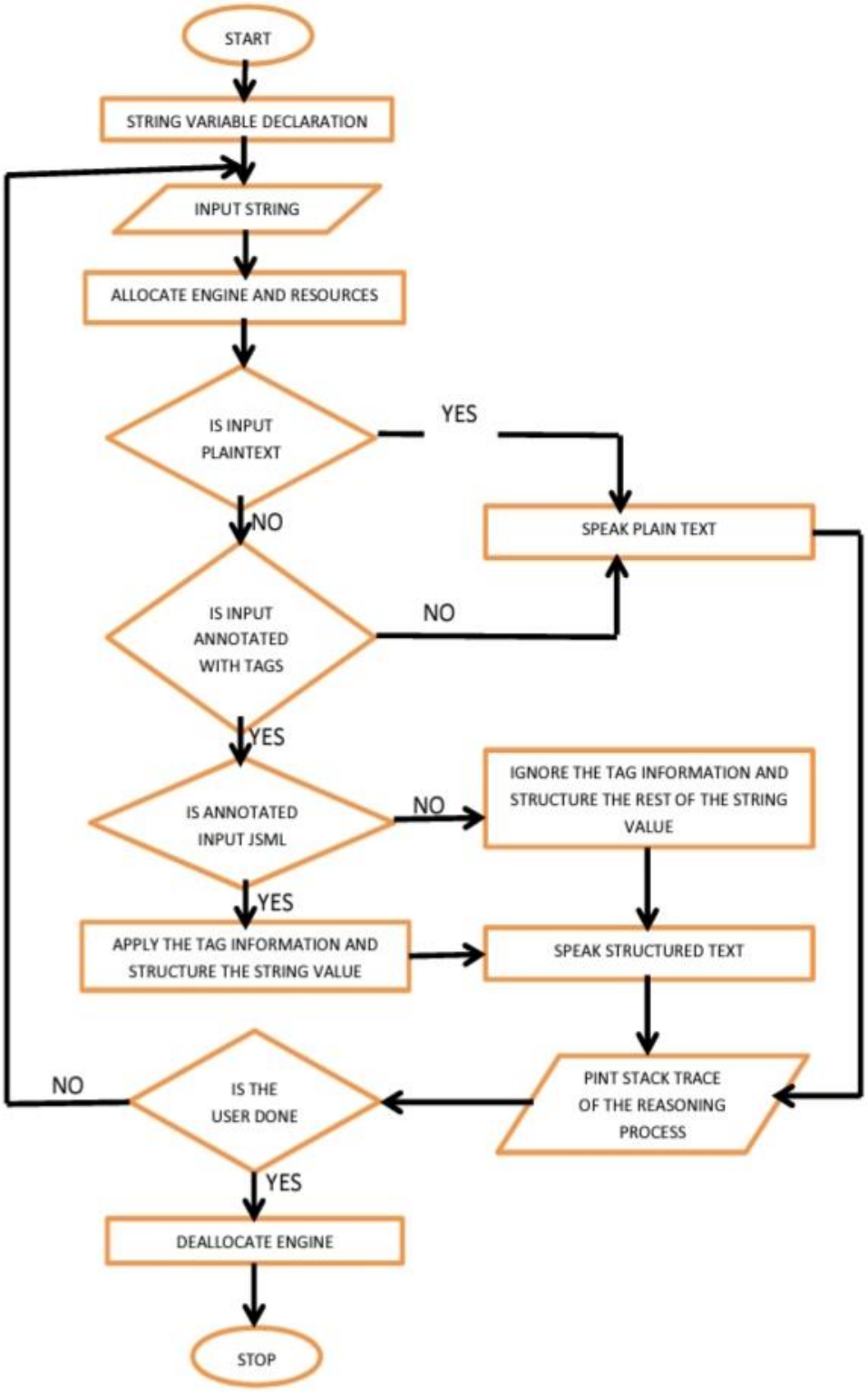

Fig 7:- Flow-chart presentation for the program

A. Choice Programming Language and Speech Engine

The speech engine also used in the new system is the FreeTTS speech engine. FreeTTS has been used as it's programmed using JAVA. This also supports Speech Application Programming Interface(SAPI) that is in the synchronism with Java Speech Application Programming Interface(JSAPI). Also,JSAPI was the standardized interface which has also been used in the new system.

FreeTTS consists an engine for vocal synthesis which supports a definite number of voices (female and male) at different frequencies. This is also recommended to use JSAPI for the interfaces with FreeTTS as JSAPI interface gives the best methods to control and to use the FreeTTS. FreeTTS engine enables full control with respect to the speech signal. This new system offers the possibility to opt a voice from three types of voices: an 8 kHz, diphone male voice named Pintu, a 16 kHz diphone named voice *Pintu16*

ISSN No:-2456-2165

and a16khz restricted domain, voice named *mantu*. The user can also establish the properties of a opted voice: the rate of speaking, the volume and finally the pitch.

A deciding factor in the selection of programming language is the special implication (JSML) provided to the program. It is a java specific mark-up language here we use to annotate spoken output to the selected construct of the user. Furthermore , there is a requirement for a language which does support third party implementation/development of program libraries for utilization in a specific situation which is not amongst the definition of the original platform. After Considering these terms or factors, the best option to choice programming language is **JAVA.** Other things which made JAVA perfect were java's dual character (implement 2 methodologies using one language), its capability to Implement absolute data hiding method (Encapsulation), its also supports for inner abstract class or the object development, and its capability to provide the ability of polymorphism; that is a key features of the program.

X. DESIGN OF THE NEW SYSTEM

Some of the few technologies included in the design of our system involves the following:

Speech Application Programming Interface (SAPI): SAPI is defined as the interface between speech technology engines and applications, both speech recognition and text to speech(Amundsen 1996). The interface permits us for multiple applications for sharing the a speech resources that is available on a computer except to program the speech engine itself. SAPI includes of three interfaces; The *voice text* interface t providing methods to begin, pause, restart, fast-forward, go back(rewind), and stop the TTS engine in the time of speech. The *attribute interface* permits us for accessing to control the primary(basic) behaviour of the TTS engine. Lastly, the *dialog interface* should be used to set and retrieve information.

Java Speech API (JSAPI): The Java Speech API is defined by a standard, easy-use and cross-platform software interface to describe the art speech technology. There are two core speech technologies carried out through Java Speech API are speech synthesis and speech recognition. Speech recognition offers computers with the capability to hear to spoken language and determine which word has been said. Speech synthesis gives the reverse process to produce synthetic speech from the text created or generated by an application, an user or an aplet. This is sometime considered to text-to-speech technology.

A. Design of Individual Objects of the Program

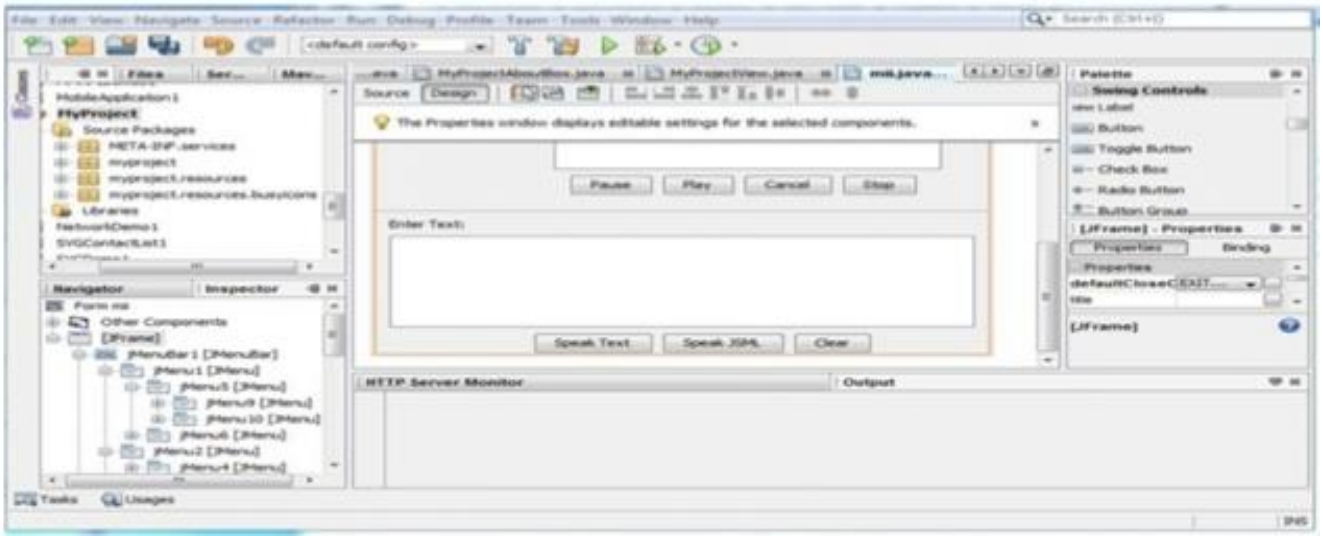

Fig 8:- Netbeans Interface & object -manipulation

ISSN No:-2456-2165

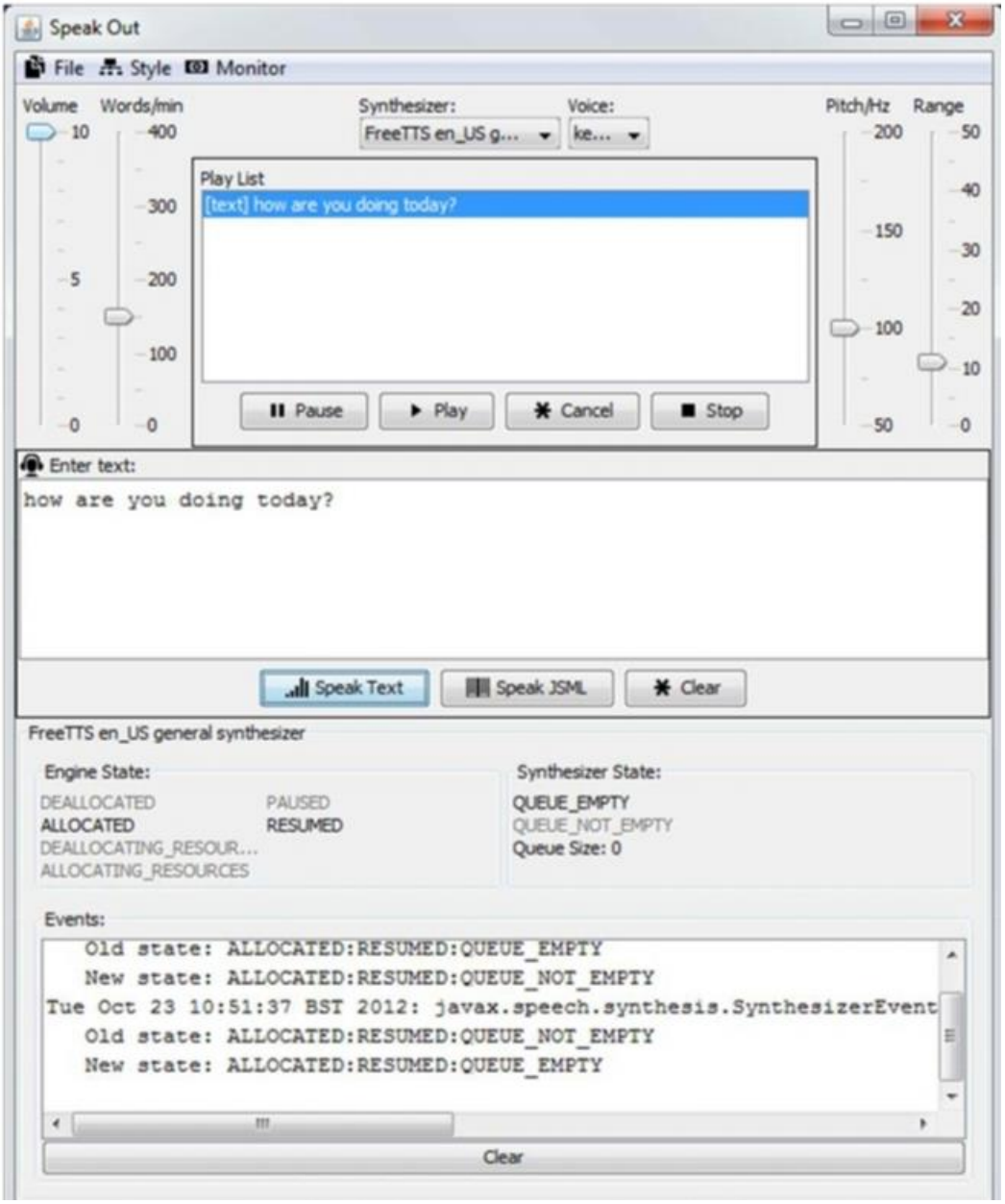

Fig 9:- Overall outlook of new system.

B. Functions of the Abstract Classes

- Menu Bar: It selects through some variables & File chooser system.
- Monitor: reasoning process by defining the allocation process and also de-allocation state.
- Voice System: It shows the several voice option given by the system
- Playable session: It maintains the timing of the speech being announced as output, and creates a speech in synchronism with rate specified.
- Playable type:It specifies the nature of text have to be spoken,if this is a text file or annotated JSML file.
- Text to Speech activator: It plays the provided text and gives an output.
- Player Model: It is a model of all the functional parts and knowledge base presentation in the program.
- Player Panel: It shows the panel and the content pane of basic objects in the program, and also specifies in which location each object is set in the system.
- Synthesizer Loader: loads the Synthesizer engine, allocates and de-allocates resources correctly.

XI. CONCLUSION AND RECOMMENDATION

Synthesizing text is a supreme and efficient technology advancement and artificial formation of speech provided a text to be spoken. With TTS(text-to-speech) synthesis, We can really mediate and can fill in the lacuna given by not fully exploiting the abilities of few handicapped individuals. It is not so easy to utilize a text-to-speech program, by only one click , our computer will able to speak any text in a clear and natural sounding voice.

So, there is requirement to use Information Technology for the problem solving .Before the utilization of the new system, perfect training would be given to the users. This training come in handy with idle tutor on how to handling JSML language and also how to utilize it for annotating text for the correct output and emphasis.

ACKNOWLEDGEMENT

I hereby wish to express my sincere gratitude and respect to Assistant Prof. Tapas Sangiri, Dept. of CSE, Bankura Unnayani Institute of Engineering, Bankura under whom I had proud privilege to work. His valuable guidance and encouragement have really led me to the path of completion of this project.

REFERENCES

[1]. Abedjieva et al. (1993): Acoustics and audio signal processing. http://www.ufh.netd.ac.za/bitstream/10353/495/1/Mhlanathesis.pdf date: 21/07/12
[2]. Allen, J., Hunnicutt, M.S., and Klatt, D. (1987). From text to speech – the MITalk system. MIT press, Cambridge, Massachusetts.
[3]. Hallahan (1996): Phonetics and Theory of Speech Production. http://www.indiana.edu/~acoustic/s702/readfull.html date: 22/07/12
[4]. I.R. Murray, J.L. Arnott, N. Alm and A.F. Newell (1991). A communication system for the disabled with emotional synthetic speech produced by rule. Proc. European Conference on Speech Communication and Technology 91.
[5]. Murray et al. (1996): Application of an analysis of acted vocal emotions. http://www.dl.acm.org/citation.cfm?id=1314912 date: 22/07/12
[6]. Fagyal, Z., Douglas, K. and Fred J. (2006) A Linguistic Introduction, Cambridge University Press, the Edinburgh Building, Cambridge CB2 2RU, UK
[7]. Amundsen (1996): Review of Speech Synthesis Technology. http://www.koti.welho.com/slemmett/dippa/dref.html date: 23/08/12
[8]. Wikipedia, the free encyclopedia: http://en.wikipedia.org/wiki.