IJCSE International Journal of Computer Sciences and Engineering Open Access
Survey Paper Vol.7, Special Issue.1, Jan 2019 E-ISSN: 2347-2693

Twee: A Novel Text-To-Speech Engine

D. Das[1], H. Hassan[2], S. Gupta[3*]

[1,2,3]Dept. of Computer Science & Engineering, University Institute of Technology, The University of Burdwan, Golapbag (North), Burdwan- 713104, West Bengal, India

**Corresponding Author: sumitsayshi@gmail.com, Tel.:+91-9830524860*

Available online at: www.ijcseonline.org

***Abstract*—** With the advancement of technology and the widespread use of smart devices, the world has witnessed that the networking and/or the connectivity horizon has broadened to an exalted level. One of the prominent researches being undertaken in this digital era is the development of Text-to-Speech (TTS) engines; which is capable enough of offering more interactivity with the prevalent smart devices. There are various TTS engines available in the market currently, but these engines lack the capability of showing the effects of human voice e.g., they fail to provide credible indications of the sentiment, mood or emotional state of mind of the speaker etc. Further speaking, presently there is no comprehensible or consummate TTS engine that could replicate human behaviour and/or mannerisms with utmost precision and accuracy. This paper proposes a novel Text-to-Speech engine named 'Twee' whose pronunciation works in sync with real world human intelligence. The proposed system is an application of the interdisciplinary field of research whereby domains such as Natural Language Processing, Artificial Intelligence and Digital Signal Processing are amalgamated to perform sentiment analysis on text through the processing of phonemes. This system works well both in mono channel mode and in stereo mode and is capable of generating varied effects on a voice depending on the type of communication.

***Keywords*—**Artificial Intelligence, Natural Language Processing, Digital Signal Processing, Phoneme, Emotion.

I. INTRODUCTION

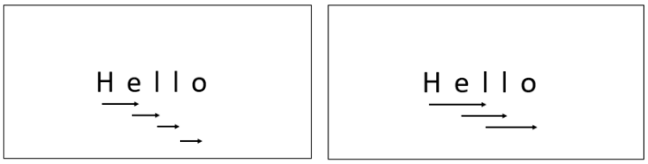
Several researchers have proposed a number of Text-to-Speech (TTS) engines in order to provide to the world an efficient system that could comprehend the actual emotional state and psyche of the speaker. But till date, there is no such system that can replicate the precise and perfect demeanour and disposition of an individual. The challenge to create a consummate TTS engine was the driving factor that paved way into proposing Twee, which is a novel Text-to-Speech engine. This proposed engine incorporates different filters and is based on novel methods to make it more interactive to the user. In addition, this system is very dynamic and capable enough to create a rich and reliable dataset. Different real-time Digital Signal Processing methods are used to change the voice effects during the time of speaking. There are mainly three techniques that are used to make the TTS engine 'Twee' more powerful when compared to other existing systems. These techniques are Natural Language Processing, Artificial Intelligence and Digital Signal Processing.

In this research work, the modified "Concatenative Synthesis" algorithms have been used. Along with it, a novel method of Phoneme's Processing namely the "Odd Mode Rendering" and the "Even Mode Rendering" have been implemented. For detecting the sentiment as implied by the text in the form of sentences or words, 10 different emotion options for each word have been used [1]. The proposed TTS engine works as follows: Firstly, based on the emotions, the system detects the sentiment conveyed by the conversational text and then changes the effect of the response. The 10 different emotions considered in this approach are Fear, Anger, Sadness, Stronger, Joy, Disgust, Surprise, Trust, Anticipation and Nothing.

The proposed TTS engine can not only work in the mono channel mode, but also depending upon the conversation, if the text is properly understood then the system can enhance the channel mode to stereo. This mode conversion feature does not exist in the other existing TTS engines and is one of the most crucial and striking feature of this system.

Three existing methods called USS (Unit Selection Speech) as a static database, Diphone and Domain-Specific Synthesis are used to make its database for dynamic mode [2]. If at any moment the USS results in false output, then out of the two known rendering techniques, the "Odd Mode Rendering" is used for odd word length and the "Even Mode Rendering" is used for even word length (see Figure 1).

It may happen that for a single alphabet there may be two different sounds generated for two different words e.g., In the words “apple” and “ate”, the starting sound of letter “a” is different in both the words. So to make it more familiar with the surrounding, an extra option of correctness has been added while speaking in the TTS engine. This correctness option enables a change (or an update) in the algorithm meant for speech formation on demand. This scheme offers a way of learning correct pronunciation using manual help or from an expert. This method for correctness does not exist in the current systems and adds embellishments to this proposed approach.



(a) (b)

Figure 1. Schematic depicting the word ‘Hello’ using (a) Even Mode Rendering (b) Odd Mode Rendering.

The rest of the paper is organized as follows, Section I contains the introduction of TTS Engines, Section II contains the architecture and the working principle of the Proposed System, Section III describes results with a comparative performance analysis, Section IV discusses the pros and cons of the proposed approach and Section V concludes the paper with future directions of improvement.

II. PROPOSED WORK

The proposed TTS engine ‘Twee’ is a standalone command line executable file which is not only used as a local file but also has its own TTS server to serve the speech synthesis in the web technology without any special dependencies on the existing web services. The following are the three main options supported by this system:

1. Normal Speech (Acts like Google TTS, Microsoft TTS etc). [3]

2. Customized Speech Synthesis (Acts like IBM speech engine). [4]

3. Artificial Intelligence (No existing engine in the market right now).

To make the dataset for TTS engine, a new “Development Panel” (see Figure 2) has been created through which more than 1000 words can be aptly synthesized in only few minutes. The whole dataset is in a raw format with an encrypted decoder to make all the synthesized files secured from any external access.

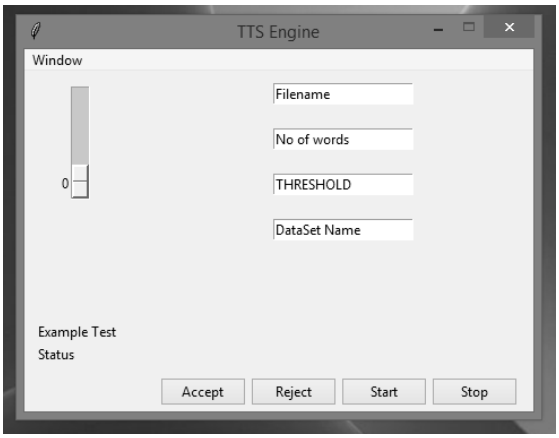


Figure 2. The TTS development Panel.

Firstly, the target sentence is inputted or fed through either a web terminal or a local console. Then the TTS engine will start its lexical phase for understanding the raw textual data. In the first phase or level (see Figure 3), the structural analysis is done to fetch the content words of the sentence which will inform about the sentiment of the specific sentence. This forms the core of Natural Language Processing (NLP).

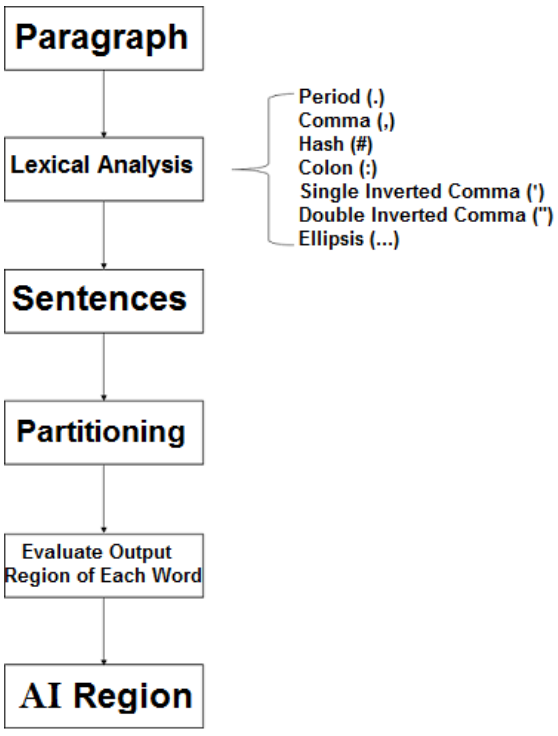


Figure 3. Flowchart depicting the First Level Filtering of Input at the NLP Level.

Next the process flow is shifted to the Artificial Intelligence (AI) Level which will gather all the emotions of words and combine them to form a unique emotion for the whole

paragraph (see Figure 4). If at any moment, the conjugation of all the emotions is not understood by the system, then it will process each sentence and each word individually. If the dataset does not contain any inputted word, then the system will use its dynamic mode of rendering to create the sound of the word. Depending upon the word length, the partition algorithm is changed to make the phonemes. Now after the generation of phonemes or the sentence emotion successfully, the process flow is shifted to the Digital Signal Processing (DSP) Level. [5][6]

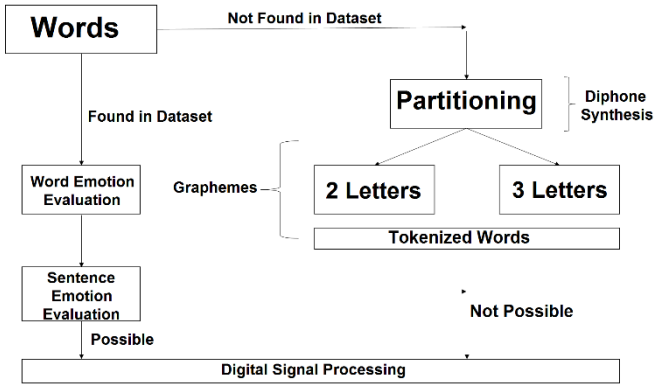


Figure 4. Flowchart depicting the Second Level Filtering of Sentences at the AI Level.

In the DSP Level (see Figure 5), there are mainly four filters for processing each of the phonemes and also the words. After the application of the filters on the phonemes, the mixing process or the domain-specific synthesis gets started to create a resultant audio file. [7][8]

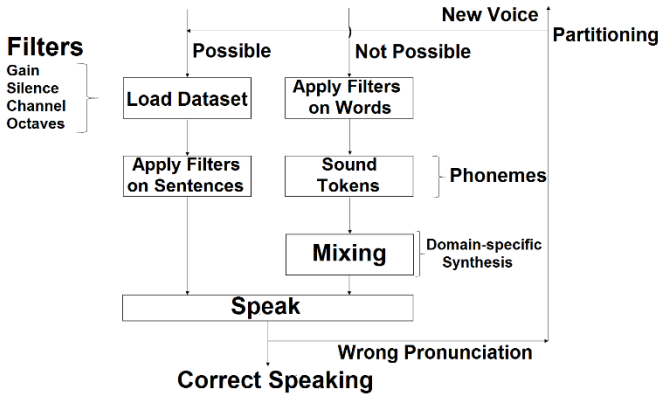


Figure 5. Flowchart depicting the Third Level Filtering of Sentences at the DSP Level.

In case of incorrect pronunciation in the speech, the additional option of correct pronunciation through user-level interaction will be used that will update or change the partitioning algorithm to create correct phonemes. [9]

III. RESULTS

Figure 6 shows the comparison of the proposed TTS engine 'Twee' with three popular existing systems viz. Google TTS, Microsoft TTS and IBM TTS on the five most important constraints like System Intelligence, Customizability, Voice Quality, Database Size and Flexibility.

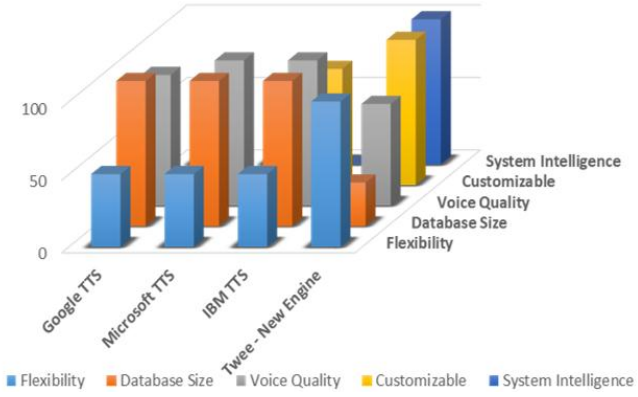


Figure 6. Comparison of Our Proposed TTS engine 'Twee' with other Existing Sytems.

Based on the vertical bar chart, it can be seen that the proposed TTS engine is much ahead in performance in terms of the parameters like System Intelligence, Customizability and Flexibility. The only parameters where the performance suffers are the limited Database Size and the not-so-good Voice Quality.

IV. DISCUSSION AND ANALYSIS

The proposed TTS engine has the following advantages when compared with the other existing TTS engines:

1. The usage of the notion of Artificial Intelligence and the analysis of sentiments.

2. The efficient utilization of DSP filters to deal with different types of voice such as Loud Voice, Soft Voice etc.

3. It offers both a standalone (offline) service as well as online service.

The limitation of the proposed system is that as it is in the development stage, so the dataset used for the purpose of training and testing is limited and currently only 10000 words have been analyzed using the Twee engine.

V. CONCLUSION AND FUTURE SCOPE

Thus, it can be concluded that the proposed TTS engine named 'Twee' offers a more interactive and a natural interface for the real world entities and due to the amalgamation of the concept of NLP, AI and DSP, this system stands out as a winner when compared with other traditional TTS engines available in the market.

As a future scope of work, the task of increasing the size of the dataset so as to comprehend the complete working of the proposed system needs to be taken up. Further, there is a provision of enhancing the quality of voice so that the proposed system can perfectly replicate the voice patterns of a real human being and thus serve the basic purpose of creating a high-end AI product.

REFERENCES

[1] A. Drahota, A. Costall, V. Reddy, *"The Vocal Communication of Different Kinds of Smile",* Speech Communication, Vol.**50,** Issue.**4**, pp.**278-287, 2007**. doi: 10.1016/j.specom.2007.10.001

[2] W.Y. Wang, K. Georgila, "*Automatic Detection of Unnatural Word-Level Segments in Unit-Selection Speech Synthesis*", In the Proceedings of the 2011 IEEE Workshop on Automatic Speech Recognition & Understanding, Waikoloa, HI, **USA**, pp.**289-294**, **2011.**

[3] R.E. Remez, P.E. Rubin, D.B. Pisoni, T.D. Carrell, *"Speech Perception without Traditional Speech Cues",* Science, New Series, Vol.**212**, Issue.**4497**, pp. **947-950**, **1981**. doi:10.1126/science.7233191

[4] J. Zhang, *"Language Generation and Speech Synthesis in Dialogues for Language Learning",* Massachusetts Institute of Technology, pp.**1-68, 2004**.

[5] S. Lemmetty, *"Review of Speech Synthesis Technology",* Helsinki Universty of Technology, pp.**1-113**, **1999**.

[6] I.G. Mattingly,*"Speech synthesis for phonetic and phonological models"*, Current Trends in Linguistics. Mouton, The Hague, Vol. **12**, pp.**2451–2487**, **1974**.

[7] FFmpeg Git, *"FFmpeg 4.0 "Wu""*, last accessed **2018-07-18**.

[8] Takanishi Lab Webpage, *"Anthropomorphic Talking Robot Waseda Talker Series"*, Retrieved from http://www.takanishi.mech.waseda.ac.jp/top/research/voice/index.htm, last accessed **2018-10-10.**

[9] Deepmind Webpage, *"WaveNet: A Generative Model for Raw Audio"*, Retrieved from https://deepmind.com/blog/wavenet-generative-model-raw-audio/, last accessed **2018-09-08**.

Authors Profile

Mr D Das is currently pursuing Bachelor of Engineering in the Department of Computer Science & Engineering, University Institute of Technology, The University of Burdwan, West Bengal, India. He is a member of IEEE. He has published more than 6 research papers in reputed international and national conferences including Springer, IETE etc. His main research work focuses on Embedded Systems, Natural Language Processing, Digital Signal Processing, Image Processing and Information Security based education.

Ms H Hassan is currently pursuing Bachelor of Engineering in the Department of Computer Science & Engineering, University Institute of Technology, The University of Burdwan, West Bengal, India. She is a blogger. She has published 2 research papers, one each in a reputed international conference and a national conference. She has also published 1 book in Amazon. Her main research work focuses on Natural Language Processing and Content Writing based education.

Mr S Gupta is currently working as Assistant Professor in the Department of Computer Science & Engineering, University Institute of Technology, The University of Burdwan, West Bengal, India. He has an Academic Experience of 6 years and a Research Experience of more than 3 years. He has published several research papers in reputed international and national conferences including IEEE, Springer, IETE etc. His main research work focuses on Data Mining, Natural Language Processing, Machine Learning and Information Security.

