A Simplified Overview of Text-To-Speech Synthesis

J. O. Onaolapo, F. E. Idachaba, J. Badejo, T. Odu, and O. I. Adu

***Abstract* — Computer-based Text-To-Speech systems render text into an audible form, with the aim of sounding as natural as possible. This paper seeks to explain Text-To-Speech synthesis in a simplified manner. Emphasis is placed on the Natural Language Processing (NLP) and Digital Signal Processing (DSP) components of Text-To-Speech Systems. Applications and limitations of speech synthesis are also explored.**

***Index Terms* — Speech synthesis, Natural Language Processing, Auditory, Text-To-Speech**

I. Introduction

A Text-To-Speech System (TTS) is a computer-based system that automatically converts text into artificial human speech [1]. Text-To-Speech synthesizers do not playback recorded speech; rather, they generate sentences using plain text as input. It is necessary to distinguish between Text-To-Speech synthesizers from Voice Response Systems. Voice Response Systems simply concatenate words and segments of sentences and are applicable only in situations where limited vocabulary is required, and pronunciation restrictions exist.

Speech synthesizers being considered in this context actually encapsulate models of the human vocal tract to produce a synthetic human voice output corresponding to input text. Since it is impracticable to store pre-recorded audio clips of all words of a language, automatic 'pronunciation' of words and sentences is necessary in Text-To-Speech systems. *eSpeak* [2] and *SpeakVolumes* [3] are good examples of speech synthesis software. *eSpeak* is free, open-source software while *SpeakVolumes* is commercial.

It must however be noted that Text-To-Speech systems do not sound perfectly natural due to audible glitches [4]. This is because speech synthesis science is yet to capture all the complexities and intricacies of human speaking capabilities.

II. Machine Speech

Text-To-Speech system processes are significantly different from live human speech production (and language analysis). Live human speech production depends of complex fluid mechanics dependent on changes in lung pressure and vocal tract constrictions. Designing systems to mimic those human constructs would result in avoidable complexity.

In general terms, a Text-To-Speech synthesizer comprises of two parts; namely the Natural Language Processing (NLP) unit and the Digital Signal Processing (DSP) unit. The Natural Language Processing unit handles phonetization and intonation along with rhythm and it outputs a phonetic transcript of the input text [5]. The Digital Signal Processing unit transforms the phonetic transcript it receives into machine speech [1]. Fig. 1 illustrates a simplified representation of a Text-To-Speech synthesizer. The following sections explore Natural Language Processing (NLP) and Digital Signal Processing (DSP) components in detail.

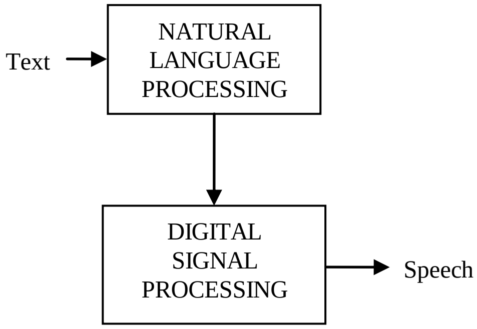

Fig. 1. A simplified diagram of a Text-To-Speech synthesizer [1]

A. *Natural Language Processing (NLP)*

The Natural Language Processing (NLP) module comprises

J. O. Onaolapo is a graduate student in University College London, Gower Street WC1E 6BT London. (phone: +44.777.653.1714; e-mail: jonaolapo@gmail.com).

F. E. Idachaba is with the Department of Electrical and Information Engineering, Covenant University, PMB 1023 Ota, Nigeria (e-mail: francis.idachaba@covenantuniversity.edu.ng).

J. Badejo is with the Department of Electrical and Information Engineering, Covenant University, PMB 1023 Ota, Nigeria (e-mail: joke.badejo@covenantuniversity.edu.ng).

T. Odu is with the Department of Electrical and Information Engineering, Covenant University, PMB 1023 Ota, Nigeria (e-mail: tiwalade.majekodunmi@covenantuniversity.edu.ng).

O. I. Adu is with the Department of Electrical and Information Engineering, Covenant University, PMB 1023 Ota, Nigeria (e-mail: damilola.adu@covenantuniversity.edu.ng).

of text analyzer, phonetization and prosody generation modules. Prosody refers to rhythm, stress and intonation of speech. More details of the above listed modules can be found in the following sections. Fig. 2 shows a block diagram of the NLP module.

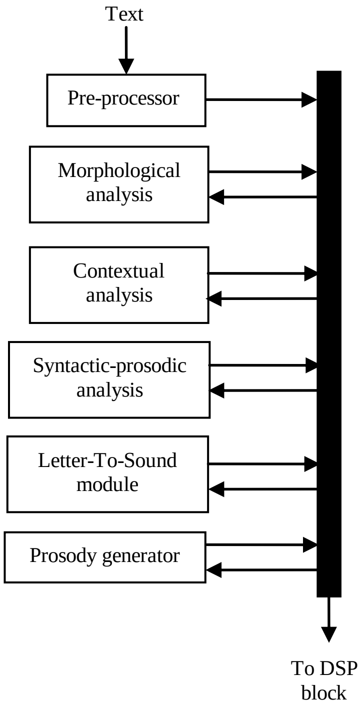

Fig. 2. Block diagram of Natural Language Processing module [1]

Text analyzer

The text analyzer comprises of four basic parts, namely: a pre-processing block, a morphological analysis block, a contextual analysis block and a syntactic-prosodic parser.

The pre-processing block converts abbreviations, numbers and acronyms into full text when necessary. It also breaks input sentences into groups of words.

The morphological analysis block categorizes each word in the sentence being analyzed into possible parts of speech, on the basis of the word's spelling. Compound words are decomposed into their basic units in this module.

The contextual analysis module streamlines the list of possible parts of speech of words in sentences, by considering the parts of speech of neighbouring words.

The syntactic-prosodic parser locates the text structure (in terms of clause and phrase constituents) that tends more closely to the prosodic realization of the input sentence.

Phonetization

A Letter-To-Sound (LTS) module is used for phonetic transcription of incoming text. Worthy of note here however, is the fact that this transcription is beyond a dictionary look-up operation. This is because most words have different phonetic transcriptions depending on context. Also, pronunciation dictionaries do not account for morphological variations in words. In addition, pronunciations of words in sentences differ from pronunciation of those same words when they are isolated. Furthermore, not all words are present in a phonetic dictionary. As a result, phonetization can be dictionary-based or rule-based (based on a collection of letter-to-sound rules).

Prosody generation

As stated earlier, prosody refers to rhythm, stress and intonation of speech. Prosody directs focus to specific parts of a sentence, such as emphasis laid on a specific syllable, thus attributing special importance or contrast to that part of the sentence. Prosodic features also help to segment sentences into chunks comprising of groups of words and syllables and also to identify the relationships between such chunks. The prosody generator is responsible for prosody generation. Generation of a natural-sounding prosody is one of the biggest challenges faced in the design of Text-To-Speech systems.

B. *Digital Signal Processing (DSP) component*

The Digital Signal Processing (DSP) component handles the actual machine 'pronunciation' of words, phrases and sentences, analogous to human speech articulation (based on input to the Digital Signal Processing component). This component can be implemented in two ways, namely rule-based synthesis and concatenative synthesis. The two approaches are explored in detail below.

Rule-based synthesizers

Rule-based synthesizers, which are usually formant synthesizers, generate speech via the dynamic modification of several parameters. Such parameters as fundamental frequency, voicing and noise levels are modified over time to create an artificial speech waveform. Many formant-based speech synthesis systems generate unnatural speech (not sounding human). The large number of parameters involved introduces complications during analysis and as such some errors are introduced. Combating such errors makes development of rule-based synthesizers very time-consuming. Despite that, rule-based synthesizers are common with phoneticians and phonologists (for instance, Klatt synthesizer). Rule-based synthesizers usually have programs that are smaller than concatenative synthesizers (since rule-based synthesizers do not rely on database of speech samples).

Concatenative synthesizers

Concatenative synthesizers string together pieces of recorded speech extracted from a database of speech samples. As a result, concatenative synthesizers generate the most natural-sounding artificial speech. Concatenative synthesizers actually possess a very limited knowledge (phonetics-wise) of the data they work on. Audible glitches are sometime observed in the output of concatenative synthesizers due to amplitude and timbre mismatches between concatenated samples. A process called equalization is used in alleviating the effects of amplitude mismatches.

While synthesizing speech, concatenantive synthesizers produce a sequence of concatenated segments, retrieved from its speech sample database. The prosody of the segments is adjusted to match the values deduced from the output of the Natural Language Processing module.

Synthesizing speech

A sequence of appropriate segments is first computed from the output of the Natural Language Processing module (which serves as input to the Digital Signal Processing module). Prosodic characteristics are then imposed on the individual segments. Afterwards, segments are matched to one another by smoothing out discontinuities. The stream of parameters derived is then used to produce synthesized speech. For best results, the number and length of segments used should be small as possible.

C. Applications of Text-To-Speech synthesis

Areas of application of Text-To-Speech systems include the following:

In telecommunications, Text-To-Speech systems made it possible to listen to text read by machines, for instance, from a database (instead of human operators). Queries to such databases can be transmitted via user speech (speech recognition systems) or telephone keypad (DTMF systems).

In multimedia, Text-To-Speech synthesis has made possible the existence of talking books, toys and interactive games. Around 2007 for instance, Animo Limited developed a software package which is able to generate narration and lines of dialogue. That software package was developed based on Animo's speech synthesis software, FineSpeech [6].

In addition, for individuals with speech impairment, speech synthesis has provided an artificial voice hence simplifying communication with others. The famous physicist Stephen Hawking [7] delivered his lectures using a speech synthesizer (he lost the ability to speak following an operation). Individuals with sight impairment also benefit from software such as screen readers and paper document readers (OCR) which are based on speech synthesis technology.

D. Conclusion

This paper has explored the workings of TTS synthesis in a simplified manner. In TTS systems, the Natural Language Processing module has been shown to be the module that actually 'reads' and 'understands' the input text, while the Digital Signal Processing module 'vocalizes' the input content. Some applications of text-to-speech synthesis have also been examined.

REFERENCES

[1] T. Dutoit, "High-quality text-to-speech synthesis: An overview," *JOURNAL OF ELECTRICAL AND ELECTRONICS ENGINEERING AUSTRALIA*, vol. 17, pp. 25–36, 1997.

[2] "eSpeak: Speech Synthesizer." [Online]. Available: http://espeak.sourceforge.net/. [Accessed: 30-Mar-2014].

[3] "Text-to-Speech | TTS on Demand | Speech Synthesis on Demand | TTS Web Service | SpeakVolumes by innoetics." [Online]. Available: http://www.speakvolumes.eu/demo.php. [Accessed: 30-Mar-2014].

[4] D. Jurafsky and H. James, "Speech and language processing an introduction to natural language processing, computational linguistics, and speech," 2000.

[5] A. Trilla, *Natural Language Processing techniques in Text-To-Speech synthesis and Automatic Speech Recognition*. Departament de Tecnologies Media Enginyeria i Arquitectura La Salle (Universitat Ramon Llull), Barcelona, Spain atrilla@salle.url.edu, 2009.

[6] "Animo: Social Solutions." [Online]. Available: http://www.animo.co.jp/EN/socialsolutions/. [Accessed: 31-Mar-2014].

[7] "Stephen Hawking - Home." [Online]. Available: http://www.hawking.org.uk/. [Accessed: 31-Mar-2014].