Articulatory Phonetics GenAI with Natural Language Prompts

Zara Shaikh, Gnaneswary Green Narayanasamy,
Emaan Qureshi, Arham Ahmed, Gina Chiodaroli

Mentor: Dr Xiatian Zhu

January 2025

1 Introduction

Text-to-Speech (TTS) technology has undergone remarkable advancements in recent years, driven by the growing demand for accessible and user-friendly applications. TTS models are pivotal in bridging the gap between written and spoken language, fostering inclusivity in content and how it is interpreted by consumers. This literature review delves into sources on this topic, including the development of these systems, the applications of this technology - particularly in the education sector - and the challenges faced when trying to achieve natural speech synthesis. By carefully analysing these sources, this review aims to identify the most effective approach and optimal methods for developing a TTS system, whilst providing a clear understanding of the key components and processes involved in TTS technology. This will help guide the creation of a website that incorporates a fully operative AI to assist students and offers a safe, ethical, and inclusive way to implement this model for the user's desired purposes.

2 Examining the evolution of Speech Synthesis

Speech synthesis has profoundly reshaped human communication, empowering people with disabilities, and enabling new applications in accessibility and everyday life. 'From Hawking to Siri: The Evolution of Speech Synthesis' [1] describes the evolution of these systems, from early models like Kratzenstein's 1730 vowel synthesiser to groundbreaking innovations such as the Speech Plus CallText 5010. This source is particularly valuable for highlighting the societal impact of TTS technology and helping us understand the historical progression of speech synthesis. The article's credibility is supported by being published on Deepgram's website - a company specialising in TTS systems - reinforcing its relevance in this field.

Sanusi describes the fascination people have always had with objects that sound like humans, dating back to medieval legends of the Brazen Head - a magical mechanism able to answer "yes" and "no" to questions. This curiosity led to the first speech synthesiser model being developed by Christian Kratzenstein in 1730, which was able to produce five vowels using resonant tools and a reed. Wolfgang von Kempelen followed suit in 1791, creating the Acoustic Mechanical Speech Machine that inspired scientists like Charles Weathstone and Alexander Graham Bell to develop machines capable of producing vowels, consonants, and a limited number of words.

The article goes on to describe the rise of electronic speech synthesis in recent times, highlighting synthesisers such as the Speech Plus CallText 5010 - a groundbreaking TTS system created by Dennis Klatt - which enabled Stephen Hawking to communicate after losing his ability to speak due to motor neuron disease. Hawking praised the system, stating, "This system allowed me to communicate much

[1] From Hawking to Siri: The Evolution of Speech Synthesis [Internet]. Deepgram. Available from: https://deepgram.com/learn/evolution-of-speech-synthesis-tts

better than I could before ... I have written a book, and dozens of scientific papers ... I think that this is in large part due to the quality of the speech synthesiser." The article considers how this technology has supported and benefited millions of people with disabilities through integrations into technologies such as Siri, navigational systems, and real-time text-to-speech generation. Hawking's statement supports its use in communication and his praise emphasises the potential of electronic speech synthesisers.

Moreover, the article examines advancements in deep learning-based speech synthesis, such as Hidden Markov Models (HMM), WaveNet, and Tacotron 2, all of which map linguistic to acoustic features, overcoming earlier limitations to produce more natural, human-like speech. The article elaborates on HMMs and how they work: models that enable a TTS system to generate an optimal parameter sequence from subword HMMs. They are flexible with changing voice characteristics; though the acoustic features are sometimes over-smoothed, making the generated voice sound muffled. Deep learning synthesis uses neural networks trained on large amounts of labeled data which is used in both speech recognition and synthesis. The difference between deep learning and the HMM is that it can directly perform mapping from linguistic to acoustic features with deep neural networks, which have proven extraordinarily efficient at learning inherent features of data. The article references several models based on the Deep Learning approach such as the Restrictive Boltzmann Machines and End-To-End speech synthesis, providing a strong foundation for developing TTS systems from a deep learning based approach. It also outlines various applications of TTS models, from Siri to Speech Plus CallText 5010, offering inspiration and practical examples for implementing a customized TTS system.

Sanusi's article is an informative resource on the evolution of speech synthesisers, exploring their potential and gravitas on communication for millions of individuals with disabilities. It references the ethical importance of TTS systems in empowering voices otherwise silenced by physical limitations such as Hawking's Motor Neuron disease. While the material is made accessible to a broader audience by simplifying technical processes, it reduces its utility for us seeking a deeper understanding. We addressed this by referring to another source afterwards that provides a more in-depth look at the technical aspects behind the models. Despite its simplifications, the article offers valuable insights into the applications and advancements of TTS technology, making it a strong starting point for understanding the field and inspiring further exploration.

3 Analysing the core components and uses of TTS models

The development of TTS technology has transformed the way we interact with digital content, enabling applications in education, accessibility, and smart devices. 'Planning the development of TTS synthesis models and datasets with dynamic deep learning' [2] provides a detailed exploration of these systems, detailing their core components, synthesis techniques, applications, and future directions. By breaking down their fundamental processes, the paper has offered us a clear framework for understanding how TTS models function. Written by Hawraz Ahmad and Tarik Rashid, both established experts in software engineering and informatics with PhDs from the Salahaddin University-Erbil and University College Dublin, this source is a credible foundation for exploring the intricacies of TTS technology. However, while the source provides a broad overview of the field, several limitations, such as the limited focus on emerging techniques must be considered when applying its insights to our project.

Speech Synthesis is the process of converting text into speech that mimics the natural human voice. Central to this process are the core components of TTS systems: text processing, prosody generation, and speech waveform creation. Text processing transforms non-standard words such as numbers and abbreviations into spoken form using methods like rule-based approaches, which handle vast vocabularies, and dictionary-based approaches, which preserve phonological details for restricted vocabularies. Prosody generation further improves speech by adding features like intonation, stress,

[2] Ahmad HA, Rashid TA. Planning the development of text-to-speech synthesis models and datasets with dynamic deep learning. Journal of King Saud University - Computer and Information Sciences. 2024 Sep 1;36(7):102131–1.

and pitch; though it struggles with replicating our natural pitch variations. Lastly, speech waveform creation converts phonetic symbols into audio output.

The authors continue by describing the distinct approaches and advantages of the various synthesis techniques used to generate speech. Formant synthesis uses vocal tract rules to create artificial speech waves, focussing on pitch, voicing, and noise levels for synthetic speech. In contrast, concatenative synthesis combines pre-recorded speech segments and balances unit and database size. Statistic parametric models map speech parameters compactly, with methods such as Hidden Markov Models (HMM), allowing flexibility in voice generation. More recently, deep learning methods have revolutionised the field, utilising neural networks to achieve unparalleled naturalness and adaptability. Examples include WaveNet, which generates raw audio waveforms for realistic output, and Tacotron which assembles character-based and context-aware speech output.

The applications of TTS technology span a wide range of fields, from assistive technology to smart devices. In the case of assistive technology, TTS systems play a crucial role in aiding visually impaired individuals and improving language learning by teaching pronunciation and spelling. Intelligent assistants like Siri and Google Assistant and powered by TTS, enabling seamless interactions between users and technology. Looking ahead, the future of TTS lies in improving the naturalness of speech and context understanding, and developing support for underrepresented languages.

In summary, TTS technology has revolutionised accessibility, education, and smart devices by enabling natural speech synthesis. Ahmad and Rashid's source provides valuable direction for understanding core system components and deep learning methods, whilst also highlighting areas of improvement like support for diverse languages. These insights offer guidance to us for developing models and aligning them with emerging needs, ensuring inclusivity in a rapidly evolving digital landscape.

4 Exploring the world of open-source TTS models

As demand for more accessible voice solutions increases, advancements in TTS models have followed suit, leading to a variety of solutions catering to different needs. Sherlock Xu's article: 'Exploring the world of open-source TTS models' [3], provides a comprehensive overview of a selection of these open-source models, detailing their strengths, weaknesses, and commercial uses. This source was chosen to help guide our selection of the most suitable TTS model based on the purpose of our website, while also highlighting potential challenges. Xu's strong background in technology adds credibility to his writing, though the subjectivity of the evaluation and the lack of in-depth technical analysis limits its applicability for more technical decision making. Additionally, the fast-paced nature of TTS development means that the source may not fully account for the most recent advancements. Building on this overview, the following section delves into a more detailed analysis of the strengths and weaknesses of the five TTS models mentioned in the article and how they relate to our project.

XTTS-v2's multilingual capabilities and expressive speech synthesis make it a convincing option for an educational platform designed to help students. With support for 17 languages, this model is well-suited for accommodating students from diverse backgrounds. Furthermore, its ability to replicate emotion and speaking style with minimal input could enhance engagement for students listening to text-based materials. However, the model's non-commercial license limits its applicability for a public website, and with the project shut down, reliance on the open-source community may result in slow updates and a lack of long-term support. Another promising candidate is Chat TTS, which excels in high-quality speech synthesis and is optimised for dialogue tasks. This makes it a strong contender for interactive features such as summarising lengthy text and answering students' questions. With extensive training on 100,000 hours of data, Chat TTS ensures clear narration for PDFs

[3] Exploring the World of Open-Source Text-to-Speech Models. https://www.bentoml.com/blog/exploring-the-world-of-open-source-text-to-speech-models. Accessed 6 Jan. 2025

or other text-based materials. Unlike XTTS-v2, this model has a restriction on language support, only offering English and Chinese. This, plus stability issues common with autoregressive models, might limit its effectiveness for a platform aiming to support a diverse student population and handle lengthy materials seamlessly.

Melo TTS stands out for its broad language support, even offering a range of English dialects, such as British, American, and Australian. Its MIT license for commercial use is ideal for implementation on a public website without legal restrictions. OpenVoice v2 is also licensed under this. It requires minimal input to instantly clone the voice and tone of the speaker. This model's ability to control voice attributes like emotion and pauses make it a highly customisable option for creating engaging and personalised content for students. Compared to Melo TTS, OpenVoice supports fewer languages and may not sound as natural, which could affect usability for multilingual audiences. Another model which offers a similar granular voice style control via text prompts is Parler-TTS. The model's detailed control over voice characteristics and flexible model sizes make it a versatile option for an educational platform. The 'Mini' model is efficient for quick speech generation, while the 'Large' - which may be impractical for us to use due to the high computational resources requirement - allows for more expressive speech synthesis.

5 Investigating TTS applications as a resource for learning

TTS softwares have had remarkable breakthroughs and developments in recent years and their uses have varied in multitude, from Google Translate to Microsoft Applications. An article, written by Professor Reima Al-Jarf [4], a respected professor in ESL, linguistics, and translation studies affiliated with the University of Iowa, explores the integration of TTS software in undergraduate interpreting training and delves into its use in an educational setting and studies how beneficial it can be for students. The article, 'TTS Software as a resource for independent practice', was published in the International Journal of Translation and Interpretation Studies, helping solidify the validity of this article. The work proposes a model for using TTS tools to enhance students' independent practice, detailing its definition, advantages, and practical implementations. This article defines TTS models as converting written text into spoken words, offering flexibility, ease of use, and customizable features like voices, accents, and speeds which helps accrue the exact specifications a TTS model entails and provides a guideline on what the final-end product should be able to achieve.

The article explores the integration of the TTS models with interpretation training by students which entailed of instructional stages that include a pre-task phase where instructors introduce TTS tools, a task phase focused on independent practice in interpreting modes such as simultaneous and sight interpreting, and a post-task phase for reflection and peer discussion. The article emphasizes the instructor's role as a facilitator and highlights how TTS can support interpreting practice and be implemented into real-life applications such as in eBooks, Google Docs, and web pages. All possible implementations offer insight on and help to define the uses and practicality of the TTS software and all valid interpretations and uses of the TTS software can be masterfully taken inspiration from and implemented in our own TTS model. Limitations of the source are that it is limited and specific to one use and practicality of TTS models,as the interpretation practice used by students in this specific article may not be applicable to the use of TTS model if we are not making the same considerations in mind with students when designing and modelling our TTS. Moreover, TTS tools may not replicate the cultural and linguistic nuances critical for real-life interpreting scenarios.

Nonetheless, the article provides practical guidance and insight on adapting TTS technologies to different contexts, helping refine and evaluate TTS model development for our own project and influence us in the purpose and use of the TTS technology.

[4] Al-Jarf R. Text-to-Speech Software as a Resource for Independent Interpreting Practice by Undergraduate Interpreting Students. International Journal of Translation and Interpretation Studies. 2022 Aug 14;2(2):32–9.

6 Considering the ethical implications of TTS models

As the use of artificial intelligence has become more widespread and controversial, it is important to consider the ethical implications of creating and developing text-to-speech systems. These systems involve generative AI, a key aspect of which includes the emotive responses they elicit. When considering the ethical implications of TTS models, it is crucial to address their potential for misuse, such as in the creation of deep fakes or the propagation of false information. By evaluating and reviewing the unethical aspects of TTS models, we can reduce the likelihood of incorporating these risks into the final project and adopt approaches free from negative stigma.

The article "AI Ethics is Everywhere: It's Time to Pay Attention" by Assaf Asbag [5]explores the ethical challenges and biases that often emerge in AI development. With over 15 years of experience in technology and data sciences, Asbag is a credible authority in the field. His expertise is evident in his clear articulation of the biases that may arise during the creation of TTS systems. The article highlights key principles for ethical AI development—fairness, safety, transparency, accountability, and respect for privacy.

One of the issues Asbag emphasizes is fairness, particularly the lack of voice synthesis options that incorporate a wide range of accents, tones, and speaking styles. This exclusion marginalizes underrepresented groups, limiting their access to materials tailored to their needs. Moreover, a lack of transparency poses another major ethical dilemma for TTS. There could be instances in which it is unclear whether an AI or a human is communicating. Asbag argues that this lack of clarity is deeply detrimental, as it "erodes trust."

The article confirms the importance of AI ethics in ensuring that technology benefits humanity while minimising harm. For example, Amazon's recruitment tool, which was biased against women due to flawed training data, and the ElevenLabs voice cloning incident highlight the critical need for safeguards like identity verification. Asbag also points to a notable example of a "fabricated audio clip of U.S. President Joe Biden making controversial statements." These real-world examples underscore the importance of incorporating ethics into AI development to prevent harm and build trust.

While the source explores a wide range of factors contributing to the unethical aspects of TTS, it does not delve deeply enough into how these issues can be practically addressed. Asbag suggests incorporating ethics into development processes through measures like testing for bias or adding explainability. However, these suggestions lack detailed guidance on implementation, which could be a limitation for those new to TTS development, such as our team.

As a whole, Asbag's article is a valuable resource for understanding the ethical implications of TTS software. It effectively communicates the disparities that could arise during development and explains why these issues are unethical. When implementing TTS software, we must consider potential exploitation of accents and cultural appropriation, avoid misrepresentation, and detect harmful content to prevent the creation of deep fakes or the spread of misinformation. Overall, the source provides meaningful insights to help developers identify and mitigate unethical factors that may be unknowingly built into their AI models while considering potential challenges.

7 Conclusion

Our exploration in this literature review has provided a comprehensive foundation for understanding the history, development, and uses of TTS technology. By examining varied sources, this review highlighted the wide range of applications and the potential challenges we could face during this project. Historical insights from Sanusi's work enriched our understanding of how TTS systems have evolved

[5] Assaf Asbag. AI Ethics is Everywhere: It's Time to Pay Attention - aiOla [Internet]. aiOla. 2025 [cited 2025 Jan 9]. Available from: https://aiola.com/the-tech-talk/ai-ethics-and-safety/

to meet societal demands. Additionally, detailed analysis of open-source models by Xu guided our selection of a model that prioritised accessibility and naturalness to align with the website's educational objectives. Several sources reviewed various TTS architectures such as the Hidden Markov model and the Tacotron, providing us with a strong technical base to develop and model our own TTS off and techniques to integrate. Al Jarf's article explored the multitude of TTS implementations, underscored the value of students using TTS models for learning practice, guiding our project's focus objective to be centered around students and to be implemented as an educational tool allowing students to upload documents and have them read aloud.

After thorough discussion about the source and direction from our mentor, we decided to explore MetaVoice for our project. MetaVoice is a 1.2B parameter base model trained on 100,000 hours of data. With zero-shot cloning, the model can generate high quality speech from 30 seconds of reference audio while mimicking the speaking style and unique characteristics of the voice. MetaVoice is released under the Apache 2.0 license, allowing for straightforward integration into the website. While this model was not discussed in Xu's article, the detailed exploration of the features and limitations of the ones mentioned helped us define criteria for selecting a model that meets the goals of our website.

Through the culmination of sources and thorough evaluation discussed in this review, we extracted the most valuable insights and will apply them to the implementation of our final text-to-speech model.

8 Bibliography

1. From Hawking to Siri: The Evolution of Speech Synthesis [Internet]. Deepgram. Available from: https://deepgram.com/learn/evolution-of-speech-synthesis-tts

2. Ahmad HA, Rashid TA. Planning the development of text-to-speech synthesis models and datasets with dynamic deep learning. Journal of King Saud University - Computer and Information Sciences. 2024 Sep 1;36(7):102131–1.

3. Exploring the World of Open-Source Text-to-Speech Models. https://www. bentoml.com/blog/exploring-the-world-of-open-source-text-to-speech-models. Accessed 6 Jan. 2025.

4. Al-Jarf R. Text-to-Speech Software as a Resource for Independent Interpreting Practice by Undergraduate Interpreting Students. International Journal of Translation and Interpretation Studies. 2022 Aug 14;2(2):32–9.

5. Assaf Asbag. (2025) AI Ethics is Everywhere: It's Time to Pay Attention. aiOla. https://aiola .com/the-tech-talk/ai-ethics-and-safety/