ISSN: 2634 - 8853

Journal of Engineering and Applied Sciences Technology

SCIENTIFIC Research and Community

Research Article

Open Access

Text-To-Speech in Voice Assistants: Challenges and Mitigation Strategies

Ashlesha Vishnu Kadam

Amazon.com, LLC, Amazon Music, City Seattle, State WA, USA

ABSTRACT

Voice Assistants (VAs) have grown rapidly from technological novelties to integral parts of our daily lives to perform tasks like streaming music or news, setting alarms or responding to questions. These virtual conversational agents rely on an intricate combination of technologies, and one of the pivotal components is Text-to-Speech (TTS) synthesis. In this paper, we delve into the technical intricacies of TTS in voice assistants, addressing challenges, solutions, and future directions. VAs like Alexa, Siri and Google Assistant have transformed human-computer interactions. The underpinning TTS technology is crucial for converting text-based information into spoken language, making the interaction more natural and accessible. The synthesis of human-like speech from textual data is a complex and interdisciplinary domain, encompassing fields such as speech signal processing, natural language processing, deep learning, and linguistics. This paper aims to contribute a detailed analysis of TTS in voice assistants, emphasizing not only the theoretical aspects but also the practical implementation and real-world implications. The paper will examine the challenges associated with TTS, considering its technical, linguistic, and user-centric dimensions. The paper will also present mitigation strategies for these challenges. In a world where voice-driven interactions are becoming commonplace, a deep understanding of TTS is vital. By delving into the depths of this technology, we can unlock its full potential and ensure that voice assistants continue to enrich our lives and technical domains.

***Corresponding author**

Ashlesha Vishnu Kadam, Amazon.com, LLC, Amazon Music, City Seattle, State WA, USA.

Received: December 12, 2022; **Accepted:** January 16, 2023; **Published:** February 20, 2023

Keywords: Artificial Intelligence, Automatic Speech Recognition, Natural Language Understanding, Text-to-Speech, Voice Technology

Introduction

Voice Assistants (VAs) have become household names. Voice Assistants (VAs) are an application of Artificial Intelligence (AI) and Natural Language Processing (NLP) that recognize and understand human speech and respond to it in a way that is understandable by humans. Voice lends a natural and intuitive interface for humans to interact with technology. This coupled with the ease with which voice assistants are accessible on mobile phones (e.g., Siri on iPhone) and smart speakers (e.g., Alexa on Echo devices) has helped with the adoption of VAs [1]. A critical part of the voice interaction is TTS, or text-to-speech, i.e., the response of a conversational agent to the user in a verbal manner. TTS is a research topic sits at the intersection of technology, linguistics and artificial intelligence. Over time, there have been several technological and linguistic advancements in the TTS space. For example, the application of deep learning (DL) has led to transformative models like WaveNet, Tacotron and GPT-3 to enhance the naturalness of the TTS [2,3]. These models leverage neural networks and massive datasets to generate speech that approaches human-like quality [4]. End-to-end TTS models, such as Tacotron 2 and FastSpeech, streamline the process by directly mapping text to speech waveforms, minimizing intermediate steps and enhancing efficiency. Linguistic advancements include code-switching models (i.e., switching between 2 or more languages) and ability to handle multi-lingual and cross-lingual TTS. However, deeper challenges exist in TTS handling in voice assistants, as discussed in subsequent sections in this paper.

End-To-End Working of Voice Assistant

Before diving into ASR-specific challenges, it is important to review the end-to-end working of voice assistants. VAs can be activated using their specific wake words. For example, users of Amazon's VA say "Alexa" while those of Apple's VA say "Hey Siri" to invoke the respective VA. Post this invocation, the next task is translating the speech of the user to text tokens, i.e. the Automatic Speech Recognition (ASR) stage that does speech-to-text (STT) [5]. The next stage is taking the output of the ASR stage, i.e., a string of tokens, and parsing them in order to understand the syntactic and semantic interpretation of this text string. This stage is called Natural Language Understanding (NLU), and the outcome is an understanding of the intention of the user [6]. Once the VA understands what the user is looking for, it can use multiple systems in the back-end to retrieve the right response to this query, including but not limited to the internet, a cloud platform connecting to specific servers in the back-end, an application, and more. When the response is retrieved, the VA again converts the response back into a format that is understandable to the user, like text to speech (TTS) of a response, text displayed in case of a multi-modal interface, or simply the desired action being executed (e.g., switching off the lights). See Figure 1 for an overview of how a hypothetical voice assistant, Nova's, end to end functionality could be.

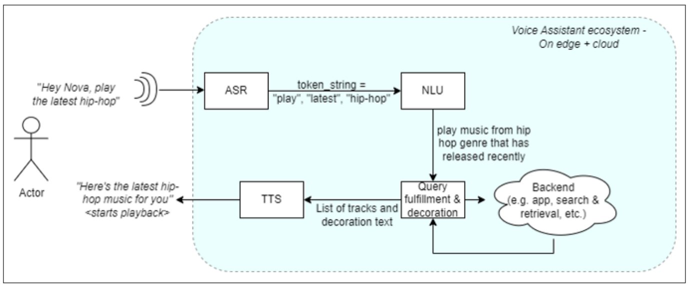
Voice Assistant ecosystem - On edge + cloud
"Hey Nova, play the latest hip-hop"
ASR
token_string = "play", "latest", "hip-hop"
NLU
play music from hip hop genre that has released recently
Actor
"Here's the latest hip-hop music for you" <starts playback>
TTS
List of tracks and decoration text
Query fulfillment & decoration
Backend (e.g. app, search & retrieval, etc.)

Figure 1: End to end overview of how a hypothetical voice assistant Nova might look like [6]

Challenges in Text-To-Speech

This section focuses on the key challenges when it comes to TTS on voice assistants.

Making Speech Sound Natural

A key challenge in TTS technology of voice assistants is the ability to make the speech that is produced sound natural and human-like. This includes appropriate intonation of the speech, emphasis on the right parts of speech, imperfections like pauses and "filler" expressions like "umm", "ah", and emotion in how the speech sounds. This becomes important because it can influence and affect the listener's ability to understand, be interested in and engage with the content that the voice assistant is presenting. There are several factors that make the sound of the generated TTS appear to be not as natural or human-like as one would desire. Top challenges are (1) Lack of or incorrect prosody – Prosody refers to the natural rhythm, intonation, stress and other features of a speech that result in the speech sounding natural. An example is when someone speaks something at a rapid pace, which could indicate excitement or hurry, but the same words at a slow pace could indicate boredom or sadness. Another example is how the intention behind saying something can completely change depending on whether a sentence is said with a cheerful tone or a sarcastic tone. (2) Poor pronunciation, especially of rare words – While it is challenging for the TTS to always pronounce words the right way, the problem is exacerbated in case of rare or unusual words like proper nouns of non-popular places (e.g., Ljubljana), or quirky names of brands or artists (e.g., the artist Chvrches) [7]. (3) Lack of emotional expression – TTS systems like the ones in voice assistants can have difficulty expressing emotions in the generated speech [8]. Most TTS systems do not convey the omnipresent emotional contexts that exist in human-to-human interactions.

Adapting to a Specific Speaker's Voice

While TTS systems can produce high quality speech that is synthesized, the similarity of this speech to the speaker is lower compared to the scenario of a single speaker model [4]. There are several challenges when it comes to the voice assistant adapting to every specific speaker, (1) Inadequate Training Data – TTS training requires a large corpus of data during the training stage to build a baseline of TTS capability. However, adapting to a specific speaker's voice needs an even higher amount of data than needed for baseline functionality if the TTS system is expected to adapt to a speaker's voice [6]. (2) Variability in Speakers – Individuals have different vocal characteristics, like their tone, accent, pitch, volume, and other speech idiosyncrasies. These variables make it challenging for the TTS system to adapt to every individual's voice [9]. (3) Tendency to Overfit – Overtraining a TTS system based on training data from specific speaker(s) can result in overfitting, where the TTS system becomes too specialized to the training data and performs poorly on new data, i.e., a new speaker's voice or new or rare vocabulary [10]. (4) Challenges in Fine-tuning – Being able to fine-tune a TTS system to a specific speaker's voice can be challenging because it requires adjusting the parameters of an initial model. This process can be time-consuming and requires expertise in speech synthesis [4].

Handling Multilingual Speech

In order to perform well when it comes to multi-lingual input, TTS models need to be trained on a diverse corpus of speech samples from various languages. There are several challenges when it comes to a voice assistant handling multi-lingual speech deftly (1) Inadequate quantity and quality of language resources for certain languages (e.g., Pashto, Swahili) are limited compared to those for popular languages (e.g., English, Spanish), making it hard to find enough data to train a TTS model adequately to handle multi-lingual scenarios. Translation of or to these languages also becomes challenging due to the same reason of paucity of training data. (2) Variations in language introduced due to various dialects, accents and regional adaptations, further add challenges for voice assistants [8]. (3) Language specific text normalization can be challenging. Text normalization is the process of converting written text into a phonetic representation that can be used by TTS systems. Different languages have different text normalization rules, which can affect the naturalness and intelligibility of synthesized speech [11].

Maintaining Low Latency

When it comes to TTS, especially on a voice assistant where users don't have a high tolerance for latencies, it is critical to have the TTS system operate with low latency. High latency can be caused by hardware and software speed, network transmission times and traffic [12]. Speaker variability can also affect the latency of TTS systems [13]. Code-switching, which is the practice of alternating between two or more languages or language varieties in a single conversation, can also introduce challenges in terms of being able to respond in an adaptive, code-switching, way.

Mitigation Strategies

The mitigation strategies mentioned below can help overcome the challenges mentioned in the previous section.

Making Speech Sound Natural

In order to mitigate the challenges mentioned previously about being able to make the voice assistant's TTS sound more natural and human like, there are several techniques can be implemented. To improve prosody in TTS, prosody modeling using deep neural networks and reinforcement learning techniques can enhance speech expressiveness. These algorithms can learn from large amounts of data and improve their accuracy over time. For example, Deep Neural Networks (DNNs) can model the relationship between the input text and the corresponding speech waveform [14]. Convolutional Neural Networks (CNNs) can extract features from the input text and convert them into a spectrogram, which is then used to generate the speech waveform. They can improve the naturalness of the speech by capturing the nuances of human speech [15]. Recurrent Neural Networks (RNNs) can model temporal dependencies in the input text and generate the corresponding speech waveform. RNNs can improve the naturalness of the speech by capturing the rhythm and intonation of human speech [16]. Generative Adversarial Networks (GANs) can generate speech that is more natural and human-like by training a generator network to create speech that is indistinguishable from real speech, and a discriminator network to distinguish between real and generated speech [17].

To correct pronunciation of rare words, Grapheme-to-Phoneme (G2P) models can be used. A grapheme is a written symbol that represents a sound, while a phoneme is the smallest unit of sound in a language. For example, in the word “cat”, the /k/ sound represented by “c” is a phoneme, while the letter “a” representing the sound /æ/ is a grapheme. The G2P models convert strings of graphemes to corresponding sequences of phonetic transcription. These models can convert out-of-vocabulary words (OOV), e.g., proper nouns, colloquial words, cultural expressions as well as heteronyms in their phonetic form to improve the quality of the synthesized text. G2P systems allow users to enforce the desired pronunciation by providing a phonetic transcript of the input. Instead of models, rule based G2P conversion is another alternative, provided linguistic specialists have created these conversion rules [18].

To enhance expression of emotion, multiple techniques can be applied. Expressive visual text-to-speech, that converts text to emotionally expressive speech to improve emotion processing ability and social communication skills in individuals with autism spectrum conditions, can also be deployed to bring expressiveness of emotion in voice assistants [19]. Controllable expressive TTS involves developing a deep learning-based TTS system that can be controlled to produce speech with different emotions [20]. Finally, fine-grained style transfer for expressive TTS synthesizes emotional speech with fine-grained styles such as sarcasm, irony, and politeness [21].

Adapting to a Specific Speaker’s Voice

To overcome the lack of training data, transfer learning can be used. Transfer learning involves pre-training a TTS model on a large dataset and then fine-tuning it on a smaller dataset of the target speaker's voice [22]. To address variability in speakers, the TTS system can leverage speaker embeddings, which are learned representations of the speaker's voice that can be used to adapt TTS systems to a specific speaker's voice [23]. To counter tendency to overfit the performance to a specific set of speakers, regularization techniques such as dropout and weight decay to prevent overfitting during training can be deployed [24]. To make fine-tuning easier, transfer learning can be leveraged, where a model is pre-trained on a large dataset and then the fine tuning is done on a smaller dataset of the target speakers’ voice.

Handling Multilingual Speech

Integrating contextual machine translation models tailored to the voice assistant's domain or session context ensures high-quality, contextually relevant translations. For example, a voice assistant used for entertainment can respond accurately to queries about music genres, instruments, cultural references, and more. Each language exhibits unique linguistic nuances, phrases, idioms, and cultural references. Multilingual pre-trained models, such as BERT and XLM-R, offer an efficient way to capture these language-specific intricacies. These models can be fine-tuned to a specific domain, enhancing their accuracy. For instance, a model fine-tuned on music-related content can provide rich responses to music-related queries, surpassing mere word-to-word translations. Users frequently switch between languages during interactions with voice assistants, posing a significant challenge. For regions with language code-switching, specialized models can be employed to understand and respond to mixed-language input. For instance, a voice assistant designed for a bilingual Canadian region can effectively handle queries in both French and English, seamlessly transitioning between the two languages.

Maintaining Low Latency

To address speaker variability, speaker-adaptive TTS systems can be used that can adapt to the speaker's voice characteristics and reduce the latency of the system [13]. To address code-switching challenges, code-switching TTS systems that can generate speech in multiple languages and recognize code-switching instances. Finally, to address speaker variability challenges, speaker-adaptive TTS systems can adapt to the speaker's voice characteristics and reduce the latency of the system.

Future Direction

As TTS becomes increasingly prevalent as voice assistant adoption continues growing, the field of TTS is ripe with opportunities for future research. Prominent areas are captured below.

Multimodal Integration

As TTS gets more and more integrated into modalities like natural language understanding (NLU), computer vision (CV) and others, it can benefit from being more context aware and utilizing the non-verbal inputs coming in from other modalities.

Emotion Synthesis

Voice assistants often sound robotic to users. Making the TTS more emotion-aware can help the voice assistant convey a broad spectrum of emotions.

Customization and Adaptation

Personalizing TTS voices and interaction styles to match individual user preferences is an area with substantial potential. Future research can explore techniques for real-time voice adaptation, enabling voice assistants to adapt their speaking style, pitch, and pacing to better suit individual users.

Applications Beyond Voice Assistants

TTS technology has applications in areas like education, accessibility, healthcare, and entertainment. Future research can explore the adaptation of TTS for these domains, leading to innovative and inclusive solutions.

Conclusion

In this paper, an overview of how voice assistants work was provided, along with a deep dive specifically into the challenges of TTS. The paper also covers mitigation strategies for these challenges and suggests future research topics that can help in improving TTS technology. Further research is needed to evolve the last mile of customer experience on voice assistants, i.e., to improve TTS.

References

1. Hoy MB (2018) Alexa, Siri, Cortana, and more: an introduction to voice assistants. Med Ref Serv Quart 37: 81-8.
2. Van den Oord A, Dieleman S, Zen H, Simonyan K, Vinyals O, et al. (2016) Wavenet: A generative model for raw audio. Computer Science https://arxiv.org/abs/1609.03499.
3. Wang Y, Skerry-Ryan R J, Stanton D, Wu Y, Weiss R J, et al. (2017) Tacotron: Towards end-to-end speech synthesis. Computer Science https://arxiv.org/abs/1703.10135.
4. Brown T B, Mann B, Ryder N, Subbiah M, Kaplan J, et al. (2020) Language models are few-shot learners. Computer Science https://arxiv.org/abs/2005.14165.

5. Luiza Stelitano (2021) A quick and easy guide to voice assistants. Artificial Intelligence https://www.miquido.com/blog/what-are-voice-assistants/.
6. Hieu-Thi Luong, Junichi Yamagishi (2019) A Unified Speaker Adaptation Method for Speech Synthesis using Transcribed and Untranscribed Speech with Backpropagation https://arxiv.org/pdf/1906.07414.pdf.
7. Felix Burkhardt, Uwe D Reichel (2016) A Taxonomy of Specific Problem Classes in Text-to-Speech Synthesis: Comparing Commercial and Open Source Performance. Telekom Innovation Laboratories, Research Institute for Linguistics, Hungarian Academy of Sciences, Hungary 744-749.
8. Kuligowska Karolina, Kisielewicz Paweł, Włodarz Aleksandra (2018) Speech synthesis systems: Disadvantages and limitations. International Journal of Engineering and Technology (UAE) 7: 234-239.
9. Cao Beiming, Alan Wisler, Jun Wang (2022) Speaker Adaptation on Articulation and Acoustics for Articulation-to-Speech Synthesis. Sensors 22: 6056.
10. Bastian Schnell, Philip N Garner (2022) Investigating a neural all pass warp in modern TTS applications. Speech Communication 138: 26-37.
11. Benoy Kurian, Puthiyidam Jiby (2020) A study of Text to Speech systems for Non-English Languages. ResearchGate https://www.researchgate.net/publication/339089496_A_study_of_Text_to_Speech_systems_for_Non-English_Languages.
12. Sunil Kumar Jang Bahadur (2022) Solving Automatic Speech Recognition Deployment Challenges https://developer.nvidia.com/blog/solving-automatic-speech-recognition-deployment-challenges/.
13. What is latency? | How to fix latency. CloudFlare https://www.cloudflare.com/learning/performance/glossary/what-is-latency/.
14. Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, et al. (2012) Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. IEEE Signal processing magazine 29: 82-97.
15. Aaron Brown (2021) Text to Speech-Lifelike Speech Synthesis Demo (Part 1) https://towardsdatascience.com/text-to-speech-lifelike-speech-synthesis-demo-part-1-f991ffe9e41e.
16. Hojjat Salehinejad, Sharan Sankar, Joseph Barfett, Errol Colak, Shahrokh Valaee (2017) Recent advances in recurrent neural networks. Neural and Evolutionary Computing https://arxiv.org/abs/1801.01078.
17. Jungil Kong, Jaehyeon Kim, Jaekyoung Bae (2020) Hifi-gan: Generative adversarial networks for efficient and high-fidelity speech synthesis. Advances in Neural Information Processing Systems 33: 17022-17033.
18. Kłosowski Piotr (2022) "A Rule-Based Grapheme-to-Phoneme Conversion System" Applied Sciences 12: 2758.
19. Cassidy SA, Stenger B, Van Dongen L, Yanagisawa K, Anderson R, et al. (2016) Expressive visual text-to-speech as an assistive technology for individuals with autism spectrum conditions. Comput Vis Image Underst 148: 193-200.
20. Tits Noé, Kevin El Haddad, Thierry Dutoit (2021) Analysis and Assessment of Controllability of an Expressive Deep Learning-Based TTS System. Informatics 8: 84.
21. RJ Skerry-Ryan, Eric Battenberg, Ying Xiao, Yuxuan Wang, Daisy Stanton, (2018) Towards end-to-end prosody transfer for expressive speech synthesis with tacotron. in ICML 4693-4702.
22. Paarth Neekhara, Jason Li, Boris Ginsburg (2022) Adapting TTS models For New Speakers using Transfer Learning. Sound https://arxiv.org/abs/2110.05798.
23. Mengzhe Geng, Xurong Xie, Zi Ye, Tianzi Wang, Guinan Li, et al. (2015) Speaker Adaptation Using Spectro-Temporal Deep Features for Dysarthric and Elderly Speech Recognition. Journal of Latex Class Files 14.
24. Tal Rosenwein. Overcoming Automatic Speech Recognition Challenges: The Next Frontier https://towardsdatascience.com/overcoming-automatic-speech-recognition-challenges-the-next-frontier-e26c31d643cc.