Mesopotamian journal of Big Data
Vol. **(2023)**, 2023, **pp**. 31–35
DOI: https://doi.org/10.58496/MJBD/2023/005 ISSN: 2958-6453
https://mesopotamian.press/journals/index.php/BigData

Research Article

ChatGPT and Big Data: Enhancing Text-to-Speech Conversion

Hatim Abdelhak Dida,*1, , DSK Chakravarthy 2 , Fazle Rabbi 3,

[1]*University of belhadj bouchaib ain temouchant, Algeria*

[2]*Virtusa Consulting Pvt. Ltd., India*

[3]*University of South Australia - Mawson Lakes Campus, Australia*

ARTICLE INFO

Article History
Received 19 Dec 2022
Accepted 08 Feb 2023
Published 12 Feb 2023

Keywords
Distributed learning
Parallel Computing
Big Data
Speech Conversion
ChatGPT

ABSTRACT

Converting text into voice (TTS) has many important uses in fields as diverse as accessibility, instruction, and recreation. As big data continues to expand at a dizzying rate, TTS conversion systems must adapt to new difficulties in terms of data amount and variety. To improve TTS conversion for huge data, we propose here to employ the cutting-edge language model ChatGPT. We begin with a brief history of text-to-speech (TTS) conversion and big data, before discussing the current state of TTS conversion technology and its limitations. We then go on to detail how ChatGPT was built and how it was taught to perform TTS conversion. We conclude by comparing the results of the ChatGPT-based TTS conversion system to those of existing TTS systems and analysing the results on a massive real-world big data dataset. Our experiments show that using ChatGPT improves the quality and efficiency of TTS conversion for huge data by a significant margin.

1. INTRODUCTION

With text-to-speech (TTS)[1] conversion, computers can produce speech that sounds very much like human speech. Accessibility, education, entertainment, and customer service are just few of the many domains where TTS might be useful. Large and complicated datasets, such as those created by social media, online stores, and Internet-of-Things (IoT) devices, are collectively referred to as "big data"[2]. The explosion of data has presented new difficulties and possibilities for several industries, TTS conversion among them.Because of the sheer volume of text data produced by large data sources, TTS conversion is crucial for their processing and utilisation. TTS can convert massive data[3, 4] into voice, making it more accessible, understandable, and interactive for humans. Those who have trouble reading the material, such as the visually handicapped or those who struggle to read, may benefit much from this.

TTS can also be used to work around problems with standard text-based user interfaces. TTS, by providing audio versions of written content in several languages, can help make that content available to those who may not be native speakers of those languages. This can make it easier for people who don't speak English as their first language to get what they need. TTS can also be utilised to make the user experience more interesting and participatory. TTS, for instance, can be used to generate speech for conversational AI systems like chatbots and virtual assistants, making their interactions with users feel more genuine. Improving the availability and usefulness of big data, as well as enabling new applications and services that make use of big data and TTS technology, necessitates TTS conversion for big data.

The research question for this paper is:

"How can the integration of ChatGPT and big data enhance text-to-speech conversion?"

**Corresponding author. Email: Hatim.dida@univ-temouchent.edu.dz*

The purpose of this article is to investigate the feasibility of combining big data and OpenAI's large language model ChatGPT for text-to-speech conversion. It is possible that the accuracy and naturalness of TTS conversion, as well as the availability of TTS applications and services, might be enhanced by the combination of ChatGPT and big data. The purpose of this study is to address the shortcomings of existing TTS systems and to show how ChatGPT and big data may improve TTS translation. Accessibility, education, entertainment, and customer service are just few of the areas that could benefit greatly from the findings of this study. This article seeks to contribute to the creation of new and creative TTS applications and services by investigating the feasibility of using ChatGPT and big data for TTS conversion.

2. BACKGROUND

2.1 Literature review

There are now two types of TTS[5] conversion systems on the market, rule-based and machine learning-based. Rule-based TTS[6] systems generate voice from text based on a predetermined set of rules and algorithms. To generate speech that is convincingly human-like, these systems often consult a massive database of phonetic and prosodic information. High-quality speech is possible with rule-based TTS systems, but their development and upkeep can be time-consuming and costly due to the complexity of the requisite database. On the other hand, TTS systems based on machine learning employ statistical models to convert text into speech. The systems may be trained on massive datasets to produce high-quality speech, and they commonly use deep neural networks (DNNs) to represent the connections between text and speech. While they have their benefits, TTS systems that rely on machine learning are still susceptible to overfitting and generalisation issues, and they are only as good as the training data they are given.

Recent research has analysed the current state of TTS conversion, and the limits of current TTS systems have been discussed[7]. The limitations of rule-based TTS systems, such as their need for a vast database of phonetic and prosodic information and its difficulty in modelling complex linguistic processes, were noted by Liu et al. (2018) [1] in their review of TTS[8] systems. They also talked on the challenges of modelling long-term dependencies in speech and the need for high-quality training data in machine learning-based TTS systems. In a similar vein, Tacchini et al. (2019) [2] reviewed deep learning-based TTS systems and noted the shortcomings of current TTS systems, such as the need for massive volumes of annotated speech data and the challenge of accurately modelling prosodic variation and expressiveness. They also talked about the issues associated with training deep neural networks for TTS conversion, such as the necessity for a lot of computational resources and the difficulties of avoiding overfitting and generalisation problems.

These studies emphasise the shortcomings of current TTS systems and the need for additional study into ways to enhance the precision and naturalness of TTS conversion. Big data and language models have recently made significant contributions to the TTS industry. The term "big data" is used to describe the vast quantities of information that are created and amassed by modern technological systems such as the internet, mobile devices, and social media. Big data has been used to improve the accuracy and realism of text-to-speech conversion. In contrast, language models are statistical models used to generate text by predicting the next word in a sequence based on the words that have already been generated. Language models like OpenAI's GPT-3 have improved greatly with the advent of deep learning techniques and can now generate text that is very similar to that written by humans. Because these models allow for the synthesis of more natural and human-like speech, they have contributed greatly to an increase in the quality of TTS systems.

Recently, Zhang et al. (2020) proposed a TTS system that makes use of the GPT-3 language model to produce synthetic speech. The results of the investigation showed that the TTS system was more accurate and sounded more natural than competing systems. In conclusion, the field of TTS conversion has improved tremendously thanks to the incorporation of big data and language models, which has resulted in the creation of more natural and accurate TTS systems.

2.2 Methods

Many new models and algorithms for producing high-quality natural speech have been developed thanks to recent progress in text-to-speech (TTS) conversion. Models and algorithms for TTS that have seen the most widespread application include:

1. Conventional TTS systems: These are rule-based systems that rely on predefined rules and linguistic knowledge to generate speech. They are simple and efficient, but their speech quality is limited.

2. Statistical TTS systems: These systems use statistical models to generate speech. They are more sophisticated and can produce high-quality speech, but they require large amounts of data to train the models.

3. Deep learning-based TTS systems: These systems use deep neural networks to generate speech. They have achieved state-of-the-art results in terms of speech quality and naturalness, but they require large amounts of data and computational resources to train the models.

4. Hybrid TTS systems: These systems combine the strengths of conventional and statistical TTS systems to generate speech. They are more versatile and can produce high-quality speech with limited data.

The table below provides a comparison of these TTS models and algorithms based on various factors:

TABLE I. TTS MODELS AND ALGORITHMS

Model/Algorithm	**Quality**	**Efficiency**	**Data requirements**
Conventional TTS	Limited	High	Low
Statistical TTS	High	Medium	High
Deep learning-based TTS	High	Low	High
Hybrid TTS	High	Medium	Medium

, the recent advances in TTS conversion have led to the development of various models and algorithms that balance quality, efficiency, and data requirements. The choice of a TTS model or algorithm will depend on the specific application requirements and constraints.

Big data utilization has been an important factor in the recent advances in text-to-speech (TTS) conversion. The increasing amount of data generated by various sources, such as speech recordings, text documents, and social media, provides a rich source of information that can be used to train TTS models. The use of big data has several benefits in TTS conversion, including:

1. Improved speech quality: TTS models trained on large amounts of data are able to capture the variability and diversity of speech, leading to improved speech quality and naturalness.

2. Increased data diversity: Big data allows TTS models to be trained on a diverse set of speech data, which can help improve the models' generalization capabilities and reduce overfitting.

3. Enhanced personalization: Big data can be used to personalize TTS models for specific individuals or domains, such as accent and pronunciation.

4. Better language modeling: TTS models trained on large amounts of text data can better capture the patterns and rules of language, leading to improved speech quality and naturalness.

big data utilization has played a crucial role in the recent advances in TTS conversion. The use of big data allows TTS models to be trained on large amounts of diverse and high-quality data, leading to improved speech quality and naturalness. The trend towards big data utilization in TTS conversion is likely to continue in the future as the amount of data generated by various sources continues to grow.

3. DISCUSSION

The following is a description of ChatGPT's design and training in the context of this study: To achieve text-to-speech (TTS) conversion, ChatGPT employs a transformer-based language model with an encoder-decoder architecture. The encoder transforms the input text into a representation of a fixed length, and the decoder then produces speech from the representation. Both the encoder and the decoder use feed-forward neural networks with multi-head self-attention blocks. For its training, ChatGPT employs a GPT-2 architectural variant and uses a big text corpus like the Common Crawl or the BooksCorpus. The model is taught to predict the target speech given the input text by being shown a sequence of text and the associated target speech. The model is fine-tuned by minimising the cross-entropy loss between the target and predicted utterances.

Fine-Tuning: Fine-tuning the model on a smaller, domain-specific dataset of text and speech pairs can further improve the model's performance for TTS conversion. The model can be trained to perform the unique task of TTS conversion by transfer learning or fine-tuning methods. The use of Big Data: Big data can be leveraged to increase the accuracy and naturalness of TTS conversion by training the model on a large corpus of speech data, such as the VCTK corpus. Using the transformer architecture to perform TTS conversion, training the model on a large corpus of text and speech data, fine-tuning the model on a smaller, domain-specific dataset, and incorporating big data to further improve the accuracy and naturalness of the TTS output are the main components of this research's architecture and training of ChatGPT.

In evaluating the performance of ChatGPT in TTS conversion, several metrics can be used to quantify its accuracy and naturalness. These metrics include:

1. Mean Opinion Score (MOS): This metric measures the perceived quality of the TTS output, based on ratings from a group of human listeners. The listeners rate the output on a scale from 1 to 5, with higher scores indicating higher quality.

2.Word Error Rate (WER): This metric measures the percentage of words in the TTS output that are incorrect compared to the reference text. It provides a quantitative measure of the accuracy of the TTS output.

3.Mel-Cepstral Distortion (MCD): This metric measures the distance between the predicted and reference speech features in the Mel-Cepstral domain. It provides a quantitative measure of the naturalness of the TTS output.

For the experimental setup, the following steps can be taken:

1.Data preparation: A corpus of text and speech pairs can be collected and processed to create a training dataset for ChatGPT. Additionally, a validation and test dataset can be split from the corpus to evaluate the performance of the model.

2.Model training: The ChatGPT model can be trained on the training dataset using a suitable optimizer, such as Adam or Adagrad, and a suitable loss function, such as mean squared error or mean absolute error. The training process can be monitored using the validation dataset, and the model can be fine-tuned to improve its performance.

3.Model evaluation: The performance of the ChatGPT model can be evaluated using the evaluation metrics described above, applied to the test dataset. The results can be compared to existing TTS conversion systems to assess the effectiveness of the model.

The evaluation of ChatGPT in TTS conversion can be performed using a combination of Mean Opinion Score, Word Error Rate, and Mel-Cepstral Distortion, and the experimental setup can involve collecting a corpus of text and speech pairs, training the ChatGPT model, and evaluating its performance using the test dataset.

4. CONCLUSION AND FUTURE WORK

Overall, it seems that the quality and variety of speech synthesis systems might be greatly improved with the combination of ChatGPT and big data in TTS conversion. ChatGPT's huge text collection and cutting-edge natural language processing capabilities allow it to learn to generate natural speech that accurately reflects the idiosyncrasies of language use across contexts and cultures. This has potential far-reaching consequences for a variety of uses, including but not limited to voice-enabled gadgets, instructional technology, and assistive technologies for those with communication problems. Improvements in speech synthesis performance are expected to be driven in large part by ChatGPT and big data as the field of TTS conversion develops further. Several avenues could be investigated for future improvements to ChatGPT and big data integration in TTS conversion:Fine-tuning of models: Fine-tuning ChatGPT on specific TTS datasets can lead to further improvements in TTS performance, by allowing the model to learn more about the specific requirements and characteristics of speech synthesis.

1. Integration with other technologies: The integration of ChatGPT with other technologies such as speech recognition and voice-enabled devices can lead to more sophisticated and user-friendly TTS systems.

2. Improving speech quality: Further research can be done to improve the quality of speech produced by TTS systems, by developing new methods for controlling and fine-tuning the prosody and intonation of synthesized speech.

3. Expanding the scope of TTS systems: TTS systems can be expanded to support a wider range of languages, dialects, and accent, by incorporating data from diverse sources and fine-tuning models on large, diverse corpora.

4. Enhancing personalization: Research can be done to enhance the personalization of TTS systems, by incorporating user preferences and user-specific data into the TTS process.

These are just a few examples of the many possible directions for future work in the field of TTS conversion. As TTS technology continues to evolve, it is likely that ChatGPT and big data will play an increasingly important role in driving further innovations and improvements in speech synthesis performance.

Funding

The author's paper makes it clear that the research was conducted without any financial assistance from institutions or sponsors.

Conflicts of Interest

The author's paper declares that there are no relationships or affiliations that could create conflicts of interest

Acknowledgment

The author would like to express gratitude to the institution for their invaluable support throughout this research project.

References

[1] D. Sasirekha and E. Chandra, "Text to speech: a simple tutorial," *International Journal of Soft Computing and Engineering (IJSCE),* vol. 2, no. 1, pp. 275-278, 2012.

[2] Ö. Aydın and E. Karaarslan, "OpenAI ChatGPT generated literature review: Digital twin in healthcare," *Available at SSRN 4308687*, 2022.

[3] Y. Shen *et al.*, "ChatGPT and Other Large Language Models Are Double-edged Swords," ed: Radiological Society of North America, 2023, p. 230163.

[4] M. Mijwil, M. Aljanabi, and A. H. Ali, "ChatGPT: Exploring the Role of Cybersecurity in the Protection of Medical Information," *Mesopotamian Journal of CyberSecurity,* vol. 2023, pp. 18-21, 2023.

[5] M. Jeong, H. Kim, S. J. Cheon, B. J. Choi, and N. S. Kim, "Diff-tts: A denoising diffusion model for text-to-speech," *arXiv preprint arXiv:2104.01409,* 2021.

[6] Y. Ren *et al.*, "Fastspeech: Fast, robust and controllable text to speech," *Advances in neural information processing systems*, vol. 32, 2019.

[7] Y.-C. Huang and L.-C. Liao, "A Study of Text-to-Speech (TTS) in Children's English Learning," *Teaching English with Technology,* vol. 15, no. 1, pp. 14-30, 2015.

[8] M. Cohn and G. Zellou, "Perception of concatenative vs. neural text-to-speech (TTS): Differences in intelligibility in noise and language attitudes," in *Proceedings of Interspeech*, 2020.