ISSN: 2634 - 8853

Journal of Engineering and Applied Sciences Technology

SCIENTIFIC Research and Community

Research Article

Open Access

Towards Intelligent Aviation: The Integration of Voice Assistants in Autonomous Aircraft

Ramona Devi

USA

ABSTRACT

This paper explores the integration of a virtual assistant into an autonomous aircraft, focusing on the utilization of cutting-edge technologies for automatic speech recognition, speech synthesis, and natural language processing. By leveraging the power of Edge AI, particularly NVIDIA Jetson devices, we demonstrate how this amalgamation of technologies creates a lightweight, mobile, and efficient voice assistant system. We introduce a novel approach to enhance the operational capabilities of autonomous aircraft, facilitating human-machine interaction and enhancing overall user experience. This research paves the way for the application of edge-based voice assistants in autonomous vehicles, offering a glimpse into the future of intelligent and responsive in-flight systems.

***Corresponding author**

Ramona Devi, USA.

Received: December 07, 2022; **Accepted:** December 16, 2022; **Published:** December 21, 2022

Keywords: ASR, Speech Synthesis, Llm, Voice Assistant

Introduction

The advent of autonomous vehicles has revolutionized the way we perceive transportation, with applications extending beyond the conventional terrestrial landscape. Unmanned aerial vehicles (UAVs) and autonomous aircraft are poised to redefine the future of air travel. However, the operational dynamics of these aircraft are intricate, necessitating advanced technologies for seamless control and interaction. This paper addresses the need for an intuitive, efficient, and versatile human-machine interface within the realm of autonomous aircraft, presenting the concept of an edge-based voice assistant as a solution.

With the proliferation of artificial intelligence (AI) and natural language processing (NLP), virtual assistants have become ubiquitous in our daily lives. The interaction between humans and machines through voice commands has grown increasingly important, offering a seamless and user-friendly way to control and manage a wide array of devices and systems. Drawing inspiration from this trend, we extend the application of virtual assistants to autonomous aircraft, thereby enhancing their functionality and accessibility.

The approach in this paper hinges on three key pillars of technology: automatic speech recognition (ASR), speech synthesis, and a large language model, LLAMA 2, for natural language processing. ASR technology converts spoken language into text, enabling the aircraft's system to understand and process user commands. Speech synthesis, on the other hand, transforms textual responses into natural-sounding speech, facilitating communication between the aircraft and its passengers or operators. LLAMA 2, a state-of-the-art large language model, serves as the backbone of our voice assistant, ensuring sophisticated natural language understanding and generation.

What sets this approach apart is the integration of these technologies into edge devices like NVIDIA Jetson, known for their robust AI capabilities and compact form factors. This not only reduces the latency in communication but also ensures that the entire system is lightweight and mobile, critical factors in the context of autonomous aircraft.

The overarching objective of this paper is to propose an edge-based voice assistant for autonomous aircraft, catering to the unique challenges and requirements of the aviation industry. It aims to provide a foundation for seamless human-machine interaction within the autonomous aircraft environment, thus enhancing safety, usability, and the overall passenger experience. Furthermore, our findings hold the promise of extending these edge-based voice assistant applications to a wider array of autonomous vehicles, ushering in a new era of intelligent and responsive transportation systems.

Speech Pipeline

Integrating speech recognition, speech synthesis, and natural language processing into a unified pipeline is instrumental in crafting an efficient and versatile voice assistant. Speech pipeline can be represented as Figure 1. This comprehensive pipeline begins with automatic speech recognition (ASR) technology, which transcribes spoken language into machine-readable text. ASR lays the foundation for understanding user input, making it a critical component in the initial interaction. Once the speech is transcribed, the natural language processing (NLP) component analyzes the text, extracting meaning, intent, and context from the user's words. It enables the system to comprehend the user's requests, even in complex, context-dependent situations. Furthermore, NLP ensures the system's ability to generate meaningful responses, as it takes into account the semantics and grammar of the language. The output goes into the control allocation block which converst this

response into the actuators action which in turn changes the state of the aircraft. The sensors sense this new state and feedback to the control system. This response is also transformed into speech by speech synthesis service.

In the final stage, speech synthesis transforms the system's text-based responses back into natural-sounding speech, making the voice assistant's interactions with the user more intuitive and human-like. The integration of these components results in a seamless, bidirectional flow of information: from the user's spoken input through ASR and NLP for understanding and interpreting, to the speech synthesis for generating articulate responses. This complete speech pipeline not only facilitates efficient human-machine communication but also ensures that voice assistants can provide valuable and context-aware assistance in a variety of applications, ranging from virtual call centers to autonomous vehicles, thus enhancing user experience and accessibility.

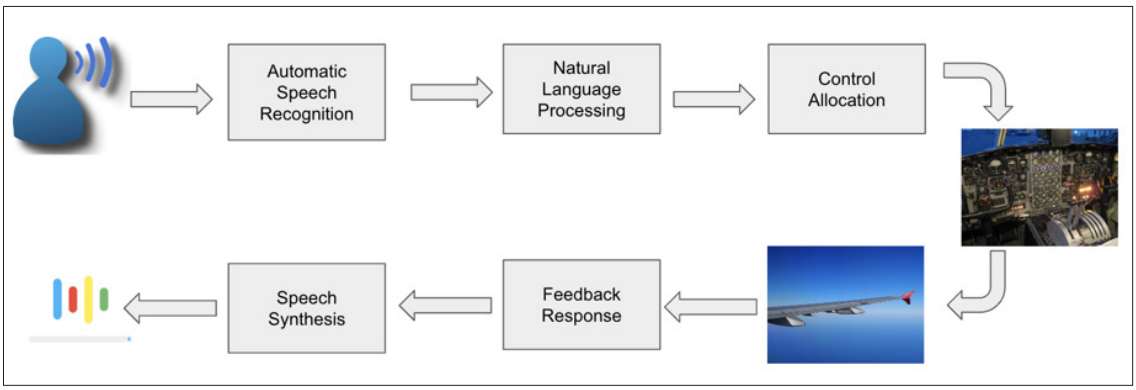

Figure 1: Speech Pipeline for Voice Assistant

ASR

The Automatic Speech Recognition (ASR) pipeline is a critical component of modern speech processing systems, designed to convert spoken language into machine-readable text [1-5]. This complex process typically involves multiple stages to accurately transcribe audio input. First, the audio signal undergoes pre-processing, which includes tasks like noise reduction, filtering, and feature extraction to make it suitable for analysis. Following this, acoustic modeling comes into play, where the system recognizes phonetic and acoustic patterns in the audio data. Language modeling then refines the output, considering the likelihood of word sequences. Many ASR pipelines further incorporate post-processing steps to correct errors and enhance transcription quality. The accuracy and performance of ASR systems have improved significantly in recent years, thanks to advancements in deep learning and the availability of large, high-quality speech datasets. The ASR functional pipeline is shown in Figure 2.

Open AI's Whisper model is a prominent example of a state-of-the-art ASR system. Built upon a deep neural network architecture, Whisper is designed to provide high-quality, versatile speech recognition capabilities. It has been trained on an extensive and diverse dataset containing 680,000 hours of multilingual and multitask supervised data. This vast training corpus ensures that the model can effectively handle various languages and domains. Whisper can be fine-tuned for specific applications, making it a versatile choice for industries ranging from transcription services to voice assistants and more. Its exceptional accuracy and robustness have made it a valuable tool for organizations seeking to harness the power of speech recognition in their applications, enabling more accessible and efficient interactions between humans and machines. Whisper is a testament to the continual advancements in ASR technology, opening up new possibilities for natural language processing and voice interaction across diverse domains.

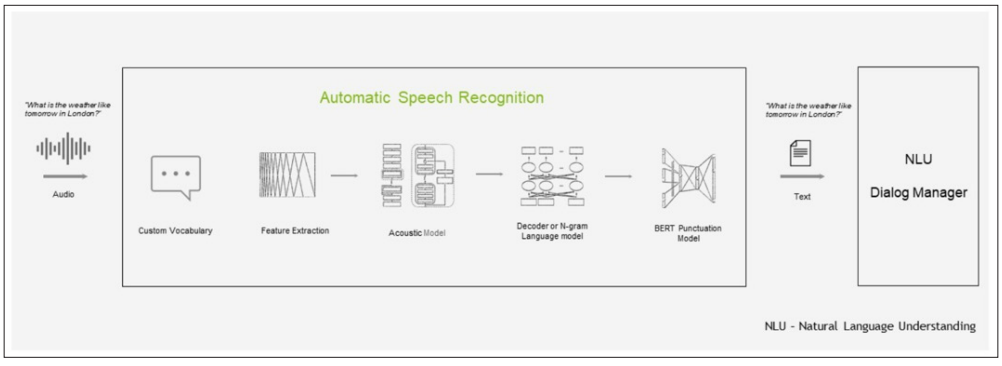

Figure 2: ASR Pipeline

Speech Synthesis

Speech synthesis, also known as text-to-speech (TTS), is a technology that converts written text into audible speech and the structure is shown in Figure 3 [6-10]. It is a crucial component in a wide range of applications, from voice assistants and audiobooks to accessibility tools for individuals with disabilities. The process of speech synthesis typically involves several key steps. First, a linguistic analysis is performed on the input text to identify phonemes, words, and sentence structures. Next, the model generates a phonetic transcription of the text, which is used to determine the pronunciation of each word and the prosody of the speech, including

intonation and stress patterns. This information is then combined with acoustic models that contain information about how different phonemes or units of speech sound. Finally, the model generates the speech waveform, which is the audible output that can be played through speakers or headphones. One notable technology for speech synthesis is Tacotron 2, which employs deep learning to produce natural-sounding speech from text.

Tacotron 2 is an advanced deep learning model designed for speech synthesis. It builds upon the success of its predecessor, Tacotron, and enhances the quality and naturalness of synthesized speech. This model employs a sequence-to-sequence architecture, which means it takes a sequence of input text (in the form of characters or phonemes) and generates a corresponding sequence of spectrogram frames that represent the speech. What sets Tacotron 2 apart is its ability to predict both the duration of phonemes and the corresponding spectral features simultaneously, ensuring more accurate prosody and natural intonation. The model then uses a neural vocoder, often WaveNet or Griffin-Lim, to convert the spectrogram frames into the final waveform, resulting in lifelike and expressive synthetic speech. Tacotron 2's advanced architecture, combined with its training on large and diverse datasets, has made it a powerful tool for various TTS applications, enabling more human-like and intelligible synthetic speech.

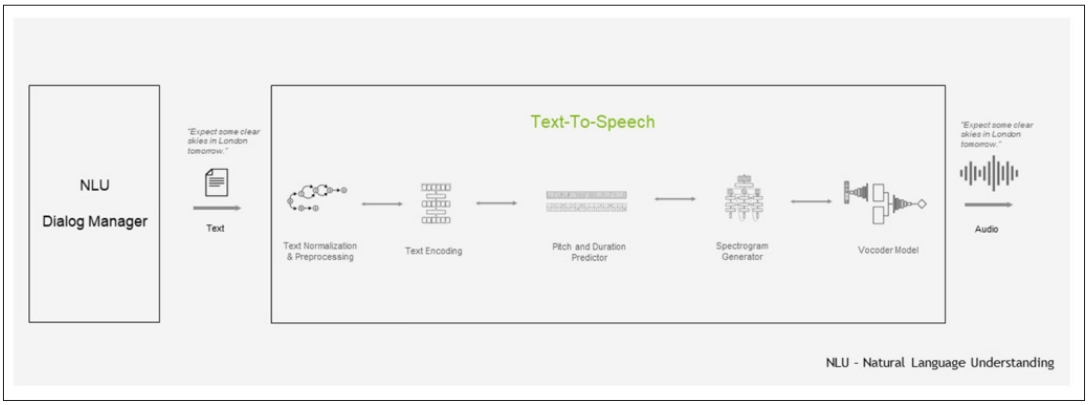

Figure 3: Speech Synthesis

NLP

Natural Language Processing (NLP) is a subfield of artificial intelligence that focuses on the interaction between humans and computers using natural language [11-15]. It encompasses a wide range of tasks, including text analysis, machine translation, sentiment analysis, chatbots, and more. NLP systems are designed to understand, interpret, and generate human language, enabling machines to communicate with users in a manner that is both meaningful and contextually relevant. The core components of NLP involve various levels of linguistic analysis, such as tokenization, part-of-speech tagging, syntax parsing, and semantic analysis. These tasks require the use of sophisticated models, algorithms, and large datasets to process and generate human language accurately and intelligently.

LLAMA 2, which stands for "Language Model for Language Modeling and Analysis," is a powerful large language model that can be effectively harnessed for a wide range of natural language processing tasks. Built upon cutting-edge deep learning techniques and trained on vast amounts of text data, LLAMA 2 excels at understanding and generating human language. It can be utilized for tasks such as text generation, summarization, translation, sentiment analysis, and more. Due to its extensive pretraining and fine-tuning capabilities, LLAMA 2 serves as a versatile tool for researchers and developers looking to tackle NLP challenges. Its sophisticated understanding of context, semantics, and syntax allows for the creation of more intelligent and contextually aware NLP applications, making it an invaluable asset in the field of natural language processing.

Hardware

Running speech recognition, speech synthesis, and natural language processing services on edge devices, like the NVIDIA Jetson, offers several advantages. Edge devices have the computational power and hardware acceleration required to perform these resource-intensive tasks locally, reducing the need for constant data transfer to remote servers. This results in significantly lower latency, ensuring real-time responsiveness and a smoother user experience, critical in applications such as autonomous vehicles or remote field operations. Moreover, edge devices provide greater privacy and data security since sensitive audio and text data can be processed locally without leaving the device. Their compact form factors make them ideal for mobility, enabling voice assistant applications in scenarios where lightweight, portable solutions are essential. Overall, the deployment of these services on edge devices optimizes performance, privacy, and mobility while enabling sophisticated voice assistant functionality in a wide range of applications.

Application

The integration of a speech recognition, speech synthesis, and natural language processing system in an autonomous aircraft offers a diverse range of applications. Passengers can benefit from voice-activated controls that enhance in-flight comfort and convenience, allowing them to communicate their preferences and requests naturally. The system can provide real-time information and assistance to the flight crew, improving operational efficiency and emergency response capabilities. Additionally, it can cater to passenger needs by offering in-flight services and entertainment through voice commands. Beyond passenger services, this technology can facilitate crew training and coordination, ensuring a smooth and safe flight experience. By providing multilingual support, enhancing accessibility, and enabling maintenance and diagnostics procedures, this integrated system holds the potential to significantly improve the overall functionality and experience within the autonomous aircraft environment.

Conclusion

In conclusion, the paper has outlined a groundbreaking approach to enhancing the capabilities of autonomous aircraft through the integration of a comprehensive speech recognition, speech synthesis, and natural language processing system. By deploying this technology, we can significantly improve the in-flight experience for passengers, streamline communication and coordination among the flight crew, and enhance operational efficiency. The potential applications of this system range from voice-activated controls and emergency response to cabin services, entertainment, and multilingual support, making it a valuable addition to the aviation industry. Furthermore, the system's ability to run on edge devices such as NVIDIA Jetson ensures low latency, robust privacy, and portability, making it an ideal fit for autonomous aircraft. As I look to the future of air travel, this integrated system holds the promise of reshaping how we interact with autonomous aircraft, setting the stage for more intuitive, safe, and efficient journeys in the sky.

References

1. Hinton G, Deng L, Yu D, Dahl GE, Mohamed AR, et al. (2012) Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. IEEE Signal Processing Magazine, 29(6), 82-97.
2. Graves A, Mohamed AR, Hinton G (2013) Speech recognition with deep recurrent neural networks. In Acoustics, speech and signal processing (ICASSP) 2013 IEEE international conference on 6645-6649.
3. Chan WY, Zhang Y, Jaitly N, Schuster M, Sivadas S, et al. (2016) Listen, attend and spell. arXiv preprint arXiv:1609.06773.
4. Amodei D, Ananthanarayanan S, Anubhai R, Bai J, Battenberg E, et al. (2016) Deep speech 2: End-to-end speech recognition in English and Mandarin. In International conference on machine learning 173-182.
5. Kim Y, Raghavan P, Alkhouli T, Jang SY (2017) On the training aspects of end-to-end speech recognition with RNN-transducer. In 2017 IEEE automatic speech recognition and understanding workshop (ASRU) 310-316.
6. Van Den Oord A, Dieleman S, Zen H, Simonyan K, Vinyals O, et al. (2016) Wavenet: A generative model for raw audio. arXiv preprint arXiv:1609.03499.
7. Shen J, Pang R, Weiss RJ, Schuster M, Jaitly N, et al. (2018) Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions. In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 4779-4783.
8. Sotelo J, Mehri S, Kumar K, Jain A, Leonard C, et al. (2017) Char2Wav: End-to-end speech synthesis. arXiv preprint arXiv:1702.07825.
9. Wang Y, Skerry-Ryan RJ, Stanton D, Wu Y, Jaitly N, et al. (2017) Tacotron: Towards end-to-end speech synthesis. arXiv preprint arXiv:1703.10135.
10. Ping W, Zhang Z, Wang H, Zen H, Wang W, et al. (2017) Deep learning for Chinese text-to-speech synthesis: From models to applications. arXiv preprint arXiv:1712.05689.
11. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, et al. (2017) Attention is all you need. In Advances in neural information processing systems 30-30.
12. Devlin J, Chang MW, Lee K, Toutanova K (2018) BERT: Bidirectional Encoder Representations from Transformers. arXiv preprint arXiv:1810.04805.
13. Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, et al. (2020) Language models are few-shot learners. arXiv preprint arXiv:2005.14165.
14. Radford A, Wu J, Child R, Luan D, Amodei D, et al. (2019) Language models are unsupervised multitask learners. Open AI Blog 1: 1-9.
15. Liu Y, Ott M, Goyal N, Du J, Joshi M, et al. (2019) RoBERTa: A robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692.