ABSTRACT

A pdf-to-audio converter is an application that converts text into spoken word, by analyzing and processing the text using Natural Language Processing (NLP) and then using Digital Signal Processing (DSP) technology to convert this processed text into synthesized speech representation of the text. Here, we developed a useful text-to-speech synthesizer in the form of a simple application that converts inputted text into synthesized speech and reads out to the user which can then be saved as an mp3.file. The development of a text to speech synthesizer will be of great help to people with visual impairment and make making through large volume of text easier.

TABLE OF CONTENTS

CHAPTER ONE

1.0 INTRODUCTION

1.1 BACKGROUND OF THE STUDY

In 1991, Adobe co-founder Dr. John Warnock propelled the paper-to-advanced digital revolution with an idea he called, The Camelot Project. The objective was to empower the growing digital users the ability to capture documents from any application, send electronic renditions of these documents anywhere, and view and print them on any machine. By 1992, Camelot had developed into PDF.

Today, it is a document format trusted by businesses/organizations around the globe. PDF, or Portable Document Format, was the first file format of its kind to have the ability to store and offer content and images in a way that would protect the formatting of the original document. Regardless of which software, hardware or platform it is being viewed on. Inspired by the idea of digitizing the contents of the Library of Congress, Warnock's Team Camelot expanded to include developers with a diverse range of coding skills to build the platform-agnostic file format. Not only would the file format need to be compatible with the most popular platforms, they also needed to have developers that had experience working with printer drivers. It was this additional element that eventually helped to boost PDF's popularity around the globe - users everywhere could choose to 'Save as PDF' instead of printing out the document. The team created the initial version of the format in roughly a year, but the public launch would wait until June 1993 - in tandem with Adobe Acrobat Version 1.0. Acrobat was released to huge fanfare, and as the first program with the ability to read the file format, it's adoption was crucial to the initial success of PDF. Inspired by the idea of digitizing the contents of the Library of Congress, Warnock's Team Camelot expanded to include developers with a diverse range of coding skills to build the platform-agnostic file format. Not only would the file format need to be compatible with the most popular platforms, they also needed to have developers that had

experience working with printer drivers. It was this additional element that eventually helped to boost PDF's popularity around the globe - users everywhere could choose to 'Save as PDF' instead of printing out the document.

1.2 STATEMENT OF THE PROBLEM

With PDF being the most used document format globally, there is a need to converts text in PDF formats into Audio signal. These can be utilized for various purposes, e.g. in the educational system, car navigation, announcements in railway stations, response services in telecommunications, and e-mail reading. Furthermore, people with vision disabilities can't view or read PDF files and this is a major setback. This research addresses the problems in converting PDF text into speech. One is how to improve the naturalness of synthetic speech in PDF-based text into an Audio system.

1.3 AIMS AND OBJECTIVE OF THE STUDY

This research aims at the Design and Implementation of a PDF to Audio System to aid accessibility and easy text to voice assimilation of documents in PDF format.

The following are the objectives of the study:

1. Develop a system that will convert PDF text to audio for easy assimilation of document.
2. A system to easily detect a PDF file and convert to audio.
3. To design a system that will assist people with reading disabilities to easily convert PDF text to audio files.
4. To design and implement a system that will assist students' reading comprehension skills.

1.4 MOTIVATION OF THE STUDY

Presenting reading material orally in addition to a traditional paper presentation format increases the inability of users to be able to decode reading material, and therefore, has the potential to prevent students with reading disabilities better comprehend written texts. There

are several different technologies for presenting oral materials (e.g., text-to-speech, reading pens, audiobooks). Already text were accessible orally through books-on-tape and through human readers. There is a need to develop and implement a text-to-speech system that will be used widely in the educational settings from elementary school through universities. With the implementation of the PDF to Audio system, they will be an improved effects of text-to-speech and related tools for oral presentation of material on reading comprehension for students with reading disabilities.

1.5 SCOPE OF THE STUDY

The scope of the research is focused on implementing a PDF to Audio system to improve the usage of PDF documents and in order to achieve a more flexible audio speech system.

1.6 LIMITATION OF THE STUDY

During the development of the research study, they following limitations were encountered;

1. Limited research material available at the school library and on the internet.
2. High cost of setting up the system as it requires a high programming language.
3. Combining school work and carrying out the research.

1.7 DEFINITION OF TERMS

PDF – Portable Document Format.

Audio - Audio most commonly refers to sound, as it is transmitted in signal form.

Speech- the expression of or the ability to express thoughts and feelings by articulate sounds.

Oral - relating to the transmission of information or literature by word of mouth rather than in writing.

Reading Disabilities - a condition in which a sufferer displays difficulty reading.

File - a collection of data treated as a unit by a computer.

Document- A document is a written, drawn, presented, or memorialized representation of thought.

CHAPTER TWO

2.0 LITERATURE REVIEW

2.1 OVERVIEW OF PDF

In the 1990s, Adobe developed the Portable Document Format (PDF) to distribute documents, including text formatting and graphics, in a method that is independent of operating systems, hardware, and software. All of the text, fonts, vector graphics, raster images, and other data necessary to show a fixed-layout flat document are included in each PDF file, which is based on the PostScript language. Because it was standardized as ISO 32000 in 2008, the usage of PDF is no longer subject to royalties. Aside from simple text and pictures, PDF files may also contain layers, rich media (including video material), three-dimensional objects using U3D or PRC, and several other data formats. Other types of content that can be included in PDF files include logical organizing components, interactive elements like annotations and form fields, rich media, and layers. In order to support workflows needing these functionalities, the PDF specification additionally includes metadata, file attachments, and digital signatures.

2.2 PDF TEXT TO SPEECH

A Text-To-Speech (TTS) synthesizer is a computer-based system that can read any text aloud, regardless of whether it was manually entered into the computer by the user or scanned and sent to an OCR system. Try to be as clear as possible.

In that we are interested in the automatic generation of new sentences, there is a fundamental distinction between the system we are about to explain and any other talking machine (as a ca ssette-player, for example). There is still room for improvement in this definition. Voice Response Systems—systems that merely combine isolated words or sentences—are only appropriate when a small vocabulary is required (typically a few hundred words) and when the sentences to be pronounced adhere to a very specific structure, as is the case, for example, with

the announcement of arrivals in train stations. To record and preserve every word used in the language would be impossible (and fortunately pointless) in the context of TTS synthesis. As a result, it would be more accurate to define Text-To-Speech as the automatic generation of speech using a grapheme-to-phoneme transcription of the sentences to speak. This activity does not initially appear to be particularly challenging. After all, a human being is capable of appropriately pronouncing a sentence, even one from his early years. We all possess a thorough understanding of the reading conventions of our mother tongue, most often unconsciously. At primary school, they were passed on to us in a condensed form, and we enhanced them over the years. To suggest that it is only a matter of time until the computer is likely to match the human being in that regard would be a bold statement. We would have to express certain reservations notwithstanding the current state of our knowledge and methods, as well as the recent advancements made in the disciplines of signal processing and artificial intelligence. In actuality, reading taps into the deepest parts of human intelligence that are frequently untapped.

2.2.1 AUTOMATIC READING

Every single and individual synthesizer is the outcome of a unique and original imitation of human reading ability that was subjected to technological and creative limitations that were typical of the era in which it was created. The idea of high quality TTS synthesis first came into existence in the middle of the 1980s as a result of significant advancements made in speech synthesis and natural language processing methods, mostly as a result of the introduction of new technologies (Digital Signal and Logical Inference Processors). It is currently essential for the extension of the speech products family.

High Quality TTS Systems have a wide range of potential uses. Here are a few instances: Communication services. Access to textual information over the phone is made feasible via TTS systems. It is worthwhile to take into account such a possibility given that only roughly 30% of phone conversations actually involve any interaction. Simple messages like "don't

miss" listings for local cultural events (cinemas, theaters) can be found in texts, as well as massive datasets that are essentially unreadable and saved as digital speech. With the use of a speech recognizer or DTMF systems, users of such systems for retrieving information might speak their queries into the system or type them into the phone's keyboard. If necessary, our artificially intelligent machines might even be able to speed up queries by offering lists of keywords or even summaries. In relation to this, AT&T recently coordinated a number of customer testing for a number of promising phone services (Levinson et al. 2008). They are Who's Calling, Integrated Messaging, Telephone Relay Service, and Automated Caller Name and Address (a computerized version of the "reverse directory"). Who's Calling allows you to hear the spoken name of your caller before connecting and hang up to reject the call. As long as the synthetic utterances could be understood, these applications have proven to be acceptable—and even popular. Most of the time, naturalness wasn't a big deal.

Language education. High Quality TTS synthesis can be coupled with a Computer Aided Learning system, and provide a helpful tool to learn a new language. To our knowledge, this has not been done yet, given the relatively poor quality available with commercial systems, as opposed to the critical requirements of such tasks.

Assistance for people with disabilities. Voice impairments result from mental or motor/sensation disorders, and machines can be a huge help in the latter case: with the aid of a specially designed keyboard and a quick sentence assembling program, synthetic speech can be produced in a matter of seconds to address these impediments. Astrophysician Stephen Hawking gives all of his lectures in this manner, and the Telephone Relay Service is another example. Blind people also frequently use computersMass-market synthesizers that come with sound cards will soon encroach on the market for speech synthesis for blind computer users. Even though it isn't currently in a format that is helpful to blind people, DEC speak (TM) is already included with the most recent SoundBlaster (TM) cards.

books and toys that talk. Speech synthesis has already had an impact on the toy industry. The ingenious "Magic Spell" from Texas Instruments served as the inspiration for the development of several talking toys. The low quality of the products on the market unavoidably limits their potential for educational use. This may change if high-quality syntheses were available at reasonable costs. Vocal Monitoring. In some cases, oral information is more efficient than written messages. The appeal is stronger, while the attention may still focus on other visual sources of information. Hence the idea of incorporating speech synthesizers in measurement or control systems.

Multimedia, man-machine communication. In the long run, the development of high quality TTS systems is a necessary step (as is the enhancement of speech recognizers) towards more complete means of communication between men and computers. Multimedia is a first but promising move in this direction.

Fundamental and applied research. TTS synthesizers possess a very peculiar feature which makes them wonderful laboratory tools for linguists: they are completely under control, so that repeated experiences provide identical results (as is hardly the case with human beings). Consequently, they allow to investigate the efficiency of into native and rhythmic models. A particular type of TTS systems, which are based on a description of the vocal tract through its resonant frequencies (its formants) and denoted as formant synthesizers, has also been extensively used by phoneticians to study speech in terms of acoustical rules. In this manner, for instance, articulatory constraints have been enlightened and formally described.

2.2.2 HOW DOES A MACHINE READ?

From now on, it should be clear that a reading machine would hardly adopt a processing scheme as the one naturally taken up by humans, whether it was for language analysis or for speech production itself. Vocal sounds are inherently governed by the partial differential equations of

fluid mechanics, applied in a dynamic case since our lung pressure, glottis tension, and vocal and nasal tracts configuration evolve with time. These are controlled by our cortex, which takes advantage of the power of its parallel structure to extract the essence of the text read: its meaning. Even though, in the current state of the engineering art, building a Text-To-Speech synthesizer on such intricate models is almost scientifically conceivable (intensive research on articulatory synthesis, neural networks, and semantic analysis give evidence of it), it would result anyway in a machine with a very high degree of (possibly avoidable) complexity, which is not always compatible with economic criteria. After all, flies do not flap their wings!

Figure 1 introduces the functional diagram of a very general TTS synthesizer. As for human reading, it comprises a Natural Language Processing module (NLP), capable of producing a phonetic transcription of the text read, together with the desired intonation and rhythm (often termed as prosody), and a Digital Signal Processing module (DSP), which transforms the symbolic information it receives into speech. But the formalisms and algorithms applied often manage, thanks to a judicious use of mathematical and linguistic knowledge of developers, to short-circuit certain processing steps. This is occasionally achieved at the expense of some restrictions on the text to pronounce, or results in some reduction of the "emotional dynamics" of the synthetic voice (at least in comparison with human performances), but it generally allows to solve the problem in real time with limited memory requirements.

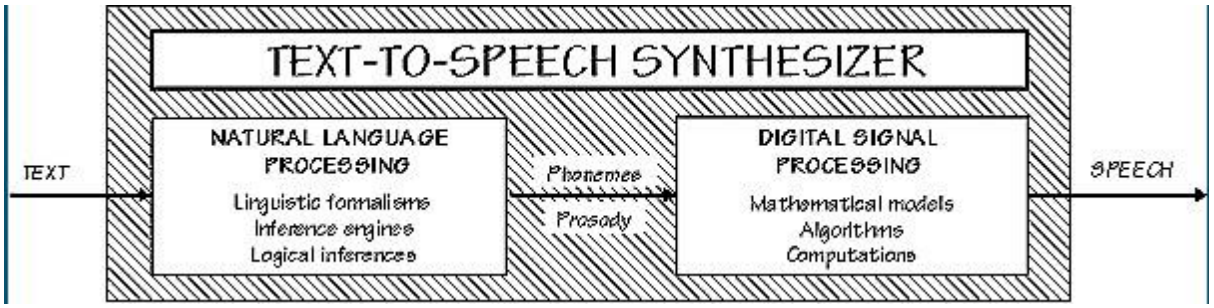


Figure 1. A simple but general functional diagram of a TTS system.

Speech synthesis can be described as artificial production of human speech (Suendermann et al., 2010). A computer system used for this purpose is called a speech synthesizer, and can be implemented in software or hardware. A text-to-speech (TTS) system converts normal language text into speech (Allen et al., (1987). Synthesized speech can be created by concatenating pieces of recorded speech that are stored in a database. Systems differ in the size of the stored speech units; a system that stores phones or diphones provides the largest output range, but may lack clarity. For specific usage domains, the storage of entire words or sentences allows for high-quality output. Alternatively, a synthesizer can incorporate a model of the vocal tract and other human voice characte0ristics to create a completely "synthetic" voice output (Rubin et al., 1981). The quality of a speech synthesizer is judged by its similarity to the human voice and by its ability to be understood. An intelligible text-to-speech program allows people with visual impairments or reading disabilities to listen to written works on a home computer.

A text-to-speech system (or "engine") is composed of two parts: (Van Santen et al., 1997). A front-end and a back-end. The front-end has two major tasks. First, it converts raw text containing symbols like numbers and abbreviations into the equivalent of written-out words. This process is often called text normalization, preprocessing, or tokenization. The front-end then assigns phonetic transcriptions to each word, and divides and marks the text into prosodic units, like phrases, clauses, and sentences. The process of assigning phonetic transcriptions to words is called text-to-phoneme or grapheme-to-phoneme conversion. Phonetic transcriptions and prosody information together make up the symbolic linguistic representation that is output by the front-end. The back-end—often referred to as the synthesizer—then converts the symbolic linguistic representation into sound. In certain systems, this part includes the computation of the target prosody (pitch contour, phoneme durations), (Van Santen et al, 1997) which is then imposed on the output speech. 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 Concatentive 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 (Wasala, 2006).

- **Domain-specific Synthesis:** Domain-specific synthesis concatenates pre-recorded words and phrases to create complete utterances. It is used in applications where the variety of texts the system will output is limited to a particular domain, like transit schedule announcements or weather reports (Lamel et al., 2006). The technology is very simple to implement, and has been in commercial use for a long time, in devices like talking clocks and calculators. The level of naturalness of these systems can be very high because the variety of sentence types is limited, and they closely match the prosody and intonation of the original recordings. Because these systems are limited by the words and phrases in their databases, they are not general-purpose and can only synthesize the combinations of words and phrases with which they have been pre-programmed. The blending of words within naturally spoken language however can still cause problems unless many variations are taken into account. For example, in nonrhotic dialects of English the "r" in words like "clear" /ˈklɪə/ is usually only pronounced when the following word has a vowel as its first letter (e.g. "clear out" is realized as/ˌklɪəɹˈʌʊt/) (Van Truc et al., 2013). Likewise in French, many final consonants become no longer silent if followed by a word that begins with a vowel, an effect called liaison. This alternation cannot be reproduced by a simple word-concatenation system, which would require additional complexity to be context-sensitive. 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.

- **Unit Selection Synthesis:** Unit selection synthesis uses large databases of recorded speech. During database creation, each recorded utterance is segmented into some or all of the following: individual phones, diphones, half-phones, syllables, morphemes, words, phrases, and sentences. Typically, the division into segments is done using a specially modified speech recognizer set to a "forced alignment" mode with some manual correction afterward, using visual representations such as the waveform and spectrogram (Black, 2002). An index of the units in the speech database is then created based on the segmentation and acoustic parameters like the fundamental frequency (pitch), duration, position in the syllable, and neighboring phones. At runtime, the desired target utterance is created by determining the best chain of candidate units from the database (unit selection). This process is typically achieved using a specially weighted decision tree. Unit selection provides the greatest naturalness, because it applies only a small amount of digital signals 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 the point of concatenation to smooth the waveform. The output from the best unit-selection systems is often indistinguishable from real human voices, especially in contexts 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 the gigabytes of recorded data, representing dozens of hours of speech (Kominek et al., (2003). Also, unit selection algorithms have been known to select segments from a place that results in less than ideal synthesis (e.g. minor words become unclear) even when a better choice exists in the database (Zhang, 2004).
- **Diphone Synthesis:** Diphone synthesis uses a minimal speech database containing all the diphones (sound-to-sound transitions) occurring in a language. The number of diphones depends on the phonotactics of the language: for example, Spanish has about 800 diphones,

and German about 2500. In diphone synthesis, only one example of each diphone 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 (Kominek & Black, 2003). or MBROLA. (Dutoit et al., 1996). The quality of the resulting speech is generally worse than that of unit-selection systems, but more natural-sounding than the output of formant synthesizers. Diphone synthesis suffers from the sonic glitches of concatenative synthesis and the robotic-sounding nature of formant synthesis, and has few of the advantages of either approach other than small size.

2.3 STRUCTURE OF A TEXT-TO-SPEECH

2.3.1 SYNTHESIZER SYSTEM

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. From the information now available, it can produce a speech signal. The structure of the text-to-speech synthesizer can be broken down into major modules:

- **Natural Language Processing (NLP) module:** It produces a phonetic transcription of the text read, together with prosody.
- **Digital Signal Processing (DSP) module:** It 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 (“m” in “Maximum”). In addition, several letters can correspond to a single phoneme (“ch” in “rich”). There are two strategies to determine pronunciation:
- In dictionary-based solution with morphological components, as many morphemes (words) as possible are stored in a dictionary. Full forms are generated by means of inflection, derivation and composition rules. Alternatively, a full form dictionary is used in which all possible word forms are stored. Pronunciation rules determine the pronunciation of words not found in the dictionary.
- In a rule based solution, pronunciation rules are generated from the phonological knowledge of dictionaries. Only words whose pronunciation is a complete exception are included in the dictionary. The two applications differ significantly in the size of their dictionaries. The dictionary-based solution is many times larger than the rules-based solution’s dictionary of exception.
- **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). The output of the NLP module is passed to the DSP module.

This is where the actual synthesis of the speech signal happens. In concatenate 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.

2.3.2 THE NLP COMPONENT

Figure 2 introduces the skeleton of a general NLP module for TTS purposes. One immediately notices that, in addition with the expected letter-to-sound and prosody generation blocks, it comprises a morpho-syntactic analyser, underlying the need for some syntactic processing in a high quality Text-To-Speech system. Indeed, being able to reduce a given sentence into something like the sequence of its parts-of-speech, and to further describe it in the form of a syntax tree, which unveils its internal structure, is required for at least two reasons:

1. Accurate phonetic transcription can only be achieved provided the part of speech category of some words is available, as well as if the dependency relationship between successive words is known.

2. Natural prosody heavily relies on syntax. It also obviously has a lot to do with semantics and pragmatics, but since very few data is currently available on the generative aspects of this dependence, TTS systems merely concentrate on syntax. Yet few of them are actually provided with full disambiguation and structuration capabilities.

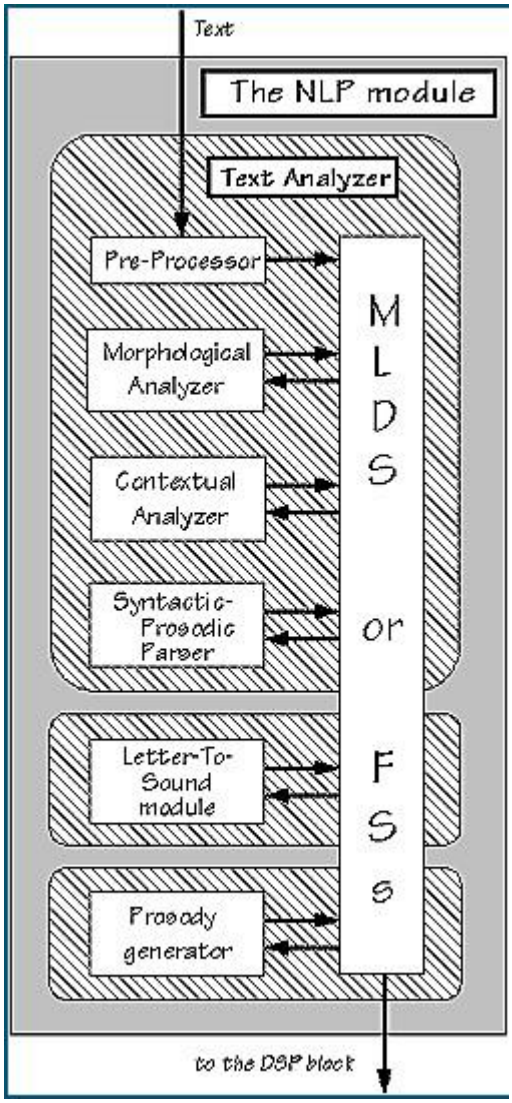


Fig 2. The NLP module of a general Text-To-Speech conversion system.

2.4 TEXT ANALYSIS

The text analysis block is itself composed of:

- A pre-processing module, which organizes the input sentences into manageable lists of words. It identifies numbers, abbreviations, acronyms and idiomatics and transforms them into full text when needed. An important problem is encountered as soon as the character level: that of punctuation ambiguity (including the critical case of sentence end detection). It can be solved, to some extent, with elementary regular grammars.

- A morphological analysis module, the task of which is to propose all possible part of speech categories for each word taken individually, on the basis of their spelling. Inflected, derived, and compound words are decomposed into their elementary graphemic units (their morphs) by simple regular grammars exploiting lexicons of stems and affixes (see the CNET TTS conversion program for French (Larreur, 1989), or the MITTALK system (Allen et al., 1987).

- The contextual analysis module considers words in their context, which allows it to reduce the list of their possible part of speech categories to a very restricted number of highly probable hypotheses, given the corresponding possible parts of speech of neighbouring words. This can be achieved either with n-grams (Kupiec, 1992), which describe local syntactic dependences in the form of probabilistic finite state automata (i.e. as a Markov model), to a lesser extent with mutli-layer perceptrons (i.e., neural networks) trained to uncover contextual rewrite rules, as in (Benello et al,. 1989), or with local, non-stochastic grammars provided by expert linguists or automatically inferred from a training data set with classification and regression tree (CART) techniques [Sproat et al. 92, Yarowsky 94].

- Finally, a syntactic-prosodic parser, which examines the remaining search space and finds the text structure (i.e. its organization into clause and phrase-like constituents) which more closely relates to its expected prosodic realization (see below).

2.3.3. AUTOMATIC PHONETIZATION

A poem of the Dutch high school teacher and linguist G.N. Trenite surveys this problem in an amusing way. It desperately ends with :

> Finally, which rimes with "enough", Though, through, plough, cough, hough, or tough?
> Hiccough has the sound of "cup", My advice is ... give it up!

The Letter-To-Sound (LTS) module is responsible for the automatic determination of the phonetic transcription of the incoming text. It thus seems, at first sight, that its task is as simple as performing the equivalent of a dictionary look-up! From a deeper examination, however, one quickly realizes that most words appear in genuine speech with several phonetic transcriptions, many of which are not even mentioned in pronunciation dictionaries. Namely:

1. Pronunciation dictionaries refer to word roots only. They do not explicitly account for morphological variations (i.e. plural, feminine, conjugations, especially for highly inflected languages, such as French), which therefore have to be dealt with by a specific component of phonology, called morphophonology.

2. Some words actually correspond to several entries in the dictionary, or more generally to several morphological analyses, generally with different pronunciations. This is typically the case of heterophonic homographs, i.e. words that are pronounced differently even though they have the same spelling, as for 'record' (/rekoùd/ or /rIkoùd/), constitute by far the most tedious class of pronunciation ambiguities. Their correct pronunciation generally depends on their part-of-speech and most frequently contrasts verbs and non-verbs , as for 'contrast' (verb/noun) or 'intimate' (verb/adjective), although it may also be based on syntactic features, as for 'read' (present/past)

3. Pronunciation dictionaries merely provide something that is closer to a phonemic transcription than from a phonetic one (i.e. they refer to phonemes rather than to phones). As denoted by (Withgott and Chen 1993): "while it is relatively straightforward to build computational models for morph phonological phenomena, such as producing the dictionary pronunciation of 'electricity' given a base form

'electric', it is another matter to model how that pronunciation actually sounds". Consonants, for example, may reduce or delete in clusters, a phenomenon termed as consonant cluster simplification, as in 'softness' [sofnIs] in which [t] fuses in a single gesture with the following [n].

4. Words embedded into sentences are not pronounced as if they were isolated. Surprisingly enough, the difference does not only originate in variations at word boundaries (as with phonetic liaisons), but also on alternations based on the organization of the sentence into non-lexical units, that is whether into groups of words (as for phonetic lengthening) or into non-lexical parts thereof (many phonological processes, for instance, are sensitive to syllable structure).

5. Finally, not all words can be found in a phonetic dictionary: the pronunciation of new words and of many proper names has to be deduced from the one of already known words.

2.3.4 PROSODY GENERATION

The term prosody refers to certain properties of the speech signal which are related to audible changes in pitch, loudness, syllable length. Prosodic features have specific functions in speech communication. The most apparent effect of prosody is that of focus. For instance, there are certain pitch events which make a syllable stand out within the utterance, and indirectly the word or syntactic group it belongs to will be highlighted as an important or new component in the meaning of that utterance. The presence of a focus marking may have various effects, such as contrast, depending on the place where it occurs, or the semantic context of the utterance.

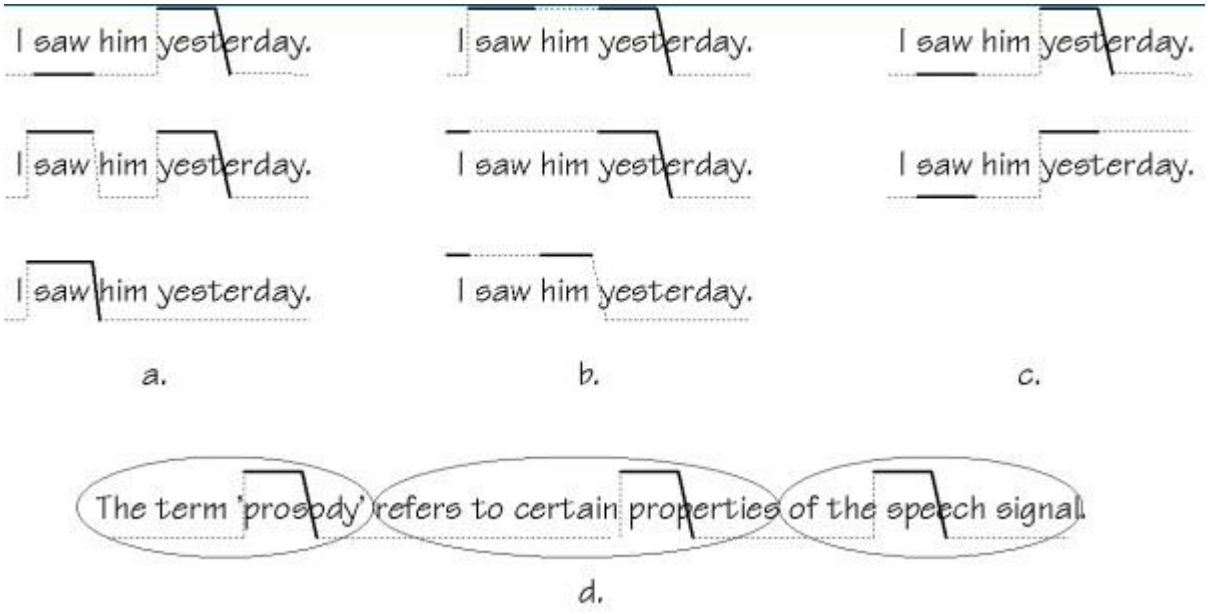


Fig. 3. Different kinds of information provided by intonation (lines indicate pitch movements; solid lines indicate stress).

- Focus or given/new information;
- Relationships between words (saw-yesterday; I-yesterday; I-him)
- Finality (top) or continuation (bottom), as it appears on the last syllable;
- Segmentation of the sentence into groups of syllables.

Although maybe less obvious, there are other, more systematic or general functions.

2.3.5. RULE-BASED SYNTHESIZERS

Rule-based synthesizers are mostly in favour with phoneticians and phonologists, as they constitute a cognitive, generative approach of the phonation mechanism. The broad spreading of the Klatt synthesizer (Klatt, 2011), for instance, is principally due to its invaluable assistance in the study of the characteristics of natural speech, by analytic listening of rule-synthesized speech.

2.4 RELATED WORK

2.4.1 SPEECHIFY

Speechify is a popular text-to-speech app available for iOS, Android, and desktop computers that can read PDF files out loud and convert them into MP3 or different audio formats. Speechify can read from a variety of file types, including PDF documents, and offers several voice and speed options. In addition to the natural-sounding voices Speechify offers, you can also use celebrity voices like Gwyneth Paltrow's voice, to read your documents aloud. Speechify integrates with cloud storage services such as Dropbox and Google Drive are one of our many top-rated and highly-sought after features. This allows members to access their PDF documents from anywhere and convert them to different audio formats using Speechify.

2.4.2 NATURAL READER

Natural Reader is a text-to-speech software that can read PDF files out loud and convert them into an audio format such as MP3 or WAV. It can be used on both Windows and Mac computers, and is also available as an iPhone or iPad app. Natural Reader can read out loud from a wide range of file formats including PDF documents, EPUB, HTML, TXT, RTF, and Microsoft Office files such as Excel, PowerPoint, and Word. Natural Reader offers multiple voices to choose from, and the ability to adjust the voice speed and volume. It also has a toolbar integration with Microsoft Word, which enables users to read out loud their documents directly from Word.

2.4.3 ADOBE ACROBAT PRO DC

Adobe Acrobat Pro DC is a widely used PDF editor, but it also includes a PDF-to-audio conversion feature. It can convert PDF files to various audio file formats, such as MP3, M4A, WMA, OGG, and WAV. Adobe Acrobat Pro DC is available for both Mac and Windows operating systems, and can also be used on mobile devices. Adobe Acrobat Pro DC is not just

an audio converter, but also a powerful PDF editor. It allows users to edit PDFs and add audio annotations to specific parts of the document. This is particularly useful for those who want to add audio notes or feedback to a PDF file.

2.4.4 VOICE DREAM READER

Voice Dream Reader is an app available for iOS and Android devices that provides a range of text-to-speech features. It can convert PDF files to MP3 audio, and also supports other file types such as TXT, EPUB, and HTML. The app includes different customization options such as voice and speed adjustments, and it can even turn PDF documents into audiobooks. The Voice Dream Reader app is specifically designed for reading and listening to content. It includes a feature that allows users to create playlists of their favorite articles and stories, and can even turn long PDF documents into audiobooks for easy listening.

2.4.5 READSPEAKER

ReadSpeaker is another online service that can convert PDF files to audio format, including MP3 and WAV. The service offers a high level of customization, allowing users to adjust the voice, speed, and other settings to suit their preferences. ReadSpeaker offers a plugin for Chrome that can read out loud any online PDF document. This is really useful for those who frequently read online PDF documents and want to listen to them rather than read them.

2.4.6 BALABOLKA

Balabolka is a text-to-speech software that can convert PDF files to various audio file formats, such as MP3, WAV, and OGG. It is available for Windows computers and supports a wide range of languages. Balabolka offers OCR functionality, which can recognize text in images and JPG files within the PDF document. This means that even if a PDF contains images with text, Balabolka can still convert them to audio format.

CHAPTER THREE

3.0 METHODOLOGY

The first step in the system development life cycle is the identification of a need. This is a user's request to change, improve or enhance an existing system. Because there is likely to be a stream of such requests, standard procedures must be established to deal with them. The objective of project selection is to determine whether the request is valid and feasible before a recommendation is reached to do nothing, improve or modify the existing system or build a new one.

3.1 PROPOSED SYSTEM

In this current busy routine people do not find time to read a book, or to convert the PDF file into an MP3 player using third-party applications or web applications. Even I have a directory at which I store pdf books that I plan on reading, but I never do. So, I thought hey, why do not I make them audiobooks and listen to them while I do something else ! In this system, we are developing a GUI application using python to convert the PDF file into audio format and read it out to the user. The application is more user-friendly as it does not require any audio file or MP3 player. The user will have to select the PDF file which the user wants to listen to.

3.2 GENERAL DESCRIPTION OF THE EXISTING SYSTEM

In today's world where most information is shared digitally, visually impaired persons always require their reading glasses to have access to this information, in a situation where they somehow forgot their reading glasses, they won't be to have access this information. But with text to speech system digital information can be read out to a visually impaired person.

3.3 FACT FINDING METHODS USED

There are two main sources of data collection in carrying out this study, information was basically obtained from the two sources which are:

(a) Primary source and

(b) Secondary source

3.3.1 PRIMARY SOURCE

Primary source refers to the sources of collecting original data in which the researcher makes use of empirical approach such as personal interview, questionnaires or observation.

In my research, I used a method of observation were I was attentive to how contact are being operated and saved using a manual method.

3.3.2 SECONDARY SOURCE

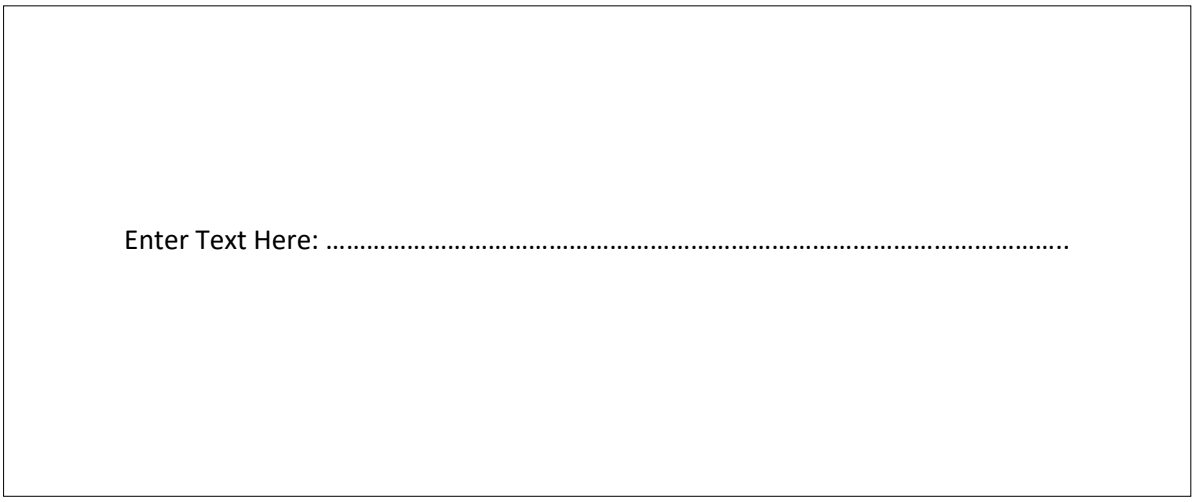
The need of the secondary sources of data for this kind of project cannot be over emphasized. The secondary data were obtained by me from the library source and most of the information from the library research has been covered in my literature review in the previous chapter of this project.

3.5 ORGINAZATIONAL STRUCTIONAL STRUCTURE

There is no organizational frame work since I was using an observation method. The organizational frame work can only be explained.

3.5 INPUT ANALYSIS

The system has only one input structure which the text input form.



3.6 PROCESS ANALYSIS

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. From the information now available, it can produce a speech signal.

The structure of the text-to-speech synthesizer can be broken down into major modules:

Natural Language Processing (NLP) module: It produces a phonetic transcription of the text read, together with prosody.

Digital Signal Processing (DSP) module: It 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 (“m” in “Maximum”). In addition, several letters can correspond to a single phoneme (“ch” in “rich”).

3.7 OUTPUT ANALYSIS

The output from the system designed is generated from the system inputs. The system format for this system audio format

CONCLUSION

PDF to audio converter is a rapidly growing aspect of computer technology and is increasingly playing a more important role in the way we interact with the system and interfaces across a variety of platforms. We have identified the various operations and processes involved in text to speech synthesis. We have also developed a very simple and attractive graphical user interface which allows the user to type in his/her text provided in the text field in the application. Our system interfaces with a text to speech engine developed for American English. In future, we plan to make efforts to create engines for localized Nigerian language so as to make text to speech technology more accessible to a wider range of Nigerians. This already exists in some native languages the Vietnamese synthesis system and the Telugu language. Another area of further work is the implementation of a text to speech system on other platforms, such as telephony systems, ATM machines, video games and any other platforms where text to speech technology would be an added advantage and increase functionality.