International Journal of Emerging Engineering Research and Technology
Volume 3, Issue 8, August, 2015, PP 214-224
ISSN 2349-4395 (Print) & ISSN 2349-4409 (Online)

Assisting the Speech Impaired People Using Text-to-Speech Synthesis

[1]Ledisi G. Kabari [2]Ledum F. Atu

[1,2]*Department of Computer Science, Rivers State Polytechnic, Bori, Nigeria*

ABSTRACT

Many people have some sort of disability which impairs their ability to communicate, thus with all their intelligence they are not able to participate in conferences or meeting proceedings and consequently inhibit in a way their contributions to development of the nation. Work in alternative and augmentative communication (AAC) devices attempts to address this need. This paper tries to design a system that can be used to read out an input to its user. The system was developed using Visual Basic6.0 for the user interface, and was interfaced with the Lemout & Hauspie True Voice Text-to-Speech Engine (American English) and Microsoft agent. The system will allow the users to input the text to be read out, and also allows the user to open text document of the Rich Text Formant (.rtf) and Text File format (.txt) that have been saved on disk, for reading.

Keywords: Text-To-Speech (TTS), Digital Signal Processing (DSP), Natural Language Processing (NLP), Alterative and Augmentative Communication (AAC), Rich Text file Format (RTF), Text File Format (TFF).

INTRODUCTION

The era has come and it is now here, where the mind and reasoning of all are needed for the development of a nation. Thus people who have some sort of disability which impairs their ability to communicate cannot be left behind. Children with this kind of impairment need to be supported in order to tape what is in them.

There is a need for a system that can convert computer text to audible and comprehensible speech for people who are too busy to read through some documents or some who don't even know how to read well and understand. People with low or impaired vision who are reading text document on a computer need to know if they are reading correctly i.e. need audible confirmation of what they are reading, hence the need for a system that reads such document together with the human reader i.e. text-to-speech system. Some visually impaired users may be unable to read a text on a computer screen. And thus need to rely on such a system as the sole means of communicating the content of a computer based text document, hence for such users the need for a text-to-speech system cannot be over emphasized.

Speech impairment is characterized by difficulty in articulation of words. Examples include stuttering or problems producing particular sounds. Articulation refers to the sounds, syllables, and phonology produced by the individual. Voice however, may refer to the characteristics of the sounds produced specifically, the pitch, quality, and intensity of the sound. Often, fluency will also be considered a category under speech, encompassing the characteristics of rhythm, rate, and emphasis of the sound produced

Information and communication technology is rapidly evolving as an effective tool for making information wide spread and available online to several communities. The industrial society is turning towards information society. The increased use of information technology is enabling people across the world to participate in the knowledge network; however visually impaired people and people with learning disabilities are being deprived of the benefits of the use of ICT and the computer system. One

**Address for Correspondence:*

ledisigiokkabari@yahoo.com

of the main reasons for this is lack of suitable human computer interface and the software designed and developed to meet local needs. In many parts of the world, people and organizations are developing commercial software like content management system and financial software etc., however due to current market needs they do not recognize the needs of text to speech (TTS) converter. There is a great need to develop a text to speech converter tool with simple human computer interface to meet needs of visually impaired people and also help people with learning disability and to put foundation for side applications. The TTS conversion tool can effectively address ICT needs of visually impaired people and generally everybody in the country**.**

The aim of this paper is to develop a system that allows users to enter text as input either directly or by typing a text file on disk, and will generate synthesized speech representation of the text i.e. read out the text as output. The usefulness of the study include among others to enable users who may be concentrating on other tasks to be alerted of important events and information. It will be able communicate the content of a document to users who may be too busy to read through them or may not be very proficient at reading and to reduce the stress or boredom of going through documents line by line to acquire important knowledge or information. To provide an efficient proof reading and error checking tool to users who are creating documents. Create a data driven voice technology for a high conversion of speech for speech provided by text-to speech conversion system. Create Simple TTS interface for the individuals of every class, age and status.

The work will indeed enhance learning environment by providing a form of encouragement for students with learning disability. It will enable users to focus their sight in other tasks while communicating the content of text document across to them. The TTS system provides accessibility to visual impaired computer users. Users that have low vision may rely on text-to-speech as their sole or alternative means of feedback from computer. The TTS software can also be utilized for reading e-books, newsletters, online newspapers etc., to avoid eye concentration on computer display which potentially leads to various eye diseases.

RELATED WORKS

Text-To-Speech Synthesizer

Text-to-speech synthesis (TTS) is the automatic conversion of a text into speech that resembles, as closely as possible, a native speaker of the language reading that text. A closer examination of the problem of converting from text into speech reveals that there are two rather well defined sub problems. The first is text-analysis, or alternatively linguistic analysis. The task here is to convert from an input text, which is usually represented as a string of characters, into a linguistic representation. This linguistic representation is usually a complex structure that includes information on grammatical categories of words, accentual or tonal properties of words, prosodic phrase information, and of course word pronunciation. The second problem is speech synthesis proper, namely the synthesis of a (digitally coded) speech waveform from the internal linguistic representation.

The artificial production of speech-like sounds has a long history, with documented mechanical attempts dating to the eighteenth century. The synthesis of speech by electronic means is obviously much more recent, with some early work by Stewart [1] that allowed the production of static vowel sounds.

The three modern approaches to synthesis, namely *articulatory* synthesis, formant synthesis, and concatenative synthesis, have a roughly equally long history: the earliest rule-based articulatory synthesizers include [2], [3] and [4]; the earliest rule-based formant synthesizer was [5] and the concept of dip hone concatenation was discussed in[6] with the first dip hone-based synthesizer reported being[7]. We discuss these different approaches in somewhat more detail in the next section.

Text-analysis for speech synthesis has a much shorter history. One of the first systems for full text-to-speech conversion for any language — and certainly the best known — was the MITalk American English system. Although a full description was only published in 1987[8], the work was actually done in the seventies. Text-to-Speech (TTS) synthesis is a process through which computer text is rendered as digital audio and then "spoken". Most text-to-speech engines can be categorized by the method that they use to translate phonemes into audible sound.

Concatenated Word

Although Concatenated Word systems are not really synthesizer, they are one of the most commonly used text-to-speech systems around. In a concatenated word engine, the application designer provides recording for phrases and individual words. The engine pastes the recordings together to speak out a sentence or phrase.

Synthesis

A text-to-speech engine that uses synthesizer generates sounds similar to those created by the human vocal cords and applies various filters to simulate throat length, mouth cavity, lip shape and tongue position. The voice produced by dip hone synthesis technology tends to sound less human than a voice produced by dip hone concatenation, but it is possible to obtain different qualities of voice by changing a few parameters.

Sub Word Concatenation

A text-to-speech engine that uses sub word concatenation links like short digital-audio segments together and performs intersegment smoothing to produce a continuous sound. In dip hone concatenation for example, each segment consist of two phonemes, one that leads into the sound and one that finishes the sound.

The conversion from written text to speech can be broken down into three major tasks: linguistic analysis, prosodic modeling and speech synthesis. Speech synthesis transforms a given linguistic representation; say a chain of phonetic symbols enriched by information on phrasing, intonation and stress, into artificial, machine-generated speech by means of an appropriate syntheses method. Text analysis modules compute the linguistic representation form written text. The figure1 illustrates a general functional diagram of a TTS system.

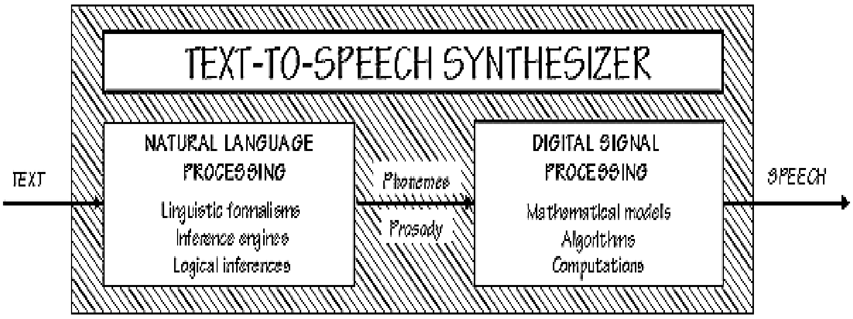


Figure1. *A simple but general functional diagram of a TTS system.(source[9])*

Speech Synthesis

Speech synthesis is the artificial production of human speech. A computer system used for this purpose is called a speech synthesizer, and can be implemented in software or hardware. There are different ways to perform speech synthesis. The choice depends on the task they are used for, but the most widely used method is Concatenative Synthesis, because it generally produces the most natural-sounding synthesized speech. Concatenative synthesis is based on the concatenation (or stringing together) of segments of recorded speech. There are three major sub-types of concatenative synthesis:

Domain-Specific Synthesis

This involves recording the voice of a person speaking the desired words and phrases. This is useful if only the restricted volume of phrases and sentences is used and the variety of texts the system will output is limited to a particular domain e.g. a message in a train station, whether reports or checking a telephone subscriber's account balance. The technology is very simple to implement, and has been in commercial use for a long time in devices like clocks and calculators. The level of naturalness of these systems can be very high because of the variety of sentence types is limited, and they closely match the prosody and intonation of the original recordings.

Unit Selection Synthesis

Unit selection synthesis uses large database of recorded speech. During database creation, each recorded utterance is segmented into some or all of the following: individual phonemes, syllables, morphemes, words, phrases and sentences. Typically the division into segments is done using a specially modified speech recogniser set into a "forced alignment" mode with some manual corrections afterwards, using visual representation such as the waveform and spectrogram. An index of the unit in speech synthesis is then created based on the segmentation and acoustic parameters like the fundamental frequency (pitch), duration, position in the syllable, and neighbouring phone. At runtime, the desired target utterance is created by determining the bet chain of candidate units from the database (unit selection). This process provides the greatest naturalness. Because it applies only to small amount of digital signal processing (DSP) to the recorded speech.DSP often makes recorded speech sound less natural, although some systems use a small amount of signal processing at a point of concatenation to smoothen the waveform. The output from the best unit-selection systems is often indistinguishable from the real human voices, especially in context for which the TTS system has been tuned. However, maximum naturalness typically require unit-selection speech databases to be very large, in some systems ranging into gigabytes of recorded data, representing dozens of hours of speech.

Dip hone Synthesis

Dip hone synthesis uses a minimal speech database containing all the Dip hone (sound-to-sound transition) occurring in language. The number of Dip hones depends on the phonotactics of the language: for example Spanish has about 800 Dip hones and German about 2500, while English has about 400. In dip hone synthesis only one example of each dip hone is contained in the speech database. At runtime, the target prosody of a sentence is superimposed on these minimal units by means of digital signal processing techniques such as linear predictive coding. PSOLA or MBROLA. The quality of the resulting speech is generally worse than that of unit-selection, but more natural-sounding thank the output of formant synthesizer. Dip hone Synthesis suffers from the sonic glitches of concatenative synthesis and robotic-sounding nature of formant syntheses, and has a few of the advantage of either approach other than small size. As such, it use in commercial applications is declining, although it continues to be used in research because there are a number of freely available software implementation.

Some other speech synthesis method includes:

- **Formant Synthesis**: This does not use human speech samples a runtime. Instead, the synthesized speech output is created using an acoustic model. Parameters such as fundamental frequency, voicing and noise levels are varied over time to create a waveform of artificial speech. This method is sometimes called rules-based synthesis; however, many concatenative systems also have rules-based components.

- **Articulaory synthesis:** This refers to computational techniques for synthesising speech, based on models of the human vocal tract and the articulation processes occurring there.
- **HMM-bases synthesis:** This is a synthesis method based on hidden Markov models. In this system the frequency spectrum (vocal tract). Fundamental frequency (vocal source) and duration (prosody) of speech are modelled simultaneously by HMMs. Speech wave forms are generated from HMMs themselves on maximum likelihood criteria.
- **Sine wave Synthesis:** Is a technique for synthesizing speech by replacing the formants (main bands of energy) with pure tone whistles.

TYPES OF SPEECH SYNTHESIS

Concept-to-Speech Synthesis

This involves a generation component that generate a textual expression from semantic, pragmatic and discourse knowledge. The speech signal can then be generated from this expression. Concept-to-speech synthesis can be used in dialog systems for example but anywhere the input is already in textual form, text-to-speech synthesis will be used.

Text-to-Speech (TTS)

This converts normal language text into speech. In text-to-speech synthesis, the text to be spoken is provided, not generates by the system. It must however be analysed and interpreted in order to convey the proper pronunciation and emphasis (e.g. to produce a question instead of a statement)

Structure of a Speech-to-Speech Synthesizer System

Text-to-speech software is used to convert words from a computer document (e.g. word processor document) into audible speech spoken through the computer speaker. It is thus suitable to define text-to-speech as the automatic production of speech, through grapheme-to-phoneme transcription of the sentences to utter.

The most important qualities of a text to speech system are naturalness, which describes how closely the output of a systems sounds like human speech; and intelligibility, which is the ease with which the output is understood. The ideal TTS synthesizer is both natural and intelligible ant TTS synthesis systems usually try to maximize both characteristics.

Text-to-speech synthesis takes place in several steps. The TTS systems get a text as input, which it first must analyze and then transform into a phonetic description. Then in a further step it generates the prosody. Form the information now available, it can produce a speech signal.

The structure of the text-to-speech synthesizer can be broken down into two major modules. These are Natural Language Processing (NLP) module which produces a phonetic transcription of the text read, together wuth prosody and Digital Signal Processing (DSP) module which transforms the symbolic information it receives from NLP into audible and intelligible speech.

The major operations of the NLP module are as follows

- **Text Analysis:** First the text is segmented into tokens. The token-to-word conversion creates the orthographic form of the token. For the token "Mr" the orthographic form "Mister" is formed by expansion, the token "12" gets the orthographic form "twelve" and "1997" is transformed to "nineteen ninety seven".

- **Application of Pronunciation Rules:** After the text analysis has been completed, pronunciation rules can be applied. Letters cannot be transformed 1:1 into phonemes because correspondence is not always parallel. In certain environments, a single letter can correspond to either no phoneme (for example, “h” in “caught”) or several phoneme (“”n” in “Maximum”). In addition, several letters can corresponds to a single phoneme (“ch” in “rich”). **Prosody Generation:** after the pronunciation has been determined, the prosody is generated. The degree of naturalness of a TTS system is dependent on prosodic factors like intonation modelling (phrasing and accentuation), amplitude modelling and duration modelling (including the duration of sound and the duration of pauses, which determines the length of the syllable and the tempos of the speech).

Figure2 illustrates a summary of the operation of the natural Language processing module of a TTS synthesizer

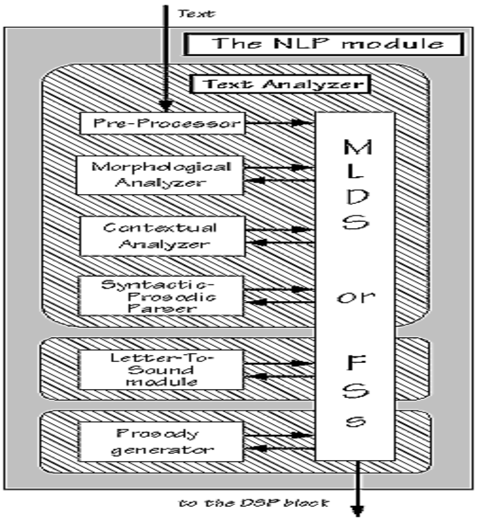


Figure2. *Natural Language Processing Module of TTS (source:[9])*

The output of the NLP module is passed to the Digital Processing Module (**DSP).** This is where the actual synthesis of the speech signal happens. In concatenated synthesis the selection and linking of speech segments take place. For individual sounds the best option (where several appropriate options are available) are selected from a database and concatenated.

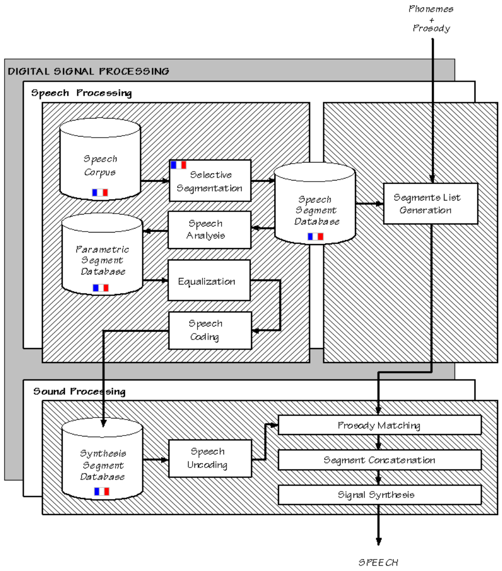


Figure3. *The DSP component of a general concatenation-based synthesizer. (source*:[9])

SYSTEM ANALYSIS AND DESIGN

Development Methodology

The TTS system as a whole comprises of two subsystems; the Interface part and the Text to speech conversion engine. The system will be developed using Visual Basic for the user interface, and will be interfaced with the Lernout & Travois Text-to-Speech Engine (American English) and Microsoft agent. The system will allow the users to input the text to be read out, and also allows the user to open text document of the Rich Text Formant (.rtf) and Text File (.txt) format that have been saved on disk, for reading. With various types of given text the TTS conversion tool will be tested for naturalness and accuracy and examined by linguistic experts to achieve more correct pronunciation of English words. The outcomes of these examinations shall be incorporated to the TTS.

System Requirements

For good implementation of this application it is necessary to choose hardware and software required for the application to run effectively in other to achieve the desired goals. The system requirements can be broken down into hardware requirements software requirements.

Hardware Requirements: The minimum hardware requirements to run the application are: Pentium III 400 MHZ, 40 Gigabyte hard disk, CD Rom/RW Drive, 128 MB RAM, A USB for flash drive users, Printer, 14" SVGA Color Monitor, Multimedia Facilities.

Software Requirements: The minimum software requirements needed to run the application are: Operating system (OS) (windows 2000, XP, and Vista), Word or Note Pad for text files.

Design Tools

In the system analysis and design stage there are some tools that are used in order to construct the model. The tools that would be used to develop this application are the Unified Modeling Language. (UML) Several different notations for describing object oriented designs were proposed in the 1980s and 1990s. The Unified Modeling Language is an integration of these notations. It describes notations for a number of different models. The Unified Modeling Language (UML) is used to specify, visualize, modify, construct and document the artifacts of an object-oriented software-intensive system under development. [FOLDOC (2001). Unified Modeling Language]. Unified Modeling Language (UML) is a standardized general-purpose modeling language in the field of software engineering. The standard is managed, and was created by, the Object Management Group.

UML includes a set of graphic notation techniques to create visual models of software-intensive systems UML offers a standard way to visualize a system's architectural blueprints, including elements such as activities, actors, business processes, database schemas, (logical) components, programming language statements, reusable software components.[15][Chonoles, Michael Jesse; James A. Schardt (2003). UML 2 for Dummies. Wiley Publishing. ISBN 0-7645-2614-6.]

The UML models that were used in designing the text to speech software are: Sequence diagram, Use case diagram and Flow charts / Data flow diagrams.

Behaviors Diagram

Behavior diagrams emphasize what must happen in the system being modeled. Since behavior diagrams illustrate the behavior of a system, they are used extensively to describe the functionality of software systems. An example is the use case diagram which was used in this paper. This model describes the functionality provided by a system in terms of actors, their goals represented as use cases, and any dependencies among those use cases. Use case diagrams graphically depict the interactions between the system and external system and users. In other words, they graphically decide who will use the system and in what ways the user expects to interact with the system. Figure4 shows the use case diagram in this work.

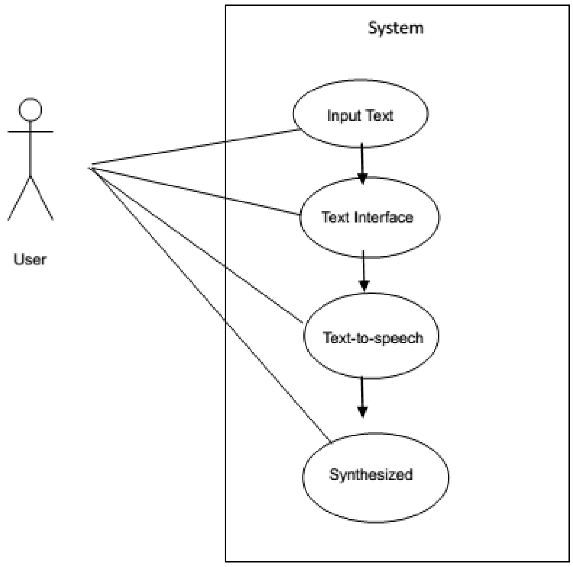


Figure4. *A Use Case diagram for the Text to Speech system*

From the case diagram the user is required to type in / input his/her preferable text in the right text format, or can even load a file from his/her computer or a storage device. The precondition is that user must type in a valid text document for the interface to convert the text to synthesized speech. The post conditions are either a success end where an audible speech is produced or a failure end where the system display error message and prompts the user to input a new text. The user can do either of the following:

- Type the text and click read, the system responds by converting the typed text to audible synthesized speech.
- Open a text file in the text interface, the system responds by converting the opened text to audible synthesized speech.

Interaction Diagram

Interaction diagrams, a subset of behavioral diagrams, emphasize the flow of control and data among the things in the system being modeled. An example is sequence diagram. Sequence diagram is an example interaction diagram that shows how objects communicate with each other in terms of a sequence of messages. Also indicates the lifespan of objects relative to those messages. It graphically depicts how objects interact with each other via messages in the execution of a use case or operation. They illustrate how messages are sent and received between objects and in what sequence is shown in figure5.

Reasons for Using Visual Basic Language

Visual Basic 6.0 programming language provides forms, command buttons and other controls that you can easily drag and drop, thus making programming very easy. Visual Basic 6.0 enables the programmer to compose the instructions that tell the computer how to perform the task required by the program under construction inside command buttons. It provides a set of programming tools, which enable the programmer to assembly and rearrange the component of the program. The tools provided by visual basic 6.0 makes programmer create good graphical user interface with ease.

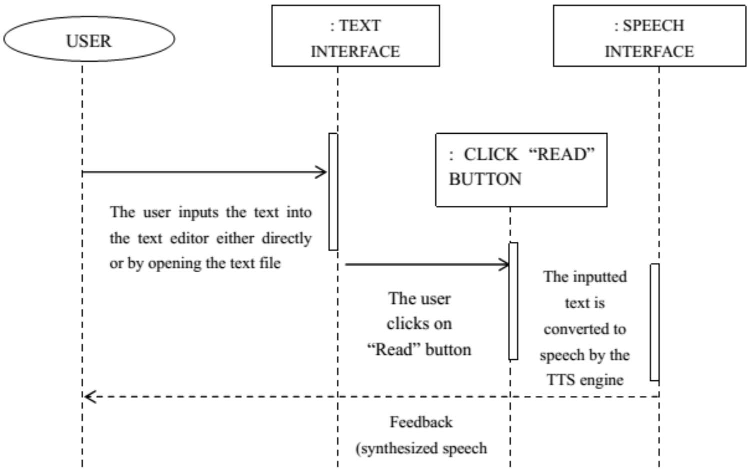

USER
: TEXT INTERFACE
: SPEECH INTERFACE
: CLICK “READ” BUTTON
The user inputs the text into the text editor either directly or by opening the text file
The user clicks on “Read” button
The inputted text is converted to speech by the TTS engine
Feedback (synthesized speech

Figure5. *Sequence diagram for the Text to Speech system.*

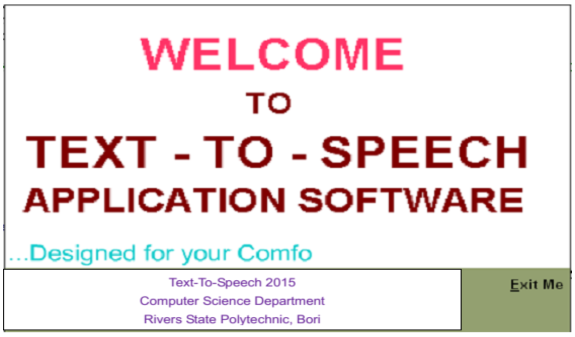
RESULTS AND DISSCUSSION

The result of the process was the development of Text-To-Speech application software whose sample outputs are shown in figure6, Figure7 and fiure8. Figure6 shows the welcome environment, figure7 shows the running of the program with the environment ready accept text input that will consequent converted to speech. Figure8 shows the text already dropped, awaiting the user to press the “Read this” button. The user can stop the reading at any time by pressing “Shut up” button.

CONCLUSION

This paper in its own way tries to contribute in assisting the speech impaired disabled to participate in meeting proceedings by typing their opinions and can be read out. Prosthetic devices of this sort must be usable in a great variety of settings. They should enable the user to be a full participant in ordinary conversations, to lead transactional encounters and to prepare speech for more formal occasions. It will go a long way to assist the general public in several other ways. The extent to which this is possible depends on a number of factors, both physical and cognitive. The speech impairment may be due to a physical disability which has no effect on the person's linguistic ability, or it may be due to a cognitive, language impairment. Often, some combination of physical and cognitive disabilities is involved. Other communication aids include systems designed for deaf users and tools for tutoring and rehabilitation for people with language impairments

Intuitively, the operations involved in the Text-To-Speech systems are the computer analogue of dynamically controlling the articulatory muscles and the vibratory frequency of the vocal folds so that the output signal matches the input requirements. In order to do it properly, the Digital Signal Processing (DSP) module should obviously, in some way, take articulatory constraints into account, since it has been known for a long time that phonetic transitions are more important than stable states for the understanding of speech.


WELCOME

TO

TEXT - TO - SPEECH

APPLICATION SOFTWARE

...Designed for your Comfo

Text-To-Speech 2015
Computer Science Department
Rivers State Polytechnic, Bori

Figure6. *Sample output showing welcome environment*

The Federal Government of any nation should as a matter of urgency pay more attention to this area of research as to tap the hidden potential in most speech disable people.

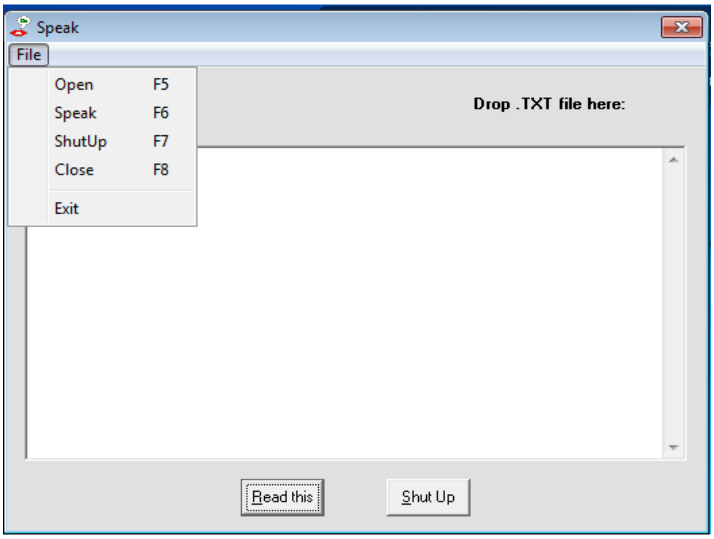
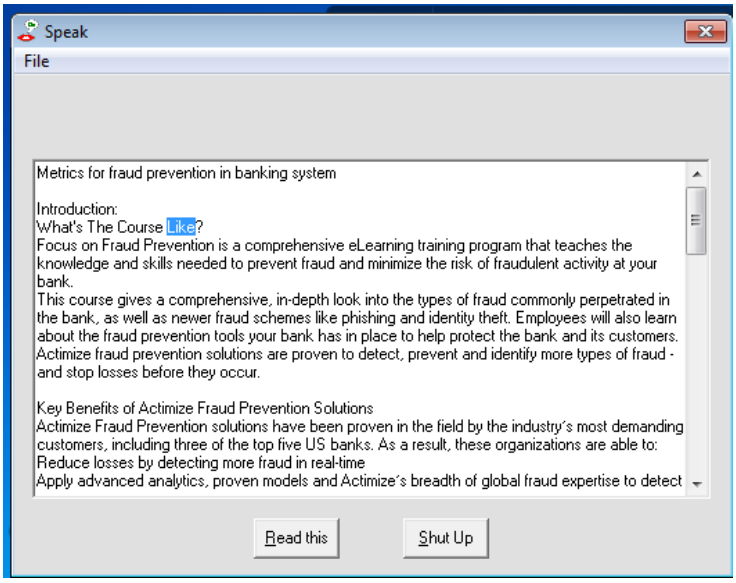


Figure7. *Sample output showing environment ready for text drop*



Metrics for fraud prevention in banking system

Introduction:
What's The Course Like?
Focus on Fraud Prevention is a comprehensive eLearning training program that teaches the knowledge and skills needed to prevent fraud and minimize the risk of fraudulent activity at your bank.
This course gives a comprehensive, in-depth look into the types of fraud commonly perpetrated in the bank, as well as newer fraud schemes like phishing and identity theft. Employees will also learn about the fraud prevention tools your bank has in place to help protect the bank and its customers.
Actimize fraud prevention solutions are proven to detect, prevent and identify more types of fraud - and stop losses before they occur.

Key Benefits of Actimize Fraud Prevention Solutions
Actimize Fraud Prevention solutions have been proven in the field by the industry's most demanding customers, including three of the top five US banks. As a result, these organizations are able to:
Reduce losses by detecting more fraud in real-time
Apply advanced analytics, proven models and Actimize's breadth of global fraud expertise to detect

Figure8. *Sample output showing text to be read*

REFERENCES

[1] J. Stewart. "Kindergarten students' awareness of reading at home and in school" Journal of Educational Research, 86(2), pp. 95-104. (1992)

[2] K. Nakata and T. Mitsuoka, "Phonemic Transformation and Control Aspects of Synthesis of Connected Speech," Journal of Radio Research Laboratories, Vol. 12, 1965, pp. 171-186.

[3] W. L. Henke, "Preliminaries to Speech Synthesis Based upon an Articulatory Model" Proceedings of the IEEE Conference on Speech Communication Process, 1967, pp. 170-182, IEEE.

[4] C. Coker, "Speech Synthesis with a Parametric Articulatory Model". In Speech Symposium. (Kyoto), 1968. Reprinted in Flanagan, James and Rabiner, Lawrence (eds.) Speech Synthesis, Dowden, Hutchinson and Ross, Stroudsburg, PA, 1973, pp. 135-139.

[5] J. Kelly and L. Gerstman, "An Artificial Talker Driven from Phonetic Input", Journal of the Acoustical Society of America, Vol. Supplement 1 33, 1961, pp. S35.

[6] G. E. Peterson, W. S. Y. Wang and S. Eva, "Segmentation Techniques in Speech Synthesis," Journal of the Acoustical Society of America, 1985, Vol. 30, pp. 739-742.

[7] N. R. Dixon, and H.D. Maxey, " Terminal Analog Synthesis of Continuous Speech Using the Diphone Method of Segment Assembly", IEEE Transactions on Audio Electro acoustics, Vol. AU-16, 1968, pp. 40-50

[8] J. Allen, M. S. and D. H. Klatt, "From Text to Speech: The MIT talk System", Cambridge University Press, Cambridge, UK, 1987.

[9] T. Dutoit, "A short Introduction to Text-to-Speech Synthesis", TTS Research Team, TCTS Lab. [Online]. Available: http//tcts.fpms.ac.be/synthesis/intortts_old.html